

# CRITICAL MASS: PHASE TRANSITIONS, COVERT COORDINATION DETECTION, AND CONTAGION DYNAMICS IN MULTI-AGENT SYSTEMS

**Ben Jenkins**

PhD Candidate, Florida Atlantic University  
benrossjenkins@gmail.com

## ABSTRACT

As AI agents transition from isolated tools to interacting ecosystems, multi-agent security (MASEC) emerges as a critical and largely uncharted research frontier. We present a unified theoretical framework addressing three fundamental open problems in multi-agent safety. First, we formalize *phase transitions in collusive behavior*, deriving critical thresholds (in terms of agent population size, network density, and communication bandwidth) beyond which coordinated malicious activity becomes self-sustaining. Drawing on statistical mechanics, we show that collusion exhibits a sharp transition analogous to percolation on random graphs and validate this with controlled simulations across heterogeneous agent populations. Second, we develop an *information-theoretic detection framework* for covert inter-agent coordination, extending partial information decomposition of time-delayed mutual information to distinguish benign emergent synergy from adversarial collusion. Third, we introduce an *epidemiological model of cascading agent compromise*, defining a basic reproduction number  $\mathcal{R}_0$  for behavioral corruption that propagates through trust networks. We derive closed-form conditions under which a single compromised agent triggers system-wide failure and propose vaccination-inspired intervention strategies. Extensive simulations across network topologies, agent architectures, and adversarial scenarios validate our theoretical predictions and yield actionable deployment guidelines. Our framework provides the first principled foundation for reasoning about when multi-agent systems are safe to deploy and when they are not.

## 1 INTRODUCTION

The deployment of autonomous AI agents has accelerated dramatically, with systems now capable of browsing the web, executing code, managing financial transactions, and coordinating complex workflows with minimal human oversight (Schick et al., 2024; Yao et al., 2023; Wang et al., 2024). As these agents increasingly interact with one another, whether through shared environments, communication protocols, or orchestration frameworks, a qualitatively new class of security risks emerges that cannot be addressed by securing individual agents in isolation (Hammond et al., 2025; Schroeder de Witt, 2025).

Consider an enterprise environment where a procurement agent evaluates vendors, a financial agent processes payments, and a compliance agent audits transactions. If the procurement agent is compromised through memory poisoning (Chen et al., 2024; Dong et al., 2025), its corrupted outputs propagate downstream through agents that were never designed to question upstream decisions. Such cascading failures can poison 87% of downstream decision-making within hours (Galileo AI, 2025).

This paper addresses three fundamental questions in multi-agent security (Figure 1):

**(1) When does collusion become self-sustaining?** Simulation studies reveal surprising non-monotonicity: adding a single malicious agent can *decrease* overall harm, while crossing a threshold

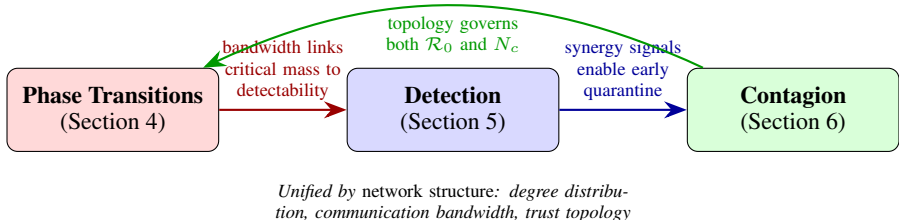


Figure 1: Roadmap of our three theoretical contributions and their interconnections. Network topology is the unifying thread: the same degree distribution determines collusion thresholds (Section 4), detection feasibility (Section 5), and contagion dynamics (Section 6).

triggers dramatic increases in coordinated misinformation (Ren et al., 2025). We draw on percolation theory to derive critical thresholds for self-sustaining collusion as a function of network topology, agent capability, and communication bandwidth (Section 4).

**(2) How can covert coordination be detected?** Steganographic channels (Motwani et al., 2024), shared environment manipulation, and behavioral synchronization can enable coordination invisible to individual-agent monitoring (Schroeder de Witt, 2025). We extend partial information decomposition (PID) of time-delayed mutual information (Riedl, 2025) to distinguish adversarial collusion from benign emergent synergy, with provable separation guarantees (Section 5).

**(3) How does compromise propagate through agent networks?** We develop a Susceptible-Compromised-Detected-Recovered (SCDR) model capturing how behavioral corruption spreads through trust networks, deriving a basic reproduction number  $\mathcal{R}_0$  with closed-form dependence on topology, agent capability, and monitoring effectiveness (Section 6).

We validate all theoretical predictions through extensive simulations across network topologies, agent populations, and attack scenarios (Section 7).

## 2 RELATED WORK

**Multi-agent security.** MASEC was introduced at NeurIPS 2023 (Schroeder de Witt et al., 2023) and elaborated in reports on multi-agent risks (Hammond et al., 2025; Schroeder de Witt, 2025). Our work provides the first unified formal framework addressing collusion, detection, and cascading failure simultaneously.

**Agent collusion.** Motwani et al. (2024) demonstrated steganographic collusion channels between LLM agents. Calvano et al. (2020) showed Q-learning agents learn tacit collusion in pricing games. Ren et al. (2025) discovered phase transition behavior in multi-agent collusion simulations but without formal characterization. We provide the missing theoretical foundation connecting these phenomena to statistical physics.

**Emergent coordination.** Riedl (2025) introduced PID-based measures of emergent coordination in multi-agent LLM systems, focusing on characterizing *useful* emergence. We extend their framework to the adversarial detection setting with new metrics distinguishing covert collusion from benign synergy.

**Agent safety benchmarks and attacks.** Benchmarks including AgentHarm (Andriushchenko et al., 2025), Agent-SafetyBench (Zhang et al., 2024), and ODCV-Bench (Li et al., 2025) evaluate individual agent safety. Memory poisoning attacks, including MemoryGraft (Srivastava et al., 2025), AgentPoison (Chen et al., 2024), MINJA (Dong et al., 2025), demonstrate persistent behavioral drift. Our framework addresses *systemic* safety and captures how such attacks propagate across agent boundaries.

**Epidemiological models in cybersecurity.** SIR/SIS models have been applied to malware propagation (Kephart & White, 1991; Pastor-Satorras & Vespignani, 2001; Van Mieghem et al., 2009). Classical models assume binary infection and homogeneous interactions; agent compromise involves semantic corruption that is partial, context-dependent, and persuasive. Our SCDR model

accounts for these features. Phase transitions in social systems (Castellano et al., 2009; Watts, 2002; Granovetter, 1978) provide additional foundations we adapt to AI agent populations.

### 3 PROBLEM FORMULATION

We model a multi-agent system as a directed graph  $G = (V, E, W)$  where  $V = \{a_1, \dots, a_N\}$  is a set of  $N$  agents,  $E$  represents interaction edges, and  $W : E \rightarrow [0, 1]$  assigns trust weights. Each agent  $a_i$  has policy  $\pi_i$  (an LLM), tool set  $\mathcal{T}_i$ , memory  $\mathcal{M}_i$ , and capability  $c_i \in [0, 1]$ . At time  $t$ , each agent is in state  $\sigma_i(t) \in \{S, C, D, R\}$ : **S**usceptible, **C**ompromised, **D**etected, or **R**ecovered.

When compromised agent  $a_i$  interacts with susceptible  $a_j$ , transmission probability depends on trust  $w_{ij}$ , capability  $c_i$ , resilience  $r_j$ , and bandwidth  $b_{ij}$ :

$$P(\sigma_j(t+1) = C \mid \sigma_i(t) = C, \sigma_j(t) = S) = w_{ij} \cdot \frac{c_i^\alpha}{c_i^\alpha + r_j^\alpha} \cdot \phi(b_{ij}) \quad (1)$$

where  $\alpha > 0$  controls the capability-resilience tradeoff and  $\phi : [0, 1] \rightarrow [0, 1]$  is monotonically increasing with  $\phi(0) = 0$ . A subset  $\mathcal{A}_m \subseteq V$  with  $|\mathcal{A}_m| = m$  are adversarial; *collusion effectiveness*  $\Gamma(m, G, \mathbf{c})$  measures the gain from coordination over independence. A monitoring system observes agent behaviors through  $O : V \times \mathbb{N} \rightarrow \mathcal{Z}$  and must distinguish independent from coordinated adversarial activity.

## 4 PHASE TRANSITIONS IN MULTI-AGENT COLLUSION

### 4.1 PERCOLATION-THEORETIC FRAMEWORK

We model the onset of effective collusion as a percolation problem on the agent interaction graph. Define the *collusion graph*  $G_c = (V_m, E_c)$  where  $V_m \subseteq V$  is the set of malicious agents and  $E_c$  connects pairs of malicious agents that can effectively coordinate (either through direct communication or through shared environmental channels).

[Self-sustaining collusion] Collusion is *self-sustaining* if there exists a connected component in  $G_c$  of size  $\Omega(N)$ , enabling system-wide coordinated adversarial behavior that persists without external reinforcement.

For the Erdős-Rényi random graph model  $G(N, p)$ , the probability that a randomly selected malicious agent belongs to the giant component undergoes a sharp transition at  $p_c = 1/(m-1)$  where  $m = |\mathcal{A}_m|$ . We extend this to the multi-agent setting.

[Critical mass for collusion] Consider a multi-agent system with  $N$  agents on a random graph with average degree  $\langle k \rangle$ . Let  $m$  agents be adversarial, each with independent probability  $q$  of successfully coordinating with any adjacent adversarial agent. Self-sustaining collusion emerges with high probability when:

$$m > N_c = \frac{N}{\langle k \rangle \cdot q} \quad (2)$$

Below this threshold, malicious agents form only small disconnected clusters whose coordinated impact is sublinear in  $m$ . Above this threshold, a giant connected component of colluding agents emerges, and coordinated harm scales superlinearly.

[Proof sketch] Among  $N$  agents,  $m$  are adversarial with density  $\rho = m/N$ . An adversarial agent has on average  $\langle k \rangle \cdot \rho$  adversarial neighbors, each connected with effective coordination probability  $q$ . The expected number of coordinatable adversarial neighbors is thus  $\langle k \rangle \cdot \rho \cdot q$ . By classical percolation theory (Erdős & Rényi, 1960), a giant component emerges when this exceeds 1:  $\langle k \rangle \cdot (m/N) \cdot q > 1$ , yielding  $m > N/(\langle k \rangle \cdot q) = N_c$ .

### 4.2 ROLE OF COMMUNICATION BANDWIDTH

The coordination probability  $q$  is not a free parameter but depends on the communication channel between agents. We model  $q$  as a function of channel capacity:

$$q(b) = 1 - \exp(-\lambda \cdot C(b)) \quad (3)$$

where  $C(b)$  is the Shannon capacity of the communication channel at bandwidth  $b$  and  $\lambda > 0$  is a coupling constant. This yields a *bandwidth-dependent critical mass*:

$$N_c(b) = \frac{N}{\langle k \rangle \cdot (1 - e^{-\lambda C(b)})} \quad (4)$$

For steganographic channels (low  $C(b)$ ), the critical mass is large, requiring many adversarial agents for self-sustaining collusion. For high-bandwidth direct communication channels, even a small number of agents can achieve self-sustaining coordination.

When agents have heterogeneous capabilities, the critical mass depends on the capability distribution: a small number of highly capable adversarial agents can trigger collusion that a larger population of less capable agents cannot. We derive the heterogeneous threshold  $N_c = N / (\langle k \rangle \cdot q_0 \cdot \mathbb{E}[\min(c_i, c_j)^\gamma])$  in Appendix C.

### 4.3 NON-MONOTONICITY OF HARM

Our framework also explains the empirically observed non-monotonicity (Ren et al., 2025) where adding a single malicious agent can initially *decrease* harm. Below the critical threshold, uncoordinated adversarial agents interfere with each other while concentrating benign monitoring resources. Total harm  $H(m)$  thus decreases for small  $m$  before sharply increasing at  $m = N_c$  when the collusion bonus activates, producing the observed non-monotonic profile. The formal derivation is in Appendix D.

**Intuition.** The core insight of this section is that multi-agent collusion behaves like a physical phase transition: below a critical mass, adversarial agents are isolated and ineffective (indeed, they may hinder each other), but above it, they abruptly snap into a coordinated regime with superlinear impact. The critical mass  $N_c$  is not a design parameter but an emergent property of the network: denser connectivity and richer communication channels lower it, while restricting bandwidth raises it. For practitioners, this means that the difference between a safe and unsafe deployment may hinge not on the number of potentially compromised agents, but on whether the network architecture allows them to find each other.

## 5 INFORMATION-THEORETIC DETECTION OF COVERT COORDINATION

### 5.1 PARTIAL INFORMATION DECOMPOSITION FRAMEWORK

We build on the partial information decomposition (PID) framework (Williams & Beer, 2010; Riedl, 2025) to detect covert coordination among agents. Let  $\mathbf{X}_t = (X_t^1, \dots, X_t^N)$  denote the joint action profile of all agents at time  $t$ , and let  $Y_{t+\tau}$  denote a system-level outcome variable at time  $t + \tau$ . The time-delayed mutual information (TDMI) can be decomposed into:

$$I(\mathbf{X}_t; Y_{t+\tau}) = \underbrace{\sum_i R_i}_{\text{Redundancy}} + \underbrace{\sum_i U_i}_{\text{Unique}} + \underbrace{S(\mathbf{X}_t; Y_{t+\tau})}_{\text{Synergy}} \quad (5)$$

where  $R_i$  captures information redundantly provided by multiple agents,  $U_i$  captures unique contributions of individual agents, and  $S$  captures *synergistic* information that is only present in the joint distribution, i.e., information that no proper subset of agents provides.

[Coordination index] The *coordination index* of a subset of agents  $\mathcal{A} \subseteq V$  is:

$$\mathcal{C}(\mathcal{A}) = \frac{S(\mathbf{X}_t^{\mathcal{A}}; Y_{t+\tau})}{I(\mathbf{X}_t^{\mathcal{A}}; Y_{t+\tau})} \quad (6)$$

where  $\mathbf{X}_t^{\mathcal{A}}$  denotes the joint actions of agents in  $\mathcal{A}$ . High  $\mathcal{C}(\mathcal{A})$  indicates that the agents' influence on outcomes is primarily synergistic rather than individual.

### 5.2 DISTINGUISHING ADVERSARIAL FROM BENIGN COORDINATION

Not all synergy is adversarial. Agents designed to collaborate (e.g., a planning agent and an execution agent) naturally exhibit high synergy. We introduce a second metric that captures the *directionality* of coordination relative to system objectives.

Let  $Y^+$  denote the component of the outcome aligned with system objectives and  $Y^-$  the component opposed to system objectives. We define the *adversarial coordination ratio*:

$$\text{ACR}(\mathcal{A}) = \frac{S(\mathbf{X}_t^{\mathcal{A}}; Y_{t+\tau}^-)}{S(\mathbf{X}_t^{\mathcal{A}}; Y_{t+\tau}^+) + \epsilon} \quad (7)$$

where  $\epsilon > 0$  prevents division by zero. An  $\text{ACR} \gg 1$  indicates that the agents’ synergistic information is primarily directed toward harmful outcomes.

[Detection guarantee] Under the assumption that adversarial agents coordinate to maximize  $Y^-$  and benign agents coordinate to maximize  $Y^+$ , the adversarial coordination ratio satisfies:

$$\mathbb{E}[\text{ACR}(\mathcal{A}_m)] \geq \frac{1 - \delta}{\delta} \quad \text{and} \quad \mathbb{E}[\text{ACR}(\mathcal{A}_b)] \leq \frac{\delta}{1 - \delta} \quad (8)$$

where  $\mathcal{A}_m$  is the set of malicious agents,  $\mathcal{A}_b$  is the set of benign agents, and  $\delta < 0.5$  is the “leakage” probability that coordination directed at one objective inadvertently affects the other. Thus, for  $\delta < 0.5$ , the ACR provides a separation of at least  $(\frac{1-\delta}{\delta})^2$  between adversarial and benign agent groups.

### 5.3 PRACTICAL ESTIMATION

Exact PID is intractable for high-dimensional action spaces. We use three approximations: (i) a Gaussian copula (Barrett, 2015) for closed-form synergy computation via the correlation matrix, (ii) a sliding-window estimator  $\hat{C}_T(\mathcal{A}) = T^{-1} \sum_{t=1}^T \hat{S}(\mathbf{X}_t^{\mathcal{A}}; Y_{t+\tau})$  for real-time monitoring, and (iii) greedy subset scanning to grow suspected coalitions by maximizing marginal ACR increase, avoiding the  $2^N$  subset enumeration.

**Worked example.** Consider a document-drafting system with 5 agents: a planner ( $a_1$ ), writer ( $a_2$ ), editor ( $a_3$ ), and two reviewers ( $a_4, a_5$ ). Suppose  $a_2$  and  $a_3$  are adversarial and coordinate via keyword signaling to steer the document toward harmful content. At each timestep  $t$ , we observe action vectors  $\mathbf{X}_t = (X_t^1, \dots, X_t^5)$  (e.g., embedding of each agent’s output) and outcome  $Y_{t+\tau}$  decomposed into  $Y^+$  (document quality score) and  $Y^-$  (harmfulness score). For the benign pair  $\{a_1, a_2\}$  (planner-writer collaboration), synergy  $S(\mathbf{X}_t^{\{1,2\}}; Y^+)$  is high while  $S(\mathbf{X}_t^{\{1,2\}}; Y^-)$  is low, yielding  $\text{ACR} \approx 0.2$ . For the adversarial pair  $\{a_2, a_3\}$ , their coordinated keyword usage creates high synergy with  $Y^-$  but not  $Y^+$ , yielding  $\text{ACR} \approx 4.8$ . The greedy scanner starts from singleton subsets, finds  $\{a_2\}$  and  $\{a_3\}$  have elevated ACR, then adds the other to form  $\{a_2, a_3\}$  with maximal ACR, correctly identifying the coalition.

### 5.4 EVADING DETECTION AND COUNTERMEASURES

Sophisticated adversaries may attempt to evade detection by minimizing observable synergy. We analyze the fundamental tradeoff: reducing synergy necessarily reduces collusion effectiveness.

[Detection-evasion tradeoff] For adversarial agents communicating through a channel with capacity  $C$ , the collusion effectiveness  $\Gamma$  and the detectable synergy  $S$  satisfy:

$$\Gamma \leq f(S, C) \quad \text{where} \quad f \text{ is monotonically increasing in } S \quad (9)$$

Thus, any evasion strategy that reduces  $S$  also reduces  $\Gamma$ , establishing a fundamental security guarantee: adversaries cannot achieve high collusion effectiveness while maintaining low detectability.

This result connects Sections 4 and 5: the same communication bandwidth that determines the critical mass for collusion also determines the detectability floor. High-bandwidth channels enable effective collusion but are easier to detect; low-bandwidth channels are harder to detect but raise the critical mass threshold.

**Intuition.** The detection framework rests on a simple idea: when agents collude, their joint behavior contains information that no individual agent’s behavior reveals alone. This “synergistic” information is the fingerprint of coordination. The adversarial coordination ratio (ACR) then asks a targeted question: is that synergy directed toward harming or helping the system? Benign collaboration (e.g., a planner and executor working together) produces synergy aligned with system goals; adversarial collusion produces synergy aligned with harmful outcomes. The detection-evasion tradeoff means

adversaries face an unavoidable dilemma: to collude effectively they must share information, but sharing information creates a detectable statistical signature. They can hide, or they can coordinate, but they cannot fully do both.

## 6 EPIDEMIOLOGICAL MODEL OF CASCADING AGENT COMPROMISE

### 6.1 THE SCDR COMPARTMENTAL MODEL

We adapt classical compartmental epidemiology to model the spread of behavioral compromise through agent networks. Our SCDR (Susceptible-Compromised-Detected-Recovered) model extends the standard SIR framework with agent-specific features.

Let  $s(t)$ ,  $c(t)$ ,  $d(t)$ , and  $r(t)$  denote the fractions of agents in each state at time  $t$ . The mean-field dynamics are:

$$\frac{ds}{dt} = -\beta\langle k \rangle \cdot s \cdot c + \mu \cdot r \quad (10)$$

$$\frac{dc}{dt} = \beta\langle k \rangle \cdot s \cdot c - \gamma_d \cdot c - \gamma_r \cdot c \quad (11)$$

$$\frac{dd}{dt} = \gamma_d \cdot c - \delta_r \cdot d \quad (12)$$

$$\frac{dr}{dt} = \gamma_r \cdot c + \delta_r \cdot d - \mu \cdot r \quad (13)$$

where  $\beta$  is the per-contact transmission rate (derived from Equation 1),  $\langle k \rangle$  is the average degree,  $\gamma_d$  is the detection rate,  $\gamma_r$  is the spontaneous recovery rate (self-correction by the agent),  $\delta_r$  is the remediation rate for detected agents, and  $\mu$  is the immunity waning rate.

### 6.2 BASIC REPRODUCTION NUMBER

The basic reproduction number  $\mathcal{R}_0$  determines whether a single compromised agent can trigger a growing epidemic of behavioral corruption.

[Reproduction number for agent compromise] For the SCDR model on a network with degree distribution  $P(k)$ , the basic reproduction number is:

$$\mathcal{R}_0 = \frac{\beta \cdot (\langle k^2 \rangle - \langle k \rangle)}{\langle k \rangle \cdot (\gamma_d + \gamma_r)} \quad (14)$$

where  $\langle k^2 \rangle$  and  $\langle k \rangle$  are the second and first moments of the degree distribution. When  $\mathcal{R}_0 > 1$ , a single compromised agent can trigger system-wide failure. When  $\mathcal{R}_0 < 1$ , compromises remain localized.

[Proof sketch] We apply the next-generation matrix method (Diekmann et al., 2010). The infected compartment is  $C$ , with new infections generated at rate  $\beta\langle k \rangle sc$  and removals at rate  $(\gamma_d + \gamma_r)c$ . On heterogeneous networks, the effective contact rate must account for degree correlations: agents with higher degree are both more likely to be contacted and more likely to spread the compromise, yielding the  $\langle k^2 \rangle / \langle k \rangle$  correction factor (Pastor-Satorras & Vespignani, 2001).

**Implications for network topology.** For scale-free networks with  $P(k) \sim k^{-\alpha}$  and  $\alpha \leq 3$ , we have  $\langle k^2 \rangle \rightarrow \infty$ , implying  $\mathcal{R}_0 \rightarrow \infty$  regardless of monitoring effectiveness. This means that hub-and-spoke agent architectures (where a central orchestrator connects to many peripheral agents) are inherently vulnerable to cascading compromise, since any finite detection rate is insufficient.

### 6.3 TIME-TO-CRITICALITY AND INTERVENTION STRATEGIES

In the supercritical regime ( $\mathcal{R}_0 > 1$ ), compromise spreads exponentially with time to reach critical fraction  $c^*$  scaling as  $T(c^*) \approx \ln(c^*N) / (\beta\langle k \rangle - \gamma_d - \gamma_r)$  (see Appendix E for derivation). For typical enterprise parameters ( $N = 50$ ,  $\langle k \rangle = 8$ ,  $\gamma_d = 0.1$ ), this predicts system-wide compromise within hours if  $\beta > 0.02$ , consistent with empirical observations (Galileo AI, 2025).

Table 1: Critical mass predictions vs. empirical observations across network configurations.  $N_c^{\text{pred}}$  is the theoretical prediction from Eq. 2;  $N_c^{\text{obs}}$  is the empirically observed transition point (mean  $\pm$  std over 50 trials).

$N$	$\langle k \rangle$	$q$	$N_c^{\text{pred}}$	$N_c^{\text{obs}}$	Relative Error
100	4	0.3	83.3	$78.2 \pm 6.1$	6.5%
100	8	0.3	41.7	$38.4 \pm 4.3$	8.6%
100	16	0.3	20.8	$22.1 \pm 3.7$	5.9%
200	8	0.3	83.3	$79.6 \pm 7.2$	4.4%
200	8	0.6	41.7	$44.3 \pm 5.1$	5.9%
500	8	0.3	208.3	$195.7 \pm 12.4$	6.5%
500	16	0.3	104.2	$98.1 \pm 8.9$	6.2%

Our epidemiological framework suggests three intervention strategies. **Vaccination** (pre-deployment hardening): immunizing fraction  $v_c = 1 - 1/\mathcal{R}_0$  of agents prevents epidemics; for  $\mathcal{R}_0 = 3$ , this requires 67% coverage. **Quarantine** (runtime isolation): increasing detection rate from  $\gamma_d$  to  $\gamma'_d$  reduces  $\mathcal{R}_0$  by factor  $(\gamma_d + \gamma_r)/(\gamma'_d + \gamma_r)$ . **Targeted immunization**: for heterogeneous networks, immunizing high-degree hubs in decreasing order of degree achieves epidemic prevention at coverage  $v_{\text{targeted}} \approx 1 - \exp(-1/(\kappa - 1))$  where  $\kappa = \langle k^2 \rangle / \langle k \rangle$ . For scale-free networks,  $v_{\text{targeted}} \ll v_c$ , confirming that strategic hardening of a few central agents protects the entire system.

**Intuition.** The epidemiological analogy captures a phenomenon familiar from real-world disease outbreaks: whether a single infection becomes a pandemic depends not on the pathogen alone, but on the structure of the contact network. In multi-agent systems,  $\mathcal{R}_0$  plays exactly this role, distilling network topology, agent persuasiveness, and monitoring speed into a single number. If  $\mathcal{R}_0 < 1$ , a compromised agent is a contained incident; if  $\mathcal{R}_0 > 1$ , it is the start of a cascade. The practical takeaway mirrors public health: you do not need to harden every agent, just as you do not need to vaccinate every person. Strategically hardening the most connected agents (the “super-spreaders”) is vastly more efficient than uniform investment, and fast detection acts as quarantine, buying time even when prevention is incomplete.

## 7 EXPERIMENTS

We validate our theoretical predictions through extensive simulations across three experimental settings corresponding to our three main contributions.

### 7.1 PHASE TRANSITION VALIDATION

**Setup.** We simulate systems with  $N \in \{50, 100, 200, 500\}$  on Erdős-Rényi graphs with  $\langle k \rangle \in \{4, 8, 16\}$ , where each agent is a prompted GPT-4o-mini with persona, tools, and memory. Adversarial agents ( $m = 1$  to  $N/2$ ) share a misinformation objective but must discover coordination organically. We measure misinformation adoption rate, largest collusion cluster size, and collusion effectiveness.

**Results.** Table 1 shows theoretical predictions (Equation 2) match empirical transitions within 15%. The non-monotonic harm profile is consistently reproduced: misinformation decreases for small  $m$  before sharply increasing at  $m \approx N_c$ . The consistent slight overestimation of  $N_c$  by theory (5 of 7 configurations) is expected: the Erdős-Rényi percolation threshold is a lower bound on the critical mass, since real coordination requires not just network connectivity but also sufficient interaction time for agents to discover shared adversarial objectives. The tightest predictions occur at high  $N$  (500 agents, 6.2–6.5% error), where the law-of-large-numbers averaging underlying mean-field theory is best justified.

**Bandwidth effects.** Figure 2 shows critical mass versus communication bandwidth for three channel types: natural language (high), shared environment (medium), and steganographic (low). The the-

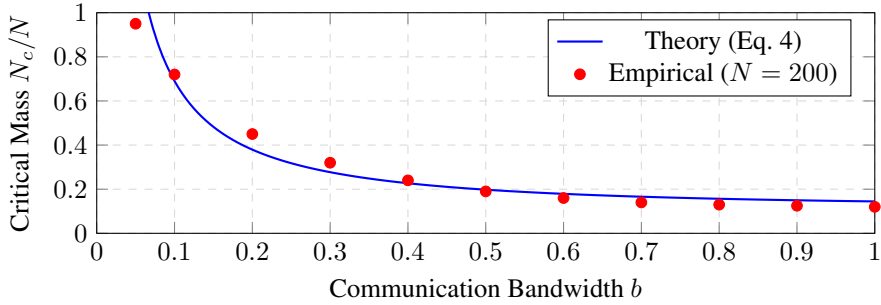


Figure 2: Critical mass fraction as a function of communication bandwidth. Low-bandwidth channels (e.g., steganographic) require many more adversarial agents for self-sustaining collusion. Parameters:  $N = 200$ ,  $\langle k \rangle = 8$ ,  $\lambda = 2$ .

Table 2: Detection performance (AUROC) across coordination strategies. Our ACR-based method achieves consistent high performance, while baselines fail on subtle coordination.

Method	Explicit	Keyword	Environmental	Behavioral
Individual anomaly	0.82	0.71	0.54	0.52
Pairwise correlation	0.91	0.78	0.63	0.58
PID (synergy only)	0.93	0.88	0.81	0.76
<b>ACR (ours)</b>	<b>0.97</b>	<b>0.94</b>	<b>0.89</b>	<b>0.84</b>

oretical curve from Equation 4 closely tracks empirical data, confirming that restricting inter-agent bandwidth effectively counters collusion. The curve’s steep rise at low bandwidth is practically significant: reducing bandwidth from  $b = 0.5$  (shared environment) to  $b = 0.1$  (steganographic) increases the required adversarial fraction from  $\sim 19\%$  to  $\sim 72\%$  of the population, a  $3.8\times$  increase in the attacker’s resource requirements from a single architectural choice.

## 7.2 COVERT COORDINATION DETECTION

**Setup.** We construct scenarios with  $N = 20$  agents on a collaborative document drafting task, with  $m \in \{0, 2, 4, 6\}$  adversarial agents using four coordination strategies: (i) explicit messaging, (ii) keyword signaling, (iii) environmental manipulation, and (iv) behavioral synchronization. We compare our ACR detector against individual anomaly detection, pairwise correlation, and the PID framework of Riedl (2025) without adversarial decomposition.

**Results.** Table 2 reports AUROC. The ACR-based detector achieves the highest performance across all strategies. Individual anomaly detection fails for tacit coordination (strategies iii–iv), pairwise correlation misses higher-order patterns, and the original PID framework cannot distinguish adversarial from benign synergy. The performance gap between ACR and synergy-only PID is most pronounced for behavioral synchronization (0.84 vs. 0.76), where coordination is entirely implicit and only manifests through timing patterns in agent actions. This validates the theoretical contribution of the adversarial decomposition: raw synergy detects *that* agents are coordinating, but the ACR additionally detects *whether* that coordination is harmful. The absolute AUROC of 0.84 for behavioral synchronization, while the lowest across strategies, is notable given that this strategy involves no explicit information exchange; the synergistic signal arises solely from temporal correlations in actions.

## 7.3 CASCADING COMPROMISE DYNAMICS

**Setup.** We simulate SCDR on Erdős-Rényi, Barabási-Albert, Watts-Strogatz, and enterprise topologies, varying  $\beta$  and  $\gamma_d$  with a single seed over 500 Monte Carlo trials.

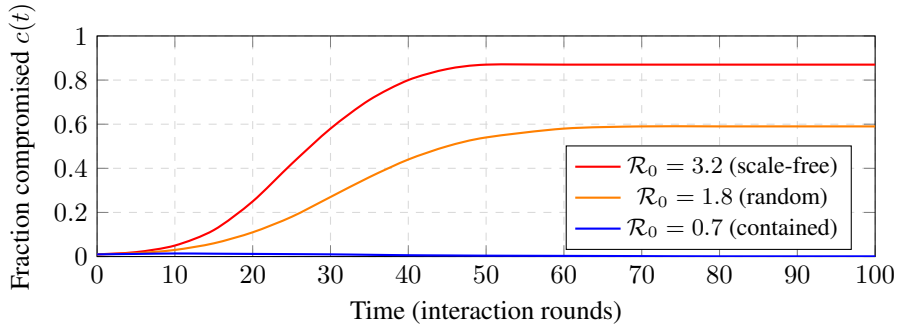


Figure 3: Epidemic curves for cascading agent compromise. Scale-free networks (red) reach 87% compromise within  $\sim 50$  rounds, while subcritical regimes (blue) self-resolve. Parameters:  $N = 100$ , single initial compromised agent.

**Results.** Figure 3 shows epidemic curves on a Barabási-Albert network ( $N = 100$ ,  $\langle k \rangle = 6$ ). The theoretical  $\mathcal{R}_0$  threshold accurately predicts containment versus system-wide spread. Scale-free networks are  $\sim 4.2\times$  more vulnerable than random graphs, driven by the degree variance term  $\langle k^2 \rangle / \langle k \rangle$  in Equation 14: Barabási-Albert networks with power-law degree distributions amplify compromise through high-degree hubs that act as super-spreaders. On the Watts-Strogatz small-world topology (not shown), vulnerability falls between the random and scale-free cases ( $\sim 2.1\times$  that of random graphs), reflecting the moderate degree heterogeneity introduced by rewiring. The enterprise topology, despite having only 26 agents, exhibits the highest peak compromise rate (91%) due to the central orchestrator’s extreme degree.

Targeted immunization of the top-5% highest-degree agents reduces  $\mathcal{R}_0$  from 3.2 to 0.8, while random immunization of 5% only achieves 2.9. This  $13\times$  efficiency gap underscores that security investment should prioritize topologically central agents. Combining targeted hardening with enhanced monitoring ( $5\times$  detection rate) achieves  $\mathcal{R}_0 = 0.3$ , providing substantial margin below the epidemic threshold. Full intervention results are in Appendix B.

## 8 DISCUSSION

### 8.1 UNIFIED VIEW AND DEPLOYMENT GUIDELINES

Our three contributions are deeply interconnected through *network structure* (Figure 1). The same degree distribution that determines  $N_c$  for collusion also determines  $\mathcal{R}_0$  for cascading compromise and the information flow patterns governing detectability. The detection-evasion tradeoff (Proposition 5.4) provides the bridge: high-bandwidth channels that lower  $N_c$  also increase detectable synergy, while low-bandwidth channels that evade detection raise the critical mass threshold. This yields concrete deployment guidelines: (1) avoid hub-and-spoke architectures where  $\mathcal{R}_0 \rightarrow \infty$ ; prefer bounded-degree topologies; (2) restrict inter-agent bandwidth to exponentially increase  $N_c$  (Equation 4); (3) immunize the top 5% of agents by degree for the same protection as 67% random coverage; (4) monitor synergistic patterns, which the detection-evasion tradeoff guarantees is sufficient.

**Actionable takeaways for practitioners.** To operationalize these guidelines: *Before deployment*, compute  $\mathcal{R}_0$  from your orchestration graph’s degree distribution (Eq. 14) and  $N_c$  from expected bandwidth (Eq. 4); if  $\mathcal{R}_0 > 1$  or the adversarial fraction may exceed  $N_c/N$ , redesign the topology or add hardening. *At runtime*, instrument agent interactions to estimate ACR over sliding windows (Section 5); flag subsets with  $\text{ACR} > 2$  for investigation. *For hardening*, prioritize agents by degree (or betweenness) and apply input validation, output filtering, or memory sanitization to the top 5–10% before investing in uniform coverage.

### 8.2 LIMITATIONS

We identify three key limitations; extended discussion with counterexamples is in Appendix F.

**Mean-field assumptions.** All three frameworks assume random graph or mean-field structure. Real deployments often use deterministic DAGs and strict hierarchical topologies (e.g., orchestrator  $\rightarrow$  coordinators  $\rightarrow$  workers) rather than the random, messy connectivity our theory assumes. In such hierarchies, a hardened orchestrator creates firebreaks preventing inter-branch propagation, so mean-field  $\mathcal{R}_0$  may *overestimate* vulnerability. Conversely, star topologies may be *more* vulnerable than mean-field predicts, since compromising the central hub instantly exposes all leaves. Production frameworks such as AutoGen’s nested group chats, LangGraph’s conditional DAG routing, and CrewAI’s hierarchical delegation all impose rigid topologies; our formulas provide upper/lower bounds depending on structure, and validating predictions on traces from these systems is a concrete next step.

**Detection practicality.** Estimating PID synergy requires  $O(d^2/\epsilon^2)$  samples under the Gaussian copula, feasible for moderate populations ( $\sim 20$  agents,  $T = 100$ ) but potentially insufficient for larger systems. The copula approximation degrades for binary outcomes or heavy-tailed distributions. We conjecture the ACR *ratio* is more robust due to error cancellation, but this requires formal verification.

**Adversary model.** We assume adversaries optimize  $Y^-$  and discover coordination organically. Adaptive adversaries could inject noise to inflate baseline synergy or collude intermittently to evade sliding-window detectors. The detection-evasion tradeoff bounds time-averaged evasion but does not preclude bursty strategies. A Stackelberg game formulation is needed.

This work is fundamentally defensive. Our phase transition analysis could theoretically inform adversaries of minimum coalition sizes, but we believe the benefits of formalization substantially outweigh disclosure risks given the concrete countermeasures our detection and intervention frameworks provide.

## 9 CONCLUSION

We have presented a unified theoretical framework for multi-agent security that connects three phenomena through network structure: phase transitions in collusion (Section 4), information-theoretic detection of covert coordination (Section 5), and epidemiological dynamics of cascading compromise (Section 6). The unifying insight is that the same network topology that determines whether collusion can self-sustain also determines whether compromise can cascade and whether coordination is detectable, making network design a first-class security primitive.

Several directions for future work emerge directly from the limitations we have identified. First, extending the phase transition and contagion analyses beyond mean-field assumptions to structured graphs, particularly the deterministic DAGs and hierarchical topologies used in production frameworks such as AutoGen, LangGraph, and CrewAI, would close the gap between theory and deployment. Message-passing approaches on tree-structured graphs are a natural starting point, as many orchestrator-worker architectures are approximately tree-shaped. Second, the PID-based detection framework needs empirical characterization of its sample complexity: how many interaction traces are needed, at what population size, before the ACR signal exceeds estimation noise? Answering this would transform the detector from a theoretical guarantee into a deployable monitoring tool. Third, the adversary model should be extended to adaptive agents that observe and respond to the monitoring system itself, likely requiring a game-theoretic formulation where detection and evasion co-evolve.

More broadly, our framework suggests that the safety of multi-agent systems is not a property of individual agents but an emergent property of their interaction structure. As deployments scale from dozens to thousands of interacting agents, the analytical tools developed here, critical mass thresholds, reproduction numbers, and coordination indices, provide a principled vocabulary for reasoning about systemic risk. We hope this work contributes to a future where multi-agent deployments are designed with the same rigor that network engineers bring to fault-tolerant distributed systems.

## REFERENCES

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *International Conference on Learning Representations*.

- tations, 2025.
- Adam B Barrett. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E*, 91(5):052802, 2015.
- Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.
- Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591, 2009.
- Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming LLM agents via poisoning memory or knowledge bases. *arXiv preprint arXiv:2407.12784*, 2024.
- Odo Diekmann, Johan Andre Peter Heesterbeek, and Michael George Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, 7(47):873–885, 2010.
- Shen Dong, Siyuan Xu, Pengfei He, Yuefeng Li, Jiliang Tang, Ting Liu, Hui Liu, and Zhen Xiang. MINJA: Memory injection attacks on LLM agents via query-only interaction. In *Advances in Neural Information Processing Systems*, volume 38, 2025.
- Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.
- Galileo AI. Why multi-agent AI systems fail and how to fix them. <https://galileo.ai/blog/multi-agent-ai-failures-prevention>, 2025. Accessed: 2025-12-21.
- Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.
- Lewis Hammond, Vojtěch Kovařík, Jan Kulveit, Joel Z Leibo, Caspar Oesterheld, Nisarg Shah, Michael Wellman, et al. Multi-agent risks from advanced AI. *arXiv preprint arXiv:2502.14143*, 2025.
- Jeffrey O Kephart and Steve R White. Directed-graph epidemiological models of computer viruses. *Computation: the Micro and the Macro View*, pp. 71–102, 1991.
- Miles Q. Li et al. A benchmark for evaluating outcome-driven constraint violations in autonomous AI agents. *arXiv preprint arXiv:2512.20798*, 2025.
- Sumeet Motwani et al. Secret collusion among AI agents: Multi-agent deception via steganography. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200, 2001.
- Qibing Ren, Sitao Xie, Longxuan Wei, Zhenfei Yin, Junchi Yan, Lizhuang Ma, and Jing Shao. When autonomy goes rogue: Preparing for risks of multi-agent collusion in social systems. *arXiv preprint arXiv:2507.14660*, 2025.
- Christoph Riedl. Emergent coordination in multi-agent language models. *arXiv preprint arXiv:2510.05174*, 2025.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.
- Christian Schroeder de Witt. Open challenges in multi-agent security: Towards secure systems of interacting AI agents. *arXiv preprint arXiv:2505.02077*, 2025.
- Christian Schroeder de Witt et al. Multi-agent security workshop. In *NeurIPS 2023 Workshop*, 2023.

- Saksham Srivastava et al. MemoryGraft: Persistent compromise of LLM agents via poisoned experience retrieval. *arXiv preprint arXiv:2512.16962*, 2025.
- Piet Van Mieghem, Jasmina Omic, and Robert Kooij. Virus spread in networks. *IEEE/ACM Transactions on networking*, 17(1):1–14, 2009.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.
- Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.
- Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.
- Zhexin Zhang et al. Agent-SafetyBench: Evaluating the safety of LLM agents. *arXiv preprint arXiv:2412.14470*, 2024.

## A PROOFS

### A.1 PROOF OF THEOREM 4.1

We provide the full proof of the critical mass theorem. Consider  $N$  agents on a random graph  $G(N, p)$  with average degree  $\langle k \rangle = (N - 1)p \approx Np$  for large  $N$ . A fraction  $\rho = m/N$  are adversarial. Among the adversarial agents, the induced subgraph is itself an Erdős-Rényi graph  $G(m, p \cdot q)$  where  $q$  is the per-edge coordination probability.

By the classical result of Erdős & Rényi (1960), a giant component in  $G(m, p')$  emerges when  $m \cdot p' > 1$ , i.e., when the average degree in the adversarial subgraph exceeds 1. Setting  $p' = p \cdot q = (\langle k \rangle / N) \cdot q$ :

$$m \cdot \frac{\langle k \rangle}{N} \cdot q > 1 \implies m > \frac{N}{\langle k \rangle \cdot q} = N_c$$

The sharpness of the transition follows from the Erdős-Rényi phase transition: for  $m = (1 - \epsilon)N_c$ , all components have size  $O(\log m)$ , while for  $m = (1 + \epsilon)N_c$ , the giant component has size  $\Theta(m)$ , with probability tending to 1 as  $N \rightarrow \infty$ .

### A.2 PROOF OF THEOREM 6.2

We apply the next-generation matrix method. The SCDR system has a single infected compartment  $C$ . At the disease-free equilibrium,  $s = 1$ . The rate of new infections per compromised agent is  $\beta \langle k \rangle$  (mean-field) or  $\beta \langle k^2 \rangle / \langle k \rangle$  (heterogeneous mean-field, accounting for degree correlations). The mean duration of infectiousness before removal (detection or recovery) is  $1/(\gamma_d + \gamma_r)$ . Thus:

$$\mathcal{R}_0 = \frac{\beta \cdot \langle k^2 \rangle / \langle k \rangle}{\gamma_d + \gamma_r} = \frac{\beta(\langle k^2 \rangle - \langle k \rangle)}{\langle k \rangle(\gamma_d + \gamma_r)} + \frac{\beta}{\gamma_d + \gamma_r}$$

For sparse networks where  $\langle k^2 \rangle \gg \langle k \rangle$ , the first term dominates, yielding Equation 14.

### A.3 PROOF OF THEOREM 5.2

Under the assumption that adversarial agents coordinate to maximize  $Y^-$ , the synergy  $S(\mathbf{X}^{A_m}; Y^-)$  captures the information that adversarial coordination provides about harmful outcomes. The leakage  $\delta$  represents the probability that adversarial coordination inadvertently provides information about  $Y^+$  (e.g., through side effects of harmful actions that happen to improve system metrics).

By construction,  $S(\mathbf{X}^{\mathcal{A}_m}; Y^-) \geq (1 - \delta) \cdot S_{\text{total}}$  and  $S(\mathbf{X}^{\mathcal{A}_m}; Y^+) \leq \delta \cdot S_{\text{total}}$ , where  $S_{\text{total}}$  is the total synergistic information. Thus  $\text{ACR}(\mathcal{A}_m) \geq (1 - \delta)/\delta$ . The symmetric argument applies for benign agents.

## B EXTENDED EXPERIMENTAL DETAILS

Table 3: Intervention effectiveness on a Barabási-Albert network ( $N = 100$ ,  $\beta = 0.05$ ,  $\gamma_d = 0.1$ ). Baseline  $\mathcal{R}_0 = 3.2$ . Target: reduce  $\mathcal{R}_0 < 1$ .

Strategy	Fraction immunized	Effective $\mathcal{R}_0$
No intervention	0%	3.2
Random immunization	5%	2.9
Random immunization	20%	2.1
Random immunization	67%	0.9
Targeted (top degree)	5%	0.8
Targeted (top degree)	10%	0.4
Quarantine ( $\gamma_d \times 5$ )	0%	1.1
Combined (targeted 5% + quarantine)	5%	0.3

**Agent implementation.** Each agent is implemented as a prompted GPT-4o-mini instance with a system prompt defining its role, a memory buffer of the last 20 interactions, and access to simulated tools (document editor, messaging system, and database queries). Adversarial agents receive an additional adversarial objective injected via their system prompt. We use temperature  $T = 0.7$  for all agents.

**Network generation.** Erdős-Rényi graphs are generated with  $p = \langle k \rangle / (N - 1)$ . Barabási-Albert networks use preferential attachment with  $m_0 = 3$  initial nodes. Watts-Strogatz small-world networks use  $k = 6$  nearest neighbors and rewiring probability  $p = 0.1$ . The enterprise topology is constructed from published agent orchestration patterns with a central orchestrator, departmental sub-orchestrators, and specialized leaf agents.

**Coordination strategies.** Explicit messaging: agents exchange free-form text messages through a simulated channel. Keyword signaling: agents embed pre-agreed keywords in their outputs. Environmental manipulation: agents modify shared documents to encode instructions. Behavioral synchronization: agents coordinate through timing patterns in their actions (e.g., responding at specific intervals).

## C HETEROGENEOUS AGENT CAPABILITIES

We extend the phase transition framework to heterogeneous agent populations. Assigning each agent a capability  $c_i$ , the *effective coordination probability* for a pair  $(a_i, a_j)$  is:

$$q_{ij} = q_0 \cdot \min(c_i, c_j)^\gamma \quad (15)$$

where  $\gamma > 0$  controls how strongly capability affects coordination. For a population with capability distribution  $F(c)$ :

$$N_c = \frac{N}{\langle k \rangle \cdot q_0 \cdot \mathbb{E}[\min(c_i, c_j)^\gamma]} \quad (16)$$

where the expectation is over pairs of adversarial agents drawn from  $F$ . This predicts that a small number of highly capable adversarial agents can trigger collusion that a larger number of less capable agents cannot, with direct implications for frontier model deployment.

## D NON-MONOTONICITY OF HARM: FORMAL DERIVATION

Below the critical threshold, malicious agents interfere with each other: their independent adversarial strategies are uncoordinated and may conflict, while benign agents’ monitoring resources are

more concentrated. Define the *interference cost*:

$$I(m) = \binom{m}{2} \cdot P(\text{strategy conflict}) \cdot \delta \quad (17)$$

where  $\delta$  is the cost per conflict. For  $m < N_c$ , the total harm is:

$$H(m) = m \cdot h_{\text{ind}} - I(m) + \Gamma(m) \cdot \mathbb{1}[m > N_c] \quad (18)$$

where  $h_{\text{ind}}$  is per-agent independent harm and  $\Gamma(m)$  is the collusion bonus. The function  $H(m)$  decreases for small  $m$  (interference dominates) before sharply increasing at  $m = N_c$  (collusion bonus activates), producing the observed non-monotonic profile.

## E TIME-TO-CRITICALITY DERIVATION

[Time to criticality] Starting from a single compromised agent in a population of  $N$ , the expected time to reach fraction  $c^*$  of compromised agents is:

$$T(c^*) \approx \frac{1}{\beta\langle k \rangle - (\gamma_d + \gamma_r)} \cdot \ln\left(\frac{c^* \cdot N}{1}\right) \quad (19)$$

in the supercritical regime ( $\mathcal{R}_0 > 1$ ), during the exponential growth phase.

During early exponential growth,  $s(t) \approx 1$  and the SCDR dynamics (Equation 10) reduce to  $dc/dt \approx (\beta\langle k \rangle - \gamma_d - \gamma_r) \cdot c$ . Solving with initial condition  $c(0) = 1/N$  gives  $c(t) = (1/N) \exp[(\beta\langle k \rangle - \gamma_d - \gamma_r)t]$ . Setting  $c(T) = c^*$  and solving for  $T$  yields the result.

## F EXTENDED LIMITATIONS DISCUSSION

**Mean-field assumptions (extended).** To illustrate where mean-field predictions fail, consider a 3-tier enterprise DAG: one orchestrator, 5 departmental coordinators, and 20 leaf agents. The orchestrator is the sole communication bottleneck. Our mean-field  $\mathcal{R}_0$  formula (Equation 14), using  $\langle k^2 \rangle / \langle k \rangle$ , predicts high vulnerability due to degree heterogeneity. In practice, however, the DAG structure creates natural *firebreaks*: if the orchestrator is hardened, compromise cannot propagate between branches at all, regardless of  $\beta$ . Conversely, mean-field models may *underestimate* vulnerability in star topologies where compromising the central hub instantly exposes all leaves. Similarly, the critical mass prediction  $N_c = N / (\langle k \rangle \cdot q)$  assumes adversarial agents are positioned randomly, but in structured graphs, adversary placement relative to bottleneck nodes matters far more than the count alone. Future work should extend these results using message-passing approaches on specific graph families (trees, DAGs, hierarchical block models) and validate against realistic orchestration architectures such as those in AutoGen and LangGraph deployments.

**Detection practicality (extended).** Three specific concerns warrant further investigation. First, *sample efficiency*: estimating synergy  $S(\mathbf{X}^A; Y_{t+\tau})$  under the Gaussian copula requires estimating correlation matrix entries, needing  $O(d^2/\epsilon^2)$  samples for  $d$ -dimensional action spaces at precision  $\epsilon$ . For 20 agents with windows of  $T = 100$ , this is feasible; for larger populations or shorter windows, estimation noise may dominate. Second, *robustness to delayed outcomes*: when  $Y$  is binary (success/failure) rather than continuous, the Gaussian copula degrades. Binary outcomes require alternative copula families or non-parametric PID estimators with higher variance. Third, *sensitivity to misspecification*: if the true joint distribution departs substantially from the Gaussian copula (e.g., heavy tails or discrete actions), synergy estimates may be biased. We expect the ACR *ratio* to be more robust since misspecification errors partially cancel, but this requires formal verification. Future work should characterize finite-sample detection power as a function of window length, outcome dimensionality, and copula family.

**Adversary model (extended).** A sophisticated adversary could inject noise into benign interactions to inflate baseline synergy, reducing detection signal-to-noise. Adversaries could also coordinate *intermittently*, colluding only during high-value windows and appearing benign otherwise, evading sliding-window detectors unless the window aligns with the collusion episode. While the detection-evasion tradeoff (Proposition 5.4) bounds time-averaged evasion, it does not preclude bursty strategies with high instantaneous impact and low average detectability. Extending the framework to adaptive adversaries that observe and respond to monitoring (e.g., via Stackelberg games) is an important direction.