Early Mission Degradation Detection in Multi-Agent Autonomous Systems

Bijan Mehralizadeh
Department of Computer Science
George Washington University
Washington, DC, USA

Tejaaswini Narendran
Department of Computer Science
George Washington University
Washington, DC, USA

***Abstract*—Multi-agent autonomous systems exhibit strong operational resilience due to redundancy and inter-agent coordination. However, the same redundancy obscures early indicators of failure, as missions often degrade gradually rather than failing catastrophically. In this work, we address the problem of early mission degradation detection under strict communication constraints. We formalize a grid-based multi-agent exploration task with enforced connectivity and introduce a temporal graph-based representation of mission evolution. Using Weisfeiler–Lehman (WL) Graph2Vec embeddings, we encode full missions into fixed-dimensional latent representations. We show that early deviations from nominal behavior emerge as smooth, separable manifolds in the embedding space well before observable mission failure. Our results demonstrate that structural graph embeddings provide a lightweight, scalable, and interpretable early-warning mechanism for resilient autonomous systems.**

***Index Terms*—Autonomous Systems, Swarm Robotics, Resiliency, Mission Degradation, Graph2Vec, Weisfeiler–Lehman, Anomaly Detection**

I. Introduction

Autonomous systems integrate sensing, sensor fusion, decision-making, planning, control, and learning-based components to execute missions with minimal or no human supervision. These systems are deployed across centralized, decentralized, hybrid, and swarm-based architectures and operate in safety-critical domains such as disaster response, environmental monitoring, surveillance, logistics, and defense.

Resiliency differs fundamentally from reliability. Reliability measures failure-free operation under nominal conditions, whereas resiliency captures the intrinsic ability of a system to withstand disruptions, recover from faults, adapt to unexpected conditions, and degrade gracefully when recovery is not possible. In multi-agent and swarm systems, individual agent failures rarely cause instantaneous mission collapse. Instead, failures propagate gradually through communication, motion coordination, and sensing distortions, producing progressive mission degradation.

This work focuses on *mid-mission degradation detection*: the identification of partial failures that distort collective behavior well before traditional binary mission failure indicators become observable.

II. Motivation

Autonomous systems consist of tightly coupled submodules including perception, sensor fusion, control, planning, learning, and firmware. Each module is vulnerable to faults, noise accumulation, and adversarial manipulation such as sensor spoofing, signal jamming, and firmware compromise. In large-scale swarms, redundancy masks these faults: the system may continue to operate while mission quality silently deteriorates.

Physical experiments in our laboratory demonstrate this phenomenon. A single quadrotor subjected to adversarial perturbation deviates from a waypoint-following task and eventually crashes. In swarm settings, analogous perturbations may not result in immediate crashes but instead distort spatial coverage, coordination patterns, and communication topology. This creates a failure mode that is *structural rather than catastrophic*. These observations motivate early mission health monitoring rather than post-mission failure analysis.

III. Problem Formulation

We formalize early mission degradation detection as a spatio-temporal inference problem over a dynamically evolving multi-agent communication graph. The key objective is to identify latent structural disruptions in collective behavior before explicit mission failure becomes observable through connectivity loss or task incompletion.

A. Agent and Environment Model

Let the agent set be defined as

$$\mathcal{V} = \{1, 2, \ldots, i\}, \quad i = 4, \tag{1}$$

where each agent is homogeneous in sensing, communication, and motion capabilities.

The environment is represented as a finite discrete occupancy grid

$$\mathcal{X} = \{0, \ldots, X\} \times \{0, \ldots, Y\}, \tag{2}$$

with all cells assumed to be obstacle-free. This abstraction isolates coordination and coverage dynamics while eliminating confounding effects from collision avoidance and continuous control.

The position of agent v at discrete time t is given by

$$S_v(t) = (i_v(t), j_v(t)) \in \mathcal{X}. \tag{3}$$

Each agent selects control actions from the finite action set

$$\mathcal{A} = \{\text{UP}, \text{DOWN}, \text{LEFT}, \text{RIGHT}, \text{STAY}\}, \tag{4}$$

which induces a controlled Markov process over joint agent configurations.

B. Inter-Agent Communication Model

Inter-agent proximity is measured using the Chebyshev distance

$$d_\infty((i_1, j_1), (i_2, j_2)) = \max\big(|i_1 - i_2|, |j_1 - j_2|\big), \quad (5)$$

which naturally models square-grid communication neighborhoods.

The induced time-varying communication graph is defined as

$$G_t = (\mathcal{V}, \mathcal{E}_t), \quad (6)$$

where

$$(u, v) \in \mathcal{E}_t \iff d_\infty(p_u(t), p_v(t)) \le r_c. \quad (7)$$

Connectivity is enforced as a *hard mission constraint*:

$$G_t \text{ must remain connected for all } t. \quad (8)$$

Violation of this constraint constitutes formal mission failure. Importantly, our objective is to detect degradation *prior* to this explicit failure condition.

C. Coverage and Mission Objective

Each agent maintains a local sensing neighborhood

$$\mathcal{S}_v(t) = \{x \in \mathcal{X} \mid d_\infty(x, S_v(t)) \le r_s\}, \quad (9)$$

with $r_s \le r_c$.

The total explored region over a mission of horizon T is defined as

$$\mathcal{C}_T = \bigcup_{t=1}^{T} \bigcup_{v \in \mathcal{V}} \mathcal{S}_v(t). \quad (10)$$

The mission optimization objective is

$$\max |\mathcal{C}_T| \quad \text{subject to } G_t \text{ remaining connected for all } t. \quad (11)$$

This formalization couples spatial exploration efficiency with global communication integrity. Degradation therefore manifests as deviations in coverage efficiency, coordination structure, and graph topology prior to explicit disconnection.

IV. Methods

Our approach consists of four major stages: (i) large-scale mission simulation and attack injection, (ii) temporal mission graph construction, (iii) Weisfeiler–Lehman relabeling and Graph2Vec embedding, and (iv) early-warning classification. A conceptual overview of this pipeline is shown implicitly through the successive model transformations defined below.

A. Mission Simulation and Policy Families

We implement a custom multi-agent simulator using the Gym and PettingZoo environments. The simulator operates on a 32×32 grid with four homogeneous agents and a maximum mission horizon of $T_{\max} = 500$ steps. At each timestep, all agents act synchronously.

To avoid representation bias to a single policy, multiple families of nominal behaviors are generated:

- Structured lawn-mower sweeping,
- Pure random walks,
- ϵ-greedy exploration with structured drift,
- Frontier-biased coverage exploration.

This diversity ensures that the learned graph embeddings encode structural coordination patterns rather than policy-specific motion artifacts.

B. Attack and Deviation Model

Adversarial perturbations are injected by selecting a random subset of agents (25% or 50%) and overriding their nominal actions according to

$$a_v(t) \sim \text{Uniform}(\mathcal{A}), \quad (12)$$

for randomly selected time windows.

The time of formal mission failure is defined as

$$t_{\text{break}} = \min\{t \mid G_t \text{ becomes disconnected}\}. \quad (13)$$

Each mission is therefore labeled as either nominal or degraded depending on whether $t_{\text{break}} < T_{\max}$. Importantly, our classifier is trained only on mission prefixes of length 150, well before t_{break} occurs.

C. Mission Graph Representation

Each mission trajectory is converted into a temporal graph

$$G^m = (\mathcal{T}, \mathcal{E}), \quad \mathcal{T} = \{1, \dots, T\}, \quad (14)$$

where nodes represent timesteps and edges reflect temporal succession.

Each node is labeled as

$$\ell(t) = \big(R(t), \mathbb{I}[G_t \text{ connected}], \deg(t)\big), \quad (15)$$

where $R(t)$ is the number of deviating agents and $\deg(t)$ is the average communication degree. This transforms each mission into a structured temporal object encoding coordination health.

D. Weisfeiler–Lehman Temporal Relabeling

We apply $H = 2$ iterations of the Weisfeiler–Lehman (WL) relabeling scheme:

$$\ell_h(t) = \text{hash}\Big(\ell_{h-1}(t), \text{sort}\{\ell_{h-1}(u) \mid u \in \mathcal{N}(t)\}\Big). \quad (16)$$

This produces a multiscale temporal fingerprint that captures both short-term deviations and accumulated historical disruptions.

E. Graph2Vec Embedding

Each WL-relabeled mission graph is embedded using Graph2Vec by maximizing

$$\max_{\theta} \sum_{w \in \mathcal{W}(G^m)} \log P(w \mid G^m), \tag{17}$$

with

$$P(w|G^m) = \frac{e^{\mathbf{z}_m^\top \mathbf{v}_w}}{\sum_v e^{\mathbf{z}_m^\top \mathbf{v}_v}}. \tag{18}$$

This produces a continuous vector $\mathbf{z}_m \in \mathbb{R}^{16}$ for each mission prefix. Crucially, no explicit temporal alignment or trajectory matching is required.

F. Early-Warning Classification

We perform early-warning degradation detection using a k-nearest neighbor classifier:

$$\hat{y} = \arg \min_{y \in \{0,1\}} \sum_{k=1}^{K} \|\mathbf{z}_m - \mathbf{z}_k\|_2. \tag{19}$$

This non-parametric choice is intentionally simple to demonstrate that strong degradation signals are already separable in the learned embedding geometry without requiring deep discriminative models.

V. Results

Each mission prefix of 150 timesteps is embedded into $\mathbb{R}^{16}$ using WL-Graph2Vec and projected to two dimensions via principal component analysis (PCA) for visualization. Each point in the PCA space represents the full structural evolution of a mission prefix encoded as a temporal graph.

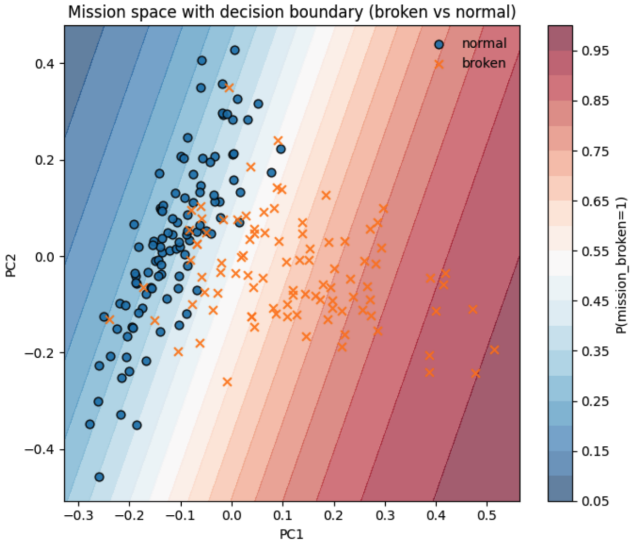

Fig. 1. Mission PCA space with kNN probability surface.

Figure 1 shows a continuous and well-structured geometry in the embedding space. Nominal missions concentrate in a compact region, while degraded missions form a smooth manifold that diverges progressively along the primary principal component. Importantly, this separation emerges even though the classifier is trained on mission prefixes that occur well before formal graph disconnection. This confirms that degradation is not a discrete event but a continuous structural process that becomes observable in the latent space far earlier than explicit failure.

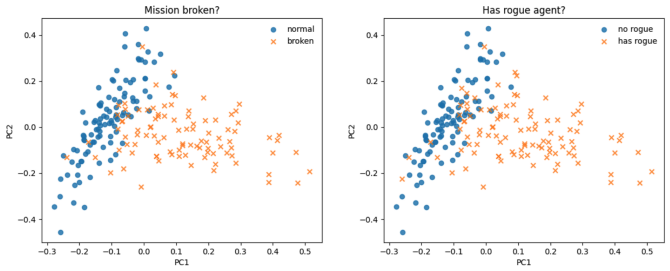

Fig. 2. Broken vs rogue agent separation.

Figure 2 compares missions containing permanently failed (broken) agents against missions containing adversarially deviating (rogue) agents. Despite both scenarios leading to mission degradation, the induced structural signatures differ measurably in the embedding space. Broken agents produce gradual loss of coordination density, while rogue agents introduce high-variance topological disruptions due to unpredictable motion. The clear clustering observed here demonstrates that the learned representation captures not only degradation severity but also the *mode* of degradation.

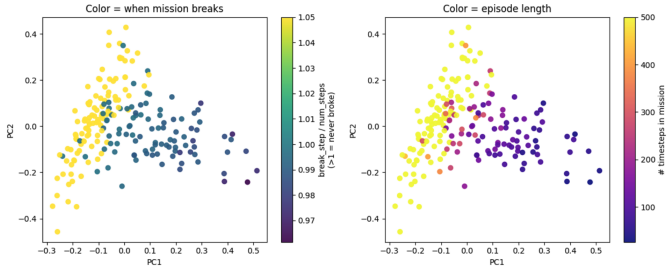

Fig. 3. Failure timing and episode length encoded in embedding space.

Figure 3 visualizes the relationship between embedding geometry, failure timing, and final episode length. Missions that fail early consistently align toward the extreme degradation region, while longer-lasting missions remain closer to the nominal cluster. This confirms that the latent distance from the nominal manifold correlates with remaining mission lifetime. As a result, the embedding provides a continuous early-warning signal rather than a binary failure indicator.

Overall, these results validate three key claims: (i) mission degradation is detectable far in advance of explicit connectivity

loss, (ii) different classes of failure induce distinct structural signatures, and (iii) degradation severity is continuously encoded in the embedding geometry. Together, these properties enable reliable early-warning detection using simple non-parametric classifiers without requiring explicit physical modeling of agent faults.

VI. Conclusion

This work establishes that early mission degradation in multi-agent systems is encoded in the latent structural evolution of communication graphs. Weisfeiler–Lehman hashing and Graph2Vec embeddings yield a continuous, interpretable degradation manifold that enables early failure detection using simple non-parametric classifiers. The approach is lightweight, communication-aware, and architecture-agnostic, making it suitable for real-time deployment on resource-constrained autonomous platforms.

References

[1] A. Narayanan, M. Chandramohan, L. Chen, Y. Liu, and S. Saminathan, "subgraph2vec: Learning Distributed Representations of Rooted Subgraphs from Large Graphs," in *Proc. of the 25th ACM International Conference on Information and Knowledge Management (CIKM)*, 2016.

[2] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler–Lehman Graph Kernels," *Journal of Machine Learning Research*, vol. 12, pp. 2539–2561, 2011.

[3] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, "graph2vec: Learning Distributed Representations of Graphs," in *Proc. of the 13th Workshop on Mining and Learning with Graphs (MLG)*, 2017.

[4] M. Grohe and M. Wiebking, "The Power of the Weisfeiler–Leman Algorithm to Decompose Graphs," arXiv:1908.05268, 2019.

[5] "[Title of the arXiv:1707.06347 paper]," arXiv:1707.06347, 2017.