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

Giselle Roman (California State University, Fullerton)

Advisors: Yugyung Lee , Dianxiang Xu , Luke Miller (University of Missouri-Kansas City), Duy Ho (California State University, Fullerton)

NSF REU AI-Empowered Cybersecurity

Introduction

Multi-agent robotic systems face growing cybersecurity threats in adversarial, unpredictable environments. Spoofing, misinformation, and communication loss can degrade trust between agents, leading to failed coordination and poor task execution. This work simulates trust-targeted attacks and evaluates defenses in both synthetic and benchmarked disaster scenarios.

- **Spoofing:** Agents assigned false trust values
- **Comms Failure:** Random loss of agent communication
- **Noise Injection:** Perturbed urgency and distance
- **Bait Tasks:** High-urgency, low-integrity decoys
- **Precedence Violations:** Broken task chains

Trust degradation is modeled dynamically across rounds. A Quadratic Unconstrained Binary Optimization (QUBO) formulation is used to optimize task allocation under uncertainty, solved via a NEAL sampler.

- **Curing:** Recovers spoofed agents based on urgency/trust
- **Decay Management:** Trust drops only from poor behavior
- **Synergy Encoding:** Enforces team-based logic
- **Urgency-Weighted Recovery:** Prioritizes valuable tasks

Results show that QUBO, combined with trust-aware countermeasures, restores coordination and outperforms greedy baselines, offering a scalable framework for resilient planning in hostile environments.

Datasets & Environments

MRTA-100 Benchmark:

- 100 curated instances from a 250K+ MRTA dataset
- Tasks mapped to hospital rooms
- Agents modeled with skill vectors
- Compared QUBO+NEAL, Greedy, and MILP under real-world constraints

Synthetic Hospital Simulation:

- 5-floor, 12×12 grid simulating post-disaster hospital
- Includes spoofing attacks, urgency decay, precedence chains, bait tasks
- Agents have limited views and experience comm failures
- Tasks may require team coordination (synergy)
- Used to test robustness under adversarial, uncertain conditions

METHODS

Simulation Setup

- Run on both a synthetic hospital environment and the MRTA-100 benchmark
- Agent-task layouts included urgency levels, spatial constraints, and coordination rules

Solvers

- **Greedy:** Assigns agents to tasks based on urgency and proximity
- **QUBO+NEAL:** Encodes trust, urgency, precedence, and synergy into a binary optimization model solved using NEAL

Trust Modeling

- Initial trust based on distance, urgency, noise, and agent role
- Trust updated each round using spoofing attacks and performance history
- Realism added through local views, communication loss, and urgency decay

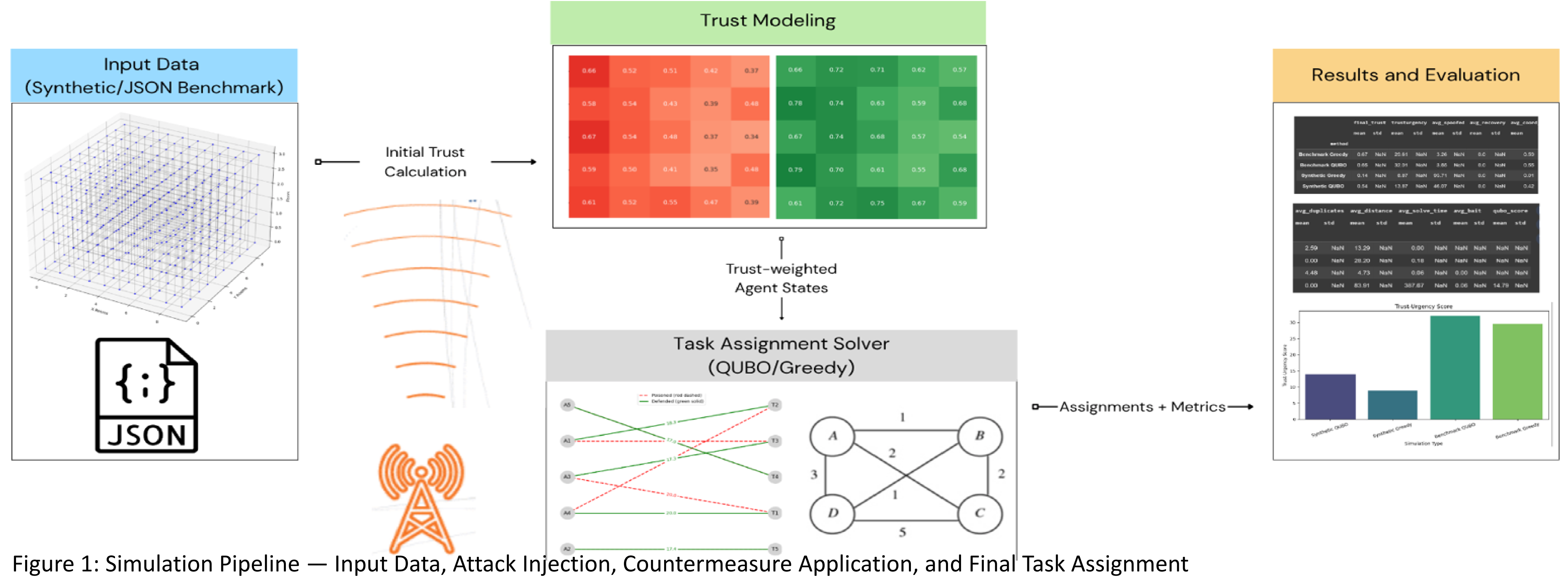

Figure 1: Simulation Pipeline — Input Data, Attack Injection, Countermeasure Application, and Final Task Assignment

RESULTS

Despite spoofing, urgency decay, and limited knowledge, QUBO agents retain high trust over time. In contrast, Greedy assignments lead to cascading trust failures and loss of coordination.

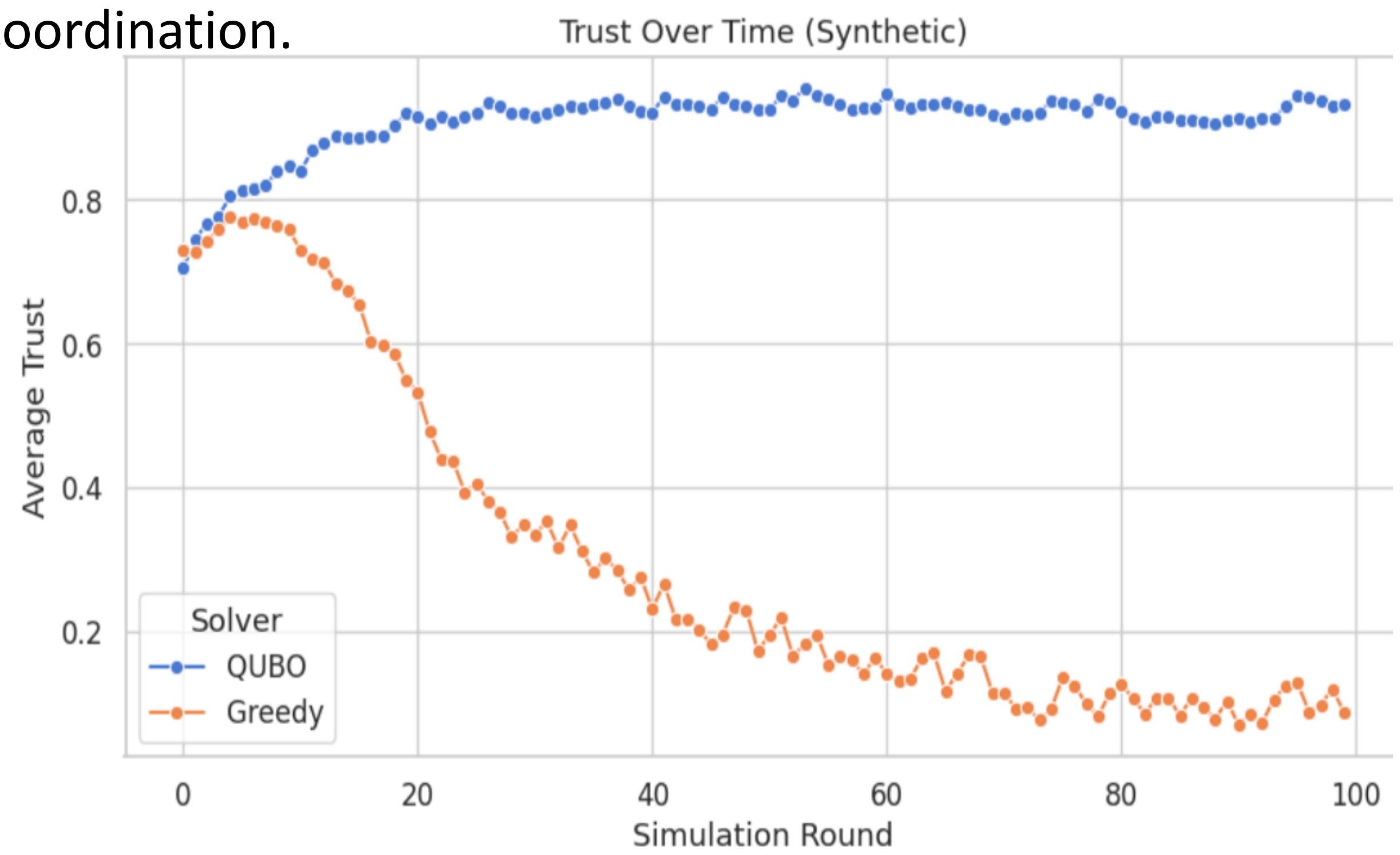

Figure 2: QUBO preserves trust under attack

Despite similar spoofing exposure, only QUBO exhibits meaningful trust recovery. Greedy fails to regain trust once lost, contributing to system collapse.

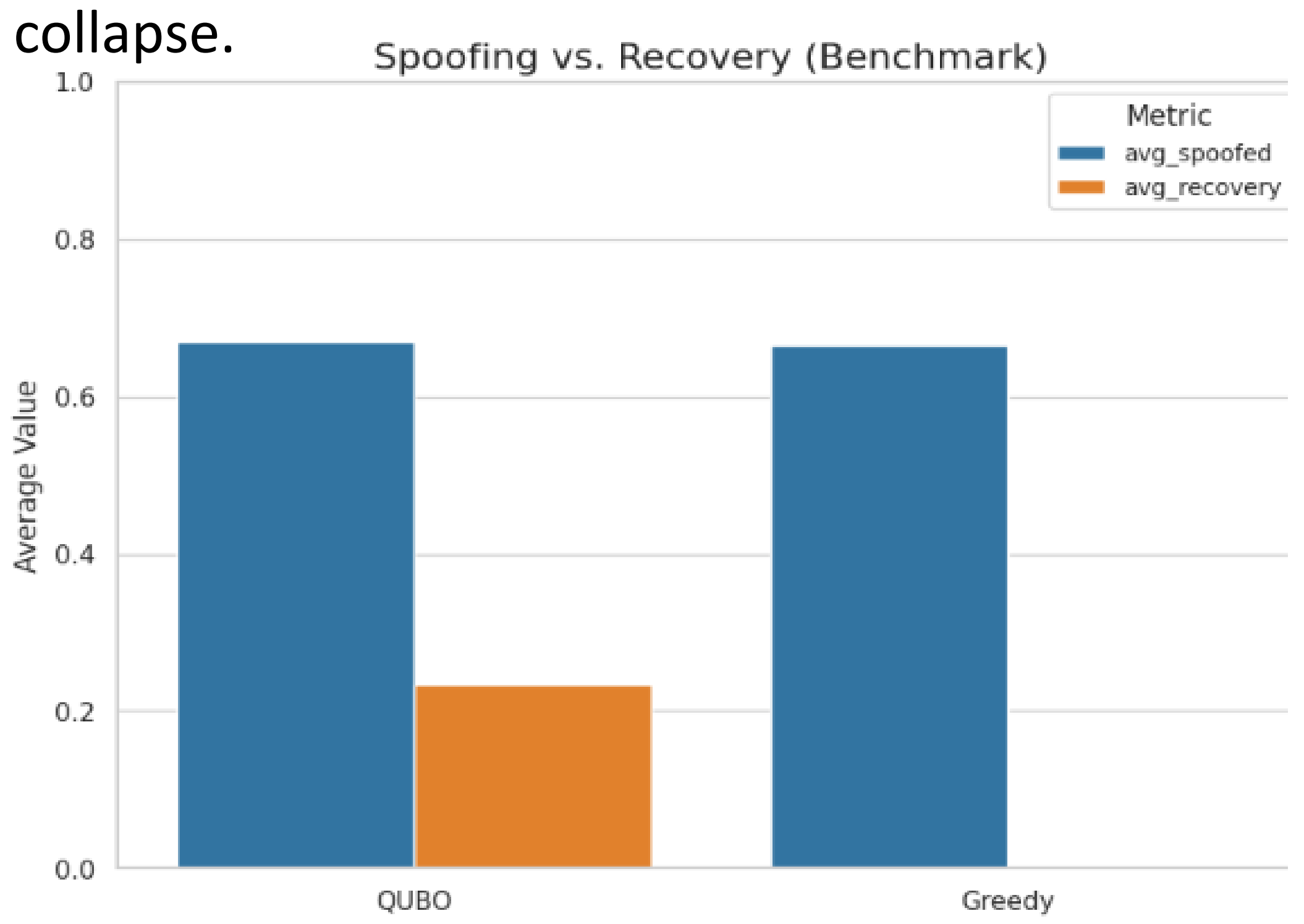

Figure 3: QUBO actively recovers trust under attack

QUBO outperforms Greedy across all critical metrics in trust stability, success rate, resilience to attacks, and constraint satisfaction, making it a reliable method for real-world post-disaster coordination.

Scenario	Solver	Initial Trust	Final Trust	Task Success Rate	Avg Spoofed Agents	Delta Trust	Reward Score
Synthetic	QUBO	0.723	0.909	0.242	0.737	0.186	0.188
Synthetic	Greedy	0.739	0.296	0	0.109	-0.443	0
Realistic	QUBO	0.726	0.568	0.712	0.67	-0.158	0.029
Realistic	Greedy	0.728	0.552	0.01	0.666	-0.176	0

Figure 4: Summary of coordination and robustness metrics under adversarial conditions

CONCLUSION

- **QUBO+NEAL based task allocation** preserves agent trust, enables coordination, and resists collapse under spoofing, communication loss, and noise.
- **Greedy solvers break down** in complex environments often failing to assign or complete tasks under adversarial conditions.
- **Simulations in both synthetic and benchmark hospital environment** show that QUBO outperforms Greedy across metrics like task success, trust preservation, and spoofing resilience.
- **Integrated defenses** (curing compromised agents, managing trust decay, enforcing team-task synergy) significantly enhance robustness.
- **QUBO scales across varied task constraints**, including precedence chains, coalition tasks, urgency decay, and limited agent views.

Future Work

- Apply **learning-based trust recovery** (reinforcement learning) to adapt defenses dynamically.
- **Enable decentralized or onboard planning** for real-time, distributed deployment with partial observability.
- Simulate **multi-adversary attacks**, including false task injection and coordinated misinformation.
- Extend to **larger-scale environments**, 3D hospital layouts, or mixed human-robot coordination.

REFERENCES

1. Bojchevski, Aleksandar, and Stephan Günnemann. “Adversarial Attacks on Node Embeddings via Graph Poisoning.” ArXiv.org, 2018, arxiv.org/abs/1809.01093.
2. Haskard, Adam, and Damith Herath. “Secure Robotics: Navigating Challenges at the Nexus of Safety, Trust, and Cybersecurity in Cyber-Physical Systems.” ACM Computing Surveys, Association for Computing Machinery, Mar. 2025, https://doi.org/10.1145/3723050.
3. SMART-LLM. (2025). *Secure Multi-Agent Resilient Tasking with LLMs: AI2-THOR Simulation Benchmark* [Data set]. TU Delft, 4TU.ResearchData. https://doi.org/10.4121/10e28ee0-9ad9-450d-8be7-6e6a91f2931f.v1

ACKNOWLEDGEMENTS

This work is based upon support from the National Science Foundation under Grant Number CNS-2349236.