

Who Gets the Blame? A Structural Theory of Scapegoating in Social Networks

Qinyang Yu*

“The best way to bring folks together is to give them a really good enemy.”

— The Wizard, in *Wicked*

Abstract

Scapegoating, the unfair attribution of blame, disproportionately targets marginalized individuals. While existing explanations emphasize identity attributes, this study develops a structural theory showing how network position shapes vulnerability to scapegoating. A leader facing systemic failure strategically decides whether to self-blame or scapegoat, while agents embedded in a network update beliefs and form opinions through social learning. Game-theoretic analysis reveals that scapegoating systematically targets peripheral nodes with low centrality. Computational simulations demonstrate that this structural vulnerability persists under stochastic heterogeneity and is further amplified by social identity bias. Together, the findings illuminate the structural mechanisms through which inequality and social injustice emerge in interconnected communities, offering implications for interventions that enhance information transparency and strengthen cross-group ties within social systems.

1 Introduction

The metaphor of the “scapegoat,” or “the goat that departs,” originates from the ritual of atonement in the Hebrew Bible, where a live goat symbolically carries away the sins of the community (Crossman, 2019; Omanson, 1991). Early anthropological and sociological studies interpreted this ritual as a universal cultural practice to remove collective guilt, restore cohesion, and sustain social order (Burkert, 1983; Frazer, 1900; Girard, 1977).

*Department of Economics and Computer Science, Duke University. E-mail: qinyang.yu@duke.edu

Over time, the notion of scapegoating has expanded beyond religion to describe the general act of shifting blame onto innocent individuals or groups (Girard, 1989). This phenomenon recurs throughout human history from European witch hunts (Arsal and Yavuz, 2014; Campbell, 2012; Levack, 2013) and antisemitic persecution (Arendt and May, 1958; Gibson and Howard, 2007) to contemporary examples during economic shocks (Luca et al., 2022; Miguel, 2005) and in organizational contexts (Dezso, 2009; Winter, 2001).

Scapegoating has been studied across political science, social psychology, and economics. Political theorists view it as an authoritarian strategy for consolidating power, deflecting accountability, and mobilizing hostility (Arendt and May, 1958; Bramoullé and Morault, 2021; Cavanaugh, 2011; Gent, 2009; Herman and Chomsky, 2021). Social psychologists emphasize its role in reducing collective anxiety (Allport, 1954; Glick, 2005; Newman and Caldwell, 2005) and its emergence through group dynamics and social contagion (Bandura, 2024; Garfinkel, 2023; Zimbardo, 2007). Economists investigate the incentive structures that make scapegoating individually rational or institutionally viable (Bauer et al., 2023; Bursztyn et al., 2022; Zussman, 2021).

Yet despite extensive research, the question of who becomes a scapegoat and why remains analytically complex. Empirically, blame tends to fall on marginalized individuals or minority subgroups who are weak, disadvantaged, and socially peripheral (Bettelheim and Janowitz, 1950; Dollard et al., 2013; Tajfel, 1979). Existing explanations, particularly identity-based theories, attribute this pattern to intergroup bias between in-group and out-group members: marginalized out-group individuals, defined by ethnicity, religion, gender, or political affiliation, are perceived as socially distant or low in power and therefore become convenient targets for in-group individuals (Brewer, 1979; Tajfel, 1979). While identity categories powerfully predict who is blamed, they leave open a deeper structural question: is identity the only determinant?

This study complements *identity-based* explanations, which focus on *who* individuals are, with a *network-based* perspective emphasizing *where* they are positioned within social structures. The two dimensions often intertwine, but by isolating structural mechanisms, this paper clarifies how social position shapes vulnerability beyond categorical identity.

Social networks shape information flow, opinion formation, and collective decision-making. Prior research on conflictual ties (Doreian and Mrvar, 1996; Heider, 1946; Labianca and Brass, 2006), bullying and victimization networks (Faris and Felmlee, 2011; Huitsing et al., 2012; Salmivalli et al., 1996), and social control (Shirado et al., 2013; Takács et al., 2008) shows that visibility and structural exposure influence who becomes a target of hostility or

exclusion. However, most network models treat sanctioning as a *decentralized* outcome of peer interaction. In contrast, real-world blame dynamics often arise in hierarchical settings where leaders intentionally redirect responsibility.

To address this gap, the present study develops a formal game-theoretic model of scapegoating as a *strategic, leader-driven* mechanism that exploits structural vulnerability for blame allocation.

The remainder of the paper proceeds as follows. Section 2 introduces the formal model and derives equilibrium conditions. Section 3 analyzes how network effects generate local and global structural vulnerability. Section 4 extends the model with stochastic heterogeneity and identity bias. Section 5 outlines a behavioral experiment linking theoretical predictions to human decision-making, and Section 6 concludes.

2 The Model

2.1 Model Setup

Suppose a community consists of a leader and a population of $|N| = n$ agents. All agents are embedded in a social network represented by an undirected graph $G = (V, E)$, where each node $v \in V$ corresponds to an individual and each edge $e \in E$ represents a social connection between two agents. The social network $G = (V, E)$ can be weighted or unweighted, and may be connected or disconnected. Let A denote its adjacency matrix, where each element A_{ij} captures the presence or strength of the tie between i and j .

The community faces a systemic crisis that affects all members. Regardless of its true origin, the situation requires that responsibility be assigned to some individual or group. Each agent i has a binary true state $\theta_i \in \{0, 1\}$, where $\theta_i = 0$ indicates innocence and $\theta_i = 1$ indicates guilt with respect to the problem. Specifically, we assume $\theta_i = 0$ for all $i \in V$, so that all agents are in fact innocent. This assumption allows us to focus on the conditions under which an innocent individual becomes the target of blame, capturing the essence of scapegoating.

The game consists of two phases: scapegoat selection and opinion dynamics.

The first phase models the leader’s decision-making process, drawing on frameworks such as the “cheap talk” information game (Crawford and Sobel, 1982) and the public broadcast model (Bloch et al., 2018). At the beginning of this phase, the leader observes the network

structure G and the true states of all agents. The leader’s strategy is defined as

$$s_L : G \rightarrow V \cup \{\emptyset\}. \quad (1)$$

If $s_L(G) = \emptyset$, the leader chooses to self-blame and the game terminates immediately; if $s_L(G) = k$ for some $k \in V$, the leader scapegoats agent k by publicly announcing $\theta_k = 1$. This announcement is observed by all agents and becomes common knowledge, and the game proceeds to the second phase.

The leader’s utility function is defined as

$$v_L(s_L) = \begin{cases} -C & \text{if } s_L(G) = \emptyset, \\ -R^{(k)} & \text{if } s_L(G) = k \text{ for some } k \in V. \end{cases} \quad (2)$$

If the leader self-blames, they bear a fixed cost $C > 0$ for assuming responsibility for the crisis. If they scapegoat agent k , they avoid the fixed cost but incur a reputational cost $R^{(k)}$ when the community does not fully accept his accusation, which will be derived in the second phase.

The second phase models the opinion dynamics among agents in the social network. Prior to the leader’s scapegoating decision, each agent i holds a prior belief about k ’s guilt, denoted by $\pi_i^{(k)}$. Trivially, we assume $\pi_k^{(k)} = 0$ for all $k \in V$.

Agents share a common belief that the leader always accuses a guilty agent with probability 1 in the interest of social justice but may accuse an innocent agent with probability $p \in [0, 1]$ due to personal motives. The parameter p captures the level of public trust: $p = 0$ indicates full trust in the leader, whereas $p = 1$ reflects complete distrust. In the benchmark case, p is assumed to be homogeneous across agents.

After the leader announces their accusation of agent k , each agent i updates their belief about k ’s guilt using Bayes’ rule. The posterior belief is

$$b_i^{(k)} = \frac{\pi_i^{(k)}}{\pi_i^{(k)} + p(1 - \pi_i^{(k)})}. \quad (3)$$

To avoid the degenerate case in which both $\pi_i^{(k)}$ and p are zero, we assume $p > 0$.

After updating their beliefs, each agent i forms a public opinion $x_i^{(k)} \in [0, 1]$ representing the perceived probability that agent k is guilty. While $b_i^{(k)}$ captures the private belief before social interaction, $x_i^{(k)}$ evolves endogenously through network-based social learning, incorporating the influence of connected peers. Each agent’s strategy is therefore to choose $x_i^{(k)}$

given their updated belief $b_i^{(k)}$ and the opinions of their neighbors.

The utility of agent i includes two components: a penalty for deviating from their private belief and a penalty for disagreeing with their neighbors, following the game-theoretic formulation of opinion dynamics in [Bindel et al. \(2015\)](#) and [Ghaderi and Srikant \(2014\)](#).

Formally,

$$u_i(x_i^{(k)}, x_{-i}^{(k)}) = \underbrace{-(x_i^{(k)} - b_i^{(k)})^2}_{\text{Penalty 1: Self fidelity}} - \underbrace{\sum_{j=1}^n A_{ij}(x_i^{(k)} - x_j^{(k)})^2}_{\text{Penalty 2: Peer conformity}}, \quad (4)$$

where $x_i^{(k)}$ is agent i 's public opinion about agent k , $b_i^{(k)}$ is their private belief, and A_{ij} is the (i, j) entry of the adjacency matrix, with $A_{ij} \neq 0$ if and only if agents i and j are directly connected. Thus, each agent is influenced only by the opinions of their neighbors.

Once the equilibrium opinions $x_i^{*(k)}$ are obtained in the second phase, the leader's reputational cost $R^{(k)}$ can be determined accordingly. This cost reflects the extent to which the community accepts the leader's accusation of agent k , and is defined as the total disagreement between the community's equilibrium opinions and the leader's announcement:

$$R^{(k)} = \sum_{i=1}^n (1 - x_i^{*(k)}). \quad (5)$$

2.2 Equilibrium Analysis

The two-stage game can be solved by backward induction. In the second phase, let the vector of agents' equilibrium opinions be $x^{(k)} = (x_1^{(k)}, \dots, x_n^{(k)})^T$, and the vector of posterior beliefs be $b^{(k)} = (b_1^{(k)}, \dots, b_n^{(k)})^T$. Let L denote the Laplacian matrix of the graph G , defined as $L = D - A$, where D is the diagonal degree matrix.

The Nash equilibrium of this opinion-formation game is characterized as follows.

Lemma 1. *The opinion game in the second phase admits a unique Nash equilibrium given by $x^{*(k)} = (I + L)^{-1}b^{(k)}$.*

Proof. In the Nash equilibrium, each agent maximizes their utility function, which is concave in $x_i^{(k)} \in [0, 1]$. The optimality condition $\frac{\partial u_i^{(k)}}{\partial x_i^{(k)}} = 0$ yields

$$(x_i^{(k)} - b_i^{(k)}) + \sum_{j=1}^n A_{ij}(x_i^{(k)} - x_j^{(k)}) = 0, \quad \forall i \in V. \quad (6)$$

Rearranging terms gives

$$\left(1 + \sum_{j=1}^n A_{ij}\right)x_i^{(k)} - \sum_{j=1}^n A_{ij}x_j^{(k)} = b_i^{(k)}, \quad \forall i \in V. \quad (7)$$

In matrix form, this can be written as $(I+D-A)x^{(k)} = b^{(k)}$, or equivalently, $(I+L)x^{(k)} = b^{(k)}$. Since the Laplacian matrix L is positive semidefinite, $I+L$ is positive definite and therefore invertible. Thus, the Nash equilibrium is uniquely given by $x^{*(k)} = (I+L)^{-1}b^{(k)}$. A similar proof appears in [Bindel et al. \(2015\)](#). \square

Since L characterizes the algebraic connectivity of G , the equilibrium opinion vector $x^{*(k)}$ reflects how individual beliefs $b^{(k)}$ aggregate through the network structure.

The equilibrium condition in the first phase follows directly from the leader's utility function. If $R^{(k)} > C$ for all $k \in V$, scapegoating becomes too costly, and the leader's best response is to self-blame. Conversely, if $R^{(k)} \leq C$ for some $k \in V$, the leader strategically selects the scapegoat that minimizes their reputational cost. Therefore, the equilibrium strategy for the leader is

$$s_L^*(G) = \begin{cases} \emptyset, & \text{if } R^{(k)} > C \text{ for all } k \in V, \\ \arg \min_{k \in V} R^{(k)}, & \text{if } R^{(k)} \leq C \text{ for some } k \in V. \end{cases} \quad (8)$$

In the case of multiple agents attaining the same minimum reputational cost, the leader is assumed to select one randomly from this set. The framework can further be extended to an l -scapegoat setting with $l \geq 2$, in which the leader blames the l agents with the lowest $R^{(k)}$. This generalization captures situations where responsibility is collectively assigned to a small minority group. For analytical clarity, however, we focus on the baseline case of $l = 1$.

Next, we substitute the equilibrium solutions from the second phase into the leader's utility function in the first phase to derive the condition for scapegoating. For analytical convenience, the following lemma simplifies this derivation.

Lemma 2 (Opinion Conservation). *For all $k \in V$, $\sum_{i=1}^n x_i^{*(k)} = \sum_{i=1}^n b_i^{(k)}$.*

Proof. Starting from the equilibrium condition $(I+L)x^{*(k)} = b^{(k)}$, multiply both sides by the vector of ones $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^n$, giving $\mathbf{1}^T(I+L)x^{*(k)} = \mathbf{1}^T b^{(k)}$. Since the Laplacian matrix L satisfies $\mathbf{1}^T L = 0$, it follows that $\mathbf{1}^T x^{*(k)} = \mathbf{1}^T b^{(k)}$, or equivalently $\sum_{i=1}^n x_i^{*(k)} = \sum_{i=1}^n b_i^{(k)}$. This completes the proof. \square

Lemma 2 highlights a fundamental property of opinion dynamics: the total sum of opinions after social learning equals the total sum of initial opinions before learning. This conservation implies that social interaction redistributes opinions across the network without creating or eliminating the overall “mass” of beliefs.

Intuitively, this principle is observable in real settings. When individuals in a group exchange views about an event, their personal opinions may shift through discussion, yet the group’s aggregate stance remains unchanged on average.

Using this conservation property, we can substitute the equilibrium opinions into the reputational cost function to obtain

$$R^{(k)} = \sum_{i=1}^n (1 - b_i^{(k)}) = \sum_{i=1}^n \left[1 - \frac{\pi_i^{(k)}}{\pi_i^{(k)} + p(1 - \pi_i^{(k)})} \right], \quad (9)$$

showing that $R^{(k)}$ is determined by the distribution of prior beliefs $\pi_i^{(k)}$ across the network. Because the leader’s decision to scapegoat depends on comparing $R^{(k)}$ with the fixed cost C , and $R^{(k)}$ itself is a function of $\pi_i^{(k)}$, the emergence of scapegoating is ultimately shaped by the social structure of prior beliefs—that is, how members of the community initially perceive one another. This link between the belief structure and scapegoating outcomes introduces the network effects analyzed in the next section, where the priors $\pi_i^{(k)}$ are specified with network-dependent heterogeneity.

3 Network Effects and Structural Vulnerability

In this section, we allow each agent’s prior beliefs about others’ types to vary based on the network structure. Specifically, agents form heterogeneous priors based on their positions in the network: trust declines as the distance between two agents increases.

Assumption 1 (Network Effects). *Let $N(i)$ denote the neighbor set of agent i , and let l_{ik} denote the geodesic distance between i and k . If k is isolated, define $l_{ik} = \infty$ for any $i \neq k$. We assume that*

$$\pi_i^{(k)} = \begin{cases} 0, & \text{if } i \in N(k), \\ \frac{1}{2}, & \text{as } l_{ik} \rightarrow \infty. \end{cases} \quad (10)$$

The limit value $\frac{1}{2}$ represents a neutral prior and may differ across communities, since some communities are generally more trusting whereas others are more skeptical. We take

$\frac{1}{2}$ as an average baseline without loss of generality. Section 4.1 later introduces additional heterogeneity in the entire prior term $\pi_i^{(k)}$.

This specification captures two mechanisms through which network topology shapes prior beliefs: (i) neighbor-based trust, reflecting that direct connections and social ties foster mutual understanding (Coleman, 1988; Granovetter, 1973); and (ii) distance-based skepticism, where limited interaction and information about distant agents increase uncertainty. By assuming that agents observe only their local neighborhoods, the model mirrors realistic information constraints: individuals know their close contacts well but rely on broader network processes, such as rumors or leader broadcasts, to form opinions about others.

We next consider two formulations for how priors vary with distance: (i) discrete network effects and (ii) decay network effects. The resulting belief distributions lead to distinct scapegoat selection conditions, which we analyze in Sections 3.1 and 3.2.

3.1 Discrete Network Effects

We first consider the simpler case where skepticism toward non-neighbors changes discretely.

Assumption 2 (Discrete Network Effects). *A prior belief $\pi_i^{(k)}$ exhibits discrete network effects if*

$$\pi_i^{(k)} = \begin{cases} 0, & \text{if } i \in N(k), \\ \frac{1}{2}, & \text{if } i \notin N(k). \end{cases} \quad (11)$$

This assumption represents step-function beliefs with a clear threshold at network distance one. Agents fully trust their neighbors but regard all others as strangers, assigning them a neutral prior of $\frac{1}{2}$. Substituting the discrete network effects from Assumption 2 into the reputational cost function yields

$$R^{(k)} = \sum_{i \notin N(k), i \neq k} \left(1 - \frac{1}{1+p}\right) + \sum_{i \in N(k)} (1-0) + (1-0) = \frac{1+np + \deg(k)}{1+p}. \quad (12)$$

This monotonicity directly leads to the following equilibrium characterization.

Proposition 1 (Local Structural Vulnerability). *Given discrete network effects, if there exists some $k \in V$ such that $R^{(k)} \leq C$, then the equilibrium strategy of the leader is*

$$s_L^*(G) = \arg \min_{k \in V} \deg(k). \quad (13)$$

This result implies that the leader scapegoats the agent with the lowest degree centrality. Low-degree agents, having limited connections and weaker influence, are less capable of mobilizing support to contest the accusation. Their structural isolation keeps public distrust in the leader’s accusation low, thereby minimizing the leader’s reputational cost. This mechanism illustrates a divide-and-conquer logic: the leader exploits local structural vulnerability to ensure successful blame attribution by securing collective acceptance while avoiding widespread opposition.

Figure 1 shows simulation results under the discrete network effects assumption across four canonical structures: Erdős–Rényi (ER), Watts–Strogatz (WS), Barabási–Albert (BA), and Stochastic Block Model (SBM), each with $n = 30$ nodes and public trust parameter $p = 0.3$. Red nodes mark the agents selected as scapegoats minimizing the reputational cost $R^{(k)}$.

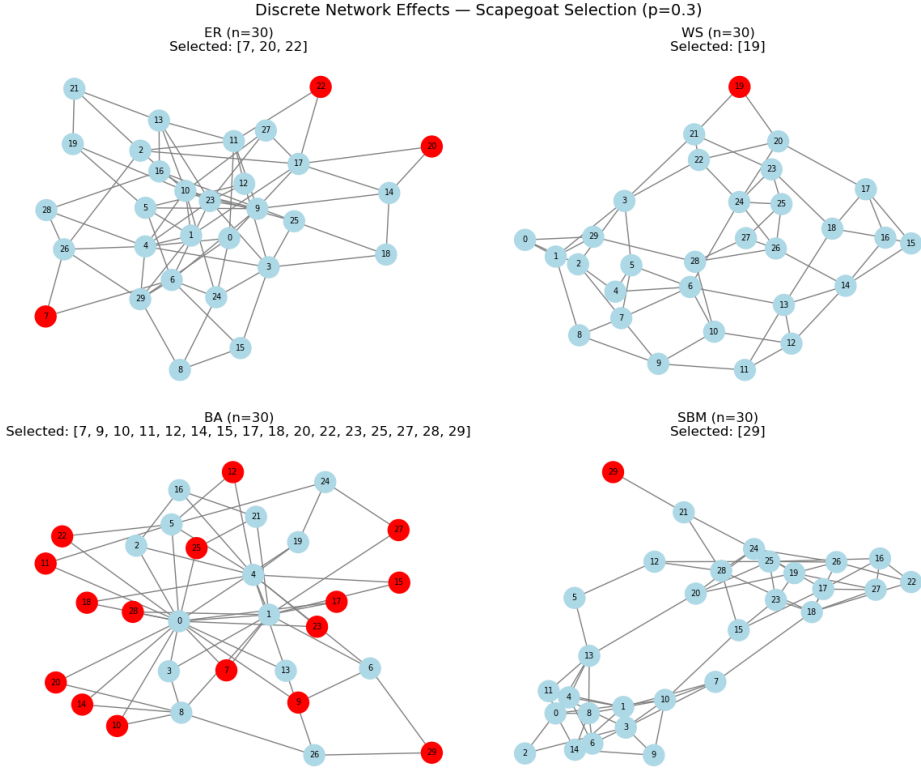


Figure 1: Scapegoat selection with discrete network effects ($p = 0.3$) across four network topologies (ER, WS, BA, SBM). Red nodes denote selected agents.

Consistent with the theoretical prediction, scapegoats mostly occupy low-degree, weakly connected positions, indicating that local structural vulnerability drives blame attribution. In ER, WS, and SBM networks, scapegoat selection concentrates on sparse nodes. However,

in the BA network, where degree distribution is highly skewed, the simulation performs poorly: nearly all locally peripheral nodes are selected, indicating excessive sensitivity to degree heterogeneity.

This limitation suggests that the discrete formulation overemphasizes local connectivity while neglecting broader relational influence. To address this, the next section relaxes the step-function assumption by introducing decay network effects.

3.2 Decay Network Effects

We now consider a more realistic specification where trust decays exponentially with distance.

Assumption 3 (Decay Network Effects). *A prior belief $\pi_i^{(k)}$ exhibits decay network effects if*

$$\pi_i^{(k)} = \begin{cases} 0, & \text{if } i = k, \\ \frac{1}{2} - \left(\frac{1}{2}\right)^{l_{ik}}, & \text{if } i \neq k. \end{cases} \quad (14)$$

This formulation refines the discrete case by allowing beliefs to decay continuously with distance, yielding a smoother and more realistic transition. The marginal decay is stronger for nearby nodes and weaker for distant ones, reflecting that trust declines sharply beyond close connections but stabilizes at low levels for remote individuals. This continuous pattern better captures how trust and skepticism diffuse through real social networks.

Substituting the decay-based prior into the reputational cost function yields a distance-weighted measure of aggregate distrust, defined as $D_k = \sum_{i \in V} (1 - b_i^{(k)})$, where $b_i^{(k)} = \frac{\pi_i^{(k)}}{\pi_i^{(k)} + p(1 - \pi_i^{(k)})}$ and $\pi_i^{(k)} = \frac{1}{2} - \left(\frac{1}{2}\right)^{l_{ik}}$. The leader then selects the agent with the lowest D_k , corresponding to the minimal reputational cost.

Proposition 2 (Global Structural Vulnerability). *Given decay network effects, if there exists some $k \in V$ such that $R^{(k)} \leq C$, the leader's equilibrium strategy is*

$$s_L^*(G) = \arg \min_{k \in V} D_k. \quad (15)$$

Intuitively, the aggregate distrust D_k provides a stricter measure than the classical decay centrality $C_k(\alpha) = \sum_{i \in V} \alpha^{l_{ik}}$, as it penalizes longer distances l_{ik} more heavily and amplifies their effects through its convex form. This makes scapegoat selection more sensitive to an agent's structural position in the network. Nevertheless, D_k remains closely related to $C_k(\alpha)$: it is a component-wise convex transformation of $C_k(\alpha)$ with $\alpha = \frac{1}{2}$, and when $p = 1$, D_k becomes an exact affine transformation of $C_k(\frac{1}{2})$.

Figure 2 empirically confirms this association: the Spearman correlation $\rho(C_k, D_k)$ approaches one as the trust parameter p increases, across all network topologies and sizes. Since we assume $p > 0$, D_k maintains a very high association with C_k , indicating that standard decay centrality serves as an effective predictor for approximating scapegoating outcomes under decay network effects.

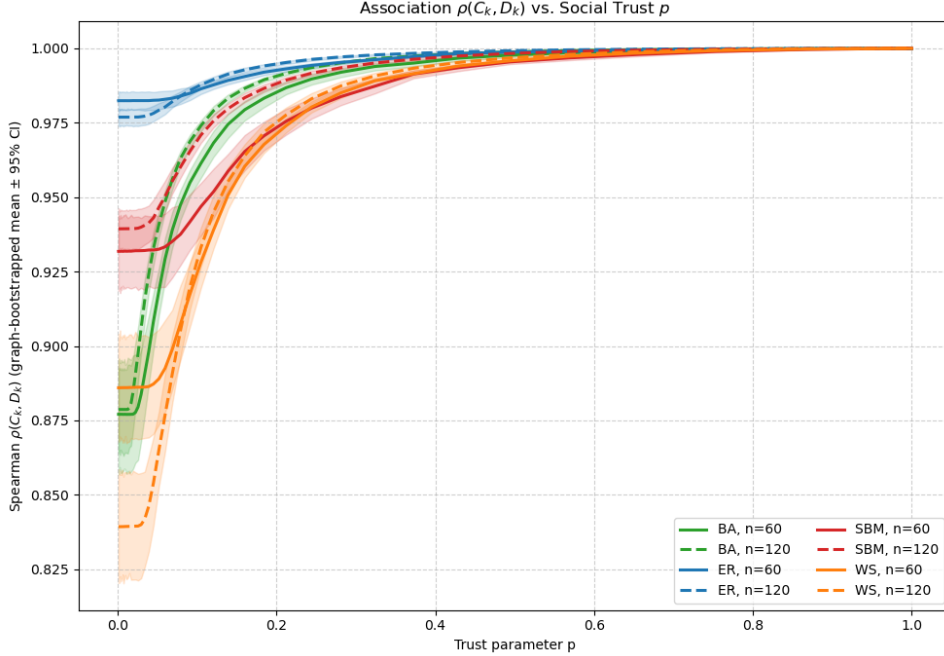


Figure 2: Mean Spearman correlation $\rho(C_k, D_k)$ with 95% bootstrapped confidence intervals across networks of different types and sizes.

Proposition 2 generalizes the result of Section 3.1 by replacing local connectivity, captured by degree centrality, with a distance-weighted measure of global connectivity. Under decay network effects, the leader no longer targets only weakly connected nodes but rather those globally distant within the trust network. The underlying mechanism thus shifts from local structural vulnerability to global structural vulnerability, linking scapegoating behavior to the broader topology of social trust.

Figure 3 shows simulation results under decay network effects using the same network configurations as in Figure 1 ($n = 30, p = 0.3$). Across all topologies, the scapegoat coincides with the node attaining the lowest D_k and, equivalently in this setting, the lowest decay centrality.

Unlike the discrete case, which yields multiple low-degree candidates, the decay formulation mitigates over-sensitivity and produces a unique, stable prediction. The model thereby captures the joint influence of local sparsity and global remoteness: agents who are struc-

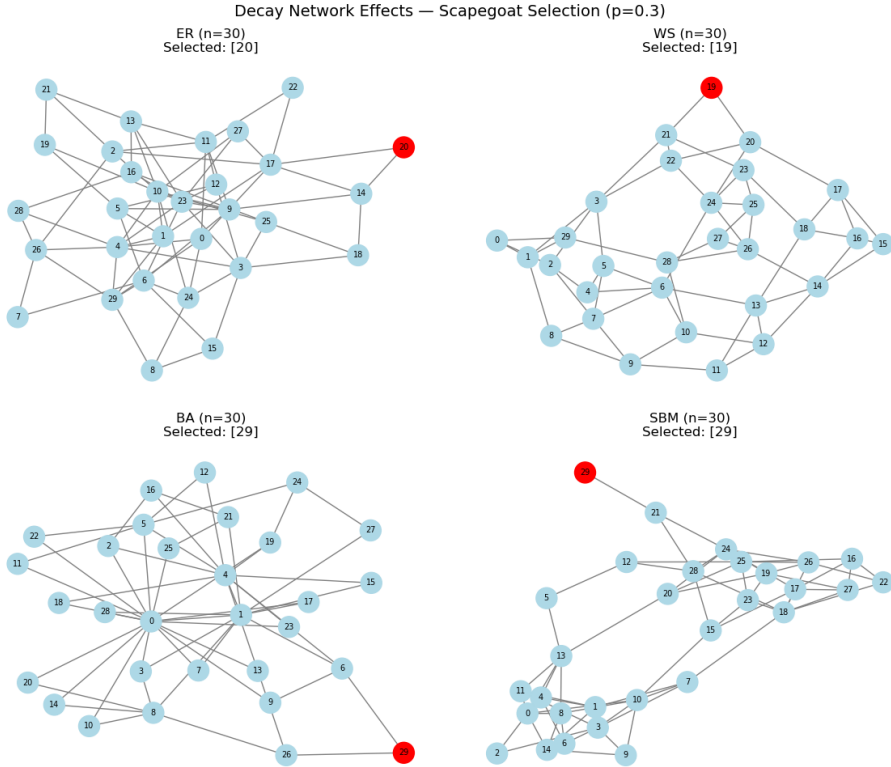


Figure 3: Scapegoat selection with decay network effects ($p = 0.3$) across four network topologies (ER, WS, BA, SBM). Red nodes denote selected agents.

turally isolated not only through limited local connections but also through long paths and restricted access to trusted clusters become the most vulnerable to blame, offering a more realistic account of blame attribution in social networks.

4 Extensions

4.1 From Equilibrium to Complexity

The preceding sections characterize scapegoating as a deterministic equilibrium outcome, linking network effects to local and global structural vulnerability. Real-world social systems, however, operate and evolve under complexity: belief formation, conformity, trust, and judgment differ across individuals and are subject to pervasive uncertainty.

To incorporate these realistic sources of heterogeneity and randomness while preserving the equilibrium framework, we extend the model through four minimal relaxations.

Assumption 4 (Stochastic Network Effects). *A prior belief $\pi_i^{(k)}$ exhibits stochastic network*

effects if

$$\pi_i^{(k)} = \sigma(\beta_0 + \beta_1 l_{ik} + \varepsilon_i), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\pi^2), \quad (16)$$

where $\sigma(\cdot)$ is the logistic function $\sigma(z) = \frac{1}{1+e^{-z}}$; β_0 is the baseline skepticism; $\beta_1 > 0$ is the slope of distance-based skepticism; and σ_π is the noise level.

Assumption 5 (Heterogeneous Public Trust). *An agent's trust toward the leader's accusation follows*

$$p_i = \sigma(\mu_p + \eta_i), \quad \eta_i \sim \mathcal{N}(0, \sigma_p^2), \quad (17)$$

where μ_p is the mean level of public trust and σ_p is the dispersion of trust heterogeneity.

Assumption 6 (Heterogeneous Conformity). *Each agent i balances self fidelity and peer conformity according to an individual conformity (or stubbornness) parameter $\phi_i \in [0, 1]$:*

$$u_i(x_i^{(k)}, x_{-i}^{(k)}) = -(1 - \phi_i)(x_i^{(k)} - b_i^{(k)})^2 - \phi_i \sum_j A_{ij}(x_i^{(k)} - x_j^{(k)})^2, \quad \phi_i \sim U(0, 1). \quad (18)$$

Under this assumption, the corresponding equilibrium opinion vector satisfies

$$(\text{diag}(1 - \phi) + \text{diag}(\phi)L)x^{*(k)} = (1 - \phi) \odot b^{(k)}. \quad (19)$$

Assumption 7 (Stochastic Decision). *The leader's scapegoating rule is probabilistic with bounded rationality, following a soft-min decision function*

$$P(k) \propto \exp[-\lambda(R^{(k)} - \min_j R^{(j)})], \quad (20)$$

where $\lambda > 0$ is the decision-temperature parameter; as $\lambda \rightarrow \infty$, the decision converges to deterministic minimization.

Together, Assumptions 4-7 introduce informational, cognitive, behavioral, and decisional stochasticity and heterogeneity into the baseline framework. We then conduct agent-based simulations to examine whether structural vulnerability persists under these stochastic and heterogeneous conditions.

To provide an initial glimpse of the results, we simulate using the same networks as in Section 3. Figure 4 visualizes scapegoat selection frequencies under stochastic network effects. The previously deterministic scapegoat selections are now represented as frequency heatmaps. Even with randomness introduced, darker nodes, indicating higher selection probabilities, remain concentrated in structurally peripheral regions across all four topologies

(ER, WS, BA, SBM). This preliminary pattern suggests that structural vulnerability persists as a statistical tendency.

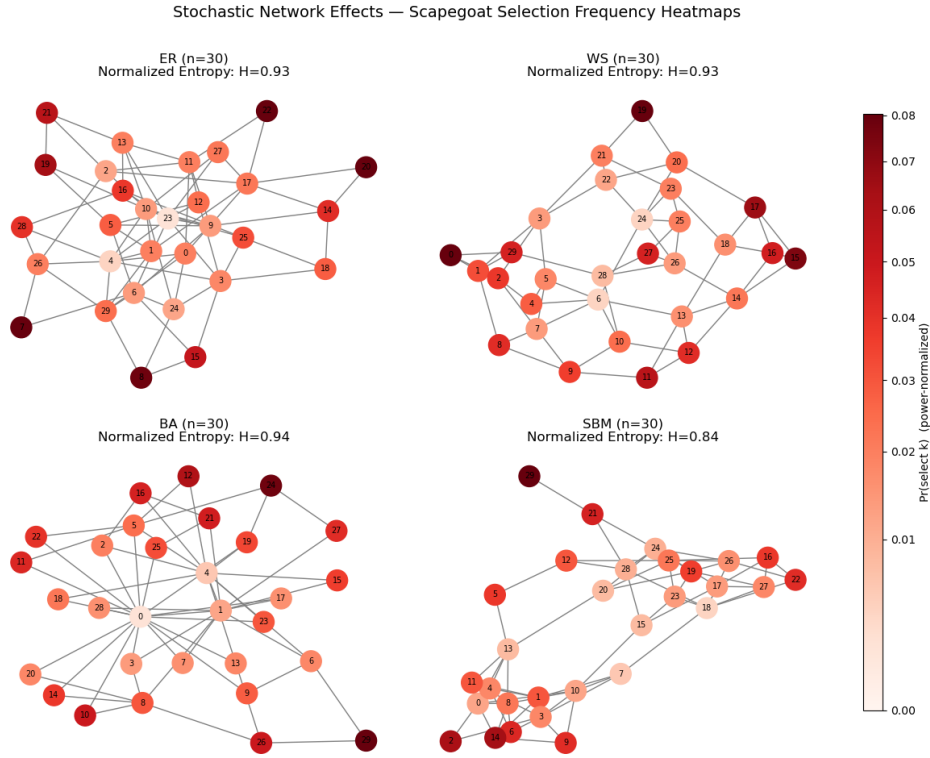


Figure 4: Scapegoat selection frequency heatmaps with stochastic network effects across four network topologies (ER, WS, BA, SBM). Darker nodes indicate higher empirical selection probability. Simulations are based on 500 Monte Carlo runs per parameter combination ($\beta_0 = -2.0$; $\beta_1 \in \{0.2, 0.3, 0.6\}$; $\sigma_\pi \in \{0.5, 1.0, 1.5\}$; $\mu_p = 0$; $\sigma_p = 1.5$; $\lambda = 15$). Reported H^* denotes the normalized entropy of the empirical selection distribution.

We formally verify this pattern by averaging results over five random graphs per topology and introducing two additional centrality measures, betweenness and eigenvector, alongside degree and decay centralities.

Figure 5 plots the scapegoat selection probability against the percentile ranks of each centrality. Across all four network types, the relationship remains predominantly negative: nodes with lower centralities exhibit substantially higher scapegoating probabilities. This confirms that structural vulnerability persists as a statistical regularity under stochastic and heterogeneous conditions, generalizing the equilibrium results from Section 3.

Some deviations appear, however, most notably the irregular pattern in the Barabási–Albert network, which arises from its highly skewed degree distribution. For example, the curve for degree centrality is consistent with the earlier observation that discrete formulations tend

to overemphasize local connectivity, and a similar non-monotonic pattern is observed for betweenness centrality due to its correlation with degree in hub-dominated networks.

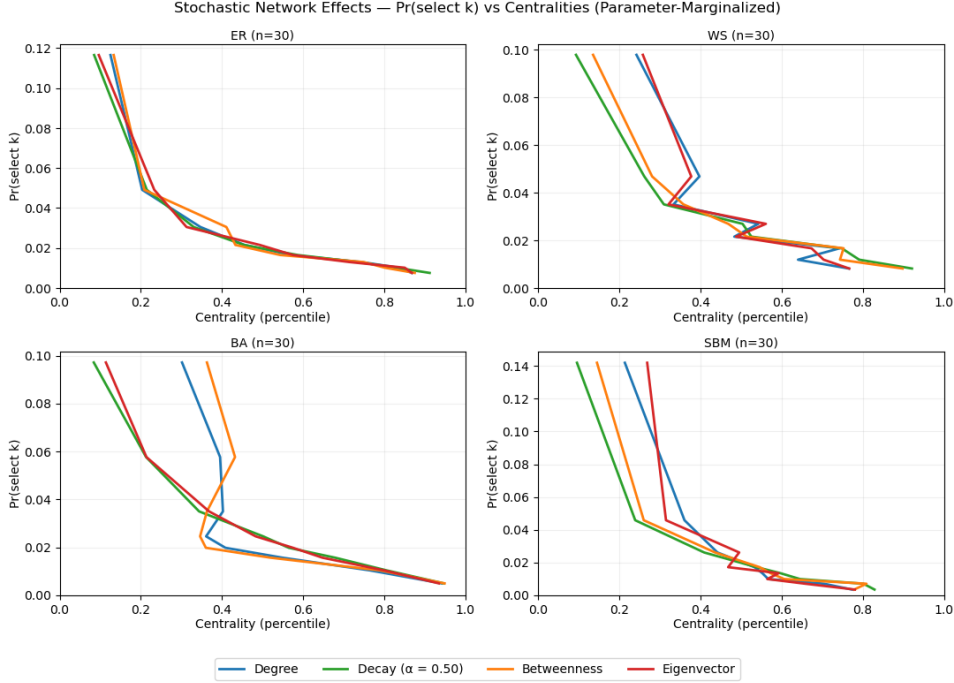


Figure 5: Relationship between scapegoat selection probability and centrality percentile for four measures: degree, decay ($\alpha = 0.5$), betweenness, and eigenvector centrality. Each panel averages results across five random graphs ($n = 30$) per topology with 500 Monte Carlo runs per parameter combination ($\beta_0 = -2.0$; $\beta_1 \in \{0.2, 0.3, 0.6\}$; $\sigma_\pi \in \{0.5, 1.0, 1.5\}$; $\mu_p = 0$; $\sigma_p = 1.5$; $\lambda = 15$).

Finally, we examine the system-level robustness of the scapegoating pattern by tracking the normalized entropy of the empirical selection distribution, denoted H^* . Mathematically, H^* is defined as

$$H^* = -\frac{1}{\log n} \sum_{k=1}^n P_k \log P_k, \quad (21)$$

where $P_k = \Pr(\text{select } k)$ is the empirical probability that agent k is scapegoated. A lower H^* indicates that scapegoating is concentrated on a few specific nodes, while a higher H^* implies a more diffuse blame attribution across the network.

Figure 6 plots H^* over the parameter space (β_1, σ_π) for all four network types. Here, the distance slope β_1 and the prior noise σ_π are the key parameters governing stochastic network effects. Across models, H^* declines as β_1 increases and σ_π decreases, indicating that when beliefs are more distance-sensitive and less noisy, the system converges toward a more deterministic scapegoat pattern. These results show that structural vulnerability is

an emergent system-level property, strengthened by distance sensitivity and weakened by stochastic disorder in agents’ priors.

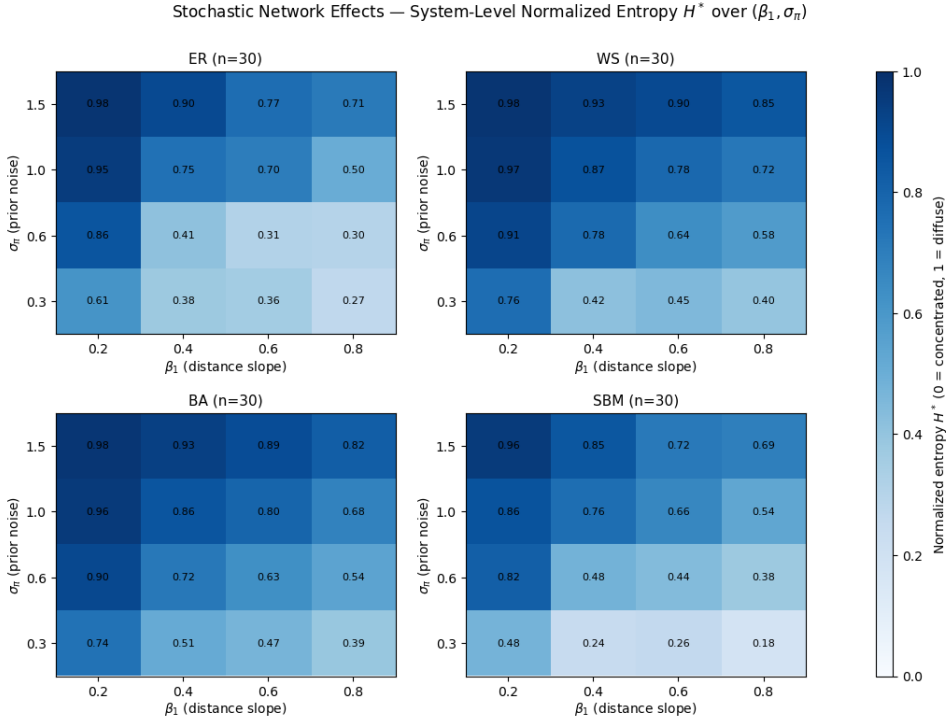


Figure 6: System-level normalized entropy H^* of the scapegoat selection distribution across the parameter space (β_1, σ_π) under stochastic network effects. Each cell reports the mean H^* across ten random graphs ($n = 30$) per topology with 1000 Monte Carlo runs per parameter combination ($\beta_0 = -2.0$; $\beta_1 \in \{0.2, 0.4, 0.6, 0.8\}$; $\sigma_\pi \in \{0.3, 0.6, 1.0, 1.5\}$; $\mu_p = 0$; $\sigma_p = 1.5$; $\lambda = 15$). Darker shades indicate higher entropy (more diffuse scapegoating) whereas lighter shades indicate lower entropy (more concentrated scapegoating).

4.2 Identity-Structure Interaction

After establishing the structural framework of scapegoating, we next examine how it interacts with social identity bias, a central mechanism emphasized in traditional scapegoating theory.

To introduce identity in the simplest form, all agents are randomly assigned to two groups: an in-group and an out-group, denoted by a group label $g_i \in \{\text{in}, \text{out}\}$. We will show that even when identity and network structure are treated as independent dimensions, their joint effects are already substantial, and would only intensify under endogenous identity–structure interactions.

To incorporate intergroup perception, agents’ priors are distorted by identity-based bias: in-group members are inherently more suspicious of out-group ones, while other relations

remain unaffected.

Assumption 8 (Identity Bias). *A prior belief $\tilde{\pi}_i^{(k)}$ exhibits social identity bias if*

$$\text{logit}(\tilde{\pi}_i^{(k)}) = \text{logit}(\pi_i^{(k)}) + \gamma \mathbf{1}\{g_i = \text{in}, g_k = \text{out}\}, \quad (22)$$

or equivalently,

$$\tilde{\pi}_i^{(k)} = \begin{cases} \frac{e^{\gamma \pi_i^{(k)}}}{(1 - \pi_i^{(k)}) + e^{\gamma \pi_i^{(k)}}}, & \text{if } g_i = \text{in}, g_k = \text{out}, \\ \pi_i^{(k)}, & \text{otherwise,} \end{cases} \quad (23)$$

where $\text{logit}(p) = \log \frac{p}{1-p}$; $\tilde{\pi}_i^{(k)}$ is the biased prior belief; and $\gamma > 0$ measures the strength of the identity distortion.

This specification integrates social identity bias into the structural framework, allowing network position and group membership to jointly shape vulnerability to scapegoating.

We next conduct simulations to test this joint effect. Figure 7 plots the scapegoat selection probability against the percentile ranks of centrality measures, comparing in-group and out-group members under the same stochastic and heterogeneous specifications as in Section 4.1. Overall, the core structural pattern persists across all network topologies: scapegoating probability decreases monotonically with network centrality.

However, the presence of identity bias introduces a systematic asymmetry: out-group members are more sensitive to structural disadvantage, exhibiting substantially higher scapegoating probabilities in peripheral positions. In contrast, in-group members are comparatively insulated from structural vulnerability, benefiting from an additional layer of identity-based protection. The same network position can thus imply different levels of vulnerability depending on group membership, meaning that identity bias effectively amplifies the inequality embedded in the network.

These results suggest that group identity and network position jointly shape vulnerability: structural isolation and identity bias reinforce one another to produce persistent asymmetries in collective attribution. When additional endogenous identity-structure interactions are introduced, such as when out-group membership itself correlates with structural disadvantage, this asymmetry is expected to further expand and deepen social hierarchy.

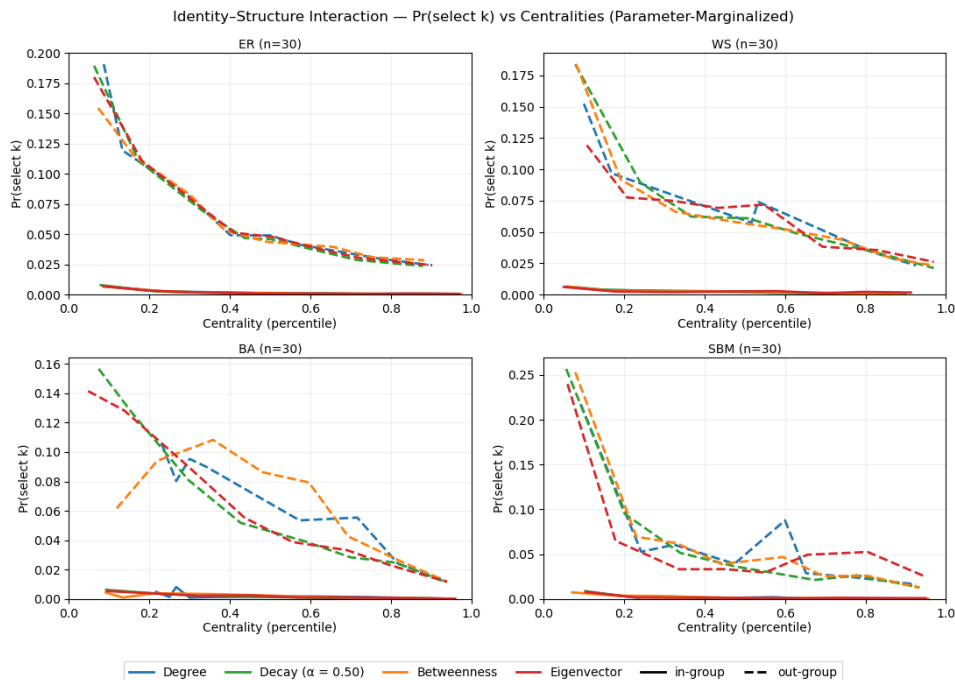


Figure 7: Relationship between scapegoat selection probability and centrality percentile under identity bias, comparing in-group (solid) and out-group (dashed) members. Each panel averages results across five random graphs ($n = 30$) per topology with 500 Monte Carlo runs per parameter combination ($\beta_0 = -2.0$; $\beta_1 \in \{0.2, 0.3, 0.6\}$; $\sigma_\pi \in \{0.5, 1.0, 1.5\}$; $\mu_p = 0$; $\sigma_p = 1.5$; $\lambda = 15$).

5 Behavioral Experiment

To provide empirical evidence for the theoretical framework, a preregistered 2×2 between-subjects experiment is planned and will be implemented.

As shown in Figure 8, the target’s network position (central vs. peripheral) serves as the primary independent variable, and the identity group (in-group vs. out-group) acts as a moderator. Participants will play the role of team leaders evaluating responsibility after a collective failure in an organizational setting. They will read a short vignette describing the situation and the team’s communication network, then rate (1-7 scale) how likely they would be to assign blame to the designated target, as well as the anticipated reputational cost of doing so. The conceptual mediation pathway is also summarized in Figure 8.

The experiment tests three main predictions consistent with the simulation results. First, leaders will be more likely to scapegoat peripheral than central members. Second, they will be more likely to scapegoat out-group than in-group targets. Third, the two effects will interact: the positional gap in scapegoating will be larger for out-group targets because identity

bias amplifies sensitivity to structural disadvantage. The expected results are illustrated in Figure 9.

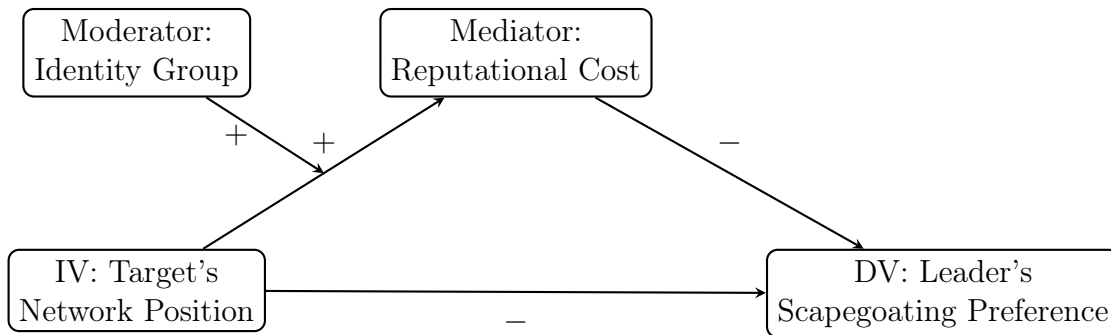


Figure 8: Conceptual framework of identity-structure interaction

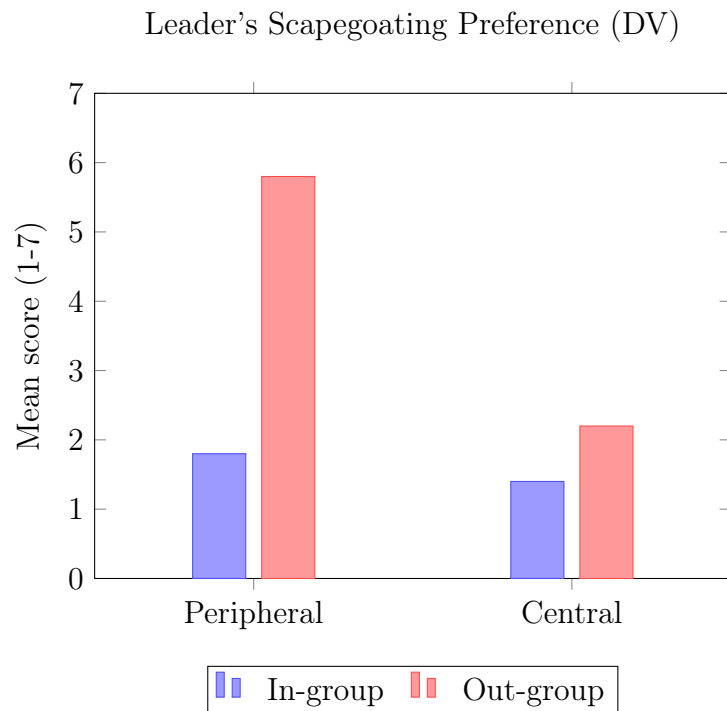


Figure 9: Predicted scapegoating pattern across network position and identity conditions

6 Discussion and Conclusion

This study develops a structural theory of scapegoating within a formal and computational framework, bridging insights from political science, social psychology, organizational behavior, network economics, and complexity science. By modeling blame allocation in social

networks, it shows that scapegoating arises not merely from individual prejudice or identity bias but from structural vulnerability embedded in the network itself. Individuals who are locally or globally peripheral, having fewer or more distant connections, bear a disproportionate risk of being blamed. This pattern persists under stochastic uncertainty and heterogeneity, revealing structural vulnerability as a robust, system-level property. When social identity bias is introduced, the asymmetry deepens: out-group members become more sensitive to structural disadvantage, while in-group members gain implicit protection, together reproducing persistent inequality in collective blame.

The significance of this study lies in distinguishing two intertwined dimensions of inequality: structural and categorical, and showing how their interaction shapes blame attribution. Structural inequality, rooted in network position, can operate independently of but also reinforce identity-based divisions. This distinction offers both theoretical and practical implications. Beyond normative prescriptions, informational interventions (e.g., enhancing transparency and communication credibility) may reduce leaders' incentives to redirect blame, while structural interventions (e.g., strengthening cross-group ties and reducing isolation) can effectively alleviate structural vulnerability and, in turn, limit its amplification through identity bias.

Future research can extend this framework in several directions. First, introducing endogenous identity-structure interactions would capture feedback loops through which social identity and structural position coevolve, potentially entrenching hierarchy. Second, incorporating dynamic leadership adaptation would allow leaders and followers to learn from past blame episodes, generating evolving equilibria over time. More broadly, linking behavioral experiments with large-scale network and field data would deepen understanding of how identity and structural factors jointly shape blame dynamics in real-world social systems.

Glossary

Term	Definition
social network	A set of individuals connected by social ties, represented as a graph.
undirected graph	A graph in which all edges have no direction and connections are mutual.
adjacency matrix	A matrix A whose entry A_{ij} records the presence or strength of the tie between nodes i and j .
neighbor set	The set $N(i)$ of nodes directly connected to node i .
degree	The number of neighbors of a node.
geodesic distance	The length of the shortest path between two nodes in the network.
diagonal degree matrix	The diagonal matrix D whose i th diagonal entry equals the degree of node i .
laplacian matrix	The matrix $L = D - A$ capturing the network's connectivity structure.
algebraic connectivity	The second-smallest eigenvalue of the Laplacian, measuring how well connected the network is.
degree centrality	A local centrality measure based on a node's degree.
decay centrality	A distance-weighted centrality $C_k(\alpha)$ that discounts more distant nodes.
betweenness centrality	A centrality measure based on how often a node lies on shortest paths between others.
eigenvector centrality	A centrality measure where a node is influential if it is connected to other influential nodes.
Erdős–Rényi (ER) network	A random graph model where each pair of nodes connects independently with a fixed probability.
Watts–Strogatz (WS) network	A small-world network model combining high clustering with short average path length.
Barabási–Albert (BA) model	A scale-free network model generated via preferential attachment.
Stochastic Block Model (SBM)	A random network model with block structure, where link probabilities depend on community membership.
opinion dynamics	The process through which individuals update opinions via social interaction.
structural vulnerability	Susceptibility to being scapegoated arising from an agent's network position.
local structural vulnerability	Vulnerability driven by local connectivity, such as low degree.
global structural vulnerability	Vulnerability driven by global remoteness in the network.
discrete network effects	A belief specification where trust changes discretely with network distance.

Term	Definition
decay network effects	A belief specification where trust decays smoothly with geodesic distance.
stochastic network effects	A belief specification where distance-based beliefs are perturbed by random noise.
in-group & out-group	A binary identity distinction dividing agents into different group labels.
normalized entropy	A rescaled entropy measure of the scapegoat selection distribution indicating concentration versus dispersion.

Notation

Symbol	Meaning
N	Set of agents in the community, with $ N = n$.
n	Number of agents, $n = N $.
$G = (V, E)$	Undirected social network (graph) with node set V and edge set E .
V	Set of nodes (agents) in the network.
E	Set of undirected edges representing social connections.
$v \in V$	Generic node (agent) in the network.
$e \in E$	Generic edge (social tie) in the network.
A	Adjacency matrix of G .
A_{ij}	Entry of A capturing the presence or strength of the tie between i and j .
θ_i	True state of agent i , $\theta_i \in \{0, 1\}$ (0 = innocent, 1 = guilty).
s_L	Leader's strategy $s_L : G \rightarrow V \cup \{\emptyset\}$.
\emptyset	Self-blame action (no agent is scapegoated).
k	Index of the agent scapegoated by the leader.
$v_L(s_L)$	Leader's utility under strategy s_L .
C	Fixed cost of self-blame borne by the leader.
$R^{(k)}$	Leader's reputational cost from accusing agent k .
$\pi_i^{(k)}$	Prior belief of agent i about agent k 's guilt.
$\pi_k^{(k)}$	Prior belief of agent k about own guilt (assumed to be 0).
p	Probability that the leader accuses an innocent agent; inverse public trust.
$b_i^{(k)}$	Posterior belief of agent i about k after the accusation.
$x_i^{(k)}$	Public opinion of agent i about k 's guilt.
$x_{-i}^{(k)}$	Profile of the opinions of all agents other than i .
$u_i(x_i^{(k)}, x_{-i}^{(k)})$	Utility of agent i in the opinion-formation game.
$x_j^{(k)}$	Public opinion of neighbor j about k 's guilt.
$x_i^{*(k)}$	Equilibrium opinion of agent i about k 's guilt.
$x^{(k)}$	Opinion vector $(x_1^{(k)}, \dots, x_n^{(k)})^T$.
$b^{(k)}$	Posterior belief vector $(b_1^{(k)}, \dots, b_n^{(k)})^T$.

Symbol	Meaning
D	Diagonal degree matrix of G .
L	Graph Laplacian $L = D - A$.
I	Identity matrix.
$s_L^*(G)$	Equilibrium scapegoating strategy of the leader on network G .
l	Number of scapegoats when extended to an l -scapegoat setting ($l \geq 2$).
$\mathbf{1}$	All-ones vector $(1, \dots, 1)^T \in \mathbb{R}^n$.
$N(i)$	Neighbor set of agent i .
l_{ik}	Geodesic (shortest-path) distance between i and k in G .
∞	Value assigned to l_{ik} when k is isolated and $i \neq k$.
$\deg(k)$	Degree (number of neighbors) of node k .
D_k	Aggregate distrust in the leader's accusation of k , $D_k = \sum_{i \in V} (1 - b_i^{(k)})$.
$C_k(\alpha)$	Decay centrality of k , $C_k(\alpha) = \sum_{i \in V} \alpha^{l_{ik}}$.
α	Distance-decay parameter in decay centrality.
$\rho(C_k, D_k)$	Spearman correlation between $C_k(\alpha)$ and D_k across nodes.
$\sigma(z)$	Logistic function $\sigma(z) = \frac{1}{1 + e^{-z}}$.
β_0	Baseline skepticism parameter in stochastic network effects.
β_1	Slope of distance-based skepticism in stochastic network effects.
ε_i	Idiosyncratic prior shock of agent i , $\varepsilon_i \sim \mathcal{N}(0, \sigma_\pi^2)$.
σ_π	Standard deviation of the prior-noise term in Assumption 4.
p_i	Individual trust parameter of agent i .
μ_p	Mean level of public trust in Assumption 5.
η_i	Idiosyncratic trust shock, $\eta_i \sim \mathcal{N}(0, \sigma_p^2)$.
σ_p	Standard deviation of trust heterogeneity.
ϕ_i	Conformity (or stubbornness) parameter of agent i in $[0, 1]$.
ϕ	Vector $(\phi_1, \dots, \phi_n)^T$ of conformity parameters.
$U(0, 1)$	Uniform distribution on the interval $[0, 1]$.
$\text{diag}(\cdot)$	Diagonal matrix formed from the entries of a vector.
\odot	Elementwise (Hadamard) product of vectors.
λ	Decision-temperature parameter in the leader's stochastic rule.
$P(k)$	Soft-min probability weight assigned to choosing scapegoat k .
$\min_j R^{(j)}$	Minimum reputational cost across all agents $j \in V$.
H^*	Normalized entropy of the empirical scapegoat selection distribution.
P_k	Empirical probability that agent k is selected as scapegoat, $P_k = \Pr(\text{select } k)$.
g_i	Group label of agent i (in-group or out-group).
$\tilde{\pi}_i^{(k)}$	Identity-biased prior belief of agent i about k 's guilt.
$\text{logit}(p)$	Logit transform $\text{logit}(p) = \log \frac{p}{1-p}$.
γ	Strength of identity-based distortion in prior beliefs.

References

- Allport, G. W. (1954). The nature of prejudice. *Reading/Addison-Wesley*.
- Arendt, H. and May, N. (1958). The origins of totalitarianism.
- Arsal, S. Y. and Yavuz, S. (2014). The portrait of the 'scapegoat woman' as witch. *Synergies Turquie*, (7).
- Bandura, A. (2024). Social learning analysis of aggression. In *Analysis of delinquency and aggression*, pages 203–232. Routledge.
- Bauer, M., Cahlíková, J., Chytilová, J., Roland, G., and Želinský, T. (2023). Shifting punishment onto minorities: Experimental evidence of scapegoating. *The Economic Journal*, 133(652):1626–1640.
- Bettelheim, B. and Janowitz, M. (1950). Dynamics of prejudice.
- Bindel, D., Kleinberg, J., and Oren, S. (2015). How bad is forming your own opinion? *Games and Economic Behavior*, 92:248–265.
- Bloch, F., Demange, G., and Kranton, R. (2018). Rumors and social networks. *International Economic Review*, 59(2):421–448.
- Bramoullé, Y. and Morault, P. (2021). Violence against rich ethnic minorities: A theory of instrumental scapegoating. *Economica*, 88(351):724–754.
- Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological bulletin*, 86(2):307.
- Burkert, W. (1983). *Homo necans: The anthropology of ancient Greek sacrificial ritual and myth*. Univ of California Press.
- Bursztyń, L., Egorov, G., Haaland, I., Rao, A., and Roth, C. (2022). Scapegoating during crises. In *AEA Papers and Proceedings*, volume 112, pages 151–155. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.
- Campbell, C. (2012). *Scapegoat: A history of blaming other people*. Abrams.
- Cavanaugh, W. T. (2011). The myth of religious violence. *The Blackwell companion to religion and violence*, pages 23–33.
- Coleman, J. S. (1988). Social capital in the creation of human capital. *American journal of sociology*, 94:S95–S120.
- Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451.
- Crossman, A. (2019). Definition of scapegoat, scapegoating, and scapegoat theory. ThoughtCo.

- Dezso, C. L. (2009). Scapegoating and firm reputation. *Robert H. Smith School Research Paper No. RHS*, pages 06–105.
- Dollard, J., Miller, N. E., Doob, L. W., Mowrer, O. H., Sears, R. R., Ford, C. S., Hovland, C. I., and Sollenberger, R. T. (2013). *Frustration and aggression*. Routledge.
- Doreian, P. and Mrvar, A. (1996). A partitioning approach to structural balance. *Social networks*, 18(2):149–168.
- Faris, R. and Felmlee, D. (2011). Status struggles: Network centrality and gender segregation in same-and cross-gender aggression. *American Sociological Review*, 76(1):48–73.
- Frazer, J. G. (1900). The golden bough: A study in magic and religion (vol. 10 of 12). *Revue des Traditions Populaires*, 15:471.
- Garfinkel, H. (2023). Studies in ethnomethodology. In *Social Theory Re-Wired*, pages 58–66. Routledge.
- Gent, S. E. (2009). Scapegoating strategically: Reselection, strategic interaction, and the diversionary theory of war. *International Interactions*, 35(1):1–29.
- Ghaderi, J. and Srikant, R. (2014). Opinion dynamics in social networks with stubborn agents: Equilibrium and convergence rate. *Automatica*, 50(12):3209–3215.
- Gibson, J. L. and Howard, M. M. (2007). Russian anti-semitism and the scapegoating of jews. *British Journal of Political Science*, 37(2):193–223.
- Girard, R. (1977). *Violence and the Sacred*. Johns Hopkins University Press, Baltimore.
- Girard, R. (1989). *The scapegoat*. JHU Press.
- Glick, P. (2005). Choice of scapegoats. *On the nature of prejudice: Fifty years after Allport*, pages 244–261.
- Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.
- Heider, F. (1946). Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107–112.
- Herman, E. S. and Chomsky, N. (2021). Manufacturing consent. In *Power and inequality*, pages 198–206. Routledge.
- Huitsing, G., Veenstra, R., Sainio, M., and Salmivalli, C. (2012). “it must be me” or “it could be them?": The impact of the social network position of bullies and victims on victims’ adjustment. *Social Networks*, 34(4):379–386.
- Labianca, G. and Brass, D. J. (2006). Exploring the social ledger: negative relationships and negative asymmetry in social networks in organizations. *Academy of Management Review*, 31(3):596–614.

- Levack, B. P. (2013). *The witch-hunt in early modern Europe*. Routledge.
- Luca, M., Pronkina, E., and Rossi, M. (2022). Scapegoating and discrimination in times of crisis: Evidence from airbnb. Technical report, National Bureau of Economic Research.
- Miguel, E. (2005). Poverty and witch killing. *The Review of Economic Studies*, 72(4):1153–1172.
- Newman, L. S. and Caldwell, T. L. (2005). Allport’s “living inkblots”: The role of defensive projection in stereotyping and prejudice. *On the nature of prejudice: Fifty years after Allport*, pages 377–392.
- Omanson, R. L. (1991). The new revised standard version with apocrypha.
- Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., and Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, 22(1):1–15.
- Shirado, H., Fu, F., Fowler, J. H., and Christakis, N. A. (2013). Quality versus quantity of social ties in experimental cooperative networks. *Nature communications*, 4(1):2814.
- Tajfel, H. (1979). An integrative theory of intergroup conflict. *The social psychology of intergroup relations/Brooks/Cole*.
- Takács, K., Janky, B., and Flache, A. (2008). Collective action and network change. *Social Networks*, 30(3):177–189.
- Winter, E. (2001). Scapegoats and optimal allocation of responsibility. *American Economic Review*, forthcoming.
- Zimbardo, P. G. (2007). *The Lucifer Effect: Understanding How Good People Turn Evil*. Random House, New York.
- Zussman, A. (2021). Scapegoating in evaluation decisions. *Journal of Economic Behavior & Organization*, 186:152–163.