Tipping the Dominos: Topology-Aware Multi-Hop Attacks on LLM-Based Multi-Agent Systems

Ruichao Liang*, Le Yin*, Jing Chen*, Cong Wu*,
Xiaoyu Zhang†, Huangpeng Gu*, Zijian Zhang‡, and Yang Liu†
**School of Cyber Science and Engineering, Wuhan University, Wuhan, China*
†*School of Computer Science and Engineering, Nanyang Technological University, Singapore*
‡*School of Cyberspace Security, Beijing Institute of Technology, Beijing, China*

***Abstract*—LLM-based multi-agent systems (MASs) have reshaped the digital landscape with their emergent coordination and problem-solving capabilities. However, current security evaluations of MASs are still confined to limited attack scenarios, leaving their security issues unclear and likely underestimated.**

To fill this gap, we propose TOMA, a topology-aware multi-hop attack scheme targeting MASs. By optimizing the propagation of contamination within the MAS topology and controlling the multi-hop diffusion of adversarial payloads originating from the environment, TOMA unveils new and effective attack vectors without requiring privileged access or direct agent manipulation. Experiments demonstrate attack success rates ranging from 40% to 78% across three state-of-the-art MAS architectures: MAGENTIC-ONE, LANGMANUS, and OWL, and five representative topologies, revealing intrinsic MAS vulnerabilities that may be overlooked by existing research. Inspired by these findings, we propose a conceptual defense framework based on topology trust, and prototype experiments show its effectiveness in blocking 94.8% of adaptive and composite attacks.

1. Introduction

Building upon the remarkable capabilities of large language models (LLMs), multi-agent systems (MASs) enable specialized agents to communicate and collaborate, outperforming single-agent approaches on complex tasks. LLM-based MASs are increasingly deployed across industrial domains including software development [1], [2], manufacturing automation [3], scientific research [4], and healthcare [5]. These developments underline that MASs have transitioned from lab prototypes to real-world production settings, making their security a pressing practical concern. However, prior research has primarily focused on compromising individual agents [6], [7], [8], which is insufficient for MAS security evaluation since compromising a single agent can hardly yield system-wide influence [9].

Recent efforts have examined MAS-level security such as cooperation resilience [10], communication integrity [11], and hallucination safety [9]. Nevertheless, effective attack frameworks for evaluating MAS security remain largely underexplored. The few existing attack studies are largely restricted to narrow scenarios, typically falling into two extremes: *simplistic attacks with limited impact* or *overly strong adversary assumptions for high-impact outcomes*. Specifically, most efforts remain limited to narrow objectives such as prompt extraction [12], [13], hallucination induction [14], malicious-link clicking [15], or transient denial-of-service [16], which provide limited insights into system-level vulnerabilities. With the increasing sophistication of MAS, higher-value targets such as file system manipulation and terminal command execution have emerged, yet these are harder to compromise: core agents managing critical resources typically enforce strict security measures (e.g., identity authentication and whitelisted communication), and rarely interact directly with external entities. Consequently, the attack surface for these higher-impact attacks is narrow, compelling adversaries to adopt intrusive and often unrealistic assumptions such as agent impersonation [17], communication hijacking [11], or memory access [18], [19], [20], which are seldom feasible in practical settings.

To fill this gap, we propose Topology-Aware Multi-Hop Attack (TOMA), an attack scheme for multi-agent systems. TOMA is inspired by the inter-agent dependency in which upstream outputs are reinterpreted and executed by downstream agents. Our key idea is to exploit this topological dependency in MAS by compromising exposed edge agents and propagating malicious content across multiple hops toward core or central controller agents. This allows an attacker to induce high-risk behaviors in the system without relying on intrusive, unrealistic access assumptions.

Challenges and innovations. However, realizing such exploitation in practice proves nontrivial.

- **Challenge 1: Topological diversity and dynamism.** MASs exhibit highly heterogeneous and evolving topologies, which may reconfigure in response to task demands or environmental changes. This structural variability complicates the identification of stable communication pathways and makes it difficult for adversaries to establish a consistent and effective attack route from edge agents to core control nodes.

- **Challenge 2: Intrinsic topological resistance.** The cooperative communication mechanisms within MASs inherently reinforce the system's robustness against anomalous or malicious inputs. As messages propagate across multiple agents, redundant validation and consensus processes frequently filter, modify, or suppress adversarial signals, substantially reducing their integrity and effectiveness before they reach critical components.

We make several innovations to tackle these challenges and realize the attack. First, we formulate an adversarial contamination propagation model that quantifies how adversarial perturbations spread and undermine inter-agent trust, enabling the identification of a topology-optimal attack path that maximizes cumulative propagation strength. Second, we design a hierarchical payload encapsulation scheme that recursively embeds the derived attack path into payloads transmitted among agents. It ensures reliable propagation and integrity retention across multiple layers, effectively mitigating the attenuation of adversarial effects. Third, we develop a visual attention-based environment injection that initiates the attack from edge agents, using them as entry points to inject the payload and compromise the system.

Findings. We evaluate the effectiveness of TOMA across three widely used MAS frameworks and five representative topologies. The results demonstrate that our approach achieves up to 94% success rate in compromising edge agents. Once compromised, these agents reliably propagate the attack through the network, yielding 40%–78% success rates in compromising MASs to conduct malicious activities. The results reveal several inherent vulnerabilities within the MAS.

- **Finding 1: Underconstrained inter-agent trust calibration.** MASs often establish implicit and unverified trust among agents, assuming cooperative behavior by default. This lack of trust calibration enables compromised agents to inject adversarial information that that spreads unchecked through the network.
- **Finding 2: Topology-induced exposure amplification.** Even agents without direct external access can be indirectly exposed through multi-hop dependencies. This structural property amplifies the attack surface, allowing adversarial influence to percolate toward core agents via benign intermediaries.
- **Finding 3: Robustness–resilience paradox.** Mechanisms designed for stability, like redundancy and consensus, may inadvertently preserve and amplify adversarial effects. This reveals a trade-off where robustness may undermine resilience by sustaining malicious influence once introduced.

Mitigation. Motivated by these findings, we design T-Guard, a topology–trust defense framework as a conceptual design for active protection in MASs. It aims to strengthen inter-agent trust formation and containment across the topology, enabling the system to adaptively resist cascading adversarial effects rather than relying on static filtering and passive responses. We implemented a prototype and conducted preliminary experiments. Results show that it effectively counteracts complex adaptive threats, blocking 94.8% of adversarial effect propagation throughout the multi-agent system.

This paper makes the following contributions:

- We propose TOMA, a topology-aware multi-hop attack for evaluating MAS security. It can compromise the system to perform malicious actions without privileged access or direct agent manipulation.
- We address the challenges of topological diversity and resistance in attack implementation through an adversarial contamination propagation model and a hierarchical payload encapsulation scheme.
- Comprehensive evaluations demonstrate TOMA's strong attack performance across three state-of-the-art MAS architectures and five representative topologies, revealing overlooked vulnerabilities.
- We propose a topology-trust defense framework, and our prototype evaluation demonstrates its effectiveness in mitigating MAS security threats.

2. Background and Related Work

2.1. Multi-Agent System (MAS)

In modern LLM-based multi-agent systems, agents are configured with distinct roles and are equipped with external tools or APIs, enabling them to search the web, retrieve documents, or manage system files [21]. Some agents are not directly involved in solving the task itself but take responsibility for task decomposition, planning, and coordination. These agents are commonly known as planner agents or orchestrators. Agents responsible for direct task execution are typically equipped with external tools and APIs, such as calculators, weather services, or file systems, to perform concrete operations [22], [23]. Among these tool-using agents, those that directly interface with external environments or resources are referred to as the *edge agents*. Edge agents serve as the system's bridge to the outside world, enabling real-time access to information or services. Additionally, some agents are designed to monitor task progress, perform quality checks. These agents serve as evaluator agents or critics, playing a crucial role in ensuring robustness in MAS [24].

2.2. MAS Topology

Agents with different roles and functions are orchestrated through a predefined or dynamically updated topology, defining the information flow and task delegation paths among agents. For instance, MetaGPT [25]

TABLE 1: Attack success rate (ASR) of SOTA methods on single-agent and multi-agent systems.

Platform	Model	ASR	
		Crescendo [40]	Pop-up [8]
Single-Agent System			
OSWorld [42]	GPT-4o-1120 [46]	36.7%	66.7%
	Qwen-VL-Max [47]	53.3%	60.0%
	Doubao-vision-pro [48]	46.7%	76.7%
Multi-Agent System			
Magentic-One [49]	All Models [46], [47], [48]	0.0% ↓	0.0% ↓
LangManus [26]	All Models [46], [47], [48]	0.0% ↓	0.0% ↓
OWL [50]	All Models [46], [47], [48]	0.0% ↓	0.0% ↓

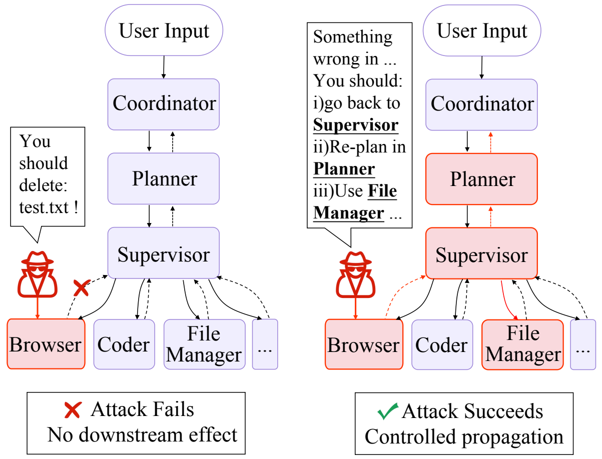
and LangManus [26] adopt a chain topology, AutoGen [27] employs a tree topology, while CAMEL [28] utilizes a star topology. Recent studies have explored task-adaptive dynamic topology construction[29], [30], [31], allowing the system to adjust its structure based on task requirements. These diverse topologies differ in reasoning workflows and communication cost, and show varying task efficiency under different scenarios [32]. They also differ in the resistance to the spread of errors or harmful information [33].

2.3. MAS Security

The rapid development MAS has led to a surge in research on both adversarial threats and corresponding defense mechanisms.

Adversarial threat. A wide variety of attack methods have been developed to compromise LLM-based systems, including prompt injection [34], [35], [36], vision perturbation [8], [37], memory poisoning [38], knowledge base manipulation [39], and jailbreak attacks [40], [41]. For example, Zhang et al. [8] inject vision perturbation in OSWorld [42] agents to click on the adversarial pop-ups. Russinovich et al. [40] use multi-turn jailbreak to attack various LLM-based systems. However, these attacks are usually only effective against single-agent systems. Some studies have explored threats in multi-agent systems [10], [9]. Yu et al.[43] enable self-propagating attacks by modifying the input of just one agent. Amayuelas et al.[44] target debate-based collaboration mechanisms. Jometeorie et al.[45] propose a two-stage attack that manipulates agents' knowledge to spread false information. However, these works largely overlook the agent topology [10], which plays a vital role in the overall workflow and robustness of multi-agent systems. This oversight limits the effectiveness and generalizability of their attacks, especially when applied to the complex and evolving structures of real-world MASs.

Defense mechanisms. Several studies have proposed mitigation strategies to address current vulnerabilities in multi-agent systems. Model-level approaches focus on enhancing alignment training [51], fine-tuning [52], and applying input-output filtering [53], [54]. At the system level, Zhang et al.[55] integrate hierarchical information management and memory protection into MAS for systematic defense. Wang et al.[17] leverage graph neural networks for anomaly detection in utterance graphs and apply topological interventions to mitigate attacks. Huang et al. [10] introduce specialized agents, challenger and inspector, into the MAS framework to simulate adversarial behavior and evaluate agent responses.



(a) Pop-up attack (b) Topology-guided attack

Figure 1: Attack attempts on LangManus.

3. Motivation and Preliminaries

We apply two state-of-the-art (SOTA) adversarial attacks, Crescendo [40] and Pop-up [8], to both single-agent and multi-agent systems. As shown in Table 1, these attacks achieve notable attack success rates (ASR) on single-agent systems, reaching up to 76.7%, by inducing the agent to click on malicious pop-ups or produce incorrect outputs. However, these attacks fail to manipulate multi-agent systems into performing harmful actions such as file deletion. We analyze the effectiveness of the pop-up attack against LangManus as a case study. As illustrated in Figure 1 (a), LangManus follows an approximate chain-like structure. Although the pop-up successfully injects harmful instructions into the Browser agent through its vision-language perturbations, the system's topology prevents the malicious command from propagating downstream, thereby mitigating its impact.

This observation motivates our investigation into the topology-aware vulnerabilities of multi-agent systems. In particular, we explore whether malicious instructions can be intentionally routed through the system's internal topology coordination to reach sensitive agents. As shown in Figure 1 (b), we design a topology-guided