Experimental Design: LLM Algorithmic Fidelity in Opinion Dynamics

Research Objective

Evaluate whether LLMs faithfully reproduce classical opinion dynamics models (DeGroot, Friedkin-Johnsen) in multi-agent simulations.

Core Research Questions

1. **Convergence Fidelity**: How well do LLM-based opinion dynamics preserve convergence properties?
2. **Bias Analysis**: Where and why do systematic biases occur in LLM opinion formation?
3. **Symmetry Violations**: Do LLMs exhibit order-dependent biases when topic framing is reversed?
4. **Calibration**: How well do LLM opinion distributions match human polling baselines?

Experimental Protocol

Phase 1: Mathematical Baseline

1. Run all 6 configurations with mathematical models
2. Determine convergence timesteps (opinion change $< 10^{-6}$)
3. Generate trajectories and equilibrium points

Phase 2: LLM Experiments

1. Run 6 configurations $\times$ 10 topics = 60 experiments
2. Use identical timestep counts from Phase 1
3. Record: opinion trajectories, posts, ratings, final states

Phase 3: Symmetry Testing

1. Run B vs A orientation for all experiments
2. Calculate symmetry violations
3. Identify order-dependent biases

Phase 4: Analysis

1. Compare LLM vs mathematical final states
2. Quantify systematic biases by topic/network
3. Analyze symmetry violations and calibration

Experimental Summary

Scale: 6 network topologies × 10 topic pairs × 2 orientations = 120 total experiments

Networks: Small-world consensus, scale-free influence, random baseline, echo chambers, empirical karate club, and stubborn agents

Topics: Mix of political/social issues (immigration, environment, guns) and apolitical debates (toilet paper, hot dogs) with human polling baselines

Method: LLM agents generate posts expressing opinions on [-1,1] scale, then rate neighbor posts, iterating until mathematical convergence timestep

Metrics: Algorithmic fidelity (L_2 norm), systematic bias, symmetry violations, calibration to human baselines

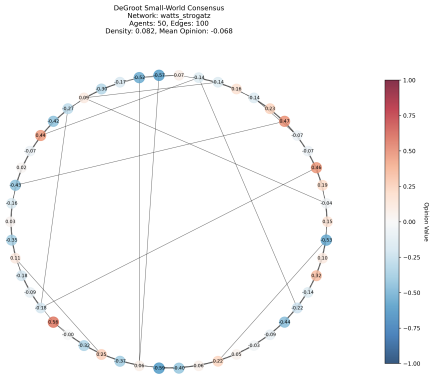
Network Configurations (6)

6 Canonical Network Configurations

We test across diverse network topologies to ensure robust findings across different social structures:

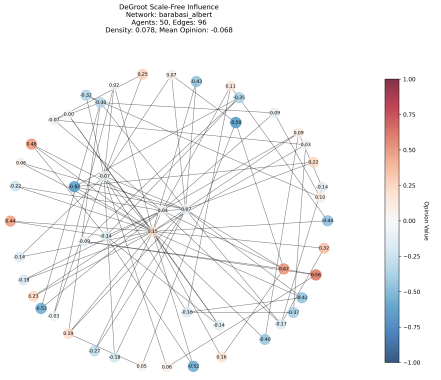
1. DeGroot Small-World Consensus

- **Network**: Watts-Strogatz small-world ($N = 50, k = 4, \beta = 0.1$)
- **Visual Structure**: Starts as a ring lattice where each agent connects to 4 nearest neighbors (2 on each side). Then 10% of edges are randomly rewired to create "shortcuts" across the network. Results in high clustering (agents have tight-knit local groups) but short average path lengths (any two agents are connected by few hops).
- **Key Properties**: Tight local groups with many neighbor-to-neighbor connections, short average path length (~3-4), degree distribution peaked around 4-6 connections per agent.
- **Model**: DeGroot
- **Justification**: Captures real-world social clustering + short paths. Parameters from Watts & Strogatz (1998), scaled for cost efficiency. Tests consensus formation in realistic social networks.



2. DeGroot Scale-Free Influence

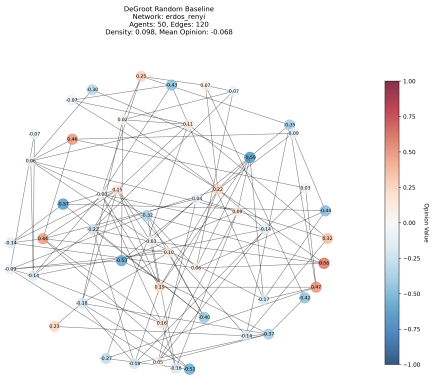
- **Network**: Barabási-Albert scale-free (N=50, m=2)
- **Visual Structure**: Starts with 2 fully connected agents, then each new agent connects to 2 existing agents with probability proportional to their current degree. Creates a "rich get richer" effect where highly connected agents become even more connected. Results in a few "hub" agents with many connections and many "peripheral" agents with few connections.
- **Key Properties**: Power-law degree distribution, high degree heterogeneity, few highly connected "influencers" (degree 10-20) and many peripheral agents (degree 2-4).
- **Model**: DeGroot
- **Justification**: Models influencer dynamics with power-law degree distribution. Parameters from Barabási & Albert (1999), scaled down. Tests how LLMs handle highly connected "influencer" agents.



3. DeGroot Random Baseline

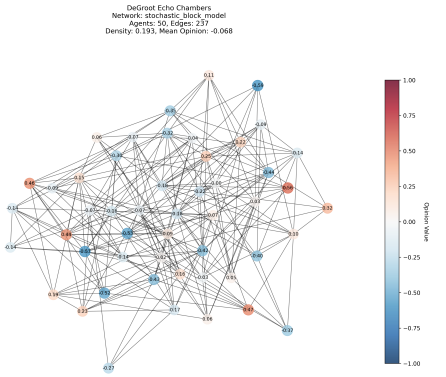
- **Network**: Erdős-Rényi random graph (N=50, p=0.1)

- **Visual Structure**: Each possible pair of agents has a 10% chance of being connected, independent of all other connections. Results in a relatively uniform, sparse network with no particular structure or clustering. Most agents have similar numbers of connections (around 5 ± 2).
- **Key Properties**: Uniform connection pattern with no local grouping, degree distribution approximately Poisson, no structural bias or communities.
- **Model**: DeGroot
- **Justification**: Provides baseline random network for comparison. Parameters from Erdős & Rényi (1959), scaled appropriately. Tests basic consensus formation without structural bias.



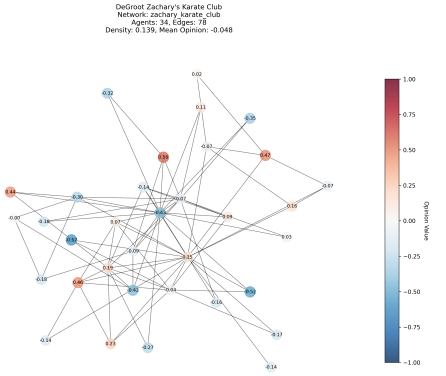
4. DeGroot Echo Chambers

- **Network**: Stochastic Block Model (N=50, 2 communities, p_intra=0.3, p_inter=0.05)
- **Visual Structure**: Two distinct clusters of 25 agents each. Within each cluster, agents have a 30% chance of being connected to each other. Between clusters, agents have only a 5% chance of being connected. Creates two "echo chambers" with dense internal connections but sparse cross-connections.
- **Key Properties**: Dense internal connections within each group, sparse connections between groups, clear community structure, potential for opinion polarization.
- **Model**: DeGroot
- **Justification**: Creates echo chambers - crucial for studying polarization. Parameters from community structure literature. Tests whether LLMs can maintain separate opinion clusters.



5. DeGroot Zachary's Karate Club

- **Network**: Empirical Zachary's Karate Club (34 nodes, 78 edges)
- **Visual Structure**: Real social network from a karate club that split into two factions. Two main clusters around the instructor (Agent 0) and administrator (Agent 33), with some bridging connections. The network represents actual friendship patterns and social influence structures from a real community.
- **Key Properties**: Empirical social structure with natural grouping patterns, two natural communities, 34 agents with varying degrees (2-17 connections), real-world influence patterns.
- **Model**: DeGroot
- **Justification**: Real empirical network validates against actual social structure. Classic benchmark in network analysis. Tests LLM behavior on real-world social topology.

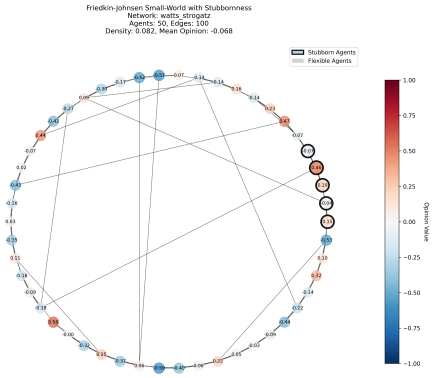


6. Friedkin-Johnsen Small-World with Stubbornness

- **Network**: Watts-Strogatz small-world ($N = 50, k = 4, \beta = 0.1$) + 10%

stubborn agents

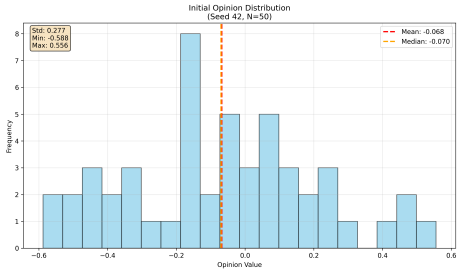
- **Visual Structure**: Same small-world structure as Configuration 1, but 5 agents (10%) are designated as "stubborn" and resist changing their opinions. These stubborn agents maintain their initial opinions throughout the simulation, while flexible agents (90%) update based on neighbor influence.
- **Key Properties**: Small-world structure + short paths, heterogeneous agent types, stubborn agents act as "anchors" preventing full consensus.
- **Model**: Friedkin-Johnsen with stubborn agents ($\lambda = 0.8$, 10% stubborn)
- **Justification**: Tests if LLMs can handle agent heterogeneity - key for realistic social simulation. Combines small-world structure with realistic stubbornness modeling.



Parameters

Opinion Initialization

- **Primary**: Normal distribution ($\mu = 0.0$, $\sigma = 0.3$) clipped to $[-1, 1]$
- **Scale**: $[-1, 1]$ for LLM interface, $[0, 1]$ for mathematical models
- **Conversion**: $X_{\text{math}} = \frac{X_{\text{LLM}}+1}{2}$
- **Seed**: Fixed seed (42) for reproducibility



LLM Interface

- **Post Generation**: Social media style (1-3 sentences, ≤ 320 chars)
- **Post Rating**: Numeric response on $[-1, 1]$ scale
- **Context**: Include neighbor posts (max 6, truncated to 220 chars each)

Topics (10 Pairs with Human Baselines)

Political & Social Issues

1. **Immigration Impact** (79% favorable) - Gallup 2025
   - A: "Immigration is a good thing for this country"
   - B: "Immigration is a bad thing for this country"
2. **Environment vs Economy** (52% environment) - Gallup 2023
   - A: "Prioritize environmental protection even if growth is curbed"
   - B: "Prioritize economic growth even if the environment suffers"
3. **Corporate Activism** (50/50 split) - Pew 2025
   - A: "Companies should make statements about political/social issues"
   - B: "Companies should not make statements about political/social issues"
4. **Gun Safety** (49% increases safety) - Pew 2023
   - A: "Gun ownership increases safety"
   - B: "Gun ownership reduces safety"
5. **Social Media Democracy** (34% good) - Pew 2024
   - A: "Social media has been good for democracy"
   - B: "Social media has been bad for democracy"

Apolitical & Cultural Debates

6. **Toilet Paper Orientation** (59% over) - YouGov 2022
   - A: "Toilet paper should go over the roll"
   - B: "Toilet paper should go under the roll"
7. **Hot Dog Sandwich** (41% yes) - YouGov 2023
   - A: "A hot dog is a sandwich"
   - B: "A hot dog is not a sandwich"
8. **Child-Free Weddings** (45% appropriate) - YouGov 2023
   - A: "Child-free weddings are appropriate"
   - B: "Child-free weddings are inappropriate"
9. **Restaurant Etiquette** (11% acceptable) - YouGov 2024
   - A: "Snapping fingers to get waiter attention is acceptable"
   - B: "Snapping fingers to get waiter attention is unacceptable"
10. **Human Cloning** (8% acceptable) - Gallup 2025
    - A: "Human cloning is morally acceptable"
    - B: "Human cloning is morally wrong"

Simulation Parameters

Convergence

- **Definition**: Converges at timestep T_{conv} when $||X[t+1] - X[t]||_2 < 10^{-6}$ for math model
- **Timestep Matching**: Run LLM simulations for exactly T_{conv} timesteps
- **No Early Stopping**: LLM simulations run for full T_{conv} regardless of opinion changes

Evaluation Metrics

- **Algorithmic Fidelity**: L_2 norm of opinion difference at convergence
  - Formula: $||X_{\text{LLM}}[T_{\text{conv}}] - X_{\text{math}}[T_{\text{conv}}]||_2$
- **Systematic Bias**: Mean difference between LLM and mathematical final opinions
  - Formula: $\frac{1}{n}\sum_i (x_{\text{LLM},i}[T_{\text{conv}}] - x_{\text{math},i}[T_{\text{conv}}])$
- **Symmetry Violation**: $|\text{mean}(A \text{ vs } B) + \text{mean}(B \text{ vs } A)|$
- **Calibration to human bias**: $|\text{LLM}_{\text{mean}} - \text{Human}_{\text{polling}}|$

Mathematical Models

We focus on two well-established opinion dynamics models from the literature:

DeGroot Model

- **Opinion Vector**: $X[k] \in [0,1]^n$
- **Weight Matrix**: $W_{ij} = \frac{A_{ij}}{\sum_k A_{ik} + \varepsilon}$
- **Update Rule**: $X[k+1] = WX[k]$
- **Properties**: Row stochastic, converges to consensus if strongly connected
- **Literature**: DeGroot (1974) - foundational model for opinion dynamics

Friedkin-Johnsen Model

- **Susceptibility Matrix**: $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_i \in [0,1]$
- **Update Rule**: $X[k+1] = \Lambda X[0] + (I - \Lambda)WX[k]$
- **Properties**: Always converges, can maintain polarization with stubborn agents
- **Literature**: Friedkin & Johnsen (1990) - extends DeGroot with stubbornness