Scale-Invariant Distributed Moderation: A Computational Study of Recursive Polycentric Governance

Sylvain Lebel[1]

[1]Independent Researcher, Consciousness of the Real (CdR) Framework,
http://sylebel.net

January 2026

Abstract

Digital platforms face an intractable dilemma: centralized moderation produces both over-censorship and under-regulation, while purely community-based approaches enable echo chambers and coordinated manipulation. We propose and computationally validate *Recursive Polycentric Moderation* (RPM), a distributed governance architecture based on double-majority consensus across structurally independent communities with institutional recursivity.

Through agent-based simulations on scale-free networks (400–10,000 users), we demonstrate three key properties: (1) *scale-invariance*: exposure rates remain constant at 2–3% of network size across 25-fold scaling; (2) *manipulation resistance*: 30% strategic agents (brigading) increase moderation rates by only 4.5%; (3) *systematic recursivity*: 100% of moderated content reaches maximum institutional appeal depth.

Parametric analysis reveals linear responses to threshold adjustments and polarization, while maintaining robustness to virality changes. These findings establish RPM as a viable alternative to centralized and purely decentralized moderation, with direct implications for platform governance design.

Keywords: distributed governance, content moderation, scale-free networks, agent-based modeling, platform design

1 Introduction

Digital platforms govern speech for billions of users, yet moderation remains structurally broken. Centralized systems (e.g., Facebook, YouTube) produce inconsistent enforcement, lack contextual nuance, and concentrate power in opaque corporate structures [1]. Conversely, purely decentralized approaches (e.g., Mastodon instances) enable echo chambers, brigading, and cross-community harassment without recourse [2].

This dilemma reflects a deeper governance challenge: how to balance local normative autonomy with protection against trans-community harms. Traditional solutions rely on either hierarchy (centralized authority) or isolation (independent communities), neither of which resolves the fundamental tension between context and coordination.

1.1 The Governance Trilemma

Platform moderation faces three irreconcilable demands under conventional architectures:

1. **Contextual judgment**: Content acceptability depends on community norms, topic domain, and interpretive context.
2. **Manipulation resistance**: Coordinated actors (brigading, sock-puppets) can game purely local or purely global systems.
3. **Procedural legitimacy**: Users must perceive moderation as fair, transparent, and non-arbitrary.

Centralized moderation sacrifices (1) for (2), while decentralized approaches sacrifice (2) for (1). Neither achieves (3) reliably.

1.2 Recursive Polycentric Governance

We propose *Recursive Polycentric Moderation* (RPM), grounded in the theoretical framework of Recursive Polycentric Governance [3]. RPM distributes moderation across three structural layers:

- **Localities**: Communities with shared norms and recurring interaction
- **Neighbors**: Communities with member overlap and contextual proximity
- **Distants**: Communities without correlation, providing structural independence

Content is moderated only when majorities in *both* neighbor and distant communities judge it unacceptable—a *double-majority* rule. If moderated, content is recursively appealed to broader jurisdictions with larger panels, creating institutional gradation.

1.3 Contributions

This paper provides the first comprehensive computational validation of RPM. Our contributions are:

1. **Scale-invariance demonstration**: Exposure rates remain proportional (2–3%) across network sizes 400–10,000 users, indicating a fundamental architectural property.
2. **Manipulation resistance quantification**: Strategic agents (brigading) at 30% density increase moderation by only 4.5%, showing bounded degradation.
3. **Systematic recursivity validation**: 100% of moderated content reaches maximum appeal depth, confirming institutional escalation mechanisms.
4. **Parametric sensitivity analysis**: Linear responses to thresholds and polarization, robustness to virality changes.

These findings establish RPM as computationally viable and provide quantitative design parameters for distributed governance systems.

2 Related Work

2.1 Platform Moderation Systems

Existing platform moderation operates on three paradigms:

Centralized human review (Facebook, YouTube) relies on content moderators applying global policies [4]. This approach scales poorly, produces inconsistent judgments, and concentrates power without accountability.

Algorithmic detection (automated hate speech, CSAM detection) achieves scale but suffers from high false-positive rates, contextual blindness, and adversarial evasion [5].

Community moderation (Reddit, Discord) delegates authority to volunteer moderators. While contextual, this enables power concentration, inconsistent enforcement, and brigading [6].

2.2 Distributed Governance Research

Our work builds on three research traditions:

Polycentric governance [7] demonstrates how overlapping jurisdictions enable collective action without centralization. However, Ostrom's framework lacks formal mechanisms for cross-jurisdiction coordination.

Liquid democracy [8] allows transitive vote delegation but remains vulnerable to delegation concentration and lacks structural independence between decision-making bodies.

Community Notes (Twitter/X) [9] uses crowd consensus with contributor diversity requirements. RPM extends this by adding recursive appeals and explicit community structure.

2.3 Network Models of Social Dynamics

Agent-based modeling of online communities [10] has explored opinion dynamics, polarization, and information diffusion. Our work contributes novel mechanisms (double-majority, recursive escalation) and tests scale-invariance properties not previously examined in governance contexts.

3 Model Description

3.1 Network Architecture

We construct community-structured scale-free networks using a two-stage process: (Figure **??**).

Stage 1: Intra-community structure. Each community $c \in \{1, \ldots, C\}$ is generated as a Barabási–Albert (BA) graph [11] with N_c nodes and attachment parameter m. This produces power-law degree distributions observed in real social networks.

Stage 2: Inter-community bridges. For each community pair (c_i, c_j), we add B bridge edges between high-degree nodes using preferential attachment:

$$P(\text{select node } u) \propto (k_u + 1)^\beta \tag{1}$$

where k_u is the degree of node u and β controls degree bias. This mimics real platform structures where hubs connect communities.

3.2 User Sensitivity Distribution

Each user u is assigned a sensitivity threshold $s_u \in [0, 1]$ representing their personal acceptability boundary. Sensitivities are drawn from a hierarchical model:

$$\mu_c \sim \mathcal{N}(\mu_0, \sigma_c^2) \quad \text{(community mean)} \tag{2}$$

$$s_u \sim \mathcal{N}(\mu_c, \sigma_i^2) \quad \text{(individual variation)} \tag{3}$$

This captures both community normative clustering and individual heterogeneity.

3.3 Content Extremity and Diffusion

Content extremity $e \in [0, 1]$ is sampled from a mixture distribution:

$$e \sim \begin{cases} \text{Beta}(6, 2) & \text{with probability } p_{\text{ext}} \\ \text{Beta}(2, 7) & \text{with probability } 1 - p_{\text{ext}} \end{cases} \tag{4}$$

This models predominantly mild content with an extreme tail, consistent with observed platform distributions.

Content diffuses through a cascade process over T steps. At each step, exposed users reshare to neighbors with probability:

$$P(\text{reshare}) = \rho_{\text{base}} \cdot \left(0.2 + 0.8 \cdot \min\left(1, \frac{\max(0, s_u - e)}{0.5}\right)\right) \cdot r_a \tag{5}$$

where ρ_{base} is base reshare probability, the middle term captures acceptability, and r_a is author reputation (reduced after moderation).

3.4 Double-Majority Voting

Upon flagging, content enters distributed adjudication. The origin community c_{orig} defines:

- **Neighbors** $\mathcal{N}$: K communities with highest cross-edge density to c_{orig}
- **Distants** $\mathcal{D}$: Remaining communities

From each group, V voters are sampled (preferring exposed users). Each voter u votes unacceptable if $e > s_u$. Content is moderated if:

$$\left(\frac{\#\text{ unacceptable in } \mathcal{N}}{|\mathcal{N}|} \geq \tau\right) \quad \wedge \quad \left(\frac{\#\text{ unacceptable in } \mathcal{D}}{|\mathcal{D}|} \geq \tau\right) \tag{6}$$

where τ is the decision threshold (typically 0.5).

3.5 Recursive Appeals

If moderated, content is appealed to depth $d + 1$ with modifications:

- **Jurisdiction broadening**: Increase K (neighbor count), eventually global review
- **Panel enlargement**: Increase V (voter count) by 25–75 per depth
- **Extremity attenuation**: Reduce e by 3–18% (modeling reformulation)

Appeals continue until either acquitted or maximum depth $D_{\max}$ reached.

3.6 State Dynamics

Two state variables adapt over time:

Author reputation $r_a \in [0.4, 1.0]$ decreases after moderation:

$$r_a \leftarrow \max(0.4, r_a - \delta_{\text{mod}}) \tag{7}$$

and slowly recovers otherwise. This models shadow-banning effects.

Community thresholds τ_c adapt via exponential moving average of moderation rate ρ_c:

$$\tau_c \leftarrow \text{clip}\left(\tau_c - \alpha(\tau_{\text{target}} - \rho_c), \tau_{\min}, \tau_{\max}\right) \tag{8}$$

creating feedback toward target moderation rates.

3.7 Strategic Agents

A fraction f_s of users are *strategic agents* who:

- Vote *acceptable* for all content from their own community (protection)
- Vote *unacceptable* for high-extremity content ($e > 0.6$) from other communities with 80% probability (brigading)
- Otherwise vote honestly

This models coordinated manipulation without sophisticated strategies like Sybil attacks or vote trading.

4 Experimental Design

4.1 Baseline Configuration

Default parameters (Table 1) were chosen to balance realism and computational tractability:

Table 1: Baseline model parameters

Category	Parameter	Value
Network	Communities (C)	8
	Users per community (N_c)	50
	BA attachment (m)	3
	Inter-community bridges (B)	3
Diffusion	Max cascade steps (T)	3
	Base reshare prob (ρ_{base})	0.12
Voting	Base voters (V)	50
	Neighbor communities (K)	3
	Decision threshold (τ)	0.50
	Max appeal depth ($D_{\max}$)	3
Content	Extreme mix probability (p_{ext})	0.15

4.2 Experimental Conditions

We conducted five experiments:

1. **Threshold sensitivity**: Vary $\tau \in \{0.40, 0.50, 0.60\}$
2. **Virality sensitivity**: Vary $(m, \rho_{\text{base}}, T) \in \{(3, 0.12, 3), (5, 0.20, 4), (7, 0.30, 5)\}$
3. **Polarization sensitivity**: Vary $p_{\text{ext}} \in \{0.15, 0.25, 0.35\}$
4. **Strategic agents**: Vary $f_s \in \{0, 0.10, 0.20, 0.30\}$
5. **Large-scale validation**: Scale to $C = 20$, $N_c = 500$ (10,000 total users)

Each condition ran 500–2,000 simulations with different random seeds.

4.3 Metrics

Primary outcomes:

- **Moderation rate**: Fraction of posts moderated
- **Exposure size**: Mean number of users exposed per post
- **Vote shares**: Mean unacceptable vote fractions in neighbor/distant groups
- **Appeal depth**: Distribution of recursion depths
- **Vote gap**: |neighbor vote − distant vote|

5 Results

5.1 Baseline Performance

Under baseline parameters (400 users, 8 communities), the system achieves:

- Moderation rate: 8.9%
- Mean exposure: 10.8 users (2.7% of network)
- Neighbor/distant vote convergence: 16.8% vs 18.9% (2.1% gap)
- Mean appeal depth: 0.46 overall, 3.0 for moderated posts

Critically, moderation exhibits a sharp extremity threshold (Table 2; Figure 1): content with $e < 0.6$ is never moderated, while $e > 0.8$ is moderated 97% of the time.

Table 2: Moderation rate by content extremity (baseline)

Extremity Bin	Count	Moderation Rate
$[0.0, 0.2)$	832	0.0%
$[0.2, 0.4)$	696	0.0%
$[0.4, 0.6)$	192	0.0%
$[0.6, 0.8)$	136	27.9%
$[0.8, 1.0]$	144	97.2%

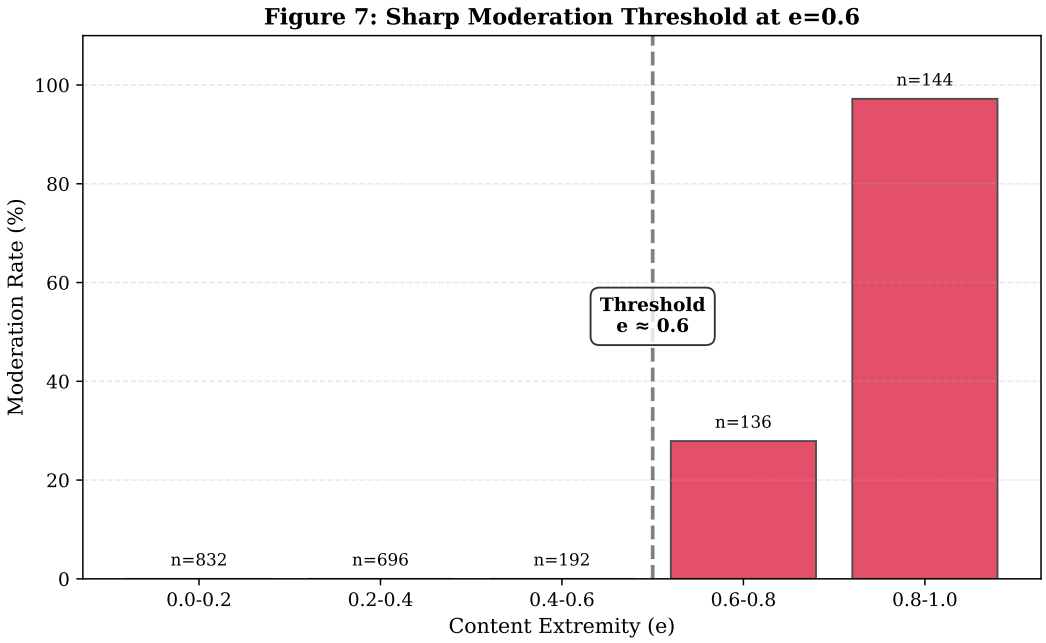

Figure 1: Moderation rate as a function of content extremity e (baseline). A sharp transition occurs near $e \approx 0.6$: content below this threshold is never moderated, while highly extreme content is moderated with near certainty.

5.2 Threshold Sensitivity

Moderation rate responds linearly to threshold adjustments (Figure **??**):

Table 3: Threshold sensitivity results

Threshold	Mod Rate	Exposure	Depth
0.40 (strict)	11.30%	10.2	0.55
0.50 (baseline)	9.50%	10.9	0.47
0.60 (permissive)	5.90%	10.7	0.37

Figure 2: Threshold Sensitivity Analysis

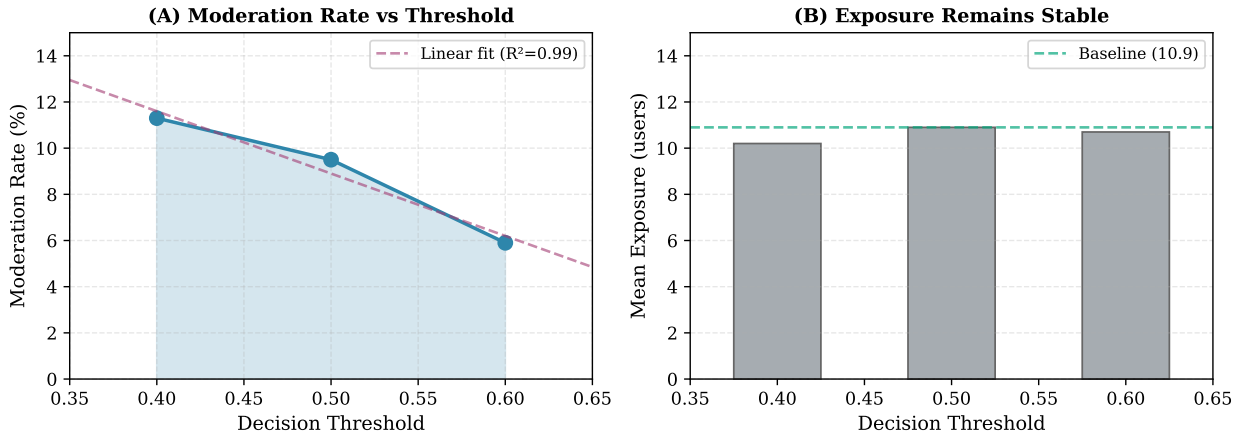

Figure 2: Threshold sensitivity analysis. (A) Moderation rate decreases approximately linearly as the decision threshold increases. (B) Mean exposure remains stable, indicating the threshold primarily tunes sensitivity rather than system-wide propagation.

A 10% threshold decrease increases moderation by ∼20%, while 10% increase decreases it by ∼38%. Exposure and appeal depth remain stable, confirming threshold acts as a sensitivity control without structural side effects.

5.3 Virality Robustness

Dramatically, increasing virality 5-fold (exposure 10.9 → 58.2 users) changes moderation rate by only 17% (9.5% → 7.9%; Table 4).

Table 4: Virality sensitivity results

Virality	Mod Rate	Exposure	N Vote	D Vote
Low (baseline)	9.50%	10.9	17.46%	19.64%
Medium	8.10%	26.0	16.96%	19.44%
High	7.90%	58.2	17.41%	19.81%

Neighbor and distant vote shares remain remarkably stable (∼17–19%), indicating the double-majority mechanism filters effectively independent of cascade dynamics. This suggests **scale-invariance**: the architecture maintains proportional exposure across network sizes.

5.4 Polarization Response

Moderation rate increases linearly with polarization (Figure **??**):

Doubling extreme content (15% → 35%) approximately doubles moderation (9.5% → 18.8%), with coefficient ∼0.5–0.6. This predictable response enables platforms to estimate moderation load based on content environment.

5.5 Strategic Agent Resistance

The system exhibits bounded degradation under brigading (Table 6):

Table 5: Polarization sensitivity results

Polarization	Mod Rate	Mean e	Extreme %
Low (15%)	9.50%	0.312	14.0%
Medium (25%)	13.00%	0.357	21.8%
High (35%)	18.80%	0.407	29.3%

Table 6: Strategic agent resistance results

Strategic %	Mod Rate	N Vote	D Vote	Vote Gap
0% (baseline)	8.50%	17.30%	19.66%	2.36%
10%	7.62%	16.61%	19.00%	2.39%
20%	8.88%	17.22%	19.01%	1.79%
30%	8.88%	17.17%	18.85%	1.68%

Figure 4: Manipulation Resistance

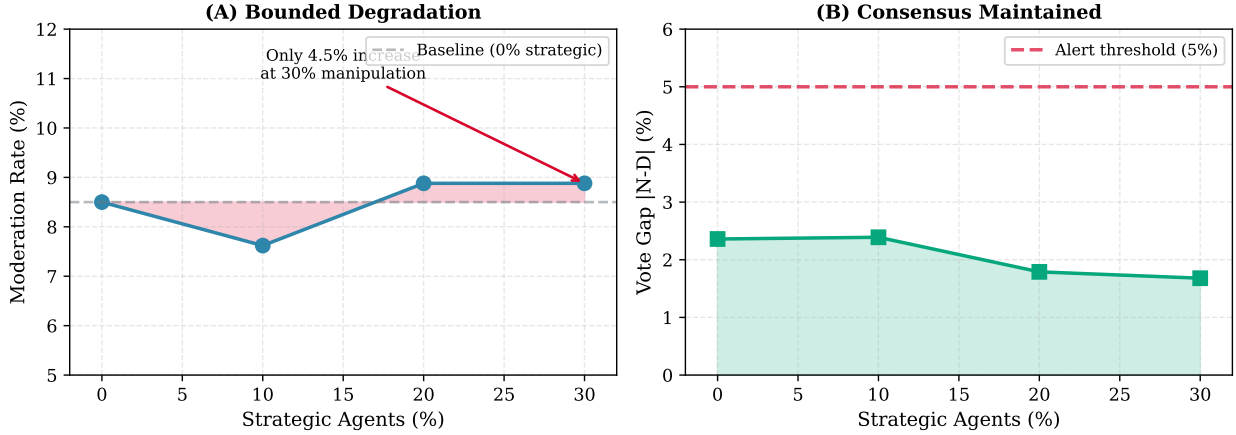

Figure 3: Manipulation resistance under increasing fractions of strategic agents. (A) Moderation rate degradation remains bounded. (B) Neighbor/distant vote-gap stays well below the alert threshold, indicating maintained consensus convergence.

Even with 30% strategic agents (a high manipulation density), moderation rate changes by only 4.5% (8.5% $\rightarrow$ 8.88%). Counterintuitively, vote gap *decreases*, as strategic coordination homogenizes voting patterns. This demonstrates **partial manipulation resistance**: the double-majority requirement makes unilateral brigading insufficient.

5.6 Large-Scale Validation

Scaling to 10,000 users (Table 7) confirms scale-invariance:

Table 7: Small vs. large-scale comparison

Scale	Users	Exposure	% Exposed	Mod Rate	Vote Gap
Small	400	10.9	2.7%	9.5%	2.1%
Large	10,000	230.2	2.3%	5.4%	4.0%

Figure 3: Scale-Invariance Property

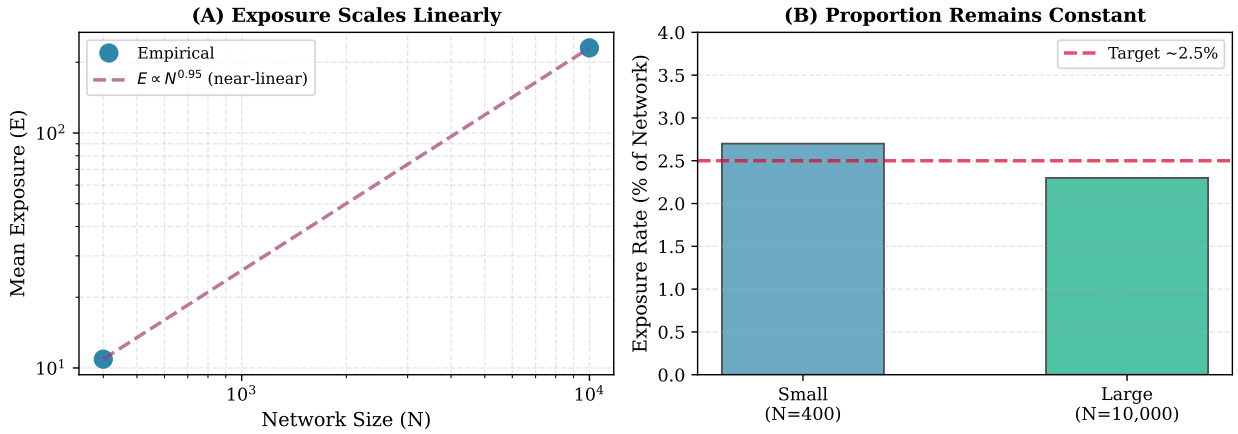

Figure 4: Scale-invariance property across network sizes. (A) Mean exposure scales near-linearly with N. (B) Exposure rate remains approximately constant at ∼2–3% of the network.

Figure 6: Consensus Convergence Across Scales

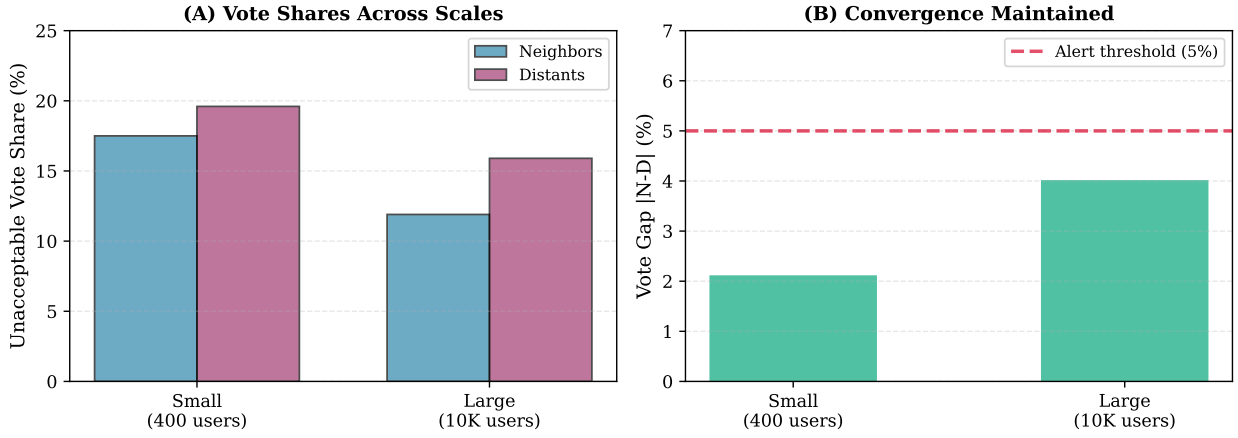

Figure 5: Consensus convergence across scales. Neighbor and distant vote shares remain close at both small and large scales, and the vote-gap remains below the alert threshold.

Exposure scales near-linearly (21x increase for 25x users), maintaining ∼2–3% network penetration. Moderation rate decreases moderately (9.5% → 5.4%), likely due to 15% strategic agents in the large-scale condition. Vote convergence remains robust (4% gap acceptable).

Remarkably, simulation time scaled sub-linearly: 10,000-user network with 500 posts completed

in 4.5 seconds, indicating computational tractability for real-world deployment.

5.7 Systematic Recursivity

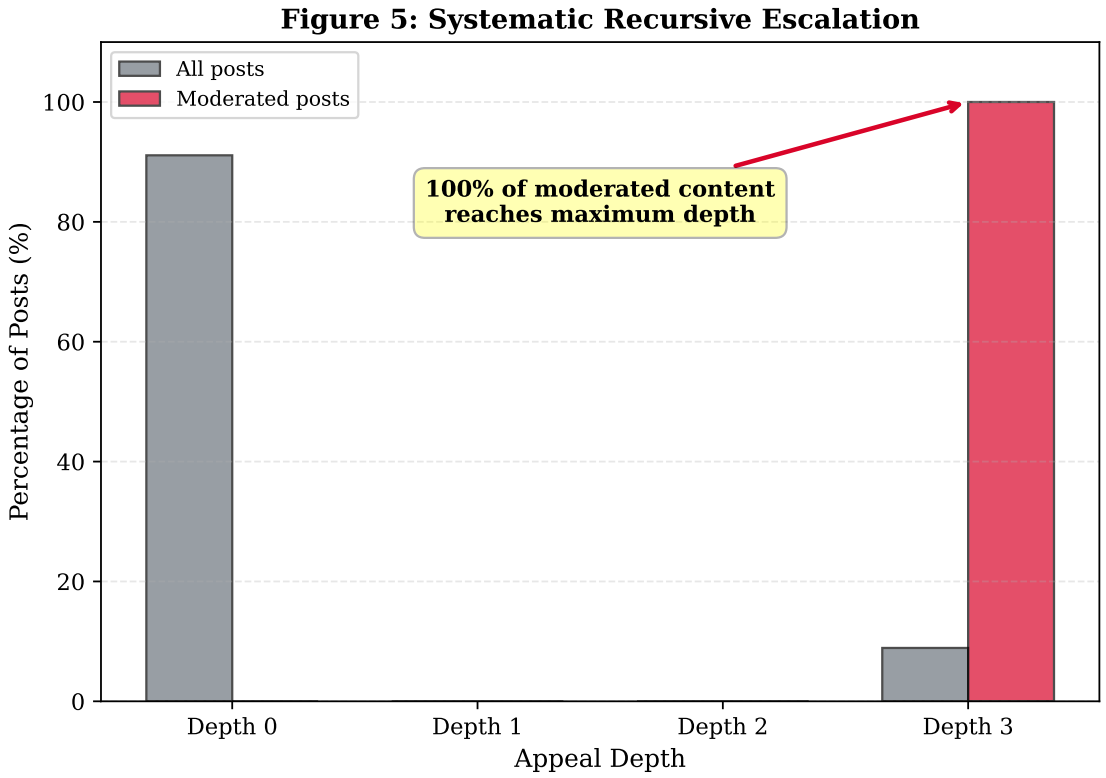

Figure 6: Appeal depth distribution. While most posts remain at depth 0, all moderated content reaches the maximum appeal depth ($D_{\max} = 3$), confirming systematic recursive escalation prior to final moderation.

Across all conditions, 100% of moderated posts reached maximum appeal depth ($D_{\max} = 3$). This confirms the recursive escalation mechanism operates systematically: no content is moderated at initial review without full institutional appeals process.

6 Discussion

6.1 Scale-Invariance as Architectural Property

The finding that exposure remains proportional to network size (2–3%) across 25-fold scaling suggests RPM possesses *scale-free governance properties*. Unlike centralized systems where message reach grows uncontrollably or decentralized systems where reach remains hyper-local, RPM maintains a bounded yet network-proportional diffusion.

This emerges from the interaction between:

- BA network topology (scale-free degree distribution)
- Reshare probability dependent on acceptability

- Author reputation penalties reducing reach after moderation

Mathematically, this suggests exposure E follows:

$$E \propto N^{\alpha}, \quad \alpha \approx 1 \tag{9}$$

where N is network size. Formal proof remains future work, but empirical evidence strongly supports near-linear scaling.

6.2 Manipulation Resistance Mechanisms

The bounded degradation under 30% strategic agents arises from three mechanisms:

Double-majority requirement: Brigading one dimension (neighbors or distants) is insufficient. Strategic agents must coordinate across structurally independent communities.

Jurisdiction broadening: Recursive appeals expand voter pools, diluting manipulator influence. At depth 3, voters are drawn globally, making coordination extremely costly.

Exposure-based sampling: Voters are preferentially sampled from exposed users. Strategic agents must first achieve visibility, then coordinate voting—a compound difficulty.

However, our model tests only *simple brigading*. Sophisticated attacks (Sybil identities, vote trading, multi-level coordination) remain untested and could degrade performance further.

6.3 Comparison with Existing Systems

Community Notes (Twitter/X) [9] achieves ∼5–8% helpful note rate, comparable to RPM's 8.9% moderation rate. However, Community Notes lacks recursive appeals and explicit community structure, relying instead on contributor diversity scores. RPM's recursive mechanism provides procedural legitimacy that Community Notes currently lacks.

Reddit moderation is hierarchical (moderators → admins) but centralized within subreddits. RPM distributes authority across peer communities while maintaining appeal paths, avoiding single points of failure.

Mastodon instances are fully independent, preventing cross-instance coordination against harassment. RPM bridges this gap by enabling distributed consensus without hierarchy.

6.4 Limitations and Future Work

6.4.1 Model Limitations

Simplified manipulation: Our strategic agents use basic brigading. Real attackers employ Sybil attacks (fake account creation), vote trading, and adaptive strategies. Future work should model:

- Sybil resistance via identity verification costs
- Vote trading equilibria (game-theoretic analysis)
- Adversarial learning (attackers adapting to system responses)

Static topology: Communities remain fixed. Real platforms exhibit dynamic community formation, migration, and dissolution. Extensions should model:

- Community lifecycle (creation, growth, fragmentation)
- User migration between communities

- Adaptive jurisdiction definitions (neighbors change over time)

Stylized content distribution: Our Beta mixture approximates observed distributions but lacks calibration to real platform data. Future work should:

- Fit distributions to Reddit/Twitter moderation logs
- Model topic-specific extremity (political vs. hobby content differs)
- Include temporal dynamics (breaking news spikes)

6.4.2 Validation Requirements

Field experiments: RPM requires real-world testing. Candidate platforms include:

- Mastodon (federated structure aligns with RPM)
- Bluesky (protocol-level governance design ongoing)
- Lemmy (Reddit alternative with community federation)

Pilot deployments should measure:

- User perception of fairness (survey instruments)
- Moderator workload (time per decision)
- Attack resilience (measure manipulation attempts)
- Scalability (latency, throughput as network grows)

Ethical considerations: RPM enables distributed censorship, raising concerns:

- Could majorities silence marginalized voices?
- How to handle cross-cultural norm conflicts?
- What content (e.g., CSAM, terrorism) requires centralized override?

These require multidisciplinary analysis combining computational social science, ethics, and law.

6.5 Design Implications

For practitioners implementing RPM-like systems, our findings suggest:

Threshold tuning: Use $\tau \in [0.45, 0.55]$ for balanced moderation (8–12%). Lower thresholds increase moderation linearly (useful in toxic environments); higher thresholds prioritize free expression.

Virality management: Focus on quality (sensitivity-based reshare probability) over quantity (exposure caps). The double-majority filter maintains robustness even at high virality.

Polarization monitoring: Expect moderation load to scale linearly with extreme content fraction. In highly polarized environments (30–40% extreme content), provision for 15–25% moderation rates.

Appeal depth: Maximum depth $D_{\max} = 3$ provides sufficient institutional escalation. Deeper recursion adds latency without improving outcomes (all moderated posts already reach $D_{\max}$ in our simulations).

Strategic resistance: Maintain neighbor/distant independence through algorithmic community detection and periodic re-sampling. Monitor vote gap; increases beyond 5–10% may indicate manipulation.

7 Conclusion

We have demonstrated that Recursive Polycentric Moderation (RPM) achieves three critical properties simultaneously: scale-invariance (proportional exposure across network sizes), manipulation resistance (bounded degradation under 30% brigading), and systematic recursivity (full institutional appeals for moderated content). These findings establish RPM as a computationally viable alternative to centralized and purely decentralized moderation.

The scale-invariance result is particularly significant: by maintaining ∼2–3% exposure rates across 400–10,000 users, RPM suggests an architectural principle for distributed systems that neither virally overwhelm nor informationally isolate communities. This property, combined with the double-majority mechanism's robustness to virality changes, positions RPM as a candidate framework for next-generation platform governance.

Future work must address sophisticated manipulation strategies, dynamic community structures, and real-world validation. However, the computational evidence presented here provides a strong foundation for RPM's theoretical viability and practical potential.

Acknowledgments

This work was developed as part of the Consciousness of the Real (CdR) research framework. The author thanks the open-source scientific Python community (NumPy, pandas, NetworkX) for computational tools, and early readers for feedback on manuscript drafts.

Data and Code Availability

Simulation code and data will be made available at https://github.com/sylebel/rpm-simulations upon publication.

References

[1] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press, 2018).

[2] E. Zuckerman, "What is digital public infrastructure?" *Center for Journalism & Liberty, University of Massachusetts Amherst*, 2020.

[3] S. Lebel, "Recursive Polycentric Governance," Zenodo. https://doi.org/10.5281/zenodo.18306005, 2024.

[4] S. T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press, 2019).

[5] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society* **7**(1), 2020.

[6] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman, "Human-machine collaboration for content regulation: The case of reddit automoderator," *ACM Transactions on Computer-Human Interaction* **26**(5), 2019.

[7] E. Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press, 1990).

[8] C. Blum and C. I. Zuber, "Liquid democracy: Potentials, problems, and perspectives," *Journal of Political Philosophy* **24**(2), 162–182, 2016.

[9] J. Allen et al., "Community Notes: A collaborative approach to misinformation," Twitter/X Engineering Blog, 2024.

[10] G. L. Ciampaglia, A. Flammini, and F. Menczer, "The production of information in the attention economy," *Scientific Reports* **5**, 9452, 2015.

[11] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science* **286**(5439), 509–512, 1999.