The Impact Of Varying Network Structures On LLM-Based Social Simulations

Anonymous ACL submission

Abstract

Generative social simulations use LLM agents to study opinion dynamics, yet the role of the underlying followership network in shaping these dynamics remains under-explored. To address this, we systematically study how different network models affect opinion evolution in a multi-agent simulator. In a controlled study of 1,136 simulation roll-outs we show that graph topology is an important driver of behavior, explaining up to 28.5% of variance in consensus change and producing large differences in topology-conditioned neighbor-alignment metrics. We further show in robustness assessments that the topology effect on neighbor alignment survives FDR correction in every condition tested, including varying the base LLM (Gemma, Llama-3.1, Qwen2.5), the survey question, the news agent, graph density. Our study provides detailed statistics on the opinion dynamics induced by seven different classical graph construction algorithms that future works may reference when designing future LLM-based social simulations.

1 Introduction

Social media platforms shape public discourse through the interplay of content and structure: *what* people say matters, but so does *who* follows *whom*. As Large Language Models (LLMs) have developed and gained more agentic capabilities, there has been a surge of interest in systems where multiple LLM agents interact autonomously, from collaborative task-solving (Luo et al., 2025) to entirely LLM-populated social-media platforms such as Moltbook (Li et al., 2026; Zhang et al., 2026). The field of generative social simulations uses LLMs to simulate humans and their interactions, with one of the end-goals being to simulate human societies and gain insights into sociological phenomena that have historically been hard to study. A key area of research within generative social simulations is social-media simulations. If LLMs can give insights into how humans behave on social media, these systems can be used as sandboxes for a broad array of research (Piao et al., 2025). While there have been several proposals for social simulation systems (Puelma Touzel et al., 2024; Yang et al., 2025; Piao et al., 2025), these lack systematic analyses of their respective components and design choices. In particular, the *followership network*, i.e., the directed graph that determines which agents observe which others' posts, is typically initialized with a single random or unspecified algorithm, with no analysis of how this structural choice affects the resulting dynamics.

Understanding the impact of such network topology choices on generative social simulations is crucial for two primary reasons. First, for researchers utilizing these systems as sociological sandboxes, disentangling the artifacts of the network structure from other components of a simulation is essential to ensure valid conclusions. Second, systematically analyzing the behavior of LLM simulations under different followership-network strategies allows us to get insights into how accurately LLMs approximate human social media behavior. Ultimately, establishing this structural baseline is a necessary foundational step in maturing LLM social simulations. By building a rigorous understanding of these underlying dynamics, these simulated environments can eventually serve as reliable, high-fidelity testbeds for studying malicious disinformation strategies and testing structural interventions, thereby helping to safeguard real-world human online ecosystems (Puelma Touzel et al., 2024).

In this work, we conduct an empirical study of network topology in LLM social simulations. Our main contributions are:

- **Showing that the followership-network topology is an important driver of opinion dynamics**. The strongest effects are observed for neighborhood-conditioned alignment metrics,

while more topology-neutral population metrics such as consensus change also show significant but smaller effects. Topology explains up to 90.6% of neighbor-alignment variance and 28.5% of consensus-change variance.

- **Detailed topology-specific statistics.** A 3-part multi-topology study totaling 1136 simulation roll-outs gives detailed statistics about the effects of various network topologies on simulations to inform future works and help address the existing validation gap (Puelma Touzel et al., 2026).
- **Robustness checks** that vary the base LLM (Gemma-3-4B, Llama-3.1-8B, Qwen2.5-7B), the survey question, the news-agent placement, and the graph density: the topology effect on neighbor-alignment shift rate is FDR-significant in every condition tested.

2 Related Works

Our work lies at the intersection of two primary areas of research: social simulations and network topology algorithms.

Social Simulators and Opinion Dynamics Given the promise of social simulations, there have been several works proposing social simulation architectures. These include Oasis (Yang et al., 2025) that scales social simulations to 1,000,000 agents and provides some initial work on the evaluation of opinion dynamics such as herding, polarization, and information propagation. Other works, like AgentSociety (Piao et al., 2025) and SandboxSocial (Puelma Touzel et al., 2024), also touch on or analyze phenomena such as polarization and opinion evolution. However, none of these works systematically vary the followership network structure, typically relying on a random (or unspecified) network initialization. While there has been some initial work on exploring the effect of network initialization algorithms on opinion dynamics, these have been limited by their scale and design choices. For example, Zheng and Tang (Zheng and Tang, 2024) provide a preliminary analysis by focusing on the difference between scale-free and small-world networks, but they only conduct studies on a system of 15 agents. Our study expands on such work in three concrete dimensions: scale (up to 4,096 agents), breadth of topologies (seven graph models), and statistical methodology (permutation ANOVA with FDR correction across 1136 total runs).

Network Initialization Algorithms To understand the impact of topology, the literature traditionally relies on established mathematical baselines to model specific macroscopic network properties. The Erdős-Rényi (ER) model (Erdős and Rényi, 1959) provides a foundational baseline of uniform randomness, while Stochastic Block Models (SBM) (Abbe, 2023) are utilized to explicitly construct community structures and study potential echo chambers. Additionally, the Barabási–Albert (BA) model (Albert and Barabási, 2002), the Forest Fire (FF) model (Leskovec et al., 2005), and directed scale-free graphs (Bollobás et al., 2003) have long been the standard for capturing the preferential attachment, dynamic growth, and hub-dominated nature characteristic of real-world social platforms. While more sophisticated approaches have been proposed, our analysis focuses on these classical, widely-used algorithms, which may be expanded by future works.

Existing research lacks a systematic empirical analysis of how varying these foundational topologies specifically alters the complex, high-dimensional opinion dynamics and information propagation within societies of LLM agents. This lack of information contributes to a well-documented validation gap in *Silicon Societies* (Puelma Touzel et al., 2026; Larooij and Törnberg, 2025; Ye et al., 2026). Our work attempts to help bridge this gap, to gain stronger insights into the capabilities of LLM agent-driven social simulations today.

3 Problem Definition

In this project, we propose to answer the following research question: **How does varying the structure of the followership network in an LLM-based social simulation impact opinion dynamics?**

In our experiments, we vary our followership network initialization across seven classical mathematical graph models (see Section 5.2). For each algorithm, the followership graph is generated first, and personas are then assigned to nodes, either randomly or using a homophily-aware strategy placing like-minded models together (see Appendix A.1).

We measure opinion dynamics using several families of metrics. For echo chambers, we use network assortativity, cross-cutting edge fraction, and mean local agreement. For herd effects, we track consensus evolution and the fraction of agents

aligning with their neighbors' majority opinion (Neighbor Alignment Shift Rate). For information propagation, we introduce a post-hoc semantic influence analysis: we embed all simulation messages using TwHiN-BERT (Zhang et al., 2023) and construct a weighted influence graph where edge weights reflect cosine similarity modulated by temporal decay, allowing us to trace influence cascades and measure global semantic drift across the simulation. As a sanity check for simulation validity, we also include a stylistic realism metric using a BERT classifier.

We hypothesize that the underlying network topology will alter the simulated opinion dynamics. Specifically, we anticipate that (1) hub-dominated topologies (e.g., BA) will facilitate rapid global information propagation but suppress local neighbor-alignment effects, while community-structured topologies (e.g., SBM) will exhibit stronger local clustering of opinion. Furthermore, (2) homophily-aware persona assignment will create measurably different echo-chamber dynamics that interact with the underlying topology.

We note that while many recent works attempt to reproduce human behavior through similar simulation setups, validating claims of human-likeness remains a particularly challenging open question (Larooij and Törnberg, 2025). We will therefore not approach this problem: we are interested in the impact of varying network structure on opinion dynamics in LLM social networks, but make no claims regarding the applicability of our results to human-populated networks.

4 Dataset

Pre-existing Data Sources. The models we use within the simulator are fine-tuned on anonymized BlueSky posts and interactions from the BluePrint dataset (Bück-Kaeffer et al., 2025) (see A.11). BluePrint clusters users into 25 behavioral archetypes based on posting patterns and topical interests. For each archetype, a Low-Rank Adaptation (LoRA) adapter is trained on that archetype's subset of posts, yielding 25 adapters that capture distinct communication styles and topical tendencies (Bück-Kaeffer et al., 2026). During simulation, each agent is assigned one of these adapters, which governs its language generation behavior. Survey questions and their per-archetype answer distributions are also sourced from BluePrint. The simulation platform is built by Bück-Kaeffer et al. (2026).

Synthetically Generated Data. The interaction data on which we compute our herd effect, echo chamber, and influence metrics are generated through LLM-based social simulation roll-outs. Given the artificial nature of this dataset, we record all information available within the simulation: the followership graph, regular answers to survey questions, messages written, impressions on each message, messages seen but not interacted-with, and the full observation history per agent. To seed dynamics into the simulation, a biased news agent posts pre-generated "news" messages. These were generated using Gemini 3.1 Pro (Team et al., 2025), used from the chat interface, explicitly biased in favor of one survey option, and are loaded from a static set of 50 pre-generated posts per topic.

5 Methodology

Our study is centered on a controlled multi-topology experiment (main study, 156 runs, seven graph models, all other simulation parameters held constant), which constitutes the main contribution of the paper. Two supporting analyses surround it: a broad-exploration phase (pilot study, 595 runs, two topologies, many varied parameters) that motivated the main study with a null result, and a robustness phase (robustness study, 385 runs) that re-runs the main study design under varied base models, survey questions, news-agent presence, and graph density. The main study and the headline robustness findings appear in the main body; full pilot study results are deferred to Appendix A.8, and per-condition robustness study breakdowns to Appendix A.7.

5.1 Metrics

Designing metrics or evaluations for social simulations can be challenging. This is partly because the imitative nature of the objective makes it hard to define what a realistic simulation means. The literature has yet to converge on a set of metrics, with previous works pointing to the issue of validation remaining poorly addressed and failing to evidence operational validity (Larooij and Törnberg, 2025). To best ground our experimentation, we thus aim to cover a wide range of metrics.

Realism We compute a stylistic realism metric by training a BERT classifier to distinguish between human-generated and simulation generated

social-media threads. We use the TwHiN model, which was pre-trained on Twitter data (Zhang et al., 2023), as our base, and further train it for our use case. For this purpose, we source human data from the BluePrint dataset and LLM data from our simulation runs. Together, we curate 200000 training samples, carefully avoiding data leakage issues. For each simulation run conducted, 150 threads are used as a test set. If a thread is harder to identify as being AI-generated, we interpret it as being more realistic.

Information Dynamics Metrics All other metrics are computed using model survey responses. Considering that they lack human baselines, they **should not** be used to claim that a given set of settings is more realistic than some other. They can, however, be used to measure how each setting impacts the simulation's opinion dynamics, giving us clues regarding which parameters matter most and how they interact. The complete list of metrics is available in Table 6. We invite readers to refer to it as needed.

Influence Analysis (Main Study) In the main study, we additionally compute a heuristic post-hoc semantic influence proxy. All simulation messages are embedded using TwHiN-BERT (Zhang et al., 2023), and a weighted influence graph is constructed: for each message, candidate "parent" messages are identified from the author's observation history, and edge weights are computed as $w_{ij} \propto \cos(\mathbf{e}_i, \mathbf{e}_j) \cdot \exp(-\beta \cdot \Delta t)$, combining semantic similarity with a Hawkes-process-style temporal decay. Edges exceeding a probability threshold (0.05) are retained as "filtered" edges, enabling cascade decomposition and component analysis. Global semantic drift is measured in sliding windows across the simulation timeline. Here, we set $\beta = 0.05$.

5.2 Variables

Main Study: controlled settings. All main study runs hold base model, population, question, news agent, proportions, and survey awareness fixed: Llama-3.1-Minitaur-8B as the base model, 256 agents, the AI-copyright question (A.3), uniform proportions across 25 LoRAs, one biased news agent placed at the highest-degree node, and survey awareness off. The two varied factors are:

- **Network topology**: seven graph models—Erdős–Rényi (ER), Barabási–Albert (BA), Stochastic Block Model (SBM), cycle, power-law cluster (PLC), fully connected (FC), and Forest Fire (FF). Per-model parameter values are listed in Appendix A.4.
- **Homophily**: on or off. When on, agents with similar initial opinions are placed in spatially adjacent nodes (Appendix A.1).

The main study comprises 156 runs (15–27 per topology, 83 with homophily on, 73 off).

Robustness Study A further 385 runs replicate the main study design while varying, one factor at a time: (i) base LLM among Gemma-3-4B, Llama-3.1-8B, and Qwen2.5-7B-Instruct (Binz et al., 2025; Yang et al., 2024; Team, 2025) (210 runs); (ii) survey question among Q1 / Q2 / Q3 (105 runs); (iii) presence vs. absence of the news agent (70 runs). All the robustness study runs use density-controlled graphs ($\approx$ 16 mean degree for the five parameterizable topologies; see Appendix A.4).

Pilot study (broad exploration; in appendix). A 595-run broad-parameter exploration with two topologies (ER and a directed scale-free graph (Bollobás et al., 2003)), four base models, four network sizes (64, 256, 1024, 4096 agents), four LoRA-proportion presets, survey awareness on/off, and news agent on/off motivated the main study by producing a near-zero topology effect that we trace to between-model variance dominating the heterogeneous design. Full pilot study results, methodology, and variables are in Appendix A.8. Details on the statistical methodology used throughout the paper are available in Appendix A.6.

5.3 Experiment Setup

Our simulator is taken from (Bück-Kaeffer et al., 2026). We instantiate a set of agents to populate our simulation and create a followership network connecting them together based on the relevant parameter selected for the run. Agents get attributed an "action probability" fixed for the duration of the simulation. At each simulation step, 10 agents are selected according to this action probability. The first 9 observe a recent thread in which someone they follow participated, while the last one has a 1/3 probability of creating a new thread and a 2/3 probability of replying to an already existing one. The simulations run for a total of 2500 steps, with agents surveyed every 250 steps starting at 0 (yielding 11 survey snapshots). Details regarding our choices of parameters are available in the appendix (see A.5).

6 Results

6.1 Controlled Multi-Topology Analysis (156 Runs)

The main study holds all confounding variables constant and expands to seven graph topologies (see Section 5.2). Here, graph topology emerges as an important driver of nearly every metric.

6.1.1 Behavioral Metrics

Table 1 reports the main results. For each metric, we report the graph-type effect (η^2 with permutation p-value) alongside the homophily effect (Cohen's d with permutation p-value). Graph topology explains large and statistically significant variance fractions across all four behavioral metrics.

The most notable result is NASR: graph topology explains 90.6% of its variance ($p_{\text{perm}} < 0.001$). Table 2 reports per-topology means ($\pm$ standard deviation) across key behavioral and echo metrics. Cycle graphs produce the highest NASR (0.121 ± 0.001) and the strongest consensus erosion (NCC $= -0.150 \pm 0.034$), while Barabási–Albert produces the lowest NASR (0.068 ± 0.006) and fully connected graphs show the least erosion (NCC $= -0.064 \pm 0.034$). For net consensus change, both topology ($\eta^2 = 0.285$) and homophily ($d = +0.48$, $p_{\text{perm}} = 0.002$) are significant. We separate the Erdős–Rényi row in Table 2 because it is the default initialization used in a lot of the existing social-simulation work; it sits in the middle of the topology ordering for every metric. Full per-topology means including initial/final values, exposure diversity, same-option exposure, all echo-chamber metrics, and all temporal deltas are in Appendix Tables 15, 16, and 17.

We note, however, that NASR's construction makes it partly mechanically sensitive to topology. In a cycle graph, the minimal two-neighbor structure means any opinion shift is very likely to match one neighbor's opinion, inflating NASR relative to denser topologies. In a fully connected graph, the "neighbor majority" collapses to the global majority, making NASR and MFR conceptually equivalent. The $\eta^2 = 0.906$ for NASR therefore reflects both genuine behavioral differences and this structural sensitivity. NCC ($\eta^2 = 0.285$), which aggregates consensus at the population level without reference to neighborhood structure, provides the more conservative and topology-neutral estimate of the behavioral impact of graph structure.

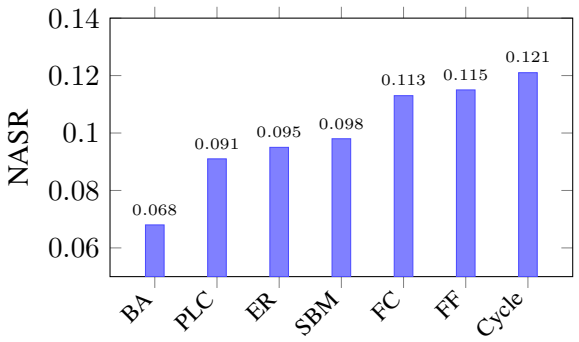

Figure 1: Per-topology NASR means in the main study, ordered by increasing NASR. BA = Barabási–Albert, PLC = Power-law Cluster, ER = Erdős–Rényi, SBM = Stochastic Block Model, FC = Fully Connected, FF = Forest Fire. The 77% relative range between BA (0.068) and Cycle (0.121) illustrates the practical impact of topology choice.

6.1.2 Echo-Chamber Metrics

In the main study, both topology and homophily produce large, significant effects on echo-chamber structure (Table 3). The temporal deltas show that homophily-initialized networks undergo substantial structural change:

Δ assortativity: -0.002 (no homophily) vs. -0.194 (homophily), $d = -1.90$

Δ local agreement: -0.066 vs. -0.144, $d = -1.35$

Δ cross-cutting edges: $+0.063$ vs. $+0.151$, $d = +1.60$

Graph topology also significantly modulates these temporal changes (all Δ metrics show graph $p_{\text{perm}} \leq 0.002$). As shown in Table 2, cycle graphs exhibit the largest assortativity decay ($\Delta = -0.225$), while fully connected graphs remain nearly stable ($\Delta = +0.002$). This indicates that LLM agents do not simply preserve the opinion clustering imposed at initialization; they shift opinions in ways that erode echo-chamber structure. Hub-dominated topologies partially resist this erosion by concentrating influence through a small number of highly connected nodes, which may act to repeatedly reinforce the dominant local opinion. Topologies with more uniform connectivity allow more diverse opinion signals to reach each agent, accelerating the breakdown of initial clustering.

6.1.3 Influence Metrics

The semantic influence analysis reveals that topology shapes information-flow (Table 4).

	Graph type		Homophily	
Metric	η^2	p_{perm}	d	p_{perm}
Net Consensus Change	0.285	<0.001***	**+0.48**	0.002**
Opinion Shift Rate	0.219	<0.001***	+0.16	0.328^{ns}
Majority Follow Rate	0.110	0.009**	+0.29	0.070^{ns}
Neighbor Alignment Shift Rate	**0.906**	<0.001***	−0.25	0.111^{ns}

Table 1: Main study behavioral metrics. Both factors are reported on consistent scales: η^2 for graph type (7 levels) and Cohen's d for homophily (2 levels). Permutation p-values use 2000 permutations. Graph topology explains large variance fractions, with NASR showing the strongest effect ($\eta^2 = 0.906$). Homophily significantly stabilizes consensus ($d = +0.48$) but does not significantly affect other behavioral metrics.

Topology	n	NCC	NASR	OSR	MFR	Final Asst.	Local Agr.
Cycle	24	−0.150±0.034	**0.121±0.001**	0.253±0.009	0.474±0.006	**0.155±0.172**	**0.582±0.090**
Forest Fire	15	−0.080±0.033	0.115±0.007	0.244±0.014	0.492±0.016	−0.003±0.002	0.527±0.017
Fully Connected	18	**−0.064±0.034**	0.113±0.002	0.245±0.003	0.488±0.005	−0.001±0.001	0.548±0.017
Stoch. Block	25	−0.121±0.036	0.098±0.006	**0.259±0.009**	0.484±0.020	0.032±0.039	0.525±0.022
PL Cluster	22	−0.112±0.038	0.091±0.006	0.255±0.011	0.488±0.012	0.001±0.014	0.528±0.020
Barabási–Albert	25	−0.132±0.039	0.068±0.006	0.255±0.008	0.485±0.020	0.024±0.029	0.538±0.026
Default in existing literature:							
Random (ER)	27	−0.106±0.062	0.095±0.007	0.255±0.010	0.488±0.020	0.017±0.026	0.531±0.048

Table 2: Per-topology means ± standard deviation for behavioral and echo-chamber metrics (main study, $n = 156$). NCC: Net Consensus Change; NASR: Neighbor Alignment Shift Rate; OSR: Opinion Shift Rate; MFR: Majority Follow Rate; Final Asst.: final assortativity; Local Agr.: final local agreement. Random (ER) is separated as the default topology used by most existing social-simulation work. All metrics show significant graph-type effects ($p_{\text{perm}} \leq 0.009$). Full metric breakdown with initial values, deltas, and other echo metrics is in Appendix A.9.

Graph topology accounts for up to 91.9% of the variance in filtered edge count (i.e. the number of high-confidence influence relationships). Homophily significantly increases the filtered-edge rate ($d = 0.92$): when agents start with similar-minded neighbors, a greater proportion of influence edges exceed the significance threshold. This could be because like-minded agents influence each-other more, but it is also possible that our measurement method misinterprets preexisting agreement for causal influence. As shown in Table 2, stochastic block and random graphs have the highest filtered-edge rates (0.0235), while fully connected graphs have the lowest (0.0212). Forest fire graphs produce the highest global semantic drift (0.111), while cycle graphs produce the lowest (0.103).

6.1.4 Cross-Domain Correlations

The influence metrics correlate meaningfully with behavioral outcomes.

Higher filtered-edge rates (more concentrated influence pathways) are associated with greater opinion shift rate ($r = 0.465$, $p < 10^{-10}$), consistent with both metrics being proxies of inter-agent influence. Higher global semantic drift correlates with net consensus change ($r = 0.482$, $p < 10^{-11}$), which suggests that our embedding-based semantic metrics do capture opinions (to a degree et least). The negative correlation between the largest influence component size and net consensus change ($r = -0.297$, $p < 10^{-4}$) is explained simply: when models do not change their minds over time, longer chains of semantically similar messages can be identified. Together, these correlations support the interpretation that topology affects behavioral outcomes in part by shaping the structure of information flow.

6.2 Robustness (385 Runs)

The robustness study re-runs the main study design under varied conditions to test the generalizability of the topology effect: 210 runs replicate the main study across three additional LLM families (Gemma-3-4B, Llama-3.1-8B, Qwen2.5-7B), 105 runs cover three survey questions (Q1, Q2, Q3), and 70 runs remove the biased news agent. All robustness study runs use density-controlled graphs (≈ 16 mean degree for the five parameterizable topologies). Across all 48 hypothesis tests,

Metric	Graph type		Homophily	
	η^2	p_{perm}	d	p_{perm}
Final assortativity	**0.363**	<0.001	+0.81	<0.001
Final local agreement	0.166	<0.001	**+1.13**	<0.001
Final cross-cutting	0.218	<0.001	−0.99	<0.001
Exposure diversity	0.308	<0.001	−0.50	0.001

Table 3: Main study echo-chamber metrics (final snapshot), reported on consistent effect-size scales. Both factors produce significant effects. Homophily effects are very large (Cohen's d up to 1.13).

Metric	Graph type		Homophily	
	η^2	p_{perm}	d	FDR q
Influence edge count	0.855	<0.001***	−0.62	<0.001***
Filtered edge count	**0.919**	<0.001***	—	0.071^{ns}
Filtered-edge rate	0.697	<0.001***	**+0.92**	<0.001***
Cascade roots	0.287	<0.001***	—	0.073^{ns}
Global drift mean	0.289	<0.001***	+0.38	0.020*

Table 4: Main study influence metrics. Graph topology is the dominant factor, explaining 85.5–91.9% of variance in edge-based influence measures.

we apply Benjamini–Hochberg FDR correction at $\alpha = 0.05$.

Table 5 summarises the headline result: the NASR topology effect survives FDR correction at $q < 0.01$ in every condition tested. Effect sizes range from $\eta^2 = 0.400$ (Qwen) to $\eta^2 = 0.835$ (no-news), all substantially lower than the main study's $\eta^2 = 0.906$ but still large. NCC topology effects do not survive FDR correction for any of the three new LLMs, consistent with our finding (also confirmed in the pilot study) that base-model identity dominates NCC variance. Within the five parameterizable topologies, density does not significantly predict NASR ($r = 0.063$, $q = 0.364$), so the topology effect within this group is structural rather than density-driven. The per-topology NASR ordering is qualitatively stable across questions (Kendall $\tau = 1.0$ for Q1 vs. Q3) and across LLM families ($\tau = 0.52$–0.62 between model pairs). Removing the news agent has no significant effect on NASR ($\Delta = +0.003$, $d = 0.18$, $q = 0.483$) but does increase NCC and OSR, consistent with the news agent acting as a persistent opinion broadcaster orthogonal to local neighbor dynamics. BERT realism does not significantly vary by topology after FDR correction ($q = 0.089$), confirming that the behavioral differences are not driven by differential text realism. Per-condition statistical breakdowns, per-topology means under each condition, and the full 48-test FDR registry are in Appendix A.7.

Condition	N	η^2_{NASR}	q	η^2_{NCC}
Main Study (Minitaur, ref.)	156	0.906	—	0.285
Gemma-3-4B	77	0.785	0.003	0.098^{ns}
Llama-3.1-8B	57	0.599	0.003	0.163^{ns}
Qwen2.5-7B	76	0.400	0.006	0.162^{ns}
Q1 (ctrl)	32	0.736	0.003	0.409^{ns}
Q2	31	0.734	0.003	0.501*
Q3	42	0.570	0.003	0.321^{ns}
No news agent	70	0.835	0.003	0.238*

Table 5: NASR and NCC topology effect sizes across the 8 robustness conditions. All NASR effects are FDR-significant ($q < 0.01$). NCC effects marked with * survive FDR correction; ^{ns} indicates the effect did not survive correction. The main study row is the reference condition (uncorrected, single battery).

7 Discussion

7.1 LLM Generative Biases as a Confound

The failure of the pilot study (Appendix A.8) to detect topology effects moderates the main study's strong detections. In the pilot study's heterogeneous design, between-model variance dominates: consensus levels range from 0.993 (Qwen on Q1) to 0.562 (Gemma on Q3), and this LLM-specific variation dwarfs any structural signal from just two topology types. The practical implication is that

LLM generative biases must be treated as a confound and controlled for before structural effects can be reliably measured. This is an inherent challenge for simulation studies that vary multiple parameters simultaneously: a null topology result in a heterogeneous design is not evidence that topology does not matter, but may simply reflect that the topology signal is masked. The main study's strong effects (η^2 up to 0.906) emerge only once this confound is removed; The robustness study then partially recovers generalizability by replicating the controlled design across three additional LLM families.

7.2 Topology as a Non-Neutral Design Choice

For researchers using LLM-based social simulations as sociological sandboxes, the attribution results in Table 2 carry a direct practical message: topology initialization is not a neutral design decision. Across otherwise identical conditions, NASR varies by almost a factor of two between the lowest (BA, 0.068) and highest (cycle, 0.121) topology, and consensus erosion varies by more than a factor of two. A simulation built on a BA graph will produce qualitatively different dynamics from one built on a cycle or fully connected graph, and reported findings will reflect this choice whether or not the researcher intends them to. Studies that do not justify and conduct sensitivity analysis over their topology choice risk drawing conclusions that are specific to that initialization.

The homophily erosion results add a related concern. Researchers who design experiments to study echo-chamber formation or persistence may find that LLM agents are structurally non-persistent: initial opinion clustering degrades substantially over the simulation ($d = -1.90$ for Δ assortativity), even though homophily leaves a residual signal in final metrics. Studies that measure only final snapshots will underestimate the degree of structural change. Whether this erosion reflects a limitation of current LLM architectures, an artifact from our setup, agents rationally updating on diverse content, or a property of the specific simulator's survey mechanism remains an open question, but it is a property researchers should be aware of when interpreting results.

7.3 Influence Analysis as a Diagnostic Tool

The post-hoc semantic influence pipeline introduced here proved capable of capturing topology-dependent information-flow patterns that correlate with behavioral outcomes. The fact that these correlations are significant ($r = 0.465$, $r = 0.482$) and topology-dependent suggests that influence-derived metrics provide complementary information to survey-response metrics. This pipeline is general: it operates on message embeddings and temporal metadata, making it applicable to other LLM simulation frameworks that record message histories. Applying it across frameworks could help identify whether topology-behavior coupling is specific to this simulator or a broader property of LLM social dynamics.

7.4 Future Work

Testing LLM-driven graph generation methods (e.g., GAG-General) would situate classical mathematical topologies relative to learned ones and test whether LLM-generated graphs introduce distinct interaction patterns. Longer simulations would clarify whether echo-chamber erosion is a transient or persistent phenomenon. Finally, introducing a recommendation algorithm as a structural layer above the followership graph would more closely replicate real social media architectures, and its interaction with static topology effects would be informative for simulation design.

8 Conclusions

We investigated how network topology affects opinion dynamics in LLM-based social simulations through a 1136-run empirical study. Our key findings are that

Topology is an important driver of opinion dynamics when confounding variables are controlled. Graph type explains up to 90.6% of variance in neighbor-alignment behavior and 28.5% of consensus change variance.

The topology effect is robust across LLM families, survey questions, and biased interference. The robustness study confirms that the NASR topology effect survives FDR correction in every condition tested ($q < 0.01$), and is not driven by graph density within the parameterizable group.

Together, these findings establish that network topology is a fundamental design parameter in LLM social simulations, not a neutral initialization choice. Future work can delve further into these phenomena and use our results to inform their design decisions.

9 Limitations

Interpretating results requires some care. Because several metrics are bounded proportions with relatively small absolute ranges, large η^2 values should be interpreted as indicating strong differentiation relative to within-condition variance rather than necessarily large absolute behavioral changes. For example, NASR spans 0.068–0.121 across topologies despite its large η^2. We therefore interpret the results primarily as evidence that topology systematically and reproducibly alters simulation dynamics, rather than as evidence of arbitrarily large behavioral magnitude shifts.

News-agent placement. The biased news agent is placed at the highest-degree node in all runs. Because degree centrality differs across topologies, the news agent's reach is not constant across conditions. However, topology effects are consistent across both homophily conditions and across metrics where the news agent has no direct causal pathway (e.g., NASR is computed from survey responses), suggesting that topology effects are not solely driven by news-agent reach.

Simulation duration. At 2500 steps with 10 agents per step, the main study's 256-agent runs yield roughly 97 actions per agent. Longer simulations might clarify whether the observed echo-chamber erosion is a transient warm-up effect or a stable long-run property. However, due to time, scope, and resource constraints, we don't conduct this analysis here.

LLM-driven topologies and Deep-Learning based topologies We do not test LLM-driven graph generation methods such as GAG-General (Peng et al., 2026), or deep-learning generated topology structures. These may show other significant and interesting variations in comparison to our setup. Our study remains limited to classical network generation algorithms.

10 Ethical Considerations

Our work studies how network structure shapes opinion dynamics in simulated social environments. The intended application is to inform the design of more rigorous LLM social simulations. In general, malicious actors typically can't control network structures in social media. As such, we don't expect strong dual-use concerns. However, we acknowledge that insights about how network topology amplifies or suppresses opinion shifts could, in principle, inform the design of more effective manipulation campaigns on real social platforms. We believe that understanding these structural dynamics is a necessary prerequisite for developing effective countermeasures, and that the controlled, artificial nature of our simulations (using LLM agents with synthetic personas, not real users) limits direct transferability. Nonetheless, researchers building on this work should consider dual-use implications and prioritize defensive applications.

References

Emmanuel Abbe. 2023. Community detection and stochastic block models. *Preprint*, arXiv:1703.10146.

Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, and 1 others. 2025. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009.

Béla Bollobás, Christian Borgs, Jennifer Chayes, and Oliver Riordan. 2003. Directed scale-free graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, page 132–139, USA. Society for Industrial and Applied Mathematics.

Aurélien Bück-Kaeffer, Je Qin Chooi, Dan Zhao, Maximilian Puelma Touzel, Kellin Pelrine, Jean-François Godbout, Reihaneh Rabbany, and Zachary Yang. 2025. BluePrint: A social media user dataset for llm persona evaluation and training. *Preprint*, arXiv:2510.02343.

Aurélien Bück-Kaeffer, Sneheel Sarangi, Maximilian Puelma Touzel, Reihaneh Rabbany, Zachary Yang, and Jean-François Godbout. 2026. The *Silicon Society* cookbook: Design space of llm-based social simulations. *Preprint*, arXiv:2605.00197.

Pál Erdős and Alfréd Rényi. 1959. On random graphs. i. *Publicationes Mathematicae Debrecen*, 6(3–4):290–297.

Maik Larooij and Petter Törnberg. 2025. Do large language models solve the problems of agent-based modeling? a critical review of generative social simulations. *Preprint*, arXiv:2504.03274.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2005. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the 11th ACM SIGKDD International Confer-*

ence on Knowledge Discovery in Data Mining (KDD '05), pages 177–187, New York, NY, USA. ACM.

Lingyao Li, Renkai Ma, Chen Chen, Zhicong Lu, and Yongfeng Zhang. 2026. The rise of ai agent communities: Large-scale analysis of discourse and interaction on moltbook. *Preprint*, arXiv:2602.12634.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiao Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Meng Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, and 7 others. 2025. Large language model agent: A survey on methodology, applications and challenges. *Preprint*, arXiv:2503.21460.

Jie Peng, Jiarui Ji, Runlin Lei, Zhewei Wei, Yongchao Liu, and Chuntao Hong. 2026. GDGB: A benchmark for generative dynamic text-attributed graph learning. In *The Fourteenth International Conference on Learning Representations*.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Maximilian Puelma Touzel, Sneheel Sarangi, Aurélien Bück-Kaeffer, Zachary Yang, Jean-François Godbout, and Reihaneh Rabbany. 2026. Position: Time to close the validation gap in llm social simulations. In *Proceedings of the 43rd International Conference on Machine Learning (ICML 2026)*. Position paper.

Maximilian Puelma Touzel, Sneheel Sarangi, Austin Welch, Gayatri Krishnakumar, Dan Zhao, Zachary Yang, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Andreea Musulan, Camille Thibault, Busra Tugce Gurbuz, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2024. A simulation system towards solving societal-scale manipulation. *Preprint*, arXiv:2410.13915.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team. 2025. Gemma 3.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, and 4 others. 2025. Oasis: Open agent social interaction simulations with one million agents. *Preprint*, arXiv:2411.11581.

Jinyi Ye, Lei Cao, Ding Chen, and Emilio Ferrara. 2026. Stop drawing scientific claims from llm social simulations without robustness audits. *Preprint*, arXiv:2605.18890.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. *Preprint*, arXiv:2209.07562.

Yunbei Zhang, Kai Mei, Ming Liu, Janet Wang, Dimitris N. Metaxas, Xiao Wang, Jihun Hamm, and Yingqiang Ge. 2026. Agents in the wild: Safety, society, and the illusion of sociality on moltbook. *Preprint*, arXiv:2602.13284.

Wenzhen Zheng and Xijin Tang. 2024. *Simulating Social Network with LLM Agents: An Analysis of Information Propagation and Echo Chambers*, pages 63–77.

A Appendix

A.1 Homophily-aware Agent Placement

When homophily is enabled, we use the spatial-opinion alignment strategy from Bück-Kaeffer et al. (2026), in which the assignment of personas to nodes is biased so that graph neighbors tend to share initial opinions. The procedure interleaves a structural ordering of the graph with an ideological ordering of the agents, then assigns them via the resulting rank.

Concretely: (i) we run a 2D Fruchterman–Reingold spring layout on the graph and project nodes onto the x-axis to obtain a one-dimensional ordering of the network's structural regions—connected nodes end up near each other on this axis; (ii) for each LoRA persona, we score its expected opinion as the dot product of its survey-answer distribution and an ordinal coding of the answer options (e.g., "Strongly Disagree" $\rightarrow$ 0, . . ., "Strongly Agree" $\rightarrow$ 4); (iii) we sort the agent population by this expected-opinion score, producing an "ideological spectrum"; (iv) we align the two orderings rank-by-rank, mapping the leftmost spatial node to the lowest-opinion agent and the rightmost to the highest. Because the spring layout colocates connected nodes in Euclidean space, this mapping

makes graph neighbors substantially more likely to share initial opinions, producing the structural homophily required to study echo-chamber dynamics. When homophily is disabled, agents are instead mapped to nodes by a uniformly random permutation seeded by the run identifier.

A.2 Metrics

A.3 Survey questions

The question we ask models in the simulations is one of the following

1. Should it be legal to use copyrighted material to train artificial intelligence models? You may only answer with 'Yes' or 'No'.

2. Is it acceptable to use genetic engineering to enhance human intelligence? You may only answer with 'Yes' or 'No'.

3. Would you rather vote for a candidate who prioritizes economic growth over environmental protection, or a candidate who prioritizes environmental protection over economic growth? You may only answer with 'Economic Growth' or 'Environmental Protection'.

These questions are taken from Bück-Kaeffer et al. (2026).

A.4 Graph-generation algorithm parameters

All graphs are generated using the parameters listed below. The main and robustness studies share these settings, except that the robustness study was explicitly designed to control the mean degree of the five parameterizable topologies (ER, BA, SBM, PL Cluster, Forest Fire) to approximately 16, so the topology effect cannot be attributed to differences in edge count. Cycle (mean degree 2) and fully connected (mean degree $N-1$) cannot be density-matched by construction.

Robustness Study (density-controlled, $\approx$ 16 mean degree).

- **Erdős–Rényi**: `nx.erdos_renyi_graph(N, p)` with $p = 1/16$; expected mean degree $= (N-1)/16 \approx 16.0$.

- **Power-law Cluster**: `nx.powerlaw_cluster_graph(N, m, p)` with $m = 8$, $p_{\text{tri}} = 0.4$.

- **Barabási–Albert**: `nx.barabasi_albert_graph(N, m)` with $m = 8$.

- **Stochastic Block Model**: 4 balanced blocks ($\approx N/4$ nodes each); $p_{\text{in}} = 0.17$, $p_{\text{out}} = 0.028$.

- **Forest Fire**: a custom undirected forest-fire generator with forward burn probability 0.21 and a cap of 4 burn visits per new node.

- **Cycle**: `nx.cycle_graph(N)`, mean degree $= 2$.

- **Fully Connected**: `nx.complete_graph(N)`, mean degree $= N-1$.

For all topologies above, $N = 257$ (256 agents + 1 news agent for the robustness study "with news" conditions; otherwise $N = 256$).

Main Study Graphs. The main study was run before density was explicitly controlled. The above parameters were also used for the parameterizable topologies, with observed mean degrees of 15.34–16.10 across ER, BA, SBM, and PL Cluster, and a wider 5.46–34.40 for Forest Fire (since the forest-fire process has stochastic spread). See Table 7 for empirical density statistics aggregated across the robustness study runs.

A.5 Parameter Justification

Using BluePrint (Bück-Kaeffer et al., 2025), we find that, on average, users post 24.3 messages per month. With 2500 steps and 10 agents acting per step, the simulation produces approximately 25,000 total actions. For 256 agents (main study), this yields roughly 97 actions per agent, sufficient for multiple opinion-revision opportunities. For the pilot study's larger networks (1024 or 4096 agents), per-agent action counts drop to roughly 24 and 6 respectively, which we acknowledge as a limitation at larger scales (see Section 7). Each simulation runs on either an A100 or an RTX8000 GPU and takes between 2 and 16 hours to complete.

A.6 Statistical Methods

In the pilot study, effects are tested using Welch's t-tests (normal approximation) with a conservative threshold of $p \leq 0.01$ to guard against multiple comparisons. In the main study, we adopt a multi-factor framework: graph-type effects (7 groups) are tested using one-way ANOVA F-statistics with permutation p-values (2000 permutations); homophily effects use a two-sample permutation test on the

Metric	Acronym	Explanation	Formula
Behavioral (Herd Effect) Metrics			
Consensus	—	The proportion of the population holding the most popular opinion at a specific time step.	$\frac{N_{\max}}{N_{\text{total}}}$
Net Consensus Change	NCC	The total convergence or divergence of the population over the entire simulation.	$C_{t_f} - C_{t_0}$
Opinion Shift Rate	OSR	The fraction of the population that changed their opinion between consecutive survey snapshots.	$\frac{N_{\text{changed}}}{N_{\text{total}}}$
Majority Follow Rate	MFR	Of agents who changed their minds, the fraction that adopted the current global majority opinion.	$\frac{N_{\text{changed}\to\text{maj}}}{N_{\text{changed}}}$
Neighbor Alignment Shift Rate	NASR	The fraction of the total population that changed their opinion to match the majority of their direct social neighbors.	$\frac{N_{\text{changed}\to\text{nb_maj}}}{N_{\text{total}}}$
Echo-Chamber (Structural) Metrics			
Assortativity	—	The tendency of agents to be connected to others with the same opinion (Network Assortativity Coefficient).	$\frac{\text{Tr}(\mathbf{e}) - \lVert\mathbf{e}^2\rVert}{1 - \lVert\mathbf{e}^2\rVert}$
Local Agreement	—	On average, what fraction of an agent's neighbors share the same opinion.	$\frac{1}{N}\sum_i \frac{\lvert\{j \in \mathcal{N}_i : o_j = o_i\}\rvert}{\lvert\mathcal{N}_i\rvert}$
Cross-Cutting Edge Fraction	—	The fraction of edges connecting agents who hold different opinions.	$\frac{\lvert\{(i,j) \in E : o_i \neq o_j\}\rvert}{\lvert E\rvert}$
Exposure Diversity	—	Mean Shannon entropy of the opinion distribution in threads each agent observed.	$\frac{1}{N}\sum_i H(\mathbf{p}_i)$
Influence Metrics (Main study only)			
Filtered-Edge Rate	—	The proportion of influence graph edges exceeding a probability threshold, indicating strong semantic influence.	$\frac{\lvert E_{\text{filtered}}\rvert}{\lvert E_{\text{total}}\rvert}$
Global Semantic Drift	—	Mean cosine-distance drift of messages from their exposure context, measured in sliding windows.	d_{window}

Table 6: Behavioral, structural, and influence metrics. Influence metrics are computed post-hoc using TwHiN-BERT embeddings (main study only).

Topology	n	deg. μ ($\pm\sigma$)	Min	Max
Cycle	54	2.00 ($\pm$ 0.00)	2.00	2.00
Forest Fire	56	15.82 ($\pm$ 8.00)	5.46	34.40
Fully Connected	54	255.84 ($\pm$ 0.38)	255.0	256.0
Stoch. Block	59	16.10 ($\pm$ 0.34)	15.36	16.84
Random (ER)	59	16.06 ($\pm$ 0.32)	15.24	17.06
PL Cluster	56	15.34 ($\pm$ 0.04)	15.26	15.40
Barabási–Albert	47	15.50 ($\pm$ 0.00)	15.50	15.50

Table 7: Empirical mean degree of generated graphs across the robustness study runs. The five parameterizable topologies are matched to a target mean degree of ≈ 16; Forest Fire has the widest spread because its generative process is stochastic.

mean difference (also 2000 permutations). Both factors are reported on consistent scales (η^2 and Cohen's d, respectively). In the robustness study, all 48 hypothesis tests are jointly corrected using Benjamini–Hochberg FDR ($\alpha = 0.05$).

A.7 Robustness Study: Per-Condition Detail

This appendix provides the per-condition breakdown of the 385 robustness study runs summarised in the main body (Table 5). All the robustness study runs hold every other main study condition fixed except the listed variable, and use density-controlled graphs (Appendix A.4).

A.7.1 Multi-Model Generalization

We replicate the main study across three additional LLM families: Gemma-3-4B, Llama-3.1-8B, and Qwen2.5-7B-Instruct (210 runs total, 7 topologies $\times$ 2 homophily conditions $\times$ $\approx$5 replications each). The NASR topology effect is FDR-significant for every model tested ($q < 0.01$; Table 5). Effect sizes range from $\eta^2 = 0.400$ (Qwen) to $\eta^2 = 0.785$ (Gemma), all substantially lower than the main study's $\eta^2 = 0.906$ but still large. The per-topology ordering is qualitatively stable across models: cycle and fully connected graphs consistently produce the highest NASR, while power-law cluster and BA consistently rank lowest (Kendall τ = 0.52–0.62 between model pairs; see Appendix A.7 for full per-topology means). NCC topology effects do not survive FDR correction for

any of the three new models, consistent with the pilot study's finding that base model dominates NCC variance. The NASR topology conclusion extends across LLM families; the NCC conclusion remains model-dependent.

A.7.2 Density Control

Within the five parameterizable topologies (ER, BA, SBM, power-law cluster, Forest Fire), all matched to approximately 16 mean degree, density does not significantly predict NASR ($r = 0.063$, $q = 0.364$). Forest Fire has the widest density range (mean degree 5.46–34.40, SD = 8.00) and still shows no within-topology density-NASR association ($r = 0.105$, $q = 0.507$). These results indicate that topology effects within the parameterizable group are driven by structural properties (clustering, community organization, preferential attachment) rather than edge count. Cycle (mean degree 2) and fully connected (mean degree $N - 1$) cannot be density-matched by construction; their behavioral extremes are consistent with structural rather than density-driven explanations. Full density measurements are reported in Appendix A.7.

A.7.3 Stylistic Realism

TwHiN-BERT classifier accuracy across topologies, using density-controlled baseline runs ($n = 32$), averages 0.902 (SD = 0.035), ranging from 0.872 (fully connected) to 0.933 (Forest Fire). After FDR correction, topology does not significantly modulate realism ($F = 2.78$, $q = 0.089$). This confirms that the behavioral differences reported across topologies are not confounded by differential text realism.

A.7.4 Cross-Question Robustness

We replicate the main study across three survey questions: Q1 (AI copyright), Q2 (genetic enhancements), and Q3 (environmental protection), totaling 105 runs. Initial conditions differ substantially: Q1 and Q2 start at moderate majority (≈ 0.56), while Q3 starts at full consensus (1.000). Despite these differences, the NASR topology effect is FDR-significant for all three questions ($q = 0.003$ each; η^2 = 0.734, 0.736, 0.570 respectively). The per-topology NASR ordering is highly consistent: the Q1–Q3 rank correlation is perfect (Kendall $\tau = 1.0$, $q = 0.008$), and Q2 pairs show $\tau = 0.714$ ($q = 0.061$). The Q3 floor effect compresses absolute NASR values (range 0.018–0.027) but does not disrupt the relative topology ordering, confirming that the NASR topology mechanism operates independently of the question's initial distribution.

A.7.5 News-Agent Ablation

We conduct 70 runs without any news agent, all other main study conditions held constant. The NASR topology effect remains large and FDR-significant ($\eta^2 = 0.835$, $q = 0.003$). Direct comparison with news-agent runs confirms that the news agent has no significant effect on NASR ($\Delta = 0.003$, $q = 0.483$, $d = 0.18$) but does significantly raise NCC ($d = 0.57$, $q = 0.031$) and OSR ($d = 0.72$, $q = 0.008$). This dissociation is expected: NASR measures alignment with direct social neighbors, a structural property independent of a single broadcaster node, while NCC and OSR are population-level aggregates sensitive to persistent opinion pressure. The most notable per-topology shift without the news agent is BA, whose NASR increases from 0.068 to 0.093, consistent with the news agent's high-degree placement monopolizing the BA hub's information environment (see Appendix A.7 for full per-topology comparison).

Robustness study batches. The three robustness study batches and their purposes are summarised in Table 8. Per-batch ANOVAs are in Tables 9, 19, and 22. Per-topology NASR means broken down by model are in Table 10. Rank-correlation diagnostics are in Tables 11 and 20. Density correlations are in Table 21. Initial-consensus differences across questions are in Table 18. News-vs-no-news direct comparisons are in Table 23. BERT topology accuracy is in Table 24.

A.8 Pilot Study: Broad Exploration (595 Runs)

This appendix details the pilot study's broad-exploration runs that motivated the controlled main study design.

Variables. The pilot study varies eight parameters simultaneously: two network topologies (Erdős–Rényi and a directed scale-free graph (Bollobás et al., 2003)); four network sizes (64, 256, 1024, 4096 agents); four base models (Llama-3.1-Minitaur-8B, Llama-3.1-8B, Qwen2.5-7B-Instruct, gemma-3-4b-pt); on/off homophily; on/off "survey awareness" (whether each agent can see its own prior survey answers); on/off biased news agent; four LoRA-proportion presets; and three survey questions (Appendix A.3). The pilot study comprises 595 simulation roll-outs with randomized

Batch	N runs	Variable
Multi-model robustness	210	Base model: Gemma-3-4B, Llama-3.1-8B, Qwen2.5-7B
Cross-question check	105	Survey question: Q1, Q2, Q3
News-agent ablation	70	No news agent (vs. 1 in main study)
Density-controlled main study analogue	32	Q1, Minitaur

Table 8: Robustness study batches.

Model	Metric	F	η^2	p (perm)	q (BH)	sig.
Gemma	NASR	42.7	0.785	<0.001	0.003	**
Gemma	NCC	1.26	0.098	0.291	0.364	ns
Gemma	OSR	1.29	0.099	0.255	0.350	ns
Gemma	MFR	2.28	0.163	0.045	0.089	ns
Llama-3.1	NASR	12.4	0.599	<0.001	0.003	**
Llama-3.1	NCC	1.63	0.163	0.159	0.231	ns
Llama-3.1	OSR	1.90	0.186	0.094	0.158	ns
Llama-3.1	MFR	2.03	0.196	0.077	0.138	ns
Qwen	NASR	7.65	0.400	<0.001	0.006	**
Qwen	NCC	2.22	0.162	0.043	0.089	ns
Qwen	OSR	4.36	0.275	0.001	0.008	**
Qwen	MFR	1.76	0.133	0.121	0.181	ns

Table 9: Per-model topology ANOVAs across all four behavioral metrics. NASR survives FDR correction for every LLM family tested; NCC and MFR do not, consistent with the broader finding that base model dominates NCC variance.

parameter combinations.

A.8.1 Behavioral Effects Are Near-Zero

Across the headline behavioral outcomes, both topology and homophily have small, non-significant effects (Table 12). With only two topology types and many confounding factors pooled, neither network topology nor homophily materially alters the magnitude of social dynamics (opinion volatility, herding, or consensus change) in cross-sectional comparisons.

A.8.2 Echo-Chamber Structure and Temporal Dynamics

Echo-chamber metrics show sensitivity to homophilic initialization but no correlation with graph type:

- Final assortativity by homophily: 0.0059 (True) vs. -0.0040 (False), $\Delta = +0.0099$, $p = 0.004$, $d = 0.24$.
- Graph type on final assortativity: $p = 0.375$ (ns).
- Final local agreement, cross-cutting edges, and same-option exposure: all graph type and homophily contrasts ns ($p \geq 0.20$).

Homophily leaves a small positive residual in final opinion assortativity (an expected sanity check: assortativity directly measures opinion homophily). The strongest pilot study effects appear in *time-series deltas* (last − first survey step), especially for assortativity (Table 14). Cell means for Δ assortativity show a non-linear interaction (Table 13): the interaction contrast is $IC = +0.0382$, confirming that power-law topology dampens the homophily-driven assortativity collapse, but only when homophily is active.

Homophily	False	True
random	$+0.0021$	-0.0671
powerlaw	$+0.0023$	-0.0286

Table 13: Pilot study: Δ assortativity by homophily and topology. Power-law topology partially dampens the homophily-driven assortativity decay.

A.8.3 Interpreting the pilot study Null

The near-zero behavioral effects admit several explanations: (a) between-model variance dominates: different base models yield very different consensus levels (e.g., Qwen on Q1 at 0.993 vs. Gemma on Q3 at 0.562), so pooling across four models masks within-topology effects; (b) limited topological contrast—with only two graph types, neither specifically designed to produce community structure, the topological variation may be insufficient; (c) per-agent action counts at the largest network

Topology	Main Study (Minitaur)	Gemma	Llama-3.1	Qwen
Cycle	**0.121**	**0.114** ($n = 9$)	0.073 ($n = 7$)	0.061 ($n = 8$)
Forest Fire	0.115	0.077 ($n = 9$)	0.062 ($n = 6$)	0.047 ($n = 17$)
Fully Connected	0.113	0.111 ($n = 8$)	**0.085** ($n = 5$)	0.049 ($n = 13$)
Stoch. Block	0.098	0.092 ($n = 9$)	0.076 ($n = 15$)	**0.070** ($n = 8$)
Random (ER)	0.095	0.094 ($n = 15$)	0.075 ($n = 5$)	0.051 ($n = 13$)
PL Cluster	0.091	0.077 ($n = 15$)	0.063 ($n = 10$)	0.048 ($n = 9$)
Barabási–Albert	**0.068**	0.078 ($n = 12$)	0.065 ($n = 9$)	0.050 ($n = 8$)

Table 10: Per-topology NASR means by base model. The ordering is qualitatively stable across models (Cycle / Fully Connected at the top, PL Cluster / BA at the bottom).

Pair	Spearman r_s	Kendall τ	p	q (BH)
Gemma vs Llama-3.1	0.750	0.619	0.051	0.094
Gemma vs Qwen	0.643	0.524	0.099	0.158
Llama-3.1 vs Qwen	0.607	0.524	0.099	0.158

Table 11: Rank-ordering consistency of the per-topology NASR mean across LLM families. With only 7 topology levels, Kendall τ requires near-perfect agreement to reach FDR significance; the point estimates are nonetheless substantive and directionally consistent.

sizes (roughly 6 actions per agent at $N = 4096$) may be too few for structural effects to propagate over the simulation horizon. These considerations directly motivated the controlled design of the main study.

A.9 Per-Topology Full Metric Breakdown (Main Study)

This appendix expands Table 2 from the main body with all behavioral, echo-chamber, and temporal-delta metrics. All values are computed from the 156 main study runs, reported as mean $\pm$ standard deviation. Initial consensus is constant (0.699 ± 0.000) across all topologies because the main study fixes the LoRA proportions and question. Tables 15, 16 and 17 list all measured metrics for every topology.

A.9.1 Cross-Question and Density Tables

Table 18 presents the initial consensus of the agent population before starting the simulation for each survey question. Table 19 reports the analysis of variance for topology on each metric for each question (i.e. how much of the variance of each metric is explained by topology, for each survey question setting). Table 21 reports correlation between some of the metrics.

A.9.2 News-Agent Ablation Tables

Table 22 reports the ANOVA when there is no biased news agent in the simulation. Table 23 directly compares runs with and without the enws agent.

A.9.3 BERT Realism

Table 24 reports BERT classifier accuracy for each topology. After FDR, the effect is not significant, meaning that our methods did not allow us to measure any stylistic improvements from varying graph topologies.

A.10 Licenses

The codebase used (Bück-Kaeffer et al., 2026) is under Apache 2.0 license. The BluePrint dataset (Bück-Kaeffer et al., 2025) is under MIT license. The Qwen2.5-7B model is under Apache 2.0 license. The Llama-3.1-8B-pt and Llama-3.1-Minitaur-8B models are under llama3.1 license. The Gemma-3-4B model is under gemma license. Research falls within intended use for all artifacts.

A.11 Blueprint dataset details

This dataset contains no Personally Identifiable Information (PII). All posts are in english, filtered via keywords to be on the topic of Canadian politics. This data comes from bluesky, meaning that it is skewed towards the average BlueSky user's profile (e.g. western, left-leaning, etc.)

A.12 AI usage disclosure

AI assistance was used during the data analysis and writing phases of this paper.

Metric	Graph type effect (random − powerlaw)	Homophily effect (True − False)
Net Consensus Change	-0.014 ($p = 0.230$, $d = -0.10$)	$+0.004$ ($p = 0.716$, $d = +0.03$)
Opinion Shift Rate	$+0.013$ ($p = 0.168$, $d = +0.11$)	-0.003 ($p = 0.775$, $d = -0.02$)
Majority Follow Rate	-0.008 ($p = 0.422$, $d = -0.07$)	-0.004 ($p = 0.725$, $d = -0.03$)
NASR	$+0.004$ ($p = 0.212$, $d = +0.10$)	-0.002 ($p = 0.601$, $d = -0.04$)

Table 12: Pilot study main effects of graph type and homophily on behavioral metrics are small and non-significant.

Delta metric	Graph type effect (random − powerlaw)	Homophily effect (True − False)
Δ assortativity	-0.0165 ($p = 0.016$, $d = -0.20$)	$\mathbf{-0.0486}$ ($p < 0.001$, $d = -0.62$)
Δ local agreement	-0.0152 ($p = 0.225$, ns)	-0.0163 ($p = 0.193$, ns)
Δ cross-cutting edges	$+0.0133$ ($p = 0.290$, ns)	$+0.0179$ ($p = 0.155$, ns)

Table 14: Pilot study temporal-delta effects. Homophily significantly reduces assortativity over time; other temporal effects are small.

Topology	n	NCC	NASR	OSR	MFR	Final cons.
Cycle	24	−0.150±0.034	0.121±0.001	0.253±0.009	0.474±0.006	0.549±0.034
Forest Fire	15	−0.080±0.033	0.115±0.007	0.244±0.014	0.492±0.016	0.619±0.033
Fully Connected	18	−0.064±0.034	0.113±0.002	0.245±0.003	0.488±0.005	0.635±0.034
Stoch. Block	25	−0.121±0.036	0.098±0.006	0.259±0.009	0.484±0.020	0.578±0.036
PL Cluster	22	−0.112±0.038	0.091±0.006	0.255±0.011	0.488±0.012	0.587±0.038
Barabási–Albert	25	−0.132±0.039	0.068±0.006	0.255±0.008	0.485±0.020	0.567±0.039
Random (ER)	27	−0.106±0.062	0.095±0.007	0.255±0.010	0.488±0.020	0.593±0.062

Table 15: Behavioral metrics per topology, Main Study (mean ± SD). Random (ER) is separated as the default initialization in existing work.

Topology	Init. Asst.	Final Asst.	Init. Loc. Agr.	Final Loc. Agr.	Init. Cross-cut.	Exposure div.
Cycle	0.381±0.375	0.155±0.172	0.739±0.158	0.582±0.090	0.261±0.158	0.429±0.013
Forest Fire	−0.002±0.002	−0.003±0.002	0.578±0.009	0.527±0.017	0.422±0.000	0.476±0.015
Fully Connected	−0.003±0.001	−0.001±0.001	0.576±0.008	0.548±0.017	0.422±0.000	0.476±0.017
Stoch. Block	0.158±0.167	0.032±0.039	0.646±0.068	0.525±0.022	0.354±0.078	0.447±0.016
PL Cluster	0.092±0.078	0.001±0.014	0.647±0.053	0.528±0.020	0.367±0.059	0.451±0.015
Barabási–Albert	0.179±0.136	0.024±0.029	0.689±0.084	0.538±0.026	0.330±0.081	0.440±0.013
Random (ER)	0.086±0.109	0.017±0.026	0.617±0.056	0.531±0.048	0.384±0.055	0.443±0.046

Table 16: Echo-chamber metrics per topology, Main Study (mean ± SD). "Init." is the first survey snapshot; "Final" is the last. Final cross-cutting equals 1− final local agreement and is omitted.

Topology	Δ Asst.	Δ Loc. Agr.	Δ Cross-cut.	Δ Same-opt. Exp.
Cycle	−0.225±0.203	−0.157±0.068	+0.157±0.068	+0.512±0.008
Forest Fire	−0.000±0.002	−0.051±0.017	+0.050±0.016	+0.532±0.019
Fully Connected	+0.002±0.000	−0.028±0.024	+0.041±0.019	+0.541±0.005
Stoch. Block	−0.126±0.138	−0.121±0.058	+0.120±0.062	+0.518±0.012
PL Cluster	−0.091±0.075	−0.119±0.049	+0.121±0.055	+0.521±0.017
Barabási–Albert	−0.155±0.128	−0.151±0.072	+0.153±0.078	+0.512±0.010
Random (ER)	−0.069±0.098	−0.086±0.059	+0.087±0.058	+0.529±0.052

Table 17: Temporal deltas (final − initial survey snapshot) per topology, Main Study (mean ± SD). Negative Δ assortativity and Δ local agreement indicate erosion of initial opinion clustering over the simulation.

Question	N	Initial consensus	Notes
Q1 (AI copyright)	32	0.555 ± 0.015	Moderate majority
Q2 (genetic enhancements)	31	0.566 ± 0.000	Moderate majority
Q3 (environmental protection)	42	1.000 ± 0.000	Full consensus at $t = 0$

Table 18: Initial consensus by survey question (robustness study question batch).

Q	Metric	F	η^2	p (perm)	q (BH)	sig.
Q1	NASR	13.9	0.736	<0.001	0.003	**
Q1	NCC	3.5	0.409	0.019	0.054	ns
Q1	OSR	1.3	0.205	0.287	0.364	ns
Q1	MFR	1.0	0.167	0.444	0.507	ns
Q2	NASR	11.0	0.734	<0.001	0.003	**
Q2	NCC	4.0	0.501	0.008	0.031	*
Q2	OSR	2.8	0.410	0.042	0.089	ns
Q2	MFR	0.67	0.144	0.685	0.730	ns
Q3	NASR	7.7	0.570	<0.001	0.003	**
Q3	NCC	2.8	0.321	0.027	0.065	ns
Q3	OSR	0.58	0.090	0.748	0.780	ns
Q3	MFR	0.93	0.138	0.514	0.574	ns

Table 19: Per-question topology ANOVAs (i.e. how much of the variance of each metric is explained by topology, for each survey question setting). NASR effects are FDR-significant for all three questions ($q = 0.003$ each).

Pair	Spearman r_s	Kendall τ	p	q (BH)
Q2 vs Q1	0.857	0.714	0.024	0.061
Q2 vs Q3	0.857	0.714	0.024	0.061
Q1 vs Q3	**1.000**	**1.000**	**0.002**	**0.008**

Table 20: NASR rank-ordering consistency across questions. The Q1/Q3 ordering is identical.

Test	r	p	q (BH)	n
FF density vs NASR	0.105	0.438	0.507	56
FF density vs NCC	0.172	0.200	0.283	56
All 5 param. topologies: density vs NASR	0.063	0.296	0.364	277

Table 21: Density–NASR correlation tests within the parameterizable topologies. None significant after FDR correction.

Metric	F	η^2	p (perm)	q (BH)	sig.
NASR	53.0	**0.835**	<0.001	**0.003**	**
NCC	3.3	0.238	0.009	**0.031**	*
OSR	3.3	0.237	0.011	**0.037**	*
MFR	1.8	0.146	0.106	0.165	ns

Table 22: No-news topology ANOVAs ($n = 70$). NASR/NCC/OSR topology effects all survive FDR correction.

Metric	With news	No news	Δ	p	q (BH)
NASR	0.109 ($n = 32$)	0.106 ($n = 70$)	+0.003	0.403	0.483
NCC	0.055 ($n = 32$)	0.027 ($n = 70$)	+0.028	0.008	0.031
OSR	0.280 ($n = 32$)	0.272 ($n = 70$)	+0.008	0.002	0.008

Table 23: Direct news vs. no-news comparison on the density-controlled Q1 baseline. The news agent leaves NASR essentially unchanged but raises NCC and OSR, consistent with its role as a population-level opinion broadcaster.

Topology	N	Mean BERT acc.	SD
Cycle	6	0.911	0.030
Forest Fire	4	0.933	0.028
Fully Connected	6	0.872	0.043
Stoch. Block	3	0.885	0.021
Random (ER)	6	0.889	0.034
PL Cluster	6	0.920	0.018
Barabási–Albert	1	0.922	—
Overall	**32**	**0.902**	**0.035**

Table 24: TwHiN-BERT human-vs-LLM classification accuracy per topology on the density-controlled Q1 baseline. Topology effect: $F = 2.78$, $p = 0.046$, $q = 0.089$ (ns after FDR). Average accuracy is uniformly high (> 0.87), so behavioral topology effects are not confounded by differential text realism.