Trust-Aware Task Allocation Under Adversarial Attacks in Multi-Agent Robotic Systems

Giselle Roman
Department of Computer Science
California State University, Fullerton
gisellejbroman@csu.fullerton.edu

***Abstract*—Multi-robot teams in disaster zones need to work together while dealing with limited info and possible cyber-attacks. This project simulates how trust between robots gets affected when some are spoofed or compromised. A trust recovery system is used to help fix these problems, along with a QUBO-based task assignment method. Using the NEAL sampler, task plans are optimized in both made-up and real datasets. QUBO's performance is compared with a basic greedy method over 100 simulation rounds. Key metrics include how well tasks are completed, how trust is regained, and how well the solver handles attacks. Results show QUBO performs much better in tough conditions, helping robots stay coordinated even during ongoing threats.**

***Index Terms*—Cybersecurity, Trust Modeling, Multi-Agent Systems, Adversarial Attacks, Trust Recovery, Defense Mechanisms, Quadratic Unconstrained Binary Optimization (QUBO), Robotic Task Allocation**

I. Introduction

In disaster areas like collapsed hospitals or unstable buildings, it's often too risky to send humans. Robots like drones or walking bots are used instead to handle rescue jobs. But for missions to work, robots need to share tasks and trust each other's data.

These robots often work with partial or incomplete info. This makes them easy targets for cyber-physical attacks like GPS spoofing or sensor jamming. Once a robot is compromised, others may stop trusting it—even if it's still capable hurting the team's overall performance.

This project looks at how trust between robots changes over time during these attacks. Robots have trust scores that go up or down depending on behavior and exposure. If a robot finishes tasks successfully, its trust improves. If it's spoofed or fails, trust drops.

To handle task assignment in this uncertain setting, the problem is turned into a QUBO model. The NEAL solver is used to make near-optimal assignments, factoring in trust, urgency, and task requirements. QUBO is tested against a greedy method that quickly assigns tasks based only on current trust and urgency.

Both synthetic hospital maps and real benchmark datasets are used. Each test runs for 100 rounds to compare performance, showing how well each solver handles attacks, adapts to changing trust, and keeps coordination going.

II. Related Work

Ensuring secure and resilient coordination between robotic systems in adversarial environments is a challenging task. Prior studies discuss the feasibility of attacks on graph-based representations for trust and the importance of defense strategies.

Graph-based learning methods like GCN and GAT are vulnerable to poisoning attacks [1]. By carefully adding or removing a small number of edges, attackers can degrade node embeddings without detection. Meta-learning-based attacks that target overall embedding quality or specific nodes show high transferability across models. This work treats the agent task trust matrix as graph-like structures that can be poisoned. This motivates the need for defense countermeasures that can detect and mitigate degradation in real-time task allocation.

The study of adversarial attacks on node embeddings [1] demonstrates how small, well-designed perturbations to a graph can significantly degrade embedding quality and downstream task performance. Their work shows that attacks are effective even with limited attacker knowledge and budget, and that defenses need to address structural vulnerabilities in graph representations. These insights highlight why robust trust modeling in multi-agent planning is essential, since our approach uses graph-like trust matrices that can be similarly attacked.

By integrating security and trust modeling, researchers have explored the intersection of cybersecurity, safety, and trust in cyber-physical systems [2]. This emphasizes that trust is not a binary concept but emerges through reliable and transparent interaction. A taxonomy of trust-relevant failures—including design, system, expectation, and user failures—explains how trust affects technical attacks and poor system designs. This motivates our research focus on real-time trust monitoring and the reallocation of tasks when anomalies are detected.

The concept of "secure robotics" [2] positions safety, trust, and cybersecurity as tightly connected elements that must be considered together when deploying robots in human environments. Their framework argues for integrating cybersecurity into system design from the start, not as an afterthought. This is relevant to our approach, which simulates trust degradation under attack while incorporating defense strategies that aim to restore trust dynamically. Their emphasis on proactive security and maintaining user trust aligns with our goal of developing

resilient task allocation under adversarial conditions.

These studies motivate the central hypothesis that integrating anomaly detection, trust modeling, and quantum-inspired optimization can create better multi-agent collaboration under adversarial conditions.

III. Methodology

This section covers how the full simulation works, from the input maps to attacks, defenses, and how tasks are assigned.

A. Input Data Sources

Two types of environments are used:

- **Synthetic Hospital Graph**:Custom-built layouts with floors, rooms, and tasks. Tasks have urgency levels and sometimes need to follow an order or be done by multiple robots.
- **MRTA Benchmark Instances**: Real-world robot task data from 100 JSON files. These include 6-agent setups with task order rules and team-based jobs.

Each setup includes agent skills, task info, urgency, and limits based on the file inputs.

B. Trust Modeling and Attack Simulation

Each robot has trust scores that update throughout the run:
Adversarial Attacks (Trust Degradation):

- **Spoofing Attacks**: Random trust drops simulate robots getting tricked or misreporting.
- **Environmental Noise**: Random changes model unclear sensor input or hardware issues.
- **Probabilistic Propagation**: Trust problems can affect nearby areas or robots.
- **Bait Tasks**: Fake-looking urgent tasks that trick greedy methods into bad choices.

Defensive Countermeasures (Trust Recovery):

- **Proportional Trust Recovery**: Successful task completions help rebuild trust.
- **Curing Agents**: Some robots are assigned to heal trust for nearby teammates.
- **Local View Filtering**: Each robot only sees a small area, mimicking real-life limits.
- **Trust-Urgency Balancing**: Assignments to low-trust agents are discouraged, even if tasks are urgent.

Together, these model a changing, risky environment with both attacks and fixes in play.

C. Task Assignment Solver (QUBO and Greedy)

The task assignment problem is converted into a QUBO model with:

- Trust-urgency tradeoffs
- Movement costs
- Task dependencies and teamwork rules
- One-task-per-agent limits

QUBO is solved using the NEAL sampler. Results are compared to a simple greedy solver that matches tasks based on current trust × urgency. Greedy is fast but easily fooled by attacks.

D. Evaluation and Output Metrics

Each round simulates attack propagation, trust evolution, and task coordination. We evaluate:

- **Trust Recovery Rate**
- **Coordination Success Rate**
- **Trust-Urgency Alignment**
- **Solver Time and Stability**
- **Precedence Violations and Task Duplicates**

All results are exported to CSV files for visualization and statistical analysis.

IV. Experimental Results and Evaluation

not good but still making good changes hopefully

V. Conclusion

...

VI. Future Work

a lot

Acknowledgments

This work was supported by the National Science Foundation under the REU program. The author thanks Dr. Yugyung Lee for mentoring and guidance throughout the project.

References

[1] A. Bojchevski and S. Günnemann, "Adversarial Attacks on Node Embeddings via Graph Poisoning," arXiv preprint arXiv:1809.01093, 2018. [Online]. Available: https://arxiv.org/abs/1809.01093

[2] A. Haskard and D. Herath, "Secure Robotics: Navigating Challenges at the Nexus of Safety, Trust, and Cybersecurity in Cyber-Physical Systems," ACM Computing Surveys, Mar. 2025. [Online]. Available: https://doi.org/10.1145/3723050