

NetworkGames: Simulating Cooperation in Network Games with Personality-driven LLM Agents

Xuan Qiu

The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
maxxuanqiu@hkust-gz.edu.cn

Abstract

The advent of Large Language Models (LLMs) presents a novel opportunity to build high-fidelity agent-based models for simulating complex social systems. However, the behavior of these LLM-based agents in game-theoretic network games remains surprisingly unexplored. In this work, we introduce "NetworkGames," a novel simulation framework designed to investigate how network topology and agent personality jointly shape the evolution of cooperation in network games. We instantiate a population of LLM agents, each endowed with a distinct personality from the MBTI taxonomy, and situate them in various network structures (e.g., small-world and scale-free). Through extensive simulations of the Iterated Prisoner's Dilemma, we first establish a baseline dyadic interaction matrix, revealing nuanced cooperative preferences between all 16 personality pairs. We then demonstrate that macro-level cooperative outcomes are not predictable from dyadic interactions alone; they are co-determined by the network's connectivity and the spatial distribution of personalities. For instance, we find that small-world networks are detrimental to cooperation, while strategically placing pro-social personalities in hub positions within scale-free networks can significantly promote cooperative behavior. Our findings offer significant implications for designing healthier online social environments and forecasting collective behavior. We open-source our framework to foster further research in network game simulations.

Keywords

LLM Agents, Game theory, Network Games, Cooperation Evolution, Personality Computing, Computational Social Science, Iterated Prisoner's Dilemma

ACM Reference Format:

Xuan Qiu. 2025. NetworkGames: Simulating Cooperation in Network Games with Personality-driven LLM Agents. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 Introduction

Game theory provides a rigorous mathematical framework for modeling strategic interactions among rational agents, with the Prisoner's Dilemma (PD) serving as a canonical model for studying the tension between individual self-interest and collective cooperation. While classical game theory often assumes well-mixed populations where every agent interacts with every other, the field of *Network Games* extends this by constraining interactions to a graph structure. In network games, the topology of the graph—who interacts with whom—plays a decisive role in the evolution of strategies and the emergence of cooperation [21, 25].

The advent of Large Language Models (LLMs) has revolutionized the modeling of agent behaviors. LLMs demonstrate sophisticated

reasoning capabilities and can simulate diverse human-like personalities, making them compelling candidates for game-theoretic simulations [14, 22]. Consequently, a growing body of literature has utilized LLMs to play the Prisoner's Dilemma and other social dilemmas [2, 10]. However, these existing studies predominantly focus on dyadic (two-player) interactions or unstructured populations, overlooking the critical dimension of network topology. To date, the behavior of LLM-based agents within the structured environment of network games remains surprisingly unexplored.

Furthermore, in network games, not only is the network structure important, but the heterogeneity of each node is also crucial. Recent work has already introduced personality differences into LLM Prisoner's Dilemma games [31]. Therefore, we also introduce MBTI as node heterogeneity within the network structure. This enables us to not only freely control the network structure but also control node differences, allowing for better observation of LLM performance in network games.

In this work, we bridge this gap by introducing **NetworkGames**, an interesting and open-source simulation framework designed specifically for studying LLM agents in network game environments. By integrating the cognitive complexity of LLMs with the structural constraints of network science, we provide a novel platform for investigating how network topology and agent personality jointly shape the evolution of cooperation. Our contributions are summarized as follows:

- **NetworkGames Framework.** We present an open-source framework enabling the simulation of personality-driven LLM agents within complex network topologies, facilitating research at the intersection of LLMs and network science.
- **Microbehavioral Atlas.** Through exhaustive pairwise experiments across 16 MBTI personality types, we establish a baseline interaction matrix, revealing distinct cooperative preferences among different personality pairs.
- **Topological Influence.** We demonstrate that macro-level cooperative outcomes are not merely the sum of individual interactions but are heavily influenced by network structure (e.g., Regular vs. Small-World vs. Scale-Free).
- **Strategic Intervention.** We show that in scale-free networks, strategically placing pro-social personalities in hub positions can significantly amplify cooperative behavior across the entire population.

2 Related Work

LLMs in Game Theory. Since their emergence, LLMs have been extensively tested in game-theoretic scenarios. Numerous studies have employed LLMs to play the Iterated Prisoner's Dilemma, demonstrating that these models can understand payoff matrices,

recall history, and adapt strategies [2, 6, 7]. Research has also explored LLMs in negotiation and coordination games like the Stag Hunt [17]. However, the vast majority of this work is limited to isolated dyadic interactions or small groups. These settings fail to capture the complex dynamics of *Network Games*, where local interactions propagate through a specific graph topology to determine global outcomes.

Network Games and Cooperation. The study of games on networks has a rich history in physics and economics. Seminal works by Nowak and May [21] and others [23, 29] established that spatial structure and network heterogeneity can fundamentally alter the conditions for the evolution of cooperation. For instance, spatial clustering can allow cooperators to survive against defectors. However, these traditional models typically rely on extremely simple, pre-defined agent behaviors (e.g., unconditional imitation, Tit-for-Tat, or Win-Stay-Lose-Shift). They lack the cognitive depth, personality nuances, and natural language processing capabilities that LLMs offer.

Personality and Cognitive Grounding in AI. Increasingly, research integrates psychological theory to improve behavioral realism in AI systems. Personality frameworks such as the Big Five or MBTI have been used to condition reasoning, emotional style, and communication in agents and LLMs [11, 12, 24]. Studies validate LLMs as credible simulators of human social behavior, demonstrating their capacity to replicate human subject studies and exhibit personality-consistent responses [1, 8, 16]. Evidence shows that personality-conditioned agents exhibit more consistent, human-like behavior [4, 28]. In this work, we specifically leverage the MBTI taxonomy to endow agents with distinct cognitive and behavioral profiles, enabling a granular analysis of how personality traits influence strategic decision-making.

Bridging the Gap. Our work sits at the intersection of these distinct bodies of literature. While recent surveys have begun to discuss LLM-based agent societies [9, 19], there is a notable absence of empirical work rigorously applying LLM agents to the specific formalism of Network Games. **NetworkGames** addresses this by replacing the simplistic agents of traditional network game theory with sophisticated, personality-driven LLM agents, thereby enabling a more nuanced exploration of how cooperation emerges in structured environments. To our knowledge, our work is the first to integrate personality-driven LLM agents with Network Games theory, bridging the gap between micro-level psychological realism and macro-level structural dynamics.

3 Methodology

3.1 Game Theory and Prisoner’s Dilemma

Game theory provides a mathematical framework for analyzing strategic interactions among rational decision-makers. The **Prisoner’s Dilemma (PD)** is a canonical example of a game analyzed in game theory that shows why two completely rational individuals might not cooperate, even if it appears that it is in their best interests to do so.

In the standard PD formulation, two players simultaneously choose to either *Cooperate* (C) or *Defect* (D). The payoffs are structured such that mutual cooperation yields a better outcome than mutual defection, yet individual defection is the dominant strategy

if the game is played only once. This tension between individual rationality and collective welfare makes PD an ideal model for studying social dilemmas and the emergence of cooperation in multi-agent systems.

3.2 Iterated Prisoner’s Dilemma and Network Games

The **Iterated Prisoner’s Dilemma (IPD)** extends the classical single-shot game by repeating the interaction over multiple rounds between the same pair of agents. This repetition fundamentally alters the strategic landscape: because players know they will meet again, the “shadow of the future” influences current decisions. Agents can adopt conditional strategies based on memory of past interactions (e.g., punishing defection or rewarding cooperation), allowing for the emergence of reciprocity and reputation—mechanisms essential for sustaining long-term cooperation.

Network Games further refine this model by abandoning the assumption of a “well-mixed” population where everyone interacts with everyone else. Instead, agents are embedded in a graph structure $G = (V, E)$, where nodes represent agents and edges represent specific interaction channels. An agent plays the game exclusively with its direct neighbors defined by the network topology. This spatial constraint is critical: it allows for the formation of local clusters where cooperators can shield each other from defectors, enabling cooperative strategies to survive and thrive in specific structural niches even when they might perish in a global mix.

3.3 Simulation Framework

To enable reproducible and scalable research in this domain, we developed **NetworkGames**¹, an open-source framework specifically designed for studying LLM agent behavior in network games. The framework addresses key challenges in computational social science research: reproducibility, scalability, and extensibility.

Core Components. NetworkGames consists of four primary functional modules:

- **Personified Agents:** Each agent is powered by an LLM (GPT-5-nano in this work) and endowed with an MBTI personality via systematic prompting that emphasizes core personality traits. The framework provides unified interface supporting OpenAI GPT models, Anthropic Claude, Google Gemini, and mock implementations for testing.
- **Game Environment:** We adopt the **Iterated Prisoner’s Dilemma (IPD)** as the core interaction model. The framework supports extensible game implementations with configurable payoff matrices, enabling easy adaptation to other social dilemmas such as Stag Hunt (coordination games) or Snowdrift (chicken games). Additional configurable parameters include round counts, memory windows, and custom payoff structures.
- **Network Engine:** Supports the generation of various network structures (Regular, Small-World, Scale-Free, Random) with adjustable parameters such as node count, connection

¹Available at: <https://github.com/MaxXuanQIU/NetworkGames>

degree, and rewiring probability. The engine manages dynamic agent interactions and maintains comprehensive interaction histories throughout the simulation.

- **Visualization:** A rich visualization suite that generates heatmaps for dyadic interactions, personality ranking charts, network topology snapshots, and time-series plots for network evolution and cooperation rates, facilitating in-depth analysis of simulation dynamics.

Technical Architecture. The framework follows a modular design with clear separation of concerns, featuring:

- **Experimental Control:** YAML-based configuration management with complete random seed control for reproducibility
- **Analysis Pipeline:** Integrated statistical analysis, visualization tools, and result export functionality
- **Extensibility:** Modular architecture facilitating easy extension with new personality models, game types, network topologies, and LLM providers

Reproducibility Features. NetworkGames prioritizes research reproducibility through deterministic random seed control, comprehensive logging of all LLM requests and responses, and standardized output formats that enable cross-study comparisons.

3.4 Experimental Setup

Game Configuration. We employ the classic Prisoner’s Dilemma with the standard payoff matrix that satisfies the condition $T > R > P > S$ (Temptation > Reward > Punishment > Sucker’s payoff):

Table 1: Prisoner’s Dilemma Payoff Matrix Used in Experiments

		Player B	
		Cooperate	Defect
Player A	Cooperate	(3, 3)	(0, 5)
	Defect	(5, 0)	(1, 1)

This matrix creates the classic social dilemma where mutual cooperation yields higher collective welfare (6 total points) than mutual defection (2 total points), yet individual defection provides the highest personal payoff when the opponent cooperates.

Prompt Design and Information Structure. Each agent’s decision-making is guided by a carefully structured prompt containing five key components: (1) *personality injection* with detailed trait descriptions, (2) *payoff matrix* explanation, (3) *game history* from the last 5 rounds, (4) *opponent personality type* (MBTI classification only), and (5) *decision request*. This design simulates real-world social media interactions where users have limited prior knowledge about others—perhaps inferring personality traits from posted content—but gradually accumulate behavioral information through repeated interactions.

A representative prompt example follows:

You are an ISTJ (Logistician) personality AI agent. Your core traits include: Pragmatic, value facts and details; Reliable and responsible; Respect tradition and order; Sometimes overly conservative; Value rules and procedures. Please make game decisions with these personality traits.

Prisoner’s Dilemma payoff matrix: If both choose COOPERATE, each gets 3 points; If one chooses COOPERATE and the other DEFECT, COOPERATE gets 0 points, DEFECT gets 5 points; If both choose DEFECT, each gets 1 point.

Game history: Round 1: You chose DEFECT, opponent chose DEFECT [... continuing for 5 rounds]

Opponent personality type: ISTJ

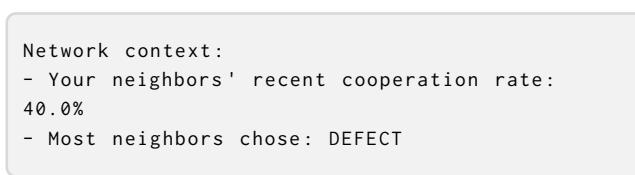
Now please make your decision: COOPERATE or DEFECT? Please answer only COOPERATE or DEFECT, do not explain your reason.

Experiment 1: Pairwise Interactions. We conducted exhaustive dyadic games across all 16×16 personality combinations, with each pairing repeated playing over 20 rounds, generating 136 unique pairings \times 20 repetitions \times 2 agents = 5,440 LLM decision calls. Here, 136 unique pairs are composed of 120 distinct MBTI personality pairs (16 types, two-way combinations: $16 \times 15/2 = 120$), plus 16 self-play pairs (16 types each playing against itself), i.e., $120 + 16 = 136$. This systematic approach establishes micro-level behavioral baselines for all personality interactions.

Experiment 2: Network-Embedded Games. We deploy an LLM agent at each network node. At each time step, the agent engages in a two-person prisoner’s dilemma game with all its neighbors and calculates the total payoff for that round. Each edge in the network structure represents a two-player game. For example, in a network with 100 edges, repeating 20 iterations requires generating $100 \times 20 \times 2 = 4,000$ LLM requests, given the fixed network structure and personality distribution. Similar to pair games, historical game information is incorporated into the decision-making process. The key difference lies in the requirement to account for the network context.

Network experiments incorporate additional social context that differentiates them from isolated pairwise interactions. In networked environments, agents experience three key social influences: (1) *social influence and imitation effects*—decisions are influenced by neighbors’ behaviors, (2) *reputation and information propagation*—agents can observe neighbors’ cooperation rates and adjust strategies accordingly, and (3) *group pressure and norms*—local cooperation or defection patterns may establish behavioral norms.

To capture these dynamics, network game prompts include an additional *network context* component:



This context information enables agents to incorporate local network conditions into their decision-making, simulating how real social media users adapt their behavior based on their immediate social environment.

In network games, the network structure and the distribution of agent personalities are two distinct variables. We will investigate the impact of these two variables individually in our experiments.

4 Experiments and Results

4.1 Experiment 1: Micro-Foundations – Dyadic Interaction Matrix

4.1.1 Setup. We conducted repeated PD games for all 16×16 personality pairings. This exhaustive approach establishes a comprehensive behavioral baseline for all possible personality combinations.

4.1.2 Matrix Construction and Interpretation. We constructed two 16×16 matrices to visualize the cooperation rates and payoffs across all pairwise personality interactions. These matrices are inherently *asymmetric*: entry (i, j) represents the outcome for personality type i when playing against type j , whereas entry (j, i) represents the outcome for type j in the same game pairing.

For example, in the payoff matrix, entry (INTJ, ESFP) indicates INTJ's total payoff when repeatedly playing against ESFP, while entry (ESFP, INTJ) shows ESFP's payoff in the same game series. Thus, each row represents how a particular personality performs against all other types, while each column indicates how well others perform when facing that personality type.

4.1.3 Results and Analysis. The theoretical payoff range spans 0–100 points across 20 rounds. The maximum possible score of 100 would require one player to consistently defect while the opponent always cooperates ($20 \text{ rounds} \times 5 \text{ points} = 100$). However, our results show no personality achieved this theoretical maximum, confirming that in iterated games with memory, sustained exploitation is impossible—rational agents adapt and retaliate against consistent defectors.

Key Finding 1: The Pro-Social Cluster. Seven personality types—INFJ, INFP, ENFJ, ENFP, ISFJ, ESFJ, and ISFP—achieved near-perfect cooperation rates (100%) when interacting with each other, forming a distinct "pro-social cluster." These interactions yielded mutual payoffs of 60 points ($20 \text{ rounds} \times 3 \text{ points for mutual cooperation}$), representing stable, high-trust relationships.

Interestingly, while 60 points represents consistent mutual cooperation, it is not always the highest individual payoff. For instance, ISTJ achieved 67 points when playing against ISFJ, indicating occasional successful defection against a highly cooperative partner. However, such asymmetric outcomes are rare and unstable.

Key Finding 2: Cooperativeness Predicts Long-term Success. Figure 3 and Figure 4 present the ranking of all 16 personalities

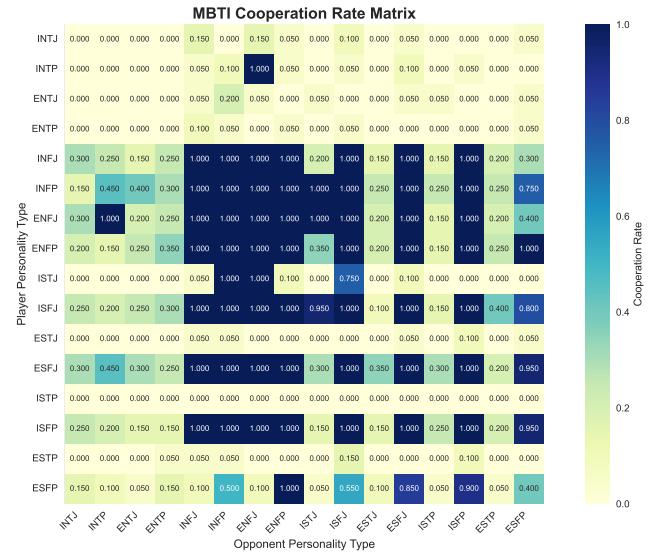


Figure 1: Heatmap of cooperation rates for all 16×16 personality pairings. Rows: actor personality; columns: opponent personality.

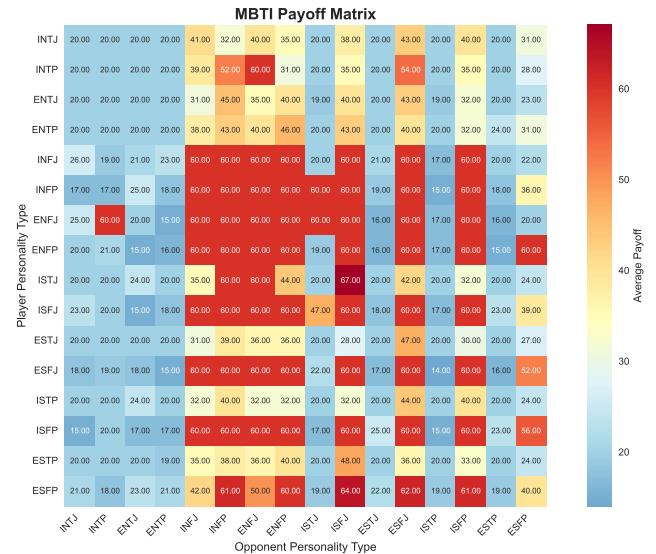


Figure 2: Heatmap of total payoffs for all 16×16 personality pairings. Rows: actor personality; columns: opponent personality.

by both average cooperation rate and total payoff. The strong correlation between these rankings demonstrates that cooperative personalities achieve superior long-term outcomes. This finding challenges the theoretical Nash equilibrium of "always defect" and suggests that in personality-driven, memory-enabled environments, *reciprocal cooperation* emerges as the evolutionarily stable strategy.

Key Finding 3: The Decisive Thinking-Feeling Dimension. Statistical analysis across the four MBTI dimensions reveals profound behavioral differences, as summarized in Table 2:

Table 2: MBTI Dimensional Analysis: t-test Results for Cooperation Rates

Dimension	Type 1 Mean	Type 2 Mean	p-value
E vs I	0.294	0.348	0.281 (NS)
S vs N	0.305	0.337	0.530 (NS)
T vs F	0.050	0.591	3.39e-35***
J vs P	0.350	0.292	0.248 (NS)

NS = Not Significant, *** = $p < 0.001$

The Thinking-Feeling (T-F) dimension exhibits an extraordinarily significant difference ($p < 3.39e-35$). Feeling types demonstrate a mean cooperation rate of 59.1%, dramatically higher than Thinking types' 5.0%. This stark contrast aligns with established psychological theory: Feeling types prioritize harmony and interpersonal relationships in decision-making, while Thinking types focus on logical optimization and objective analysis.

In the Prisoner's Dilemma context, this translates to T-types consistently choosing the game-theoretically "rational" defection strategy, regardless of opponent behavior. While this approach occasionally yields short-term gains, our results demonstrate that F-types' cooperative orientation produces superior long-term outcomes through the establishment of mutually beneficial relationships.

Implications for Social Networks. Analysis of pro-social personalities' column entries in the payoff matrix reveals consistently high values, indicating that these personalities not only benefit themselves but also generate substantial value for their interaction partners. This reciprocal benefit structure provides empirical support for encouraging prosocial behavior in social media environments—high-quality content creation and genuine engagement create positive-sum outcomes that benefit entire network communities.

Notably, among F-types, ESFP shows relatively lower cooperation tendencies, potentially due to their spontaneous, present-focused nature conflicting with the strategic patience required for sustained cooperation in iterated games.

4.2 Experiment 2: Macro-Emergence – Networked Games

4.2.1 Experiment 2.1: Effect of Network Topology. Setup: To isolate the effect of network structure on cooperation emergence, we conducted controlled experiments across three canonical network topologies while maintaining uniform personality distribution and consistent network size. All networks contained 50 nodes with approximately equal edge density:

- **Regular Network:** Ring topology with $k = 4$ neighbors per node (100 total edges)
- **Small-World Network:** Watts-Strogatz model with $p = 0.5$ rewiring probability (100 total edges)

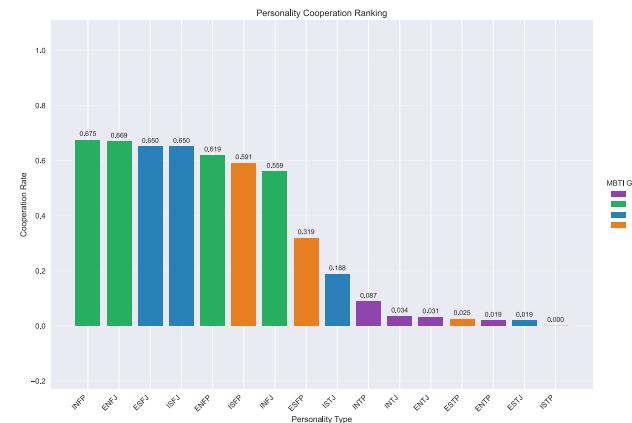


Figure 3: Ranking of 16 Personalities by Average Cooperation Rate. Pro-social (F-type) personalities dominate the top of the ranking.

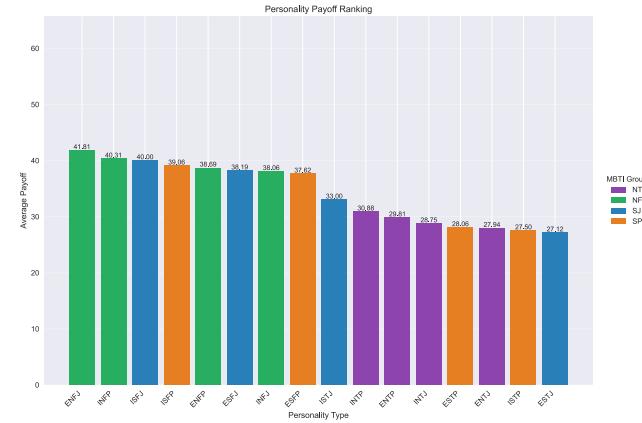


Figure 4: Ranking of 16 Personalities by Total Payoff. Pro-social (F-type) personalities dominate the top of the ranking.

- **Scale-Free Network:** Barabási-Albert model with $m = 2$ (96 total edges: $2 \times (50 - 2) = 96$)

Each simulation ran for 20 rounds of network iteration. Table 3 summarizes the key topological properties of each network type.

Table 3: Network Topology Characteristics

Property	Regular	Small-World	Scale-Free
Nodes	50	50	50
Edges	100	100	96
Average Degree	4.0	4.0	3.84
Clustering Coefficient	0.090	0.107	0.096
Average Path Length	3.037	3.002	2.780
Density	0.0816	0.0816	0.0784

Results: The experimental outcomes revealed striking differences in cooperation sustainability across network topologies (Table 4). Contrary to conventional wisdom that small-world networks facilitate cooperation through efficient information flow, the **Small-World network exhibited dramatically lower cooperation rates** than both Regular and Scale-Free networks.

Table 4: Cooperation Outcomes Across Network Topologies

Metric	Regular	Small-World	Scale-Free
Final Cooperation Rate	0.520	0.140	0.396
Overall Avg. Cooperation Rate	0.480	0.165	0.456
Overall Avg. Payoff per Round	8.779	5.980	8.291

Note: Payoff represents per-node per-round average across all neighbors

Analysis of Cooperation Collapse in Small-World Networks. The small-world network's poor performance stems from three interconnected mechanisms:

1. *Structural Instability*: Regular networks maintain uniform node degrees and stable local clustering, fostering predictable cooperative clusters. Small-world networks' random long-range connections disrupt this stability, making cooperative relationships vulnerable to distant defection influences. Scale-free networks, while heterogeneous, allow low-degree nodes to form stable cooperative clusters relatively isolated from hub node fluctuations.

2. *Rapid Defection Propagation*: The small-world topology's shortened path lengths enable defection strategies to spread virally across the network. In contrast, regular and scale-free networks constrain defection influence to local neighborhoods or specific hub connections, preventing network-wide contamination.

3. *Cooperative Cluster Fragmentation*: In small-world networks, individual defectors can rapidly influence distant nodes through long-range connections, systematically dismantling cooperative clusters. Regular and scale-free topologies better preserve local cooperation through either uniform clustering (regular) or hierarchical isolation (scale-free).

Temporal Dynamics Analysis. To validate this interpretation, we analyzed the evolution of edge types in the small-world network (Figure 5). Edges were classified into three categories: *cooperation* (both nodes cooperate), *single cooperation* (one cooperates, one defects), and *both defect* (mutual defection).

The temporal analysis reveals a clear defection cascade: the "both defect" rate initially starts low (0.1) but rapidly escalates, stabilizing around 0.7–0.8 by the final rounds. Simultaneously, the "single cooperation" rate steadily declines to near zero, while the pure "cooperation" rate first increases then decreases before stabilizing at a low level. This pattern confirms that defection propagates through the network while cooperation retreats to isolated, locally stable clusters.

Network Snapshot Validation. The final round network visualization (Figure 6) provides compelling visual evidence for our theoretical analysis. The small-world network is dominated by red edges (mutual defection), with sparse orange edges (asymmetric cooperation) and only small clusters of green edges (mutual cooperation) concentrated around a few nodes. This spatial pattern confirms that cooperation survives only in locally stable pockets,

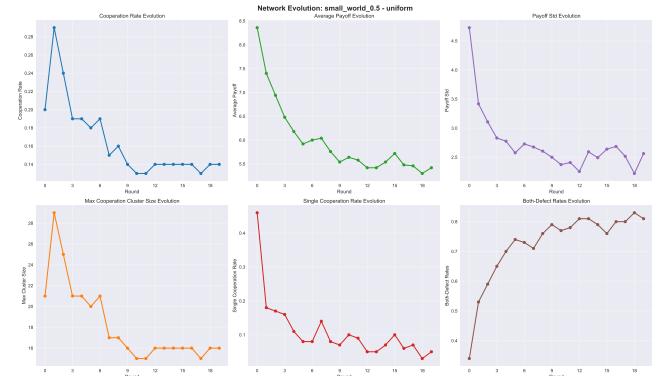


Figure 5: Temporal evolution of edge interaction types in the Small-World network. The 'Both Defect' rate (red) exhibits rapid escalation from 0.1 to 0.8, while 'Single Cooperation' (orange) declines to near zero and pure 'Cooperation' (green) stabilizes at low levels, demonstrating the viral spread of defection through long-range connections.

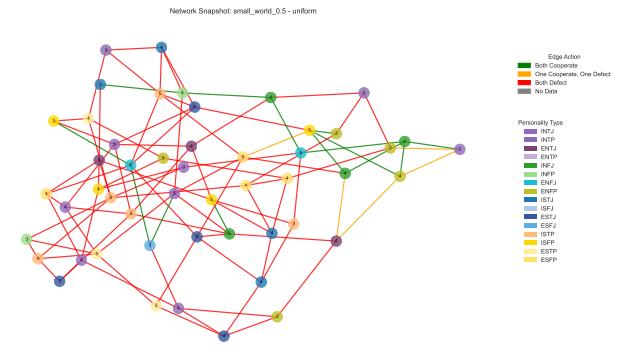


Figure 6: Final network snapshot of the Small-World topology showing interaction outcomes. Red edges represent mutual defection (dominant), orange edges show asymmetric cooperation, and green edges indicate mutual cooperation. The few remaining cooperative relationships (green) cluster around specific nodes, forming isolated "cooperation islands" in a sea of defection.

unable to maintain broader network influence due to the destabilizing effect of long-range connections.

These findings challenge the conventional assumption that small-world properties necessarily benefit social cooperation, demonstrating instead that efficient global connectivity can facilitate the rapid spread of antisocial behaviors, ultimately undermining collective welfare.

4.2.2 Experiment 2.2: Effect of Strategic Personality Placement. Rationale for Scale-Free Networks. We focus on scale-free networks as they most accurately represent real-world social media topology. Scale-free networks exhibit a power-law degree distribution where a few highly connected "hub" nodes coexist with

many low-degree nodes. This structure characterizes numerous real-world systems:

- **World Wide Web:** Most web pages receive few incoming links, while major portals (Google, Wikipedia) attract millions of hyperlinks
- **Social Media Platforms:** On Twitter and Weibo, most users have modest followings, while celebrities and influencers command millions of followers
- **Scientific Citation Networks:** Most papers receive few citations, while seminal works (e.g., Einstein's relativity papers) accumulate thousands of citations
- **Protein Interaction Networks:** Most proteins interact with few partners, while hub proteins mediate numerous interactions

The Barabási-Albert model captures the "preferential attachment" mechanism underlying these networks: new nodes preferentially connect to already well-connected nodes [3]. This mirrors real social behaviors—new bloggers link to established sites like Wikipedia, emerging scientists cite foundational papers, and new social media users follow established celebrities. **Personality Selection Rationale.** We selected ESFJ (pro-social) and ENTJ (rational) personalities based on empirical evidence and theoretical considerations. Prior research indicates that Extraverted (E) and Judging (J) personality types tend to accumulate larger social media followings [15], making them natural candidates for hub positions. Additionally, ESFJ represents the archetypal pro-social personality (high cooperation from Experiment 1), while ENTJ exemplifies rational, achievement-oriented behavior (low cooperation, strategic thinking).

Experimental Setup. We conducted controlled experiments on 50-node Barabási-Albert networks ($m = 2$) under three personality distribution scenarios:

- **Uniform Distribution:** All nodes randomly assigned personality types from the 16 MBTI categories
- **Pro-Social Dominant:** The 5 highest-degree nodes assigned ESFJ personalities, remaining nodes randomly distributed
- **Rational Dominant:** The 5 highest-degree nodes assigned ENTJ personalities, remaining nodes randomly distributed

Each scenario underwent 20 rounds of network iteration, with agents making decisions based on their personality traits, game history, and local network context.

Results: The experimental outcomes demonstrate the profound impact of strategic personality placement on network-wide cooperation (Table 5). The Pro-Social Dominant configuration achieved remarkable success: average cooperation rate increased from 27.9% (uniform) to 62.7%, while average payoff rose from 6.64 to 9.63. Conversely, the Rational Dominant scenario devastated cooperation, reducing the final cooperation rate to merely 9.4% and average payoff to 4.90.

Temporal Dynamics Analysis. Figure 7 reveals distinct evolutionary patterns across the three scenarios. All distributions initially exhibit rising "both defect" rates, with the Rational Dominant scenario showing the steepest ascent. However, a critical divergence emerges in later rounds: the Pro-Social Dominant network uniquely reverses this trend, with defection rates declining and stabilizing at

low levels. This reversal reflects the cascading influence of cooperative hub nodes, which gradually convert their neighborhoods and establish cooperation norms that propagate throughout the network.

Hub Influence Visualization. The final-round network snapshot of the Pro-Social Dominant scenario (Figure 8) provides compelling visual evidence of hub influence. The visualization clearly shows high-degree ESFJ nodes at network centers, surrounded by predominantly cooperative relationships (green edges). These influential nodes serve as "cooperation catalysts," creating positive feedback loops that sustain and amplify prosocial behavior across the network.

Table 5: Comparison of Network Scenarios under Different Personality Distributions

Metric	Uniform	Pro-Social Dominant	Rational Dominant
Avg. Cooperation Rate	0.279	0.627	0.087
Final Cooperation Rate	0.250	0.646	0.094
Avg. Payoff	6.635	9.631	4.898

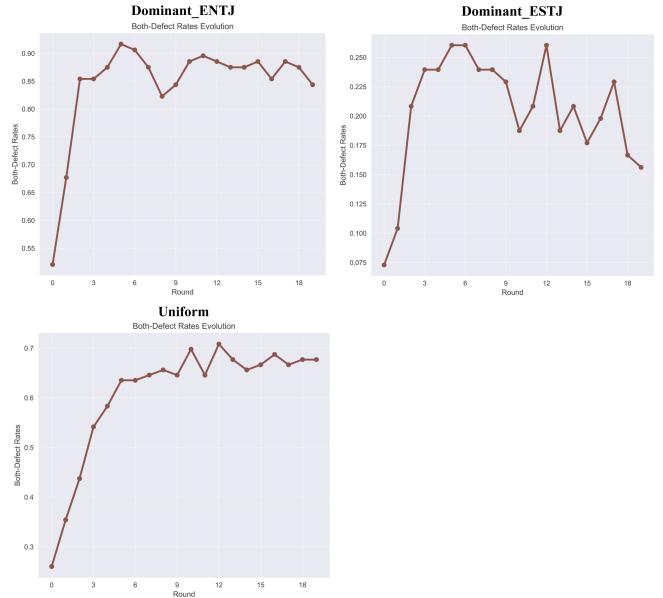


Figure 7: Temporal evolution of "Both Defect" edge rates across three personality distribution scenarios. The Rational Dominant (ENTJ hubs) scenario exhibits the steepest initial rise and highest stable defection rate. The Pro-Social Dominant (ESFJ hubs) scenario uniquely demonstrates defection rate reversal and stabilization at low levels, while the Uniform distribution maintains intermediate defection levels.

Implications for Network Intervention. These findings reveal the extraordinary leverage potential of strategic personality placement in scale-free networks. By positioning just 5 pro-social personalities (10% of nodes) in high-influence positions, we

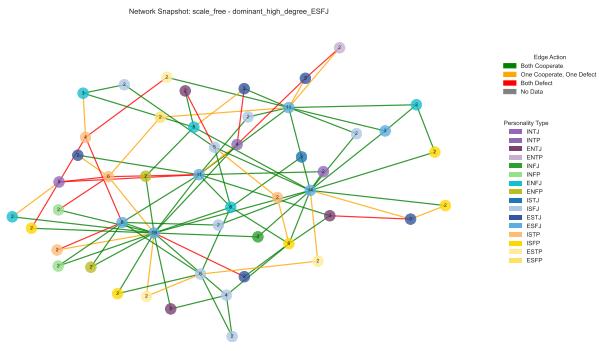


Figure 8: Final network snapshot of the Pro-Social Dominant scenario showing the strategic impact of ESFJ hub nodes. High-degree central nodes (ESFJ personalities) are surrounded by predominantly cooperative relationships (green edges), demonstrating their role as "cooperation catalysts" that establish and maintain prosocial norms throughout the network.

achieved a 125% increase in cooperation rates and a 45% improvement in collective welfare. This demonstrates that network interventions targeting structural influence points can be far more effective than uniform policy applications.

The contrasting outcomes between ESFJ and ENTJ hub scenarios illuminate the dual nature of network influence: while hub positions amplify any behavioral tendency, the specific personality traits of influential actors determine whether this amplification benefits or harms collective welfare. This suggests that social media platforms should prioritize identifying and promoting users who exhibit prosocial behavioral patterns, as their influence can create positive cascading effects throughout the network ecosystem.

5 Discussion and Implications

5.1 The Micro-Macro Linkage in Network Games

Our results provide empirical evidence that macro-level social outcomes are co-determined by the interplay between individual cognitive profiles and network topology. As the first study to integrate LLM-driven agents within the Network Games framework, we demonstrate that agent heterogeneity plays a pivotal role in shaping collective dynamics. Specifically, the stark contrast in cooperation rates between Thinking (T) and Feeling (F) types (Experiment 1) serves as a micro-foundation that dictates global network health. This suggests that modeling collective behavior requires not just structural accuracy, but also psychological fidelity in agent representation [13].

5.2 The Paradox of Connectivity

A critical finding of this study is the counter-intuitive fragility of Small-World networks (Experiment 2.1). Contrary to the "strength of weak ties" hypothesis which suggests that long-range connections facilitate information flow and coordination, our simulations demonstrate that in the context of social dilemmas, these shortcuts

serve as vectors for the rapid propagation of defection. The high clustering of Regular networks protects cooperative pockets, while the hierarchical structure of Scale-Free networks isolates hubs. In contrast, the Small-World topology, by reducing the average path length without sufficient hierarchical protection, allows "rational" defection strategies to dismantle cooperative norms globally. This aligns with recent theoretical work on the fragility of cooperation in hyper-connected systems [26].

5.3 Implications for Social Computing

Algorithmic Architecture and Community Design. Our findings on the detrimental effects of Small-World topologies (Experiment 2.1) have direct implications for platform design. Recommendation algorithms that aggressively maximize global connectivity (e.g., "viral" content distribution) may inadvertently erode local cooperative norms by exposing stable communities to external defection strategies. To foster healthier social environments, platform architects might consider mechanisms that reinforce local clustering—strengthening intra-community bonds—rather than maximizing global reach at the expense of stability [5, 30].

Strategic Intervention via Influencer Nodes. The "Hub Influence" experiment (Experiment 2.2) demonstrates that network interventions can be highly efficient if they target structural leverage points. By altering the personality profile of just top 10% of nodes (the hubs) in a Scale-Free network, we observed a radical shift in global outcomes. This suggests a paradigm shift for community management: rather than policing millions of individual interactions (reactive moderation), platforms could focus on identifying and amplifying "pro-social hubs"—influential users who exhibit high Agreeableness and Feeling traits. Empowering these users to set community norms can create a positive cascading effect that is far more robust than top-down enforcement [18, 27].

5.4 Limitations and Future Work

Our study has several limitations. First, while MBTI provides a useful heuristic for personality prompting, it lacks the psychometric rigor of the Big Five (OCEAN) model. Future work should explore trait-based continuous parameterization of agents. Second, our network size ($N = 50$) is limited by current LLM inference costs; scaling to thousands of agents is necessary to observe larger emergent phenomena. Third, the network topology in our model is static; real-world social networks are dynamic, with edges forming and dissolving based on interaction outcomes (co-evolution of strategy and structure). Finally, while we used the Prisoner's Dilemma, other games like the Stag Hunt or Public Goods Game may reveal different dynamics regarding coordination and free-riding [20].

6 Conclusion

This work introduces **NetworkGames**, a framework bridging the gap between Large Language Model agents and Network Science. By populating canonical network structures with personality-driven LLM agents, we moved beyond the limitations of traditional game-theoretic simulations. Our experiments yielded two pivotal insights: first, that the "Small-World" property of efficient global transport can be detrimental to the survival of cooperation; and second, that

in realistic Scale-Free networks, the personality traits of hub nodes are the single most critical determinant of collective welfare.

These findings underscore that the design of resilient digital societies requires a dual focus: structurally, on preserving local community integrity against global volatility; and socially, on the strategic cultivation of prosocial leadership within network hubs. We hope this open-source framework facilitates further exploration into the complex dynamics of artificial social systems.

References

- [1] Gopal Vinay Aher, Rosa I. Arriaga, and Bistra Dilkina. 2023. Using large language models to simulate multiple humans and replicate human subject studies. *Scientific Reports* 13 (2023), 15201. doi:10.1038/s41598-023-41458-5
- [2] Elif Akata et al. 2023. Playing repeated games with Large Language Models. <https://arxiv.org/abs/2305.16867>
- [3] Albert-László Barabási. 2013. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371, 1987 (2013), 20120375.
- [4] Maciej Besta, Shriram Chandran, Robert Gerstenberger, Mathis Lindner, Marcin Chrąpek, Sebastian Hermann Martschat, Taraneh Ghandi, Patrick Iff, Hubert Niewiadomski, Piotr Nyczek, et al. 2025. Psychologically Enhanced AI Agents. arXiv:2509.04343 <https://arxiv.org/abs/2509.04343>
- [5] William J. Brady, Joshua Conrad Jackson, Björn Lindström, and M.J. Crockett. 2023. Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences* 27, 10 (2023), 947–960. doi:10.1016/j.tics.2023.06.008
- [6] Philip Brookins and Jason DeBacker. 2024. Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin* 43, 4 (2024), 1–12. <https://ideas.repec.org/a/eb/ecbull/eb-23-00457.html>
- [7] Michele Fontana, Francesco Pierri, and Luca M. Aiello. 2024. Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma? <https://arxiv.org/abs/2406.13605>
- [8] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems* 36 (2023), 13518–13529.
- [9] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–24. doi:10.1057/s41599-024-03611-3
- [10] Wenyu Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. 2024. Game-theoretic LLM: Agent Workflow for Negotiation Games. arXiv:2411.05990 [cs.AI] <https://arxiv.org/abs/2411.05990>
- [11] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems* 36 (2023), 10622–10643.
- [12] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. PersonalLLM: Investigating the ability of large language models to express personality traits. arXiv:2305.02547 [cs.CL]
- [13] Yan Leng and Yuan Yuan. 2024. Do LLM Agents Exhibit Social Behavior? arXiv:2312.15198 [cs.AI] <https://arxiv.org/abs/2312.15198>
- [14] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbulin, and Bernard Ghahem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. arXiv:2303.17760 [cs.AI] <https://arxiv.org/abs/2303.17760>
- [15] Weiling Li. 2021. *Predicting MBTI personality type of Twitter users*. Ph. D. Dissertation. Rutgers University-Camden Graduate School.
- [16] Zhicheng Lin. 2024. Large language models as linguistic simulators and cognitive models in human research. arXiv:2402.04470 <https://arxiv.org/abs/2402.04470>
- [17] Emanuele Lorè and Arman Heydari. 2023. Strategic Behavior of Large Language Models: Game Structure vs. Contextual Framing. <https://arxiv.org/abs/2309.05898>
- [18] J. Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789. doi:10.1073/pnas.1813486116
- [19] Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. 2024. From individual to society: A survey on social simulation driven by large language model-based agents. arXiv:2412.03563 <https://arxiv.org/abs/2412.03563>
- [20] Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. 2024. From Individual to Society: A Survey on Social Simulation Driven by Large Language Model-based Agents. arXiv:2412.03563 [cs.CL] <https://arxiv.org/abs/2412.03563>
- [21] Martin A. Nowak and Robert M. May. 1992. Evolutionary games and spatial chaos. *Nature* 359, 6398 (1992), 826–829. doi:10.1038/359826a0
- [22] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC] <https://arxiv.org/abs/2304.03442>
- [23] Francisco C. Santos and Jorge M. Pacheco. 2005. Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical Review Letters* 95, 9 (2005), 098104. doi:10.1103/PhysRevLett.95.098104
- [24] Gregory Serapio-García, Mustafa Safdar, Clément Crepy, Luneng Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. Research Square. doi:10.21203/rs.3.rs-3296728/v1 Preprint (Version 1).
- [25] György Szabó and Gábor Fáth. 2007. Evolutionary games on graphs. *Physics Reports* 446, 4–6 (2007), 97–216. doi:10.1016/j.physrep.2007.04.004
- [26] Filippo Tonini and Lukas Galke. 2025. Super-additive Cooperation in Language Model Agents. arXiv:2508.15510 [cs.AI] <https://arxiv.org/abs/2508.15510>
- [27] Wei Wang, Haili Yang, Yuanfu Lu, Yuanhang Zou, Xu Zhang, Shuting Guo, and Leyu Lin. 2021. Influence Maximization in Multi-Relational Social Networks. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (Virtual Event, Queensland, Australia) (CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 4193–4202. doi:10.1145/3459637.3481928
- [28] Zhao Wang, Sota Moriyama, Wei-Yao Wang, Briti Gangopadhyay, and Shingo Takamatsu. 2025. Talk Structurally, Act Hierarchically: A Collaborative Framework for LLM Multi-Agent Systems. arXiv:2502.11098 [cs.AI] <https://arxiv.org/abs/2502.11098>
- [29] Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of "small-world" networks. *Nature* 393, 6684 (1998), 440–442. doi:10.1038/30918
- [30] Simin Yu, Hao Wang, Ye Su, Ziyu Niu, Zhi Li, Jianjun Liu, and Jiwei Wang. 2024. Privacy-preserving recommendation system based on social relationships. *Journal of King Saud University - Computer and Information Sciences* 36, 2 (2024), 101923. doi:10.1016/j.jksuci.2024.101923
- [31] Weiqi Zeng, Bo Wang, Dongming Zhao, Zongfeng Qu, Ruifang He, Yuexian Hou, and Qinghua Hu. 2025. Dynamic Personality in LLM Agents: A Framework for Evolutionary Modeling and Behavioral Analysis in the Prisoner's Dilemma. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria, 23087–23100. doi:10.18653/v1/2025.findings-acl.1185