Misinformation Spread on Social Media: A Game-Theoretic Approach

Felipe Manzi Cherryl Chico Olsa Berani

November 2025

Abstract

The spread of misinformation on social networks threatens societal stability by shaping public opinion and collective action. This paper develops a game-theoretic model in which agents, motivated by reputational rewards, interact with information and update their actions over time. These actions are applied to a structured social network where diffusion is governed by agents' influence and reach. By capturing how individual decisions aggregate across diverse topologies, the model quantifies contamination rates within neighborhoods after repeated updates, highlighting how utility-driven behaviors shape the velocity and extent of misinformation spread.

1 Introduction

The spread of misinformation on social media poses substantial challenges to public discourse and collective decision-making. While exposure and diffusion dynamics have been extensively studied—particularly on platforms such as Facebook, where recent evidence suggests a decline in the circulation of false content (Allcott et al., 2019), much less is known about alternative communication environments like Telegram. Unlike Facebook, Twitter/X, and other algorithm-driven platforms, Telegram does not rely on a personalized recommendation system. Information spreads primarily through channels, groups, and direct forwarding, rather than through algorithmic curation. This absence of recommendation algorithms provides a cleaner environment to observe how misinformation propagates, as diffusion is less confounded by opaque platform-level ranking mechanisms.

In this study, we aim to investigate how this distinct institutional setting shapes the propagation of misinformation. A central question guides our analysis: what motivates individuals to share false information, and how do their social relationships and reputational concerns influence this behavior?

Our work builds upon the economic foundations of information behavior. First, reputational concerns have long been recognized as fundamental drivers of human behavior: individuals care about how their actions affect how they are perceived by others (Bénabou and Tirole, 2006, 2011). This logic directly parallels models of media behavior, where information senders strategically choose what to transmit because audiences update beliefs about the sender's credibility. In particular, the reputation-driven framework of Gentzkow and Shapiro (2006)—in which media outlets condition reporting on how it affects perceived accuracy—serves as a microfoundational motivation for how we model reputation among individuals in messaging platforms.

Second, empirical research finds that news sharing is shaped not only by beliefs but also by social approval, status motives, and identity-driven incentives (Talwar et al., 2020; Wu, 2025; Vellani et al., 2024). Evidence from WhatsApp further shows that reputational mechanisms matter: forwarded-message labels influence perceptions of credibility and willingness to share (Tandoc Jr. and Foo, 2022), and exposure to fact-checking reduces subsequent forwarding (Reis et al., 2020). Work on Telegram additionally documents high-concentration misinformation ecosystems in channels shaped by tight-knit community norms (Herasimenka et al., 2022).

Finally, our approach is closely related to recent advances in misinformation modeling. Acemoglu et al. (2024) develop an equilibrium framework in which agents weigh social utility against the risk of reputational loss ("being called out"), generating endogenous filter bubbles. Dynamic network approaches such as Yilmaz and Ulusoy (2022) highlight how strategic interactions and influence-maximizing dynamics shape diffusion on evolving topologies.

While these studies provide deep insight into misinformation and strategic behavior, most focus on equilibrium characterizations, platform-level incentives, or adversarial spread. Our research departs from this by explicitly modeling multi-round diffusion where an agent's reputation and subjective beliefs evolve endogenously and shape future propagation. We focus on three primary questions:

1. How does an agent's subjective belief regarding the veracity of a message, coupled with the value they place on their personal reputation, affect the initial propensity and subsequent widespread dissemination of misinformation?
2. In a dynamic setting, how is misinformation contained within local neighborhoods, or conversely, how is it allowed to spread more widely across the network upon multiple rounds of agents updating their information and actions?
3. How does the velocity and final extent of misinformation spread vary across different, structured topologies of social networks, such as scale-free, random, or small-world networks?

To answer these research questions, our strategy employs two components: (i) a formalization of the agents' forwarding and updating decisions based on game theory, and (ii) large-scale agent-based simulations designed to trace message diffusion across social networks. Specifically, we focus on platforms such as WhatsApp and Telegram, where forwarding behavior plays a central role in information propagation. We will conduct these simulations across a spectrum of network topologies and agent types to produce quantitative results on the rate of spread and provide a comprehensive description of how information contamination evolves in various structural and behavioral scenarios.

2 Model and Environment

We study information diffusion in a social network environment similar to WhatsApp and Telegram, referred to here as WT platforms. These platforms provide a natural setting for our analysis, as they are structured around group interactions where users exchange, evaluate, and forward messages. Within this environment, the core components of our model involve: (i) the messages circulating through groups, (ii) the agents (users) who receive and act on these messages, and (iii) the relationships among agents that define group membership and interaction patterns. We formally define each of these components below.

2.1 Messages

In WT platforms, a single message m is the basic unit of information that users receive, evaluate, and decide whether to forward. Each message carries an ideological slant, reflecting the political or social bias embedded in its content, and a truth value, indicating whether the information is factually correct. Formally, each message is represented by its ideological slant and truthfulness:

$$m \sim (\beta, \theta) \tag{1}$$

$$\beta = 2x - 1, \quad x \sim \text{Beta}(\alpha_\beta, \beta_\beta) \tag{2}$$

$$\theta \sim \text{Bernoulli}(q) \tag{3}$$

We simplify ideology into two opposing poles. The parameter β captures the strength of the message's leaning, with values closer to -1 or 1 indicating stronger alignment with one of the extremes. A value of $\beta_\beta > \alpha_\beta$ skews the distribution towards 1, while $\beta_\beta < \alpha_\beta$ skews it towards -1. The truthfulness of a message is denoted by θ, where $\theta = 1$ indicates the message is true and $\theta = 0$ indicates it is false. Note that the distribution for θ is binomial distributed given the followinf dsitrbution:

$$\Pr(\theta = 1) = q, \quad \Pr(\theta = 0) = 1 - q.$$

This structure reflects how users in WT platforms often encounter messages that blend ideological bias with uncertain factual accuracy.

2.2 Agents

The set of agents N represents users in the WT network. Each agent $i \in N$ is a strategic participant in group interactions, characterized by an ideological bias $b \in [-1, 1]$ representing the agent's own slant,

and a reputation value $R \in \mathbb{R}$, reflecting their credibility based on past forwarding behavior. Thus,

$$i \sim (b, R), \qquad b_i \in [-1, 1], \quad R \in \mathbb{R}.$$

Within WT groups, agents evaluate the truthfulness of incoming messages and select an action according to expected utility. Depending on the actual truthfulness of the message and the agent's choice, their reputation updates over time. We describe each of these components below.

2.2.1 Belief Formation

In WT platforms, the truthfulness of a message is often not directly observable. Agents therefore form beliefs using two cues: the ideological slant of the message relative to their own bias, and the sender's reputation. This structure aligns with evidence that trust and credibility in group messages depend on ideological proximity and the perceived reliability of the source (Talwar et al., 2020; Vellani et al., 2024). The belief of agent i about the truthfulness of message m sent by agent j is:

$$\pi_{ij}^t = \begin{cases} \theta_t, & \text{if truth is observable} \\ \phi_{ij}(b_i, \beta_m) \cdot \sigma(R_j^t), & \text{if truth is not observable} \end{cases} \tag{1}$$

$$\phi_{ij}(b_i, \beta_m) = 1 - |b_i - \beta_m| \tag{2}$$

$$\sigma(R_j^t) = \frac{1}{1 + e^{-R_j^t}}. \tag{3}$$

Here, $\phi_{ij}(b_i, \beta_m)$ captures the *ideological proximity* between agent i and message m, while $\sigma(R_j^t)$ maps the sender's reputation into perceived credibility. Intuitively, in WT groups, agents are more likely to believe messages when they originate from ideologically aligned senders with high reputations.

2.2.2 Action Choice

After forming the belief π_{ij}^t, agent i chooses whether to forward the message to their WT contacts. The action set is

$$a \in \{0, 1\},$$

where $a = 1$ denotes forwarding and $a = 0$ denotes ignoring.

Forwarding decisions in WT platforms reflect three incentives: expressive satisfaction from sharing ideologically aligned content, reputational gains if the message is true, and reputational losses if it is false. Forwarding also carries a fixed cost, such as the effort or risk of spreading misinformation. Ignoring yields no utility. Thus, the expected utility is:

$$\mathbb{E}[u_{ij}^t] = \begin{cases} \alpha(1 - |b_i - \beta_m|) + \gamma\pi_{ij}^t - \delta(1 - \pi_{ij}^t) - k, & \text{if } a_i^t = 1 \\ 0, & \text{if } a_i^t = 0. \end{cases} \tag{4}$$

Here, α represents the benefit from forwarding ideologically aligned messages, γ the reputational gain from forwarding a true message, δ the reputational penalty from forwarding a false one, and k the fixed cost of forwarding. Because agents in WT groups often receive the same message from multiple senders, the overall utility is the average across all sources:

$$\mathbb{E}[u_i^t] = \frac{1}{|\S_i^t|} \sum_{j \in S_i^t} \mathbb{E}[u_{ij}^t],$$

where $\S_i^t$ is the set of senders from whom agent i received the message at time t. The agent forwards if $\mathbb{E}[u_i^t] \geq 0$.

2.2.3 Reputation Update

We assume reputation is the guiding principle for agents in the WT network. Empirical studies of WhatsApp and Telegram show that users rely on sender credibility, and reputational cues strongly shape forwarding behavior (Pasquetto et al., 2022; Herasimenka et al., 2022). Accordingly, our model treats

reputation as the central mechanism of credibility: it rises when agents forward true messages and falls when they forward false ones. Formally,

$$R_i^{t+1} = R_i^t + \eta \cdot \mathbf{1}\{a_i^t = 1 \wedge \theta_t = 1\} - \lambda \cdot \mathbf{1}\{a_i^t = 1 \wedge \theta_t = 0\} \tag{5}$$

where $\eta, \lambda > 0$ govern the strength of reputational gains and losses respectively. Reputation evolves only through the correctness of forwarding actions. Ignoring a message leaves reputation unchanged.

2.3 Social Network

Social networks on WT platforms are organized around groups of users who interact repeatedly. In our representation, each agent corresponds to a node, and communication ties (e.g., channel membership) are modeled as edges. Note that the edge is directed and one way to model from one agent to another mimicking the action of forwarding a message. Empirical studies show that these platforms are characterized by dense clusters of interpersonal ties within groups, alongside weaker spillovers across groups (Herasimenka et al., 2022).

We model the social network as a directed graph $G = (N, E)$, where N is the set of agents (nodes) and E is the set of directed edges representing communication ties (e.g., channel memberships). The network is generated using a stochastic block model to capture the clustered group structure typical of WT platforms. Formally, we partition agents into K latent groups $\{G_1, G_2, \ldots, G_K\}$. The probability of a directed edge from agent i to agent j depends on their group memberships:

$$\Pr((i,j) \in G) = \begin{cases} p_{\text{in}}, & \text{if } i,j \text{ share at least one group,} \\ p_{\text{out}}, & \text{otherwise,} \end{cases} \quad \text{with } p_{\text{in}} > p_{\text{out}} > 0. \tag{6}$$

Here, p_{in} is the probability of a directed edge between agents within the same group, while p_{out} is the probability of an edge between agents in different groups. This structure captures the tendency for users on WT platforms to form tightly-knit groups with frequent interactions, while still allowing for cross-group connections that facilitate broader information diffusion.

3 Data Generating Process and Simulation

Here we define the data generating process for the construction of the social network, agents, and messages. We also describe the simulation procedure used to study misinformation diffusion within this environment.

3.1 Messages

Messages are generated according to their ideological slant and truthfulness. For each round t, a message m_t is drawn as:

$$m_t \sim (\beta_t, \theta_t), \qquad \beta_t \in [-1, 1], \quad \theta_t \in \{0, 1\}.$$

3.2 Rounds

Each round proceeds as follows:

1. A message (m_t, θ_t) appears.
2. Initial recipients observe the message and, if they choose to forward, neighbors receive it.
3. Upon receiving the message from sender j, each agent i forms belief

$$\pi_{it} = \kappa_{ij}^t.$$

4. Each agent decides whether to forward ($a_{it} = 1$) or ignore ($a_{it} = 0$).
5. The true state θ_t is revealed and agents' reputations are updated.

3.3 Network Construction

We model the social network as a directed graph $G = (N, E)$, where N is the set of agents (nodes) and E is the set of directed edges representing communication ties (e.g., message forwarding). The network is generated using a stochastic block model to capture the clustered group structure typical of WT platforms. Formally, we partition agents into K latent groups $\{G_1, G_2, \ldots, G_K\}$. The probability of a directed edge from agent i to agent j depends on their group memberships:

$$\Pr((i,j) \in G) = \begin{cases} p_{\text{in}}, & \text{if } i,j \text{ share at least one group,} \\ p_{\text{out}}, & \text{otherwise,} \end{cases} \quad \text{with } p_{\text{in}} > p_{\text{out}} > 0.$$

3.4 Agent Types and Reputational Heterogeneity

Agents are partitioned into types $\tau \in \{\text{influencer}, \text{regular}, \text{bot}\}$. Each type differs in both network position and behavioral parameters.

- **Influencers / political elites** have high expected public-layer degree and substantially larger reputational parameters (γ_i, δ_i). This assumption reflects evidence that public figures face stronger reputational incentives and backlash for spreading false content.
- **Regular users** participate in several private groups and have moderate (γ_i, δ_i) values, reflecting social-image motives documented in Talwar et al. (2020); Vellani et al. (2024).
- **Bots or automated accounts** may have low reputation sensitivity and strategically chosen degree depending on the simulation goal.

3.5 Rounds

Each round proceeds as follows:

1. A message (m_t, θ_t) appears.
2. Initial recipients observe the message and, if they choose to forward, neighbors receive it.
3. Upon receiving the message from sender j, each agent i forms belief

$$\pi_{it} = \kappa_{ij}^t.$$

4. Each agent decides whether to forward $(a_{it} = 1)$ or ignore $(a_{it} = 0)$.
5. The true state θ_t is revealed and agents' reputations are updated.

3.6 Ideological Bias Distribution

Each agent receives an ideological bias:

$$b_i \sim p_L \cdot \mathcal{N}(-\mu, \sigma^2) + (1 - p_L) \cdot \mathcal{N}(\mu, \sigma^2),$$

capturing polarization and bimodality consistent with empirical ideological distributions in online environments.

3.7 Data-Generating Process (DGP) Justification

Our DGP draws on several empirical and theoretical frameworks:

- **Influence heterogeneity and hub structure** follows Aral and Walker (2012), showing that a minority of high-degree nodes disproportionately drives information cascades.
- **Overlapping group-based diffusion** follows González-Bailón et al. (2011), which models protests and online coordination via overlapping communities.
- **Telegram's hub-based misinformation channels** are motivated by Herasimenka et al. (2022), documenting high-reach public channels connected to dense follower clusters.

- **Encrypted messaging network structure** follows Bright and Ganesh (2022), describing multi-group membership and low-visibility peer networks.
- **Reputational motives** draw from Bénabou and Tirole (2006), Gentzkow and Shapiro (2006), and empirical studies of misinformation sharing (Talwar et al., 2020).

This combined structure provides a realistic environment for studying misinformation propagation across layered, overlapping, and reputationally heterogeneous networks.

3.8 Model Limitations

Several simplifying assumptions constrain the realism of the simulation:

- **Ideological bias is one-dimensional**, $b_i \in [-1, 1]$, while real preferences may be multi-dimensional.
- **Reputation is scalar and fully observable**, whereas in real encrypted networks reputation is noisy, context-dependent, and often implicit.
- **Influencer and regular-user types are discrete**, though real online influence is continuous and endogenous.
- **The network is static during a simulation window**, while real WhatsApp/Telegram groups evolve (members join/leave; admins add channels).
- **Message truthfulness is exogenous**, although real misinformation environments include strategic manipulation and coordinated campaigns.
- **Simulated DGP**: we rely on stylized random graphs and not on microdata from actual encrypted apps.

These assumptions allow tractability and clarity but should be kept in mind when interpreting results and external validity.

4 Outcome Measurement

To evaluate how misinformation propagates across different network structures and agent types, we record a series of outcome metrics at each round t of the simulation. These measures are designed to capture both the *extent* and the *dynamics* of diffusion, including whether false messages remain contained within local communities or spread widely across the network.

4.1 Reach

We define the *reach* of message m_t as the fraction of agents who received the message through any chain of forwards:

$$R_t = \frac{|\{i : \text{agent } i \text{ received } m_t\}|}{N}.$$

This metric captures how far a message travels across the network.

4.2 Forwarding Rate

Among those who received the message, we compute the *forwarding rate*:

$$F_t = \frac{\sum_{i:i \text{ received } m_t} a_{it}}{|\{i : \text{received } m_t\}|},$$

where $a_{it} \in \{0, 1\}$ denotes agent i's forwarding decision. This measure reflects the behavioral response to the message in round t.

4.3 Misinformation Contamination

To track the spread of false messages, we compute the *false-forward rate* when $\theta_t = 0$:

$$FF_t = \frac{\sum_{i:i \text{ received } m_t} a_{it} \cdot \mathbf{1}\{\theta_t = 0\}}{|\{i : \text{received } m_t\}|}.$$

We also record cumulative contamination over time as the total number of agents ever exposed to false messages.

4.4 Velocity of Spread

The *velocity* of diffusion is defined as the time required for a message to reach a given share of the population. For a threshold $\tau \in (0, 1)$, define:

$$V_\tau(t) = \min\{s \geq t : R_s \geq \tau\}.$$

A lower V_τ indicates faster spread.

4.5 Community Containment versus Spillover

Each agent belongs to one or more communities in the overlapping group structure. For each receiving agent i, let $\mathcal{C}(i)$ denote the set of communities to which i belongs. We measure the degree of within-community concentration of diffusion through:

$$C_t = \frac{1}{|\{i : \text{received } m_t\}|} \sum_{i:\text{received } m_t} \mathbf{1}\{\exists\, j \text{ sender with } \mathcal{C}(i) \cap \mathcal{C}(j) \neq \emptyset\}.$$

A high C_t indicates that diffusion remains mostly inside local group "bubbles," whereas a low C_t indicates cross-community spillovers.

4.6 Reputation Dynamics

For each agent type τ, we track the mean reputation over time:

$$\bar{R}_\tau(t) = \frac{1}{|\{i : \tau(i) = \tau\}|} \sum_{i:\tau(i)=\tau} R_i^t.$$

This allows us to study how exposure to true and false messages shapes the reputational distribution of influencers, regular users, and automated accounts.

4.7 Aggregation Across Rounds and Network Draws

For each metric $\{R_t, F_t, FF_t, C_t\}$, we compute:

- **Round-level averages:** $\frac{1}{T}\sum_{t=1}^{T} X_t$.
- **Cumulative measures:** e.g., cumulative false exposure.
- **Comparative statics:** comparison across network structures (scale-free, small-world, overlapping SBM) and agent-type parameterizations.

These aggregated measures provide a comprehensive characterization of misinformation diffusion under different structural and behavioral assumptions.

References

Acemoglu, D., Ozdaglar, A., and Villalonga, B. (2024). A model of online misinformation. *MIT Economics Working Paper*.

Allcott, H., Gentzkow, M., and Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Political Communication*, 36(2):1–21.

Aral, S. and Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

Bright, J. and Ganesh, B. (2022). Encrypted messaging and the dynamics of networked diffusion. *Journal of Communication*, 72(6):778–803.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *Quarterly Journal of Economics*, 126(2):805–855.

Gentzkow, M. and Shapiro, J. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2):280–316.

González-Bailón, S., Borge-Holthoefer, J., and Moreno, Y. (2011). The dynamics of protest recruitment through an online network. *Scientific Reports*, 1(1):1–7.

Herasimenka, A., Bright, J., and Howard, P. (2022). Misinformation and professional news on largely unmoderated platforms: The case of telegram. *Oxford Internet Institute Working Paper*.

Pasquetto, I. V. et al. (2022). Social debunking of misinformation on whatsapp: The case for strong and in-group ties. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–35.

Reis, J. C. S., Benevenuto, F., and Melo, P. (2020). Can whatsapp benefit from fact-checking to reduce misinformation? evidence from brazil. *Harvard Kennedy School Misinformation Review*, 1(3):1–12.

Talwar, S., Dhir, A., Singh, G., Virk, G. S., and Salo, J. (2020). Fake news sharing on social media: underlying psychological factors and motivations. *Information Processing & Management*, 57(1):102–131.

Tandoc Jr., E. C. and Foo, C. Y. (2022). Moving forward against misinformation or stepping back? users' responses to whatsapp's forwarded tag. *Journalism & Mass Communication Quarterly*, 99(4):1059–1081.

Vellani, K., Glickman, M., and Sharot, T. (2024). Three diverse motives for information sharing. *Communications Psychology*, 1(1):1–13.

Wu, J. (2025). Why people share misinformation on social media? a psychological framework. *Humanities and Social Sciences Communications*, 12(1):1–13.

Yilmaz, Y. and Ulusoy, A. (2022). Deep reinforcement learning for misinformation spread dynamics in social networks. *IEEE Transactions on Computational Social Systems*, 9(5):1458–1470.