Adversary Detection and Resilient Control for Multi-agent Systems

Aquib Mustafa and Dimitra Panagou

***Abstract*—This paper presents an adversary detection mechanism and a resilient control framework for multi-agent systems under spatiotemporal constraints. Safety in multi-agent systems is typically addressed under the assumption that all agents collaborate to ensure the forward invariance of a desired safe set. This work analyzes agent behaviors based on designed behavior metrics, and designs a proactive adversary detection mechanism based on the notion of the critical region for the system operation. In particular, the presented detection mechanism not only identifies adversarial agents, but also ensures all-time safety for intact agents. Then, based on analysis and detection results, a resilient QP-based controller is presented with desired safety and liveness constraints for intact agents. Finally, simulation results validate the efficacy of the presented theoretical contributions.**

***Index Terms*—Control barrier functions, adversary detection, multi-agent systems, resilient control, autonomous systems.**

I. Introduction

In recent years, research for safety-critical systems has received vast recognition as safety is one of the prime requirements for autonomous systems. For a given system, safety is accomplished by ensuring forward invariance of a safe set, which is a subset of the system's state space. The objective is to design a control law such that the closed-loop system trajectories remain always in the safe set. In the existing literature, control barrier function (CBF) based approaches that leverage quadratic programming (QP) [1]–[5] methods have shown impactful results for providing safety guarantees for both single-agent [1], [2], [6] and multi-agent systems [5], [7]–[9]. These approaches are well-suited for online implementation as QPs can be efficiently solved in real-time [8]–[11].

The aforementioned results generally consider that all agents behave normally, i.e., they apply the nominally-specified control actions. However, these systems are vulnerable to a variety of adversaries, which aim to intentionally violate desired safety or goal-reaching constraints for normally-behaving agents within given control constraints. Therefore, it is of vital importance to design a proactive adversary detection mechanism and resilient control framework that can mitigate the effect of adversarial agents while ensuring all-time safety for intact agents.

In the existing literature, several remarkable results for resilient control are presented for multi-agent systems [9], [12]–[18]. In particular, mean-subsequence-reduced (MSR) based resilient control protocols for multi-agent systems are presented in [12]–[14]. Resilient algorithms for flocking and active target tracking applications are presented in [15], [16], respectively. An adaptive control and game-theoretic resilient designs are presented in [17] and [18] to directly mitigate the effect of adversaries without identifying them. However, to our knowledge, no prior studies using CBF approaches have considered identification of adversarial agent and resilient design for multi-agent CBF with safety and goal reaching objectives. Recently, authors in [19] presented adversarial resilience for sampled-data systems under safety constraints without any adversary identification. In [9], authors used CBFs to ensure that the communication topology satisfies the r-robustness property in finite time. However, the robots in formation are assumed to apply the nominal CBF-based controller without any adversarial misbehavior. Similarly, in [20] a class of fault-tolerant stochastic CBF is presented that provide probabilistic guarantees on the safety. This work is mainly focused on solving a secure state estimation problem by deriving geometrical conditions to resolve conflicts between the constraints that may arise due to sensor faults or attacks.

In this paper, we present an adversary detection mechanism as well as a resilient CBF framework for multi-agent systems. We consider heterogeneous multi-agent systems, modeled by control-affine dynamics, where in the presence of adversarial agents, intact agents are subject to accomplish the following objectives: (i) remain inside a safe set, which can be in general time-varying, (ii) reach desired goal locations either individually or in formation, and (iii) proactively identify adversarial agents in their neighborhood and take resilient action to ensure all-time safety. To achieve these objectives, this work first analyzes agent behaviors based on metrics that act as real-time behavior monitors, and then designs a proactive adversary detection mechanism based on the notion of the critical time and critical zone for the system operation. In particular, the presented critical zone is evaluated over the critical time window under best and worst-case control actions corresponding to intact and adversarial agents, respectively. Then, the augmentation of the critical zone with the desired safety constraints provides robustness to the agent's safe set such that the presented detection mechanism not only identifies adversarial agents but also acts to ensure all-time safety for intact agents. Finally, based on the presented behavior analysis and proactive adversary detection, a resilient QP-based controller is designed to ensure all-time safety for intact agents,

This work was partially sponsored by the Office of Naval Research (ONR), under grant number N00014-20-1-2395. The views and conclusions contained herein are those of the authors only and should not be interpreted as representing those of ONR, the U.S. Navy or the U.S. Government. This work was also partially supported by the National Science Foundation (NSF) under award number 1931982.

Aquib Mustafa and Dimitra Panagou are with the Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI-48109 USA (e-mails: aquibm@umich.edu, dpanagou@umich.edu).

in the presence of adversarial agents. The overall architecture is shown in Figure 1.

The rest of this paper is organized as follows. Section II provides the notations. Section III presents the problem formulation. Section IV formulates behavior analysis and detection mechanism. Resilient CBF mechanism is presented in Section V. Simulation results are provided in Section VI. Finally, concluding remarks are discussed in Section VII.

II. Notations

$\mathbb{R}$ and $\mathbb{R}_+$ represent the sets of real numbers and non-negative real numbers, respectively. $\mathbb{R}^n$ denotes n-dimensional Euclidean space. $\|x\|$ denotes Euclidean norm of vector $x \in \mathbb{R}^n$. The set of integers greater than m is represented by $\mathbb{Z}_{>m}$. The superscript $(.)^T$ denotes transposition. The cardinality of a set S is denoted by $|S|$. The Lie derivative of a continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}$ along a vector $f : \mathbb{R}^n \to \mathbb{R}^n$ at point $x \in \mathbb{R}^n$ is represented as $L_f V(x) \triangleq \frac{\partial V(x)}{\partial x} f(x)$. We use ∂S to denote the boundary of a closed set S and $int(S)$ to denote its interior. $diag(A_1, \dots, A_n)$ represents a diagonal matrix with A_i as its diagonal entries, $\forall, i \in [1, \dots, n]$. $\wedge$ or $\bigcap$ denotes conjunction/and operator. Eventual and global temporal operators are represented by $\diamondsuit$ and $\square$.

III. Problem Formulation

Consider a group of $N \in \mathbb{Z}_{>0}$ agents, with the set of agents represented by $\mathscr{V}$ and each agent indexed $\{1, \dots, N\}$. The system dynamics of each agent $i \in \mathscr{V}$ is given by

$$\dot{x}_i(t) = f_i(x_i(t)) + g_i(x_i(t))u_i(t), \tag{1}$$

where the state vector is $x_i(t) = [p_i(t) \;\; \varphi_i(t)] \in \mathbb{R}^3$, with $p_i(t) \in \mathbb{R}^2$ and $\varphi_i(t) \in \mathbb{R}$ denoting the position vector and orientation, respectively, of agent i with respect to a global reference frame. The vector $u_i(t) \in \mathbb{R}^{m_i}$ denotes the control input of agent i, respectively. The functions $f_i \in \mathbb{R}^3$ and $g_i \in \mathbb{R}^{3 \times m_i}$ may differ among agents, but are all locally Lipschitz. We denote $f_i^p \in \mathbb{R}^2$ and $g_i^p \in \mathbb{R}^{2 \times m_i}$ the sub-matrices of f_i and g_i corresponding to the position-vector dynamics in (1). The control input constraints for each input $u_i(t)$ are represented by a nonempty, convex, compact polytope, i.e., $u_i(t) \in \mathscr{U}_i(x_i(t)) = \{u \in \mathbb{R}^{m_i} : A_i(x_i(t))u \leq b_i(x_i(t))\}$ where the functions $A_i(x_i(t)) : \mathbb{R}^3 \to \mathbb{R}^{q_i \times m_i}$ and $b_i(x_i(t)) : \mathbb{R}^3 \to \mathbb{R}^{q_i}$ are locally Lipschitz on their respective domains. Moreover, the collection of the position vectors and the control input vectors are represented as $\vec{p} = [p_1^T, p_2^T, \dots, p_N^T]^T \in \mathbb{R}^{2N}$ and $\vec{u} = [u_1^T, u_2^T, \dots, u_N^T]^T \in \mathbb{R}^{\underline{\mathrm{m}}}$ with $\underline{\mathrm{m}} = \sum_{i=1}^N m_i$, respectively.

For each agent i, the conjunction of m different safety constraints $h_n^i : \mathbb{R}^{2N} \to \mathbb{R}$, $n \in \{1, \dots, m\}$, is represented by the composite control barrier function (CBF) $h_i^s(\vec{p}) : \mathbb{R}^{2N} \to \mathbb{R}$ via Boolean AND operations using the log-sum-exp (LSE) smooth approximation to the max(.) function [5], given by

$$h_i^s(\vec{p}) = LSE[h_1^i, h_2^i, \dots, h_m^i] = ln(\sum_{n=1}^{m} exp^{h_n^i}). \tag{2}$$

We consider a set S_i^s defined as the superlevel set of a continuously differentiable function $h_i^s(\vec{p})$, given by $S_i^s = \{\vec{p} | h_i^s(\vec{p}) \leq 0\}$, $\partial S_i^s = \{\vec{p} | h_i^s(\vec{p}) = 0\}$, $\mathrm{int}(S_i^s) = \{\vec{p} | h_i^s(\vec{p}) < 0\}$. In this paper, we refer S_i^s as a safe set for agent i, and we assume that forward invariance of this set S_i^s can always be ensured [1], see Lemma 1 later on. We consider two class of safety constraints and their corresponding safe sets, namely, inter-agent and agent-to-obstacle safety constraints. We represent the conjunction of inter-agent safety constraints as $\bar{h}_i^s(\vec{p}) = \bigcap_{j \in \mathscr{N}_i} h_{ij}^s(p_i, p_j)$ with

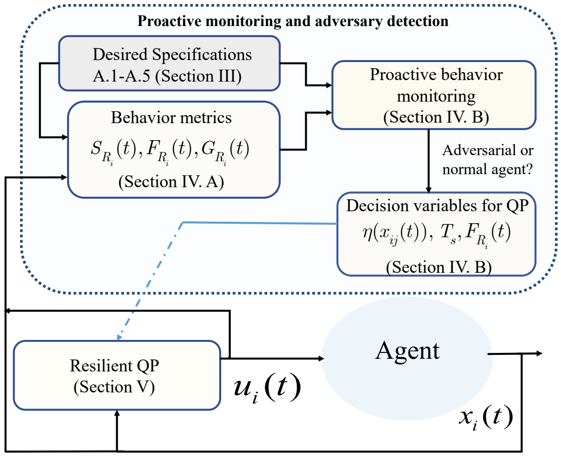

Proactive monitoring and adversary detection
Desired Specifications A.1-A.5 (Section III)
Behavior metrics
$S_{R_i}(t), F_{R_i}(t), G_{R_i}(t)$
(Section IV. A)
Proactive behavior monitoring (Section IV. B)
Adversarial or normal agent?
Decision variables for QP
$\eta(x_{ij}(t)),\ T_s, F_{R_i}(t)$
(Section IV. B)
Resilient QP (Section V)
$u_i(t)$
Agent
$x_i(t)$

Figure 1: Agents behavior: for desired specification without any adversary.

$$h_{ij}^s(p_i, p_j) = d - \|p_i - p_j\| \leq 0, \tag{3}$$

where d is a desired inter-agent safety distance, and the set $\mathscr{N}_i = \{j | \, \|p_i - p_j\| \leq R_s\}$ is the set of neighbors of agent i, where the limited sensing radius $R_s > d$. Then, the inter-agent safe set is defined as

$$\bar{S}_i^s = \{\vec{p} | \bar{h}_i^s(\vec{p}) \leq 0\}, \tag{4}$$

with $\bar{S}_i^s \supset S_i^s$. Similarly, the conjunction of agent-to-obstacle safety constraints is denoted by $\hat{h}_i^s(p_i) = \bigcap_{o_j \in \mathscr{O}_i} h_i^{o_j}(p_i)$ with

$$h_i^{o_j}(p_i) = r_{o_j} - \|p_i - c_{o_j}\| \leq 0, \tag{5}$$

where c_{o_j} and r_{o_j} denote the center and radius of spherical obstacles, and $\mathscr{O}_i$ denotes the set of the obstacles for the agent i. Then, the agent-to-obstacle safe set is defined as

$$\hat{S}_i^s = \{p_i | \hat{h}_i^s(p_i) \leq 0\}, \tag{6}$$

with $\hat{S}_i^s \supset S_i^s$. Moreover, all pairwise safety constraints, i.e., inter-agent and agent-to-obstacle safety constraints can be encoded in the augmented CBF $h_i^s(\vec{p})$ in (2) with

$$S_i^s = \bar{S}_i^s \cap \hat{S}_i^s, \tag{7}$$

where $m = |\mathscr{N}_i| + |\mathscr{O}_i|$ denotes the total number of safety constraints. Note that while in principle one can define a CBF h_i^s that is a function of both position and orientation, here we chose to define the inter-agent and agent-to-obstacle safety constraints in terms of the Euclidean distance only, as they encode that the area footprints of the agents should never intersect for any possible orientations. Note also that, for the dynamics in (1), the considered control action constraint in the QP-based control design directly provides a constraint on the rate of change of orientation.

Next, we present some preliminaries related to the performance of the agents.

Definition 1. A continuously differentiable function $V_i^g(\vec{p}) : \mathbb{R}^{2N} \to \mathbb{R}$ is called exponentially stabilizing control Lyapunov function (ES-CLF) for the positional dynamics of (1) if there exist $\beta_1, \beta_2, \beta_3 \in \mathbb{R}_+$ such that following conditions hold

$$\begin{aligned} \beta_1 \|\vec{p}\|^2 \leq V_i^g(\vec{p}) \leq \beta_2 \|\vec{p}\|^2 \\ \inf_{u_i(t) \in \mathscr{U}_i} [L_{f_i^p} V_i^g(\vec{p}) + L_{g_i^p} V_i^g(\vec{p}) u_i(t) + \beta_3 V_i^g(\vec{p})] \leq 0. \end{aligned} \tag{8}$$

where $f_i^p \in \mathbb{R}^2$ and $g_i^p \in \mathbb{R}^{2 \times m_i}$ denotes the sub-matrix of f_i and g_i corresponding to position vector dynamics in (1).

Reaching a goal location for agent i can be encoded via the following candidate Lyapunov function

$$V_i^g(p_i) = \|p_i - G_i\|^2, \, \forall i \in \mathscr{V}, \tag{9}$$

where $G_i \in \mathbb{R}^2$ denotes goal point or goal region. Similarly, goal reaching for a collaborative task among the set of agents $\mathscr{V}_f \subseteq \mathscr{V}$ in the form of formation can be encoded as

$$\bar{V}_f^g(\bar{p}) = \left\|\bar{p} - G_f\right\|^2, \tag{10}$$

where $G_f \in \mathbb{R}^2$ denotes goal point or goal region and $\bar{p} = \frac{1}{|\mathscr{V}_f|} \sum_{i \in \mathscr{V}_f} p_i$. We also consider that agents need to maintain a formation over time such that

$$h_i^{\mathscr{F}_i}(\vec{p}(t)) = \lim_{t \to \infty} \|p_i(t) - p_i^*(t)\| \to 0, \, \forall i \in \mathscr{V}_f, \tag{11}$$

where $p_i^*(t) = \frac{1}{|\mathscr{N}_i|} \sum_{j \in \mathscr{N}_i} (p_j(t) + c_{ji})$ and c_{ij} denotes inter-agent distance for the formation.

Definition 2 (Adversarial agent). We call an agent j *adversarial* if under some adversarial control action $u_j^a(t)$, either of the following holds:

1) It performs chasing to hit an intact agent i in some finite time, i.e., $\exists \, t_a < \infty$ such that $\left\|p_i(t) - p_j(t)\right\| \to 0$ as $t \to t_a$.
2) It aims to mislead agents in the set $\mathscr{V}_f$ so that they converge to a location $\bar{p}(t) = \frac{1}{|\mathscr{V}_f|} \sum_{i \in \mathscr{V}_f} p_i(t)$ such that $\left\|\bar{p}(t) - G_f\right\| \neq 0$ as $t \to \infty$.

We denote the overall set of adversarial agents by $\mathscr{A} = \mathscr{A}_s \cup \mathscr{A}_f$, where $\mathscr{A}_s$ and $\mathscr{A}_f$ represent set of adversarial agents correspond to classes 1 and 2 in definition, respectively. We called an agent *intact* if it is not adversarial.We denote the set of intact agents as $\mathscr{V} / \mathscr{A}$.

Definition 3 (Proactive adversary detection). We call an adversary detection mechanism for agent i *proactive*, if detection of the adversary happens at some time $t_d < t_a$ where t_a is given in Definition 2. In particular, adversary detection happens before any adversarial agent $k \in \{i, j\}$ violates the inter-agent safety constraint in (3) (i.e., either agent i itself or any neighbor agent j violates the inter-agent safety constraint).

We define the following as the desired objectives for the intact agent $i \in \mathscr{V} / \mathscr{A}$:

A.1 Satisfy constraints on control input, $u_i(t) \in \mathscr{U}_i(x_i(t)) = \{u \in \mathbb{R}^{m_i} : A_i(x_i(t)) u \leq b_i(x_i(t))\}$.

A.2 Ensure the forward invariance of the safe set, $p_i(t) \in S_i^s$, for all $t > 0$.

A.3 Guarantee convergence of the closed-loop trajectories to the goal region, i.e., $V_i^g(p_i(t)) \to 0$ as $t \to \infty$.

A.4 Maintain a formation over the time interval $t \in [t_a, \, t_b]$ such that $\lim_{t \to t_b} h_i^{\mathscr{F}_i}(\vec{p}(t)) \to 0, \, \forall i \in \mathscr{V}_f$.

A.5 Based on behavior metrics that capture the degree of satisfaction of the desired objectives A.1-A.4, proactively determine the set of adversarial agents $\mathscr{A} \subset \mathscr{V}$.

Please note that, in this work, we assumed that the adversarial agent has the same controller constraints on control input as mentioned in condition A.1.

Based on the objectives A.1-A.5, the problem formulation is presented as follows.

Problem 1. Consider the desired objectives A.1-A.5. Design control law $u_i(t) \in \mathscr{U}_i(x_i(t))$ for the system (1) such that objectives A.1-A.4 are satisfied for each intact agent $i \in \mathscr{V} / \mathscr{A}$.

A. Forward Invariance of a Set

In this subsection, under the assumption of no adversarial agents, i.e., $\mathscr{A} = \emptyset$, we review the necessary and sufficient conditions for guaranteeing forward invariance of a set S_i^s, known also as Nagumo's Theorem.

Lemma 1. *Let the solution of* (1) *exist and be unique in forward time. Then, for each agent $i \in \mathscr{V}$, the set S_i^s is forward invariant for the closed-loop trajectories of* (1) *for all $p_i(0) \in S_i^s$ if and only if the following condition holds:*

$$\inf_{u_i(t) \in \mathscr{U}_i} \{\mathrm{L}_{f_i^p} h_i^s(\vec{p}) + \mathrm{L}_{g_i^p} h_i^s(\vec{p}) u_i(t) \leq 0\}, \, \forall p_i(t) \in \partial S_i^s, \tag{12}$$

where ∂S_i^s represents the boundary of the safe set S_i^s.

Interested readers can refer to [21] for more details on forward invariance of sets.

To ensure the feasibility of Problem 1 when $\mathscr{A} = \emptyset$, we make the following assumption. In the existing literature, similar assumptions have been used either explicitly or implicitly (see e.g. [2]).

Assumption 1. The trajectories of each agent $i \in \mathscr{V} / \mathscr{A}$ satisfy the condition (12), for all $p_i \in \partial S_i^s$.

Assumption 2. The interior of the set $\mathscr{U}_i(x_i(t))$ is nonempty and $\mathscr{U}_i(x_i(t))$ is uniformly compact near $x_i(t)$.

Assumption 3. The functions f_i and g_i are locally Lipschitz with Lipschitz constants $b_f \in \mathbb{R}_+$ and $b_g \in \mathbb{R}_+$, $\forall i \in \mathscr{V}$, respectively.

Lemma 2. *If the initial conditions for an agent $i \in \mathscr{V}$ are such that $h_i^s(\vec{p}(0)) < 0$ and the inequality*

$$\inf_{u_i(t) \in \mathscr{U}_i} \{\mathrm{L}_{f_i^p} h_i^S(\vec{p}(t)) + \mathrm{L}_{g_i^p} h_i^S(\vec{p}(t)) u_i(t) \leq \alpha(-h_i^S(\vec{p}(t)))\}, \tag{13}$$

holds for some locally Lipschitz class-$\mathscr{K}$ function α for all $t \geq 0$, then for any $T > 0$, $h_i^s(x_i(t)) < 0$, $\forall 0 \leq t \leq T$.

B. A Quadratic Program for Safety-Control Synthesis

This subsection presents a quadratic program (QP) to compute a control input $u_i(t)$ for each agent $i \in \mathscr{V}$ to solve Problem 1 when $\mathscr{A} = \emptyset$. Let $\vec{z} = [z_1^T, z_2^T, \ldots, z_N^T]^T$ be a column vector with $z_i = [u_i, \delta_{i_1}, \delta_{i_2}, \delta_{i_3} \in \mathbb{R}^{m_i+3}]$, $\forall i \in \mathscr{V}$ as its elements. Consider the following optimization problem

$$\min_{u_i, \delta_{i_1}, \delta_{i_2}, \delta_{i_3}, \, i \in \mathscr{V}} \quad \vec{z}^T H \vec{z} + F \vec{z} \tag{14a}$$

$$s.t. \;\; A_i u_i \leq b_i, \tag{14b}$$

$$L_{f_i^p} V_i^g + L_{g_i^p} V_i^g u_i \leq -\delta_{i_1} V_i^g, \;\; \forall i \in \mathscr{V} / \mathscr{V}_f, \tag{14c}$$

$$L_{f_i^p} h_i^{\mathscr{F}_i} + L_{g_i^p} h_i^{\mathscr{F}_i} u_i \leq -\delta_{i_2} h_i^{\mathscr{F}_i}, \;\; \forall i \in \mathscr{V}_f, \tag{14d}$$

$$L_{f_i^p} h_i^S + L_{g_i^p} h_i^S u_i \leq -\delta_{i_3} h_i^S, \tag{14e}$$

where $H = diag\{H_i\}$ with $H_i = diag\{\{w_{u_l}^i\}, w_1^i, w_2^i, w_3^i\}$ denotes a diagonal matrix with positive weights $w_{u_l}^i, w_1^i, w_2^i, w_3^i >$

0 and similarly, $F = diag\{F_i\}$ with $F_i = [0_{m_i}^T \; q^i \; 0 \; 0]$ where $q^i > 0$ and $0_k \in \mathbb{R}^k$ denotes a column vector consisting of zeros. Control input constraints are encoded in (14b). Performance-based constraints, i.e., goal reaching constraints and formation are encoded in (14c) and (14d). Similarly, safety-based constraints are encoded in (14e) (which can include both inter-agent safety constraints and obstacle avoidance constraints in conjunction form, as defined in (2)). Note that, one can refer the set of trajectories corresponding to V_i^g as a subset of the safe set S_i^s in (7) corresponding to h_i^s in (12).

IV. Behavior Analysis and Detection Mechanism

A. Preliminaries

In this subsection, inspired by worst-case designs in the CBF literature [19], [22], [23], we first introduce the notion of the best and worst-case control actions for the intact and adversarial agent, respectively. Then, we present the concepts of the critical-time period and critical zone for the agent's operation, which are later required for proactive behavior monitoring, adversary detection and mitigation. In order to provide the best safety guarantee, the minimum pointwise control action by an intact agent $i \in \mathscr{V}/\mathscr{A}$ is defined as

$$u_i^{\min}(t) = \arg\min_{u_i \in \mathscr{U}_i} L_{f_i^p} h_i^S(\vec{p}) + L_{g_i^p} h_i^S(\vec{p}) u_i, \tag{15}$$

where $h_i^S(\vec{p})$ in (2) encodes all pairwise safety constraints, i.e., inter-agent and agent-to-obstacle safety constraints. $u_i^{\min}(t)$ in (15) can be determined by solving the linear program (LP)

$$\min_{u_i \in \mathscr{U}_i} m(\vec{p})^T u_i, \quad s.t. \quad A_i u_i \leq b_i, \tag{16}$$

where $m(\vec{p})^T = L_{g_i^p} h_i^S(\vec{p})$. Note that the pointwise minimum control input provides the best action towards maintaining safety, as the LP in (16) determines the pointwise minimum control action only in terms of the safety constraints $h_i^S(\vec{p})$ without considering any goal reaching objective as encoded in (14c) in terms of V_i^g. Similarly, to achieve worst-case safety behavior, the maximum pointwise control action by an adversarial agent $j \in \mathscr{A}$ is defined as

$$u_j^{\max}(t) = \arg\max_{u_j \in \mathscr{U}_j} L_{f_j^p} \bar{h}_j^s(\vec{p}) + L_{g_j^p} \bar{h}_j^s(\vec{p}) u_j, \tag{17}$$

where $\bar{h}_j^s(\vec{p})$ in (3) encodes only agent-to-agent safety constraints. In similar fashion, $u_j^{\max}(t)$ in (17) can be determined by solving the following LP

$$\max_{u_j \in \mathscr{U}_j} m(\vec{p}_j)^T u_j, \quad s.t. \quad A_j u_j \leq b_j, \tag{18}$$

with $m(p_j)^T = L_{g_j^p} \hat{h}_j^s(p_j)$. Note that in order to achieve desired adversarial chasing behavior as defined in Definition 2, adversarial agent j only needs to account for inter-agent safety. $u_j^{\max}(t)$ denotes the worst control effort by an adversarial agent $j \in \mathscr{A}$ in order to maximize $L_{f_j^p} \hat{h}_j^s(p_j) + L_{g_j^p} \hat{h}_j^s(p_j) u_j$ and achieve worst-case safety (i.e., best effort to achieve unsafe behavior). The feasibility of LP in (16) is guaranteed under Assumption 2, see [19].

B. Critical Time and Critical Zone

Now, in order to design behavior monitors and an adversary detection mechanism, we first present the concept of the critical time period and critical zone for the system (1) in this subsection. We define the critical time period T_s, and compute T_s based on the inter-agent safety constraints set $\bar{S}_i^s$ in (4). We define inter-agent distance as

$$r_{ij}(t) = \left\| p_i(t) - p_j(t) \right\|, \; \forall t > 0. \tag{19}$$

Definition 4. T_s is the **critical time period** of system (1) at the time t_k, if under the best-case control input (15) computed at the time t_k, i.e., $u_i(t) = u_i^{min}(t_k)$, $\forall t \in [t_k, t_k + T_s]$, $p_i(t_k) \in int(\bar{S}_i^s)$ implies that $p_i(t_k + T_s) \in \partial \bar{S}_i^s$.

Theorem 1. *Let Assumption* 3 *hold. Consider the agent dynamics* (1) *along with the best-case control input $u_i^{min}(t_k)$ at the time t_k given by* (15)*. Then, for the safe set $\bar{S}_i^s$ in* (4)*, the critical time period is given by $T_s = \min_{j \in \mathscr{N}_i} \{T_s^j\}$ with*

$$T_s^j \geq \frac{1}{(b_f + b_g \left\| u_i^{\min}(t_k) \right\|)} \log\Big(\frac{1}{1 - \frac{r_{ij}(t_k) - d}{k_1(t_k)}}\Big), \tag{20}$$

where $r_{ij}(t_k)$ is defined in (19)*, b_f and b_g denote the Lipschitz constant for functions $f_i(x_i(t))$ and $g_i(x_i(t))$ in* (1)*, d represents the inter-agent safety distance and*

$$k_1(t_k) = r_{ij}(t_k) + \frac{b_g \left\| p_j(t_k) \right\| \left\| u_j^{max}(t_k) - u_i^{min}(t_k) \right\|}{b_f + b_g \left\| u_i^{\min}(t_k) \right\|}. \tag{21}$$

Proof. Define the change in function for the inter-agent safety constraint defined in (3) over the time interval $[t_k, t_k + t]$ as

$$\Upsilon(t + t_k, t_k) = h_{ij}^s(p_i(t + t_k), p_j(t + t_k)) - h_{ij}^s(p_i(t_k), p_j(t_k)), \tag{22}$$

which can be written as

$$\Upsilon(t + t_k, t_k) = r_{ij}(t_k) - r_{ij}(t + t_k), \tag{23}$$

and its derivative can be computed as

$$\dot{\Upsilon}(t + t_k, t_k) = -\tfrac{(p_i(t+t_k) - p_j(t+t_k))^T}{\left\| p_i(t+t_k) - p_j(t+t_k) \right\|} (\dot{p}_i(t + t_k) - \dot{p}_j(t + t_k)), \tag{24}$$

where the term $\frac{(p_i(t+t_k) - p_j(t+t_k))^T}{\left\| p_i(t+t_k) - p_j(t+t_k) \right\|}$ is a unit vector and thus,

$$\dot{\Upsilon}(t + t_k, t_k) \leq \left\| (\dot{p}_j(t + t_k) - \dot{p}_i(t + t_k)) \right\|. \tag{25}$$

From the triangular inequality, one has

$$\begin{aligned} \dot{\Upsilon}(t + t_k, t_k) \leq & \left\| f_j^p(x_j(t + t_k)) - f_i^p(x_i(t + t_k)) \right\| \\ & + \left\| g_j^p(x_j(t + t_k)) u_j(t + t_k) - g_i^p(x_i(t + t_k)) u_i(t + t_k) \right\|, \end{aligned} \tag{26}$$

where $f_i^p \in \mathbb{R}^2$ and $g_i^p \in \mathbb{R}^{2 \times m_i}$ denotes the sub-matrix of f_i and g_i corresponding to position vector dynamics in (1). Based on Assumption 3, f_i^p and g_i^p, $\forall i \in \mathscr{V}$, are locally Lipschitz and are bounded by the Lipschitz constants $b_f \in \mathbb{R}_+$ and $b_g \in \mathbb{R}_+$, respectively. Thus, under the constant best control input evaluated at the time t_k, i.e., $u_i(t) = u_i^{min}(t_k)$, $\forall t \in [t_k, t_k + T_s^j]$, equation (26) becomes

$$\begin{aligned} \dot{\Upsilon}(t + t_k, t_k) \leq & (b_f + b_g \left\| u_i^{\min}(t_k) \right\|) \left\| (p_i(t + t_k)) - (p_j(t + t_k)) \right\| \\ & + \Delta(p_j(t_k)), \end{aligned} \tag{27}$$

with $\Delta(p_j(t_k)) = b_g \left\| p_j(t_k) \right\| \left\| u_j^{max}(t_k) - u_i^{min}(t_k) \right\|$. Now based on the defined error term $\Upsilon(t + t_k, t_k)$ in (22), one can write

$$\begin{aligned} \dot{\Upsilon}(t + t_k, t_k) \leq & -(b_f + b_g \left\| u_i^{\min}(t_k) \right\|) \Upsilon(t + t_k, t_k) \\ & + (b_f + b_g \left\| u_i^{\min}(t_k) \right\|) r_{ij}(t_k) + \Delta(p_j(t_k)). \end{aligned} \tag{28}$$

Then, based on Comparison Lemma [24], one has following solution

$$\Upsilon(t+t_k,t_k) \leq k_1(t_k)(1-e^{-(b_f+b_g\|u_i^{\min}(t_k)\|)(t-t_k)}), \tag{29}$$

with $k_1(t_k)$ in (21). Now, with $h_{ij}^s(p_i,p_j)$ defined in (3), we know that at the time instant $t = T_s^j$ from (22), one has $\Upsilon(t_k+T_s^j,t_k) = r_{ij}(t_k)-d$ as $h_{ij}^s(p_i(t_k+T_s^j),p_j(t_k+T_s^j))$ becomes zero as $p_i(t_k+T_s^j) \in \partial\bar{S}_i^s$, i.e., $d-r_{ij}(t_k+T_s^j)=0$. Then based on (29), the bound on critical time period T_s^j can be computed as

$$\frac{r_{ij}(t_k)-d}{k_1(t_k)} \leq (1-e^{-(b_f+b_g\|u_i^{\min}(t_k)\|)T_s^j}), \tag{30}$$

which finally yields (20). This completes the proof. ■

In the following theorem, we present the result for evaluation of the critical time period T_s^o for the conjunction of agent-to-obstacle safe set $\bar{S}_i^s$ in (6).

Theorem 2. *Let Assumption 3 hold. Consider the agent dynamics* (1) *along with the worst-case control input in* (17) *evaluated at the time t_k. Then, for the safe set $\hat{S}_i^s$ in* (6)*, the critical time period $T_s^o = \min_{oj\in\mathscr{O}_i}\{T_s^{oj}\}$ with*

$$T_s^{oj} \geq \frac{1}{(b_f+b_g\|u_i^{\max}(t_k)\|)}\log(\frac{1}{1-\frac{v_i^{oj}(t_k)-r_{oj}}{v_i^{oj}(t_k)+r_{oj}}}), \tag{31}$$

where $v_i^{oj}(t_k) = \|p_i(t_k)-c_{oj}\|$ and, c_{o_j} and r_{o_j} are defined in (5)*. Moreover, b_f and b_g denote the Lipschitz constant for functions $f_i(x_i(t))$ and $g_i(x_i(t))$ in* (1).

Proof. The result follows a similar argument as given in the proof of Theorem 1 with agent-obstacle pairwise safety function $h_i^{o_j}(p_i(t))$ in (5) instead of inter-agent safety function $h_{ij}^s(p_i,p_j)$ in (3). ■

Definition 5. The critical zone $\eta(p_i(t),p_j(t)) : \mathbb{R}^2 \to \mathbb{R}$ is defined as

$$\eta(p_i(t),p_j(t)) = \max_{u_i(t),u_j(t)}\left\|\int_t^{t+nT_s}(\dot{p}_i(\tau)-\dot{p}_j(\tau))d\tau\right\|, \tag{32}$$

which represents the maximum magnitude of the evolution of the difference between the position trajectories of an intact agent i and its neighbor j over the time interval $[t,t+nT_s]$, with $n\in\mathbb{Z}_{>1}$ and T_s being the critical time period provided in Definition 4.

Remark 1. *Note that nT_s in Definition 5 denotes the desired sampling time for trajectory evaluation of agent i and it can be designed such that future safety is always ensured. Based on Theorem 1, under the best-case control input, i.e., $u_i^{min}(t_k)$, the agent i reaches the boundary of the safe set $\bar{S}_i^s$ in* (4) *over the time interval $[t_k,t_k+T_s]$. That's why the critical region $\eta(p_i(t),p_j(t))$ is evaluated over horizon $[t,t+nT_s]$ with the design parameter $n\in\mathbb{Z}_{>1}$ and augmented with the inter-agent safety constraints such that it that provides robustness to actual safe region. Then, we leverage this notion to proactively detect adversarial agent without violating safety constraints and ensure all time safety for all $i\in\mathscr{V}/\mathscr{A}$ based on resiliency mechanism as presented in Section IV. C and V, respectively.*

Note that $\eta(p_i(t),p_j(t))$ can be maximized by applying best and worst-case control input $u_i^{min}(t)$ in (15) and $u_j^{max}(t)$ in (17), respectively, over the time interval $[t,t+nT_s]$. Thus, based on Definition 5, we can rewrite critical zone as

$$\eta(p_i(t),p_j(t)) = \left\|\int_t^{t+t_s}(f_i^p(\tau)+g_i^p(\tau)u_i^{min}-f_j^p(\tau)-g_j^p(\tau)u_j^{max})d\tau\right\|. \tag{33}$$

where $f_i^p\in\mathbb{R}^2$ and $g_i^p\in\mathbb{R}^{2\times m_i}$ denotes the sub-matrix of f_i and g_i corresponding to position vector dynamics in (1). For brevity, we denote $\eta(p_{ij}(t)) = \eta(p_i(t),p_j(t))$ in the rest of the paper. Also the presented formulation for the critical zone $\eta(p_{ij}(t))$ can be directly extended for the static obstacle avoidance case with obstacle at fixed position, i.e.,

$$\eta(p_{io_j}(t)) = \left\|\int_t^{t+\bar{t}_s}(f_i^p(\tau)+g_i^p(\tau)u_i^{max})d\tau\right\|, \tag{34}$$

where $\bar{t}_s = nT_s^{oj}$ with T_s^{oj} defined in (31). The designed critical zones $\eta(p_{ij}(t))$ and $\eta(p_{io_j}(t))$ are leveraged in Section IV.B for behavior monitoring and design of proactive adversary detection mechanism.

C. Behavior Metrics for Monitoring and Proactive Adversary Detection

In this subsection, we first introduce behavior metrics and then design behavior monitors for the detection of adversarial agents belongs to the set $\mathscr{A}_s$. We present the following behavior metrics: (i) Safety behavior metric, and (ii) Goal reaching behavior metric.

Safety behavior metric: To monitor the agents behavior for safe operation, the **safety behavior metric** is defined as

$$S_{R_i}(t) = \exp\left(-(\Gamma_i(p_i(t),p_j(t)))^{n_c}\right),\ \forall t\geq 0, \tag{35}$$

where

$$\Gamma_i(p_i(t),p_j(t)) = \max\{g_{ij}(p_i(t),p_j(t)), g_{io_j}(p_i(t))\}, \tag{36}$$

with

$$g_{ij}(p_i(t),p_j(t)) = \frac{d}{\|p_i(t)-p_j(t)\|},\ \forall j\in\mathscr{N}_i, \tag{37}$$

and

$$g_{io_j}(p_i(t)) = \frac{r_{o_j}}{\|p_i(t)-c_{o_j}\|},\ \forall o_j\in\mathscr{O}_i, \tag{38}$$

where c_{o_j} and r_{o_j} c_{o_j} and r_{o_j} are defined in (5). Moreover, $\mathscr{O}_i$ and $n_c\in\mathbb{Z}_{>1}$ denote the set of obstacles for the agent i and a constant design gain, respectively. Now the following proposition shows that how the designed safety behavior metric $S_{R_i}(t)$ acts as a safety monitor for an agent i.

Proposition 1. *Consider the agent dynamics* (1) *along with the safety behavior metric $S_{R_i}(t)$ in* (35)*. For the defined safe set S_i^s in* (7)*, if $p_i(t)\in int(S_i^s)$ ($p_i(t)\notin S_i^s$), then the safety behavior metric $S_{R_i}(t)\to 1$ ($S_{R_i}(t)\to 0$) for all time t.*

Proof. Note that, based on safety constraints in (3) and (5), for any $p_i(t)\in int(S_i^s)$, i.e., agent operating in safe region, one has $g_{ij}(p_i(t),p_j(t)) < 1$ and $g_{io_j}(p_i(t)) < 1$. Similarly, for any $p_i(t)\notin S_i^s$, i.e., agent operating in unsafe region, one has $g_{ij}(p_i(t),p_j(t)) > 1$ and $g_{io_j}(p_i(t)) > 1$. Thus, based on (37), (38) and (35), the safety behavior metric $S_{R_i}(t)\to 1$ ($S_{R_i}(t)\to 0$) if $p_i(t)\in int(S_i^s)$ ($p_i(t)\notin S_i^s$) with a proper design constant $n_c\in\mathbb{Z}_{>1}$. Therefore, based on the metric $S_{R_i}(t), \forall t\geq 0$, one can employ $S_{R_i}(t)$ as a monitor to evaluate the safe (unsafe) behavior of the agent. ■

Now, we leverage the presented critical zone in Section IV. B and define a metric to capture worst-case safety behavior in terms of critical zones as

$$S^w_{R_i}(t) = \exp(-\Gamma^w_i(p_i(t), p_j(t))^{n_c}),\ \forall t \geq 0, \tag{39}$$

where

$$\Gamma^w_i(p_i(t), p_j(t)) = 1 - \max\{g^w_{ij}(p_i, p_j),\ g^w_{io_j}(p_i)\}, \tag{40}$$

with

$$g^w_{ij}(p_i(t), p_j(t)) = \frac{\eta(p_{ij}(t))}{\|p_i(t) - p_j(t)\|}, \tag{41}$$

and

$$g^w_{io_j}(p_i(t)) = \frac{\eta(p_{io_j}(t))}{\|p_i(t) - c_{o_j}\|}, \tag{42}$$

$\forall i \in \mathscr{V}$, $o_j \in \mathscr{O}_i$, with $\eta(p_{ij}(t))$ and $\eta(p_{io_j}(t))$ critical zones defined in (32) and (34), respectively. We define the error between the worst-case and nominal safety metrics as

$$\gamma^S_i(t) = \|S^w_{R_i}(t) - S^n_{R_i}(t)\|, \tag{43}$$

with $S^n_{R_i}(t) = 1$ as the nominal safety behavior metric. Note that $S^n_{R_i}(t)$ represents safety behavior metric $S_{R_i}(t)$ in (35) under nominal operation of agent i, i.e., when $p_i(t) \in int(S^s_i)$.

Now, we present the first result on behavior monitoring, and show that if the safety behavior metric value is below a designed threshold, then one can always ensure the operation of an agent inside the safe set S^s_i.

Theorem 3. *Consider the agent dynamics* (1) *along with the QP based control* (14) *and the safety behavior metric* $S_{R_i}(t)$ *in* (35)*. If it holds that*

$$\|S_{R_i}(t) - S^n_{R_i}(t)\| \leq \gamma^S_i(t),\ \forall t \geq 0 \tag{44}$$

where $\gamma^S_i(t)$ *defined is* (43)*, then* $p_i(t) \in int(S^s_i)$, $\forall t \geq 0$.

Proof. If

$$\|S_{R_i}(t) - S^n_{R_i}(t)\| \leq \|S^w_{R_i}(t) - S^n_{R_i}(t)\|, \tag{45}$$

then due to the nominal safety behavior metric $S^n_{R_i}(t) = 1$, one has $S^w_{R_i}(t) \leq S_{R_i}(t)$ as $0 \leq S^w_{R_i}(t), S_{R_i}(t) \leq 1$. From $S_{R_i}(t)$ and $S^w_{R_i}(t)$ in (35) and (39), one has

$$\Gamma_i(p_i(t), p_j(t), u_i(t))) \leq \Gamma^w_i(p_i(t), p_j(t), u_i(t))), \tag{46}$$

which using (36) and (40) can be written as

$$\max\{g_{ij}(p_i, p_j),\ g_{io_j}(p_i)\} \leq 1 - \max\{g^w_{ij}(p_i, p_j),\ g^w_{io_j}(p_i)\}, \tag{47}$$

$\forall j \in \mathscr{N}_i$ and $\forall o_j \in \mathscr{O}_i$, and it further becomes

$$\max\{g_{ij}(p_i, p_j) + g^w_{ij}(p_i, p_j),\ g_{io_j}(p_i) + g^w_{io_j}(p_i)\} \leq 1, \tag{48}$$

as $g_{ij}(p_i, p_j)$, $g^w_{ij}(p_i, p_j)$, $g_{io_j}(p_i)$ and $g^w_{io_j}(p_i) \in \mathbb{R}_{>0}$ according to (37)-(38) and (41)-(42). Then, based on $g_{ij}(p_i, p_j)$, $g^w_{ij}(p_i, p_j)$, $g_{io_j}(p_i)$ and $g^w_{io_j}(p_i)$, one can write (48) as

$$\max\{\frac{d + \eta(p_{ij}(t))}{\|p_i(t) - p_j(t)\|},\ \frac{r_{o_j} + \eta(p_{io_j}(t))}{\|p_i(t) - c_{o_j}\|}\} \leq 1. \tag{49}$$

Now, from (49), one can infer that

$$(d + \eta(p_{ij}(t))) - \|p_i(t) - p_j(t)\| \leq 0,\ \forall j \in \mathscr{N}_i, \tag{50}$$

and

$$(r_{o_j} + \eta(p_{io_j}(t))) - \|p_i(t) - c_{o_j}\| \leq 0,\ \forall o_j \in \mathscr{O}_i, \tag{51}$$

as $\eta(p_{ij}(t))$ and $\eta(p_{io_j}(t)) \in \mathbb{R}_{>0}$ based on its definition in (32) and (34), respectively. Then, based on (7), it implies $h^S_i(\vec{p}) < 0$ and thus, $p_i(t) \in int(S^s_i)$, $\forall t \geq 0$. ■

Remark 2. *Note that based on Theorem 3, we know that if the safety behavior metric value is below a designed threshold, then one can always ensure the operation of an agent inside the safe set* S^s_i*. However, if the safety behavior metric violates the designed threshold, then one needs to determine why the safety of the agent is jeopardized (due to the agent's own behavior or neighbor's behavior). Thus, one needs a metric to differentiate adversarial agent trying to achieve the class 1 objective in Definition 2. Therefore, to accomplish proactive adversary identification in terms of critical time and critical zone, we also need to monitor goal reaching behavior to differentiate adversarial agent which tries to violate safety constraints. Based on the provided reasoning, we design the proactive adversary detection mechanism in Theorem 4.*

Goal reaching behavior metric: To monitor the agent's performance in terms of reaching toward the goal location, we define the **goal reaching behavior metric** as

$$G_{R_i}(t) = \exp(-(\lambda_i(p_i(t), p_i(0), G_i))^{n_c}),\ \forall t \geq 0, \tag{52}$$

with

$$\lambda_i(p_i(t), p_i(0), G_i) = \frac{\|p_i(t) - G_i\|^2}{\|p_i(0) - G_i\|^2} = \zeta V^g_i(p_i(t)),\ \forall t \geq 0, \tag{53}$$

where $\zeta = \frac{1}{\|p_i(0) - G_i\|^2}$, $V^g_i(p_i(t))$ is defined in (9) and $p_i(0)$ represents initial position of the agent i. Note that $0 \leq \lambda_i(p_i(t), p_i(0), G_i) \leq 1$ under normal or desired behavior of the agents.

In the following proposition, we analyze how the goal reaching metric defined in (53) changes for an agent when it starts deviating from the desired goal reaching behavior. As explained in Remark 3, we need the result of Proposition 2 along with behavior metrics in Theorem 4 to differentiate the adversarial agent (either agent i or its neighboring agent j) that tries to violate safety constraints.

Proposition 2. *Consider the agent dynamics* (1) *along with the QP based control* (14)*. For* $\lambda_i(p_i(t))$ *defined in* (53)*, if an agent i satisfies*

$$(L_{f^p_i}\lambda_i(p_i(t)) + L_{g^p_i}\lambda_i(p_i(t))u_i(t) \geq 0) \wedge (\lambda_i(p_i(t)) \neq 0), \tag{54}$$

for all $t \in [t_0, t_0 + t_m]$ *with some time* $t_0 \geq 0$, $t_m > 0$*, then the agent i deviates from desired goal reaching behavior.*

Proof. If condition provided in (54) holds true at some time $t \geq 0$, then under (53), one has

$$(L_{f^p_i}V^g_i(p_i(t)) + L_{g^p_i}V^g_i(p_i(t))u_i(t) \geq 0) \wedge (V^g_i(p_i(t))) \neq 0), \tag{55}$$

as $\zeta > 0$ in (53). That means the agent i has not reached the goal point and it starts deviating from the desired goal reaching behavior over time $t \in [t_0, t_0 + t_m]$ as $L_{f^p_i}V^g_i(p_i(t)) + L_{g^p_i}V^g_i(p_i(t))u_i(t) \geq 0$. ■

In the following theorem, we present the proactive adversary detection mechanism based on designed behavior metrics along with the presented critical zone and critical time in Section IV. B.

Theorem 4. *Consider the agent dynamics* (1) *along with* $S_{R_i}(t)$ *and* $\gamma^S_i(t)$ *defined in* (35) *and* (43)*, respectively. If an agent i satisfies*

$$\|S_{R_i}(t) - S^n_{R_i}(t)\| > \gamma^S_i(t),\quad \forall t \in [t_0, t_0 + nT_s], \tag{56}$$

with some time $t_0 \geq 0$ then,

1) *the position trajectories of agent i remain in the safe set for a horizon $(n-1)T_s$ ahead, i.e., $p_i(t) \in int(S_i^s)$, $\forall t \in [t_0, t_0 + (n-1)T_s]$.*
2) *If in addition $L_{f_k^p}\lambda_k(p_k(t)) + L_{g_k^p}\lambda_k(p_k(t))u_k(t) \geq 0$, $\forall t \in [t_0, t_0+(n-1)T_s]$, where $k \in \{i, j\}$, $j \in \mathscr{N}_i$, then the agent i detects any agent k (i.e., either itself i or any neighbor agent j) as adversarial at time $t = t_0 + (n-1)T_s$.*

Proof. The equation (56) under (35) and (39) with some mathematical simplification, reads

$$\Gamma_i(p_i(t), p_j(t), u_i(t))) > \Gamma_i^w(p_i(t), p_j(t), u_i(t))), \tag{57}$$

which based on (36) and (40), becomes

$$\max\{g_{ij}(p_i, p_j), g_{io_j}(p_i)\} > 1 - \max\{g_{ij}^w(p_i, p_j), g_{io_j}^w(p_i)\}, \tag{58}$$

$\forall j \in \mathscr{N}_i$ and $\forall o_j \in \mathscr{O}_i$ and thus, one has

$$\max\{g_{ij}(p_i, p_j) + g_{ij}^w(p_i, p_j), g_{io_j}(p_i) + g_{io_j}^w(p_i)\} > 1, \tag{59}$$

as $g_{ij}(p_i, p_j)$, $g_{ij}^w(p_i, p_j)$, $g_{io_j}(p_i)$ and $g_{io_j}^w(p_i) \in \mathbb{R}_{>0}$ according to (37)-(38) and (41)-(42). Then, based on $g_{ij}(p_i, p_j)$, $g_{ij}^w(p_i, p_j)$, $g_{io_j}(p_i)$ and $g_{io_j}^w(p_i)$, one can write (59) as

$$\max\{\frac{d + \eta(p_{ij}(t))}{\|p_i(t) - p_j(t)\|}, \frac{r_{o_j} + \eta(p_{io_j}(t))}{\|p_i(t) - c_{o_j}\|}\} > 1. \tag{60}$$

Thus, one can conclude at least one of the following conditions do not hold true, $\forall t \in [t_0, t_0 + (n-1)T_s]$,

$$\begin{cases} (d + \eta(p_{ij}(t))) - \|p_i(t) - p_j(t)\| \nleq 0, \ \forall j \in \mathscr{N}_i, \\ (r_{o_j} + \eta(p_{io_j}(t))) - \|p_i(t) - c_{o_j}\| \nleq 0, \ \forall o_j \in \mathscr{O}_i, \end{cases} \tag{61}$$

with $\eta(p_{ij}(t))$ and $\eta(p_{io_j}(t)) \in \mathbb{R}_{>0}$ based on its definition in (32). However, the critical zone $\eta(p_{ij}(t))$ and $\eta(p_{io_j}(t))$ are designed such that in worst-case scenario the agent can reach unsafe boundary only if $t \geq t_0 + nT_s$, i.e., when the critical zones shrink to zero. Thus, following safety conditions are satisfied, $\forall t \in [t_0, t_0 + (n-1)T_s]$,

$$\begin{cases} d - \|p_i(t) - p_j(t)\| < 0, \ \forall j \in \mathscr{N}_i, \\ r_{o_j} - \|p_i(t) - c_{o_j}\| < 0, \ \forall o_j \in \mathscr{O}_i, \end{cases} \tag{62}$$

This implies $h_i^S(\vec{p}) < 0$ and thus, the safety for agent i is guaranteed, i.e., $p_i(t) \in int(S_i^s)$, $\forall t \in [t_0, t_0 + (n-1)T_s]$. This completes the proof of part 1.

Now we prove part 2 of the theorem. Based on the condition in (56) and from the proof of part 1, one can conclude that at least one of the following conditions does not hold true:

$$\begin{cases} (d + \eta(p_{ij}(t))) - \|p_i(t) - p_j(t)\| \nleq 0, \ \forall j \in \mathscr{N}_i, \\ (r_{o_j} + \eta(p_{io_j}(t))) - \|p_i(t) - c_{o_j}\| \nleq 0, \ \forall o_j \in \mathscr{O}_i, \end{cases} \tag{63}$$

for all $\forall t \in [t_0, t_0 + nT_s]$ with $\eta(p_{ij}(t))$ and $\eta(p_{io_j}(t)) \in \mathbb{R}_{>0}$. Note that if the conditions in (63) hold true for any $j \in \mathscr{N}_i$ or $o_j \in \mathscr{O}_i$, $\forall t \geq t_0 + nT_s$, then the safety conditions (62) do not satisfy $\forall j \in \mathscr{N}_i$ and $\forall o_j \in \mathscr{O}_i$ as the designed critical zone $\eta(p_{ij}(t))$ or $\eta(p_{io_j}(t))$ in (32) and (34) becomes zero at $t = t_0 + nT_s$. Thus, $p_i(t) \notin int(S_i^s)$, $\forall t \geq nT_s$. Therefore, one needs to detect agent i or j as adversarial at $t = t_0 + (n-1)T_s$ and act to ensure safety of the intact agent. However, the violation of inter-agent safety constraints in (62) can happen either due to the behavior of agent i or neighbor agent j. Based on Proposition 2, if $L_{f_k^p}\lambda_k(p_k(t)) + L_{g_k^p}\lambda_k(p_k(t))u_k(t) \geq 0$, $\forall t \in [t_0, t_0 + (n-1)T_s]$ is satisfied, (which shows that agent $k \in \{i, j\}$ is deviating from its desired goal-reaching behavior) along with the condition (56), (which shows that the trajectory of agent i is converging to the boundary of unsafe region at $t = t_0 + nT_s$), then the agent i detects any agent k (i.e., either itself i or any neighbor agent j) as adversarial at time $t = t_0 + (n-1)T_s$. This completes the proof. ■

Algorithm 1 Identification of adversarial agents in set $\mathscr{A}_s$.

```
Initialize with design constant $n \in \mathbb{Z}_{>1}$.
procedure $\forall i \in \mathscr{V}/\mathscr{V}_f$
    At each time $t_0$, compute the critical time $T_s = \min_{j \in \mathscr{N}_i}\{T_s^j\}$ where $T_s^j$ is defined in (20).
    For all time $t \in [t_0, t_0 + (n-1)T_s]$, compute the safety behavior metric $S_{R_i}(t)$ in (35) and $\lambda_i(p_i(t))$ in (53).
    Then, evaluate the conditions in (54) and (56). If both hold true $\forall t \in [t_0, t_0 + (n-1)T_s]$, then agent $i \in \mathscr{V}/\mathscr{V}_f$ is identified as adversarial at time $t = t_0 + (n-1)T_s$, i.e., $i \in \mathscr{A}_s$.
end procedure
```

Remark 3. *Note that to ensure safety for intact neighbors, one needs to proactively detect agent i as adversarial at $t = t_0 + (n-1)T_s$ if the condition in (56) and $L_f\lambda_i(p_i(t)) + L_g\lambda_i(p_i(t))u_i(t) \geq 0$ are satisfied, $\forall t \in [t_0, t_0 + (n-1)T_s]$, and takes proactive action in time window $[t_0 + (n-1)T_s, t_0 + nT_s]$. Otherwise, the adversarial agent i violates safety with its intact neighbor, i.e., $h_i^S(\vec{p}(t)) > 0$, $\forall t > t_0 + nT_s$.*

D. Task Behavior Metric and Proactive Adversary Detection

In this subsection, we first design a metric to capture the behavior of neighboring agents in formation and present the result for detection of adversarial agent among the set of agents $\mathscr{V}_f \subseteq \mathscr{V}$. Similar to $\lambda_i(p_i(t))$ in (53), the goal reaching behavior for a collaborative task among the set of agents $\mathscr{V}_f \subseteq \mathscr{V}$ in the form of formation can also be encoded in the following metric

$$\lambda_f(\bar{p}(t), \bar{p}(0), G_f) = \frac{\|\bar{p}(t) - G_f\|^2}{\|\bar{p}(0) - G_f\|^2} = \zeta_f \bar{V}_f^g(\bar{p}(t)), \ \forall t \geq 0, \tag{64}$$

where $\zeta_f = \frac{1}{\|\bar{p}(0) - G_f\|^2}$, $V_f^g(\bar{p}(t))$ is defined in (10) and $\bar{p}(0)$ denotes the initial centroid position of agents in formation with $\bar{p}(t) = \frac{1}{|\mathscr{V}_f|}\sum_{i \in \mathscr{V}_f} p_i(t)$. Note that $0 \leq \lambda_f(\bar{p}(t), \bar{p}(0), G_f) \leq 1$ under normal behavior of the agents in the formation.

Proposition 3. *Consider the agent dynamics (1) along with the QP control (14). For $\lambda_f(\bar{p}(t))$ defined in (64), if it holds that*

$$((L_{f_i^p}\lambda_f(\bar{p}(t)) + L_{g_i^p}\lambda_f(\bar{p}(t))u_i(t)) \geq 0) \vee (\lambda_f(\bar{p}(t))) > 1), \tag{65}$$

for all $t \in [t_k, t_k + nT_s]$, then atleast one of the agent among the set of agents $\mathscr{V}_f \subseteq \mathscr{V}$ is adversarial.

Proof. The result follows a similar argument as provided in the proof of Proposition 2. ■

Task Behavior Metric: Note that the result of Proposition 3 determines that at least one of the agent among the set of agents $\mathscr{V}_f$ might be adversarial. However, in order to accomplish goal reaching for a collective task among the set of intact agents $\mathscr{V}_f/\mathscr{A}$, one needs to detect the particular adversarial agents and reject their contribution in the formation. Thus, we design a **task behavior metric** for monitoring behavior of neighboring agents in formation and present the result for detection of adversarial agent among the set of agents $\mathscr{V}_f$. The task behavior metric is defined as:

$$F_{R_{ij}}(t) = \exp\left(-(\Theta_i(p_i(t), p_j(t)))^{n_c}\right),\ \forall t \geq 0, \tag{66}$$

where

$$\Theta_i(p_i(t), p_j(t)) = \left| \left\| p_i(t) - p_j(t) \right\| - c_{ij} \right| \tag{67}$$

with desired formation distance c_{ij} between agent i and its neighbor j.

From $\Theta_i(p_i(t), p_j(t))$ in (67), in the absence of attack, once desired group of agents reach formation, then $F_{R_{ij}}(t)$ will be always close to one during nominal operation and it goes to zero only if agent i or j is adversarial. Based on the result of task behavior metric, we determine confidence value $C_i(t)$ in (69) for each agent $i \in \mathscr{V}_f$ using Algorithm 2. Note that the confidence value $C_i(t)$ in (69) represents the degree of trustworthiness of each agent i about its own information. In particular, if an agent is adversarial, then $C_i(t) \to 0$, otherwise $C_i(t) \to 1$. Now, we define

$$E_i^{\mathscr{F}_i}(p_i(t), p_j(t)) = \left| \left\| p_i(t) - p_j(t) \right\| - c_{ij} \right| - \Theta_i^w(p_i(t), p_j(t)) \leq 0, \tag{68}$$

where $\Theta_i^w(p_i(t), p_j(t)) > 0$ represents the bound on the formation error for agents without considering any adversary.

Remark 4. *Note that the formation error bound $\Theta_i^w(p_i(t), p_j(t))$ in (68) can be designed based on exponential convergence for the formation control [25]. In particular, $\bar{\Theta}_i^w(p_i(t), p_j(t)) = k_1 e^{-k_2 t}(\left\| p_i(0) - p_j(0) \right\| - c_{ij})$ where $k_1 = \sqrt{\frac{c_2}{c_1}}$ and $k_2 = \frac{\lambda_r \rho}{2c_1}$ with design constant $c_1, c_2, \rho \in \mathbb{R}_+$ and λ_r as minimum singular value of rigidity matrix corresponding to desired formation [25]. Also, for the sake of brevity, one can select an arbitrary large enough scalar value $\Theta_i^w(p_i(t), p_j(t))$ such that $\Theta_i^w(p_i(t), p_j(t)) > \bar{\Theta}_i^w(p_i(t), p_j(t))$ for all time t.*

We also define the worst-case collective behavior as

$$F_{R_{ij}}^w(t) = \exp\left(-(\Theta_i^w(p_i(t), p_j(t)))^{n_c}\right),\ \forall t \geq 0. \tag{70}$$

Then, define the error between the worst-case and nominal formation behavior metrics as

$$\gamma_i^F(t) = \left\| F_{R_{ij}}^w(t) - F_{R_i}^n(t) \right\|, \tag{71}$$

with $F_{R_i}^n(t) = 1$ as the nominal task behavior metric. From $\Theta_i(p_i(t), p_j(t))$ in (67), in the absence of any adversary, once desired group of agents reach formation, then $F_{R_i}(t) = 1$ and that shows the nominal behavior of agents in the formation.

Based on the design of $\Theta_i^w(p_i(t), p_j(t))$, $E_i^{\mathscr{F}_i}(p_i(t), p_j(t)) \leq 0$ always holds true for each $i \in \mathscr{V}_f$ if $\mathscr{A} \cap \mathscr{V}_f = \emptyset$. Thus, in the following result, based on the designed task behavior metric, one can detect the adversarial agents among the set of collaborative agents $\mathscr{V}_f$.

Theorem 5. *Consider the agent dynamics (1) along with the QP based control (14). For the task behavior metric $F_{R_{ij}}(t)$ in (66) and $\gamma_i^F(t)$ in (71), if an agent $i \in \mathscr{V}_f$ satisfies*

$$\left\| F_{R_{ij}}(t) - F_{R_i}^n(t) \right\| > \gamma_i^F, \tag{72}$$

for more than $\frac{\mathscr{N}_i(t)}{2}$ neighbors at some time $t \geq 0$, then the agent i is detected as adversarial among the set of agents $\mathscr{V}_f$.

Algorithm 2 Determination of confidence value $C_i(t)$ and identification of Adversarial Agents belong to set $\mathscr{A}_f$

```
 1: procedure $\forall i \in \mathscr{V}_f$
 2:   for $i = 1 : |\mathscr{V}_f|$
 3:     initialize $index = [\,],\ \mathscr{A}_f = [\,]$
 4:     for $j = 1 : \mathscr{N}_i(t)$
 5:       From (66) and (70), evaluate $F_{R_{ij}}(t)$ and $F_{R_i}^w(t)$.
 6:       if $F_{R_{ij}}(t) > F_{R_i}^w(t)$
 7:         $index = [index\ \ j]$;
 8:       end if
 9:     end for
10:     Using $index$ from step 7, evaluate confidence
          $C_i(t) = \exp\left(-\left(\frac{2|index|}{\mathscr{N}_i(t)}\right)^{n_c}\right),\ \forall t \geq 0,\ n_c \in \mathbb{Z}_{>1}.$   (69)
11:     if $(2|index|) > \mathscr{N}_i(t)$
12:       $\mathscr{A}_f = [\mathscr{A}_f\ \ i]$;
13:     end if
14:   end for
15: end procedure
```

Proof. Based on the statement of theorem, if (72) holds, then from (71), with $F_{R_i}^n(t) = 1$ as the nominal task behavior metric, $F_{R_{ij}}(t) > F_{R_i}^w(t)$ also holds for more than $\frac{\mathscr{N}_i(t)}{2}$ neighbors at some time $t \geq 0$. Based on task behavior metrics $F_{R_{ij}}(t)$ and $F_{R_i}^w(t)$ in (66) and (70), one has formation error in (68) as

$$E_i^{\mathscr{F}_i}(p_i(t), p_j(t)) = \left| \left\| p_i(t) - p_j(t) \right\| - c_{ij} \right| - \Theta_i^w(p_i(t), p_j(t)) > 0, \tag{73}$$

for more than $\frac{\mathscr{N}_i(t)}{2}$ neighbors. That means agent violates the formation error bound with more than half of its neighbors and this only possible if agent i itself is adversarial because we assumed at max half of neighbors can be adversarial. Thus, the agent i is detected as an adversarial agent. ■

Remark 5. *Note that, based on the result of task behavior metric, we determine confidence value $C_i(t)$ in (69) for each agent $i \in \mathscr{V}_f$ using Algorithm 2. The confidence value $C_i(t)$ captures good or bad behavior of agent. In particular, based on $C_i(t)$ in (69), if an agent is adversarial, then $C_i(t) \to 0$, otherwise $C_i(t) \to 1$. In order to design confidence value $C_i(t)$, the condition of more than 50% of neighbors is leveraged to identify intact and adversarial agents.*

Note also that the set of adversarial agents be $\mathscr{A} = \mathscr{A}_s \cup \mathscr{A}_f$ can be determined based on Algorithm 1 and Algorithm 2. The set of adversarial agents $\mathscr{A}_s$ is computed based on the results presented for safety behavior metric in Theorem 4. Similarly, the set of adversarial agents $\mathscr{A}_f$ is evaluated based on the task behavior metric in Theorem 5. After identifying the set of adversarial agents $\mathscr{A}$ and we leverage the adversary detection results along with presented behavior metrics for the design of resilient QP in the next section.

V. Resilient Controller Design

This section presents the formulation of resilient quadratic program (QP) to compute a control input $u_i(t)$ for each agent i to solve Problem 1 in the presence of the set of adversarial agent $\mathscr{A}$. Let $\vec{z}_r = [z_1^T, z_2^T, \dots, z_{N-|\mathscr{A}|}^T]^T$ be a column vector with $z_i = [u_i, \delta_{i_1}, \delta_{i_2}, \delta_{i_3}, \delta_{i_4}, \{\delta_{ij}\}]^T \in \mathbb{R}^{m_i+4+N_i}$ with $N_i = |\mathscr{N}_i|$, $\forall i \in \mathscr{V}/\mathscr{A}$ as its elements. Consider the following optimization problem

$$\min_{u_i, \delta_{i_1}, \delta_{i_2}, \delta_{i_3}, \{\delta_{ij}\},\, i \in \mathscr{V}/\mathscr{A}} \quad \vec{z}_r^T H \vec{z}_r + F \vec{z}_r \tag{74a}$$

$$s.t. \quad A_i u_i \leq b_i \tag{74b}$$

$$L_{f_i^p} V_i^g + L_{g_i^p} V_i^g u_i \leq -\delta_{i_1} V_i^g, \ \forall i \in \mathscr{V}/\mathscr{V}_f \tag{74c}$$

$$L_{f_i^p} \bar{V}_f^r + L_{g_i^p} \bar{V}_f^r u_i(t) \leq -\delta_{i_2} \bar{V}_f^r \tag{74d}$$

$$L_{f_i^p} h_i^{o_j} + L_{g_i^p} h_i^{o_j} u_i \leq -\delta_{i_3} h_i^{o_j}, \ \forall o_j \in \mathscr{O}_i \tag{74e}$$

$$L_{f_i^p} \bar{h}_i^{\mathscr{F}_i} + L_{g_i^p} \bar{h}_i^{\mathscr{F}_i} u_i \leq -\delta_{i_4} \bar{h}_i^{\mathscr{F}_i}, \forall i \in \mathscr{V}_f \tag{74f}$$

$$L_{f_i^p} \bar{h}_{ij}^S + L_{g_i^p} \bar{h}_{ij}^S u_i + \pi(\bar{h}_{ij}^S) \leq -\delta_{ij} \bar{h}_{ij}^S, \ \forall j \in \mathscr{N}_i, \tag{74g}$$

where $H = diag\{H_i\}$ with $H_i = diag\{\{w_{u_l}^i\}, w_1^i, w_2^i, w_3^i, w_4^i, \{w_{n_p}^i\}\}$ denotes a diagonal matrix with positive weights $w_{u_l}^i, w_1^i, w_2^i, w_3^i, w_4^i, w_{n_p}^i > 0$ for each $p \in \mathscr{N}_i$, and similarly, $F = diag\{F_i\}$ with $F_i = [0_{m_i}^T \ q^i \ 0_{\mathscr{N}_i+3}]$ where $q^i > 0$ and $0_k \in \mathbb{R}^k$ denotes a column vector consisting of zeros. Moreover, based on the task behavior metric $F_{R_{ij}}(t)$ in (66), the function $\bar{h}_i^{\mathscr{F}_i}(t)$ in (74f) is defined in the resilient form as

$$\bar{h}_i^{F_i}(t) = \|p_i(t) - \hat{p}_i^*(t)\|, \tag{75}$$

with $\hat{p}_i^*(t) = \frac{1}{|\mathscr{N}_i^F(t)|} \sum_{j \in \mathscr{N}_i} F_{R_{ij}}(t)(p_j(t) + c_{ji})$ and $\left|\mathscr{N}_i^F(t)\right| = \sum_{j \in \mathscr{N}_i} F_{R_{ij}}(t)$. Similarly, the function $\bar{h}_{ij}^S$ in (74g) is defined as

$$\bar{h}_{ij}^S = h_{ij}^S + \frac{\eta(p_{ij})}{n}, \tag{76}$$

and

$$\pi(\bar{h}_{ij}^S(t)) = \begin{cases} L_{f_j^p} \bar{h}_{ij}^S + L_{g_j^p} \bar{h}_{ij}^S u_j(t), \ if \ j \notin \mathscr{A}, \\ L_{f_j^p} \bar{h}_{ij}^S + L_{g_j^p} \bar{h}_{ij}^S u_j^{max}(t), \ if \ j \in \mathscr{A}. \end{cases} \tag{77}$$

where $u_j^{max}(t)$ worst case adversarial control action defined in (17). Also, based on the confidence value $C_i(t)$ in (69), the resilient CLF for collaborative goal reaching is defined as

$$\bar{V}_f^r(t) = \left\|\bar{p}^r(t) - G_f\right\|^2 \to 0 \tag{78}$$

as $t \to \infty$ with G_f as goal point or goal region, $\bar{p}^r(t) = \frac{1}{\mathscr{N}^r(t)} \sum_{i \in \mathscr{V}_f} C_i(t) p_i(t)$ and $\mathscr{N}^r(t) = \sum_{i \in \mathscr{V}_f} C_i(t)$ for collaborative goal reaching with the set of agents $\mathscr{V}_f \subseteq \mathscr{V}$ in the form of formation.

Now, in the following theorem, we present the result that solves the Problem 1 with objectives A.1-A.4 for each agent $i \in \mathscr{V}/\mathscr{A}$. Let the solution of (74) be represented by $\vec{z}_r^* = [z_1^{*T}, z_2^{*T}, \dots, z_{N-|\mathscr{A}|}^{*T}]^T$ where $z_i^*(.) = [u_i^*(.), \delta_{i_1}^*(.), \delta_{i_2}^*(.), \delta_{i_3}^*(.), \delta_{i_4}^*(.), \{\delta_{ij}^*(.)\}]^T$.

Theorem 6. *Consider the agent dynamics* (1). *Then,*

1) *the resilient QP in* (74) *is feasible for each intact agent* $i \in \mathscr{V}/\mathscr{A}$.
2) *the resilient QP in* (74) *solves Problem 1 for each intact agent* $i \in \mathscr{V}/\mathscr{A}$.

Proof. Since, we have $\bar{V}_f^r > 0$ in (78) for all $\bar{p}_i^r(t) \notin G_f$ and $t \geq 0$. One can select $u_i = u_i^* \in \mathscr{U}_\text{i}$ and define

$$\delta_{i_2} = -\frac{L_{f_i} \bar{V}_f^r + L_{g_i} \bar{V}_f^r u_i^*(t)}{\bar{V}_f^r}, \tag{79}$$

and it can be explicitly defined for all $\bar{p}_i^r(t) \notin G_f$, such that (74d) satisfies the equality condition. We know that $\bar{h}_{ij}^S < 0$ for all $p_i(t) \in int(\bar{S}_i^s)$ and for all $t \geq 0$, where the set $\bar{S}_i^s$ is defined in (4). Then, one can chose $u_i = u_i^* \in \mathscr{U}_\text{i}$ and define

$$\delta_{ij} = -\frac{L_{f_i^p} \bar{h}_{ij}^S + L_{g_i^p} \bar{h}_{ij}^S u_i + \pi(\bar{h}_{ij}^S)}{\bar{h}_{ij}^S}, \tag{80}$$

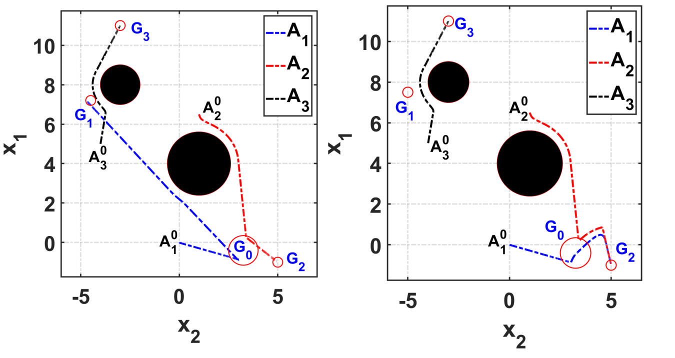
and it can be explicitly defined for all $p_i(t) \in int(\bar{S}_i^s)$, such that (74g) satisfies the equality condition. Similarly, we know that one can define slack parameters $\delta_{i_1}^*, \delta_{i_3}^*, \delta_{i_4}^*$ such that (74b) and (74d)-(74f) are satisfied with equality condition. Therefore, there exists $z_i^*(.) = [u_i^*(.), \delta_{i_1}^*(.), \delta_{i_2}^*(.), \delta_{i_3}^*(.), \delta_{i_4}^*(.), \{\delta_{ij}^*(.)\}]^T$ all constraints in (74) are satisfied and the resilient QP in (74) is feasible for each intact agent $i \in \mathscr{V}/\mathscr{A}$.

Now based on result of part 1, we present the prove of part 2. Since the resilient QP in (74) is feasible for all $i \in \mathscr{V}/\mathscr{A}$, thus there exist $z_i^*(.) = [u_i^*(.), \delta_{i_1}^*(.), \delta_{i_2}^*(.), \delta_{i_3}^*(.), \delta_{i_4}^*(.), \{\delta_{ij}^*(.)\}]^T$ which ensures (74b)-(74g) for all $t > 0$. Based on the confidence value $C_i(t)$ in (69), the resilient CLF $\bar{V}_f^r$ for collaborative goal reaching in (78) discards the adversarial agent contribution with centroid with $\bar{p}^r(t) = \frac{1}{\mathscr{N}^r(t)} \sum_{i \in \mathscr{V}_f} C_i(t) p_i(t)$ and $\mathscr{N}^r(t) = \sum_{i \in \mathscr{V}_f} C_i(t)$ as based on Algorithm 2 $C_i(t) \to 0$ for the adversarial agents. Since there exist $z_i^*(.) = [u_i^*(.), \delta_{i_1}^*(.), \delta_{i_2}^*(.), \delta_{i_3}^*(.), \delta_{i_4}^*(.), \{\delta_{ij}^*(.)\}]^T$ for all $i \in \mathscr{V}/\mathscr{A}$ which satisfaction of constraint (74d), this means centroid of intact agents among the set of agents $\mathscr{V}_f$ exponentially reaches the goal point. With similar argument constraints in (74b)-(74c) and (74e)-(74g) are satisfied and thus the resilient QP in (74) solves Problem 1 for each intact agent $i \in \mathscr{V}/\mathscr{A}$. This completes the proof. ■

Remark 6. *Note that based on the result in Theorem 6, the feasibility of the designed resilient QP in (74) depends on the optimization parameters* $\delta_{i_1}^*(.), \delta_{i_2}^*(.), \delta_{i_3}^*(.), \delta_{i_4}^*(.), \delta_{ij}^*(.)$. *These optimization parameters are similar to feasibility parameters presented in [26]–[28] to solve the conflicts among constraints and to ensure the feasibility of the CBF-CLF based QP's. Interested readers can refer to [26], [28] for more details on feasibility analysis.*

VI. Numerical Case Studies

In this section, we present two case studies to demonstrate the efficacy of the presented theoretical contributions. In the first case, we consider a multi-agent problem with objectives to visit some regions (goal reaching) while maintaining safety constraints (i.e., inter-agent and agent-to-obstacle safety), despite the presence of an adversarial agent. In the second case, we consider the multi-agent formation problem under some desired specifications, where the aim is to maintain formation among the set of collaborative agents and visit some regions



(a)

(b)

Figure 2: Agents behaviors for desired specification: (a) without any adversary. (b) when Agent 1 perform adversarial chasing toward Agent 2 for $t > 800s$.

while maintaining safety constraints with collaborative goal reaching even in the presence of adversarial agents.

1) Case 1: We consider a network of three agents with the following linearized unicycle dynamics

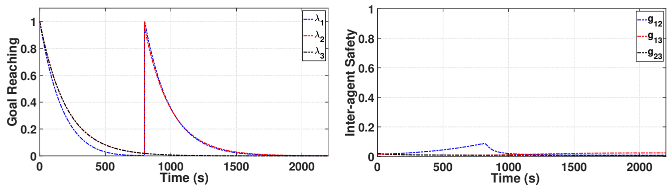
$$\begin{bmatrix} \dot{y}_1^i \\ \dot{y}_2^i \\ \dot{\theta}^i \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-\sin(\theta^i)}{b} & \frac{\cos(\theta^i)}{b} \end{bmatrix} \begin{bmatrix} u_1^i \\ u_2^i \end{bmatrix}, \ \forall i \in \{1,2,3\}$$

where $y_1^i = x^i + b\cos(\theta^i)$ and $y_2^i = y^i + b\sin(\theta^i)$ with $\begin{bmatrix} x^i & y^i \end{bmatrix}^T$ and θ^i as the position vector and orientation of agent i. For the linearized unicycle model, the control input transformation is given by

$$\begin{bmatrix} v^i \\ w^i \end{bmatrix} = \begin{bmatrix} \cos(\theta^i) & -b\sin(\theta^i) \\ \sin(\theta^i) & b\cos(\theta^i) \end{bmatrix}^{-1} \begin{bmatrix} u_1^i \\ u_2^i \end{bmatrix}, b > 0$$

Under normal operation, the multi-agent has following desired specifications or objectives $\phi = \Diamond_{[0,800]}({\Upsilon_1}^{G_0} \wedge {\Upsilon_2}^{G_0}) \wedge \Diamond_{[800,2200]}({\Upsilon_1}^{G_1} \wedge {\Upsilon_2}^{G_2}) \wedge \Diamond_{[0,2200]}({\Upsilon_3}^{G_3}) \wedge \Box_{[0,2200]}\phi_s$ with ${\Upsilon_i}^{G_r} = \|p_i - G_r\| < \delta_r, \ \forall r \in \{0,1,2,3\}$ and $\phi_s = (\|p_i - p_j\| > 0.1) \wedge (\|p_i - c_{o_j}\| > 0.4), \forall o_j \in \{o_1, o_2\}$ as desired goal reaching and safety specification for agent i, respectively. $\delta_r = 0.25, \ \forall r \in \{1,2,3\}$ and $\delta_0 = 0.75$ for goal regions. In particular, the objective for Agents 1 and 2 is to eventually reach goal location G_0 and perform some task between time duration $t \in [0,800]$, then Agent 1 and Agent 2 are supposed to reach their desired goal location G_1 and G_2 over time duration $t \in (800,2200]$ while maintaining inter-agent and agent-obstacle constraints (black eclipse in Figure 2 denotes obstacle in the environment). Similarly, Agent 3 has to reach its desired goal location G_3 over time duration $t \in [0,2200]$. Figure 2a shows the normal agent's behavior for the desired specification ϕ. Based on normalized CLF in (53) and inter-agent safety metric in (37), the goal-reaching and inter-agent safety behavior of agents are shown in Figures 3a and 3b, respectively. One can see that in absence of any adversarial agent, all intact agent follows desired behavior under control action obtained from nominal QP (14). In Figure 2a, A_1^0, A_2^0 and A_3^0 denote the initial positions of respective agents.

Now, we consider Agent 1 as adversarial and it performs adversarial chasing after $t > 800$ with aim to achieve $\|x_1(t) - x_2(t)\| \to 0$ in some finite time. Figure 2b shows the agent's behavior under adversarial chasing and violation of desired specifications ϕ. We can see in Figure 4 how the goal reaching behavior for Agent 1 starts growing after $t = 800s$ due to adversarial chasing behavior. Similarly, Figure 5 shows



(a)

(b)

Figure 3: (a) Normalized CLF for goal reaching behavior without adversary. (b) Inter-agent safety behavior without adversary.

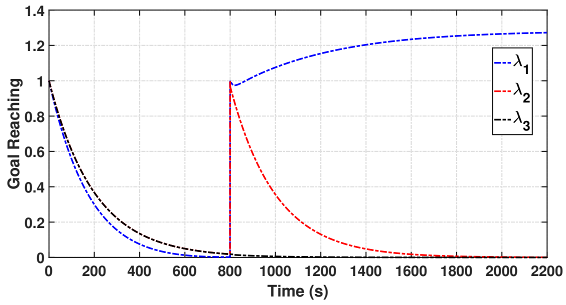
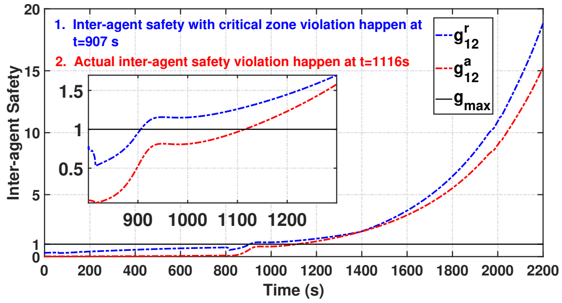


Figure 4: Normalized CLF for goal reaching behavior under adversarial chasing.


1. Inter-agent safety with critical zone violation happen at t=907 s

2. Actual inter-agent safety violation happen at t=1116s

Figure 5: Inter-agent safety behavior under adversarial chasing.

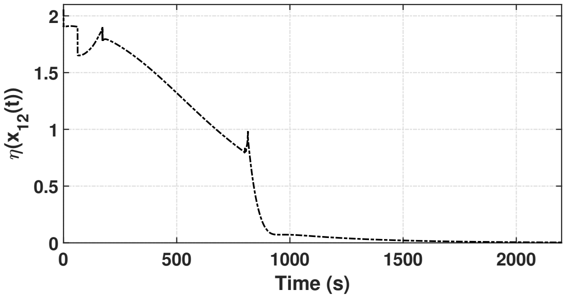


Figure 6: Critical zone between Agents 1 and 2.

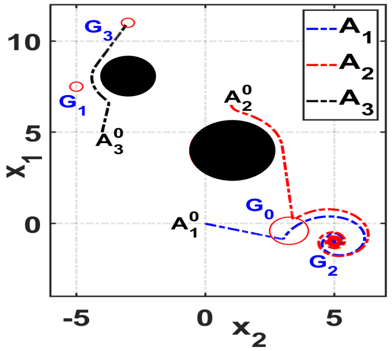


Figure 7: Agents behavior under resilient QP in (74): when Agent 1 perform adversarial chasing toward Agent 2 for $t > 800s$.

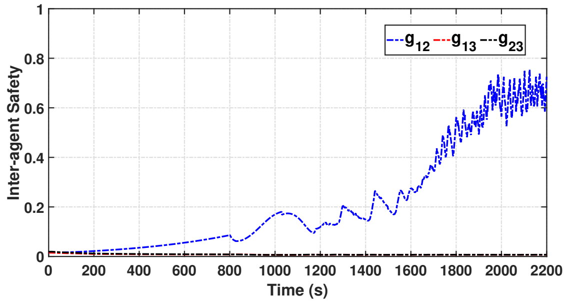


Figure 8: Inter-agent safety behavior under resilient QP in (74).

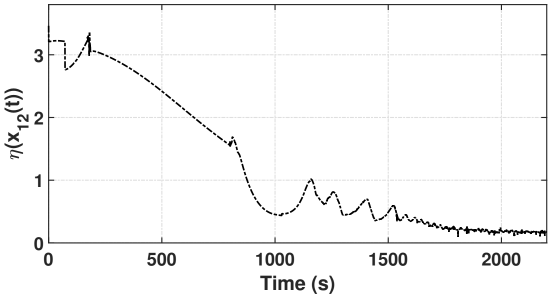


Figure 9: Critical zone between Agents 1 and 2 under resilient QP in (74).

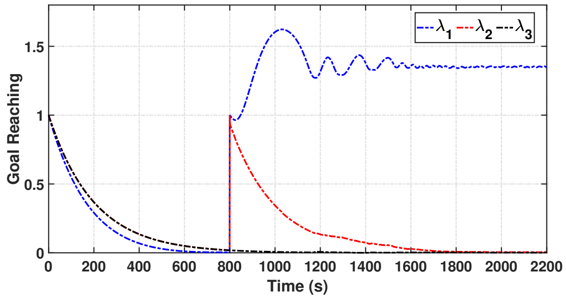


Figure 10: Normalized resilient CLF for goal reaching behavior under adversarial chasing with resilient QP in (74) .

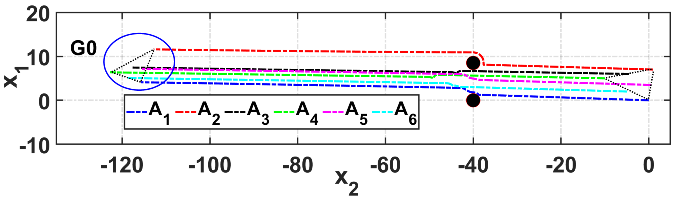


Figure 11: Collaborative goal reaching behavior of intact agents in formation without any adversarial agent.

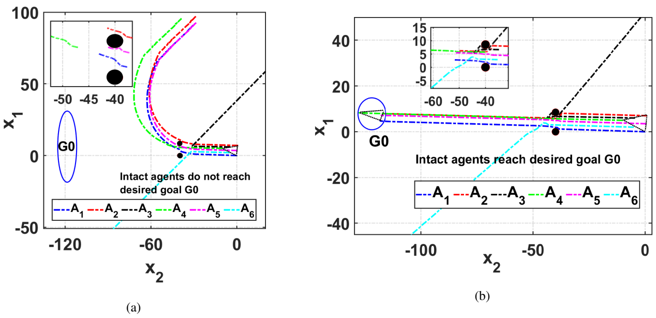


Figure 12: (a) Collaborative goal reaching behavior in the presence of adversarial Agent 3 and 6.. (b) Resilient collaborative goal reaching behavior in the presence of adversarial Agent 3 and 6.

that the inter-agent safety violation happens at $t = 1116s$ as the value of the safety behavior metric exceeds one, which implies that Agent 1 violates the safety constraint and hits the intact agent, i.e., Agent 2. Under adversarial chasing by Agent 1, one can see in Figure 6 how the critical zone shrinks as the adversarial agent reaches close to intact Agent 2, and eventually goes to zero. Figure 5 shows a proactive adversary detection happens at $t = 907s$ by leveraging the concept of the critical zone for an inter-agent safety violation and it follows the result presented in Theorem 4.

Now, based on the results of presented detection mechanism, we validate the efficacy of designed resilient QP in (74). Figure 7 illustrates the agents' behavior under adversarial chasing with resilient QP in (74); one can see that Agent 1 keeps trying to hit Agent 2, but Agent 2 escapes from adversarial chasing based on the designed resilient inter-agent safety constraint in (74g). Similarly, Figure 8 shows that the resilient QP in (74) guarantees all-time inter-agent safety for intact agents even under adversarial chasing, as the safety metric is always less then one. Also, one can see that the critical zone in Figure 9 does not shrink to zero as adversarial agent reaches close to but does not hit intact Agent 2. Figure 10 illustrates the goal reaching behavior for intact agent, i.e., Agents 2 and 3 eventually goes to zero and they reach their desired goal position G_2 and G_3. However, for Agent 1 goal reaching behavior grows after $t = 800s$ due to adversarial chasing behavior and it never reaches the specified goal location G_1.

2) Case 2: In the second case, we consider a multi-agent formation problem under some desired specifications, where the aim is to maintain formation among the set of collaborative agents and visit some regions, while maintaining safety constraints with collaborative goal reaching even in the presence of adversarial agents. In particular, we consider six agents with linearized unicycle dynamics and the desired goal location of formation centriod $G0 = [-120 \;\; 7]^T$. Figure 11 shows the collaborative goal reaching for the agents without any adversarial agents, i.e., $\mathscr{A}_f = \emptyset$. One can see in Figure 11 how agents maintain safety and reach desired goal location $G0$ over the desired time duration $t \in [0, 1000]$. Then, we consider the same scenario in the presence of multiple adversarial agents, i.e., Agents 3 and 6 act as adversarial agents for all $t > 400$ (both agents belong to class 2 type of adversarial agent as defined in Definition 2). It is shown in Figure 12a how adversarial agents mislead the collaborative goal reaching behavior and thus, intact agents do not reach the desired goal location $G0$, collectively over the time interval $t \in [0, 1000]$. Then, based on presented Algorithm 2 and Theorem 5, we detect the set of adversarial agents $\mathscr{A}_f$ and mitigate their effects in collaborative goal reaching. The Figure 12b illustrates that even in the presence of multiple adversarial agents, intact agents achieve the desired collaborative goal reaching behavior and reach the goal location $G0$ over the desired time duration $t \in [0, 1000]$.

VII. Conclusions and Future Direction

In this paper, we presented the proactive adversary detection mechanism and then designed a resilient control framework for multi-agent systems. In particular, first we analyzed agent's behaviors based on designed behavior metrics, and then designed proactive adversary detection mechanism based on the notion of the critical region for the system operation. The presented detection mechanism identified adversarial agents while ensuring all-time safety for normally behaving agents in the presence of adversarial agents. By leveraging the presented results for behavior analysis and adversary detection, we designed a resilient QP-based controller for multi-agent systems with desired safety and goal reaching constraints for intact agents, even in the presence of the adversarial agent. Finally, two case studies are presented to illustrate the efficacy of the presented theoretical contributions.

A possible direction for future work is to explore the presented framework for resilience to more sophisticated adversaries with actual adversarial actions instead of worst-case actions.

References

[1] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *2019 18th European Control Conference (ECC)*, pp. 3420–3431, IEEE, 2019.

[2] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3861–3876, 2016.

[3] M. Srinivasan, S. Coogan, and M. Egerstedt, "Control of multi-agent systems with finite time control barrier certificates and temporal logic," in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 1991–1996, IEEE, 2018.

[4] K. Garg and D. Panagou, "Control-lyapunov and control-barrier functions based quadratic program for spatio-temporal specifications," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 1422–1429, IEEE, 2019.

[5] L. Lindemann and D. V. Dimarogonas, "Barrier function based collaborative control of multiple robots under signal temporal logic tasks," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 4, pp. 1916–1928, 2020.

[6] W. Xiao and C. Belta, "Control barrier functions for systems with high relative degree," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 474–479, IEEE, 2019.

[7] L. Wang, A. D. Ames, and M. Egerstedt, "Safety barrier certificates for collisions-free multirobot systems," *IEEE Transactions on Robotics*, vol. 33, no. 3, pp. 661–674, 2017.

[8] P. Glotfelter, J. Cortés, and M. Egerstedt, "Nonsmooth barrier functions with applications to multi-robot systems," *IEEE control systems letters*, vol. 1, no. 2, pp. 310–315, 2017.

[9] L. Guerrero-Bonilla and V. Kumar, "Realization of r-robust formations in the plane using control barrier functions," *IEEE Control Systems Letters*, vol. 4, no. 2, pp. 343–348, 2019.

[10] L. Wang, A. D. Ames, and M. Egerstedt, "Safe certificate-based maneuvers for teams of quadrotors using differential flatness," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3293–3298, IEEE, 2017.

[11] D. Pickem, P. Glotfelter, L. Wang, M. Mote, A. Ames, E. Feron, and M. Egerstedt, "The robotarium: A remotely accessible swarm robotics research testbed," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1699–1706, IEEE, 2017.

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

[13] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Transactions on Automatic Control*, vol. 56, pp. 1495–1508, July 2011.

[14] J. Usevitch and D. Panagou, "Resilient leader-follower consensus to arbitrary reference values in time-varying graphs," *IEEE Transactions on Automatic Control*, vol. 65, no. 4, pp. 1755–1762, 2019.

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

[16] L. Zhou, V. Tzoumas, G. J. Pappas, and P. Tokekar, "Resilient active target tracking with multiple robots," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 129–136, 2018.

[17] A. Mustafa and H. Modares, "Attack analysis and resilient control design for discrete-time distributed multi-agent systems," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 369–376, 2019.

[18] M. Pirani, E. Nekouei, S. M. Dibaji, H. Sandberg, and K. H. Johansson, "Design of attack-resilient consensus dynamics: a game-theoretic approach," in *2019 18th European Control Conference (ECC)*, pp. 2227–2232, IEEE, 2019.

[19] J. Usevitch and D. Panagou, "Adversarial resilience for sampled-data systems under high-relative-degree safety constraints," *arXiv preprint arXiv:2102.05014*, 2021.

[20] A. Clark, Z. Li, and H. Zhang, "Control barrier functions for safe cps under sensor faults and attacks," in *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 796–803, IEEE, 2020.

[21] F. Blanchini, "Set invariance in control," *Automatica*, vol. 35, no. 11, pp. 1747–1767, 1999.

[22] S. Prajna, A. Jadbabaie, and G. J. Pappas, "A framework for worst-case and stochastic safety verification using barrier certificates," *IEEE Transactions on Automatic Control*, vol. 52, no. 8, pp. 1415–1428, 2007.

[23] A. J. Taylor, A. Singletary, Y. Yue, and A. D. Ames, "A control barrier perspective on episodic learning via projection-to-state safety," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 1019–1024, 2020.

[24] k. Hassan, *Nonlinear systems*, vol. 3. New Jersey, USA: Prentice hall Upper Saddle River, 2002.

[25] Z. Sun, S. Mou, B. D. Anderson, and M. Cao, "Exponential stability for formation control systems with generalized controllers: A unified approach," *Systems & Control Letters*, vol. 93, pp. 50–57, 2016.

[26] J. Zeng, B. Zhang, Z. Li, and K. Sreenath, "Safety-critical control using optimal-decay control barrier function with guaranteed point-wise feasibility," *arXiv preprint arXiv:2103.12375*, 2021.

[27] M. J. Powell and A. D. Ames, "Towards real-time parameter optimization for feasible nonlinear control with applications to robot locomotion," in *2016 American Control Conference (ACC)*, pp. 3922–3927, IEEE, 2016.

[28] W. Xiao, C. Belta, and C. G. Cassandras, "Sufficient conditions for feasibility of optimal control problems using control barrier functions," *arXiv preprint arXiv:2011.08248*, 2020.

Aquib Mustafa (S'17) received the Master's degree from the Indian Institute of Technology Kanpur, Kanpur, India, in 2016. and the PhD degree from the Michigan State University, East Lansing, Michigan, USA, in 2020. He is currently working as a Postdoctoral Research Fellow with the Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI, USA. His research interests include the Resilient control, Safety-critical systems, Reinforcement learning, and Multi-agent systems.

Dimitra Panagou (Senior Member, IEEE) received the Diploma and Ph.D. degrees in mechanical engineering from the National Technical University of Athens, Athens, Greece, in 2006 and 2012, respectively.

She is currently an Associate Professor with the Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI, USA. Prior to joining the University of Michigan, she was a Postdoctoral Research Associate with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Champaign, IL, USA, a Visiting Research Scholar with the GRASP Lab, University of Pennsylvania, Philadelphia, PA, USA, and a Visiting Research Scholar with the Mechanical Engineering Department, University of Delaware, Newark, DE, USA. Her research interests include the fields of multiagent planning, control and estimation, with applications in safe and resilient robotic networks, autonomous multivehicle systems, and human–robot interaction.

Dr. Panagou was a recipient of the NASA 2016 Early Career Faculty Award, the AFOSR 2017 Young Investigator Award, and the NSF CAREER Award in 2020. She is a senior member of the AIAA.