Simulating trustworthy content delivery on social media by manipulating algorithms and architecture

Mohammad Hammas Saeed[1], Manan Suri[2], David Broniatowski[1], Erica Gralla[1], Haneen Al-Rashid[1], Joseph Simons[3†], and Giovanni Luca Ciampaglia[2]

[1] George Washington University, Washington DC, USA
{ms190,broniatowski,egralla}@gwu.edu
[2] University of Maryland, College Park MD, USA {manans,gciampag}@umd.edu
[3] US Department of Health and Human Services
jsimons09@gmail.com
[†]The views expressed are those of the author and do not reflect the official position of the U.S. Department of Health and Human Services, or the United States

Abstract. This research investigates the trustworthiness of social media platforms in providing access to reliable information. With millions relying on social media for timely and accurate information, poor quality information significantly erodes trust. We examine the interplay between two key elements of social media platforms that are responsible for content delivery: the algorithms that dictate content visibility and the architectural frameworks that govern how platform components and users relate to each other, which in turn shapes how users interact and the content that they see from other users. We develop a simulation model that evaluates how different classes of social media platforms (as defined by their algorithms and architectures) handle the spread of information of varying levels of quality. We find that TikTok shows the fastest and widest information spread, driven by its highly connected architecture, while Twitter, Reddit, and Facebook diffuse more slowly. Moreover, algorithmic design further shapes outcomes, with engagement-driven ranking amplifying virality and producing long-tail popularity effects.

Keywords: architecture · algorithms · trustworthy content delivery.

1 Introduction

The rapid diffusion of information through social media platforms has reshaped how individuals engage with online content, such as news and public events. While this connectivity enables collective awareness at unprecedented scale, it also exposes critical vulnerabilities in the digital information ecosystem. The design of social media platforms has implications for which information gets amplified, potentially eroding trust in legitimate content. These phenomena are not merely consequences of user behavior, but emerge from the deep entanglement of two factors. The first factor is the design of the platform — specifically, its "architecture" — which governs relationships between the platform's components and

users' connections with each other. The architecture, in turn, shapes the affordances of the platforms themselves (e.g., broadcast vs. unicast channels, group and page structures). The second factor is set of the algorithmic mechanisms that govern what content is produced, recommended, and reshared.

Our work presents a simulation framework that models information diffusion as a dynamic, agent-based process shaped by both architectural and algorithmic constraints. Ultimately, this work seeks to provide a foundation for understanding trustworthy information delivery as a system-level property, one that cannot be fully understood without accounting for the co-evolution of users, content, algorithms, and the architectures of the infrastructures that connect them.

2 Related Works

The architecture of a system, meaning the abstract relationships among the system's constituent components, imposes both a structural and organizational logic [19, 18, 3]. Scholars have hypothesized the existence of tradeoffs between the flexibility of an architecture (how easily a system can accommodate changes to its configuration imposed by decision makers) and its controllability (how easily decision makers can impose their will on the system). This in turn shapes how information diffuses, how users are able to behave (i.e., with whom and what they can form linkages), and the degree to which interventions to control the system succeed or fail [7, 6]. Across social platforms, recent work reveals that both architectural constraints and algorithmic choices mediate not just what spreads, but how, to whom, and with what downstream effects. At a structural level, Wei et al. [28] draw on Moses's theory of generic architectures to quantify verticality and laterality in networks, offering a basis for comparing how different architectures govern flow and control. This lens complements Broniatowski and Moses [7], who introduce formal metrics (e.g., descriptive complexity) that allow us to evaluate trade-offs across system types (e.g., trees vs. teams) and later work by Broniatowski [4] shows that blending architectural principles like decomposition and abstraction yields resilient systems.

However, architectures are not static. Platforms layer algorithms on top of structural affordances, which reconfigure user exposure and behavior. Zhang et al. [30] formalize this interaction through the Form-From model, mapping content and source architectures (threaded vs. flat; networks vs. spaces) and Li et al. [16] complement this with the Affordances for Discursive Opportunities (ADO) framework, highlighting how design differences shape activism, what content is visible, legitimate, or resonant, across Twitter, Reddit, and Facebook. The work by Smaldino et al. [24] highlights the potential of information architectures (IAs) to reinforce biases or enable prosocial outcome. Similarly, algorithms are also central to shaping both exposure and experience. A growing body of work examines how recommender systems and ranking algorithms amplify or suppress content. Wang et al. [26] show that algorithmic timelines (vs. chronological ones) on Twitter/X surface more reliable yet less extreme content, albeit without shifting user perceptions. In contrast, Piccardi et al. [21] find that ex-

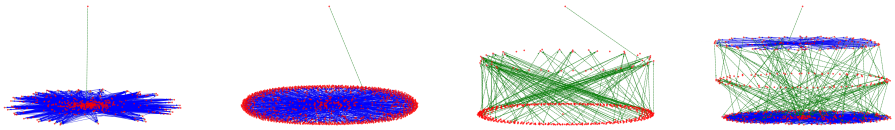

Fig. 1. Architecture for Twitter (left-most), TikTok, Reddit and Facebook (right-most)

posure to polarizing content increases affective hostility and Baumann et al. [2] identify a trade-off frontier between engagement and content diversity, suggesting algorithmic designs can optimize for both.

Simulation Models. Simulation models have been used by past works to deepen our understanding of algorithmic and structural interplay. For example, recent work [20, 12, 22] shows how personalization and link recommendation algorithms fuel echo chambers, reinforcing existing beliefs. Wang et al. [27] extends the simulation research by revealing how relevance-based link suggestions can unintentionally escalate social conflict. These findings parallel Tokita et al. [25], who show that information cascades in polarized architectures can harden ideological boundaries.

Real-world Interventions. Beyond abstract models, real-world interventions reveal how platform governance and moderation interface with architecture. For example, Cima et al. [11] evaluated Reddit's "Great Ban" and found that while most users reduced toxicity, some became more toxic. Jhaver et al. [15] propose a multi-level governance model, arguing that decentralization, mirroring polycentric governance approaches, may offer scalable moderation without collapsing into chaos. Platform architecture also mediates public accountability and social pressure. Forestal et al. [13] compares how online public shaming (OPS) plays out across Reddit, Twitter, and Wikipedia, finding that closed or nested architectures produce more deliberative outcomes, while open, flat networks breed spectacle. Recent work by Broniatowski et al. provides evidence that suggests Facebook's moderately flexible (and less controllable) layered structure, and legacy Twitter's highly flexible (and even less controllable) "network" structure, may have contributed to challenges with content moderation on each respective platform during the Coronavirus pandemic [8, 9]. Similarly, Van der Linden et al. [17] critiques the logic of engagement-driven design, showing how algorithmic amplification replaces epistemic value with virality, distorting civic and scientific discourse. Models of misinformation and countermeasures reflect this tension. Bak-Coleman et al. [1] demonstrate that viral falsehoods cannot be controlled by single-shot interventions. Ciampaglia et al. [10] and Sasahara et al. [23] show how seemingly benign algorithmic biases or user actions (like unfollowing) accelerate polarization and content degradation. Importantly, Zhou et al. [31] model real-world diffusion using power-law probabilities and heterogeneous delays, offering a more empirically grounded picture of cascade formation.

3 Methods

We construct a simulation framework that integrates multiple platform architectures, algorithms, and agent-level characteristics. Our choice of platform architectures deliberately captures the full range of the flexibility - controllability tradespace, while allowing us to study its interaction with different classes of algorithms in common use. This approach allows for controlled comparisons across scenarios that mirror real-world platforms such as Facebook, Reddit, Twitter, and TikTok.

Figure 1 highlights the structure of architectures we represent in our study.

Twitter (or X). As shown in Figure 1, X is a relatively flat social network architecture organized through a directed follower model. Users follow others to see their content in a chronologically or algorithmically ordered timeline. Unlike Reddit, there are no hard community boundaries, and unlike Facebook, there are fewer multi-entity types. All users exist in a shared space, and content can traverse widely via retweets. According to Moses' theory, X is expected to be flexible but not very controllable compared to most of the platforms in this study (with the notable exception of TikTok).

TikTok. TikTok operates as a fully-connected user "team" architecture, where the primary content discovery mechanism is the For You algorithmic feed. Rather than relying on explicit social ties, TikTok's architecture assumes latent connectivity, any content from any user can be shown to any other user. This means that content propagation is not limited by prior relationships but is determined dynamically by engagement patterns. According to Moses' theory, TikTok should be the most flexible and least controllable of the platforms included in this study.

Reddit. Reddit's architecture roughly resembles a tree-structure, which is organized into topically bounded communities called `subreddits`. Users participate across subreddits (violating the tree structure), but content is localized by design. According to Moses' theory, Reddit is likely to be the least flexible and most controllable of the platforms considered in this study.

Facebook. Facebook exhibits a multi-layered hierarchical architecture. The primary layers include, `users`, where individual accounts can produce, share, and engage with content. Next is `groups` which is an interest-based collectives where membership gates access and visibility. Lastly, there are `pages` (e.g., for brands or influencers) that broadcast to large audiences. The edges in this system are not uniform, there are friendships, group memberships, and page subscriptions each distinct relationship types. According to Moses' theory, Facebook should be moderately flexible and moderately controllable compared to the other platforms considered in this study.

Next, we model several algorithms in our simulations.

LIFO (Last-In, First-Out). The LIFO algorithm prioritizes the most recently received message. This mechanism emphasizes recency and is often used in real-time communication platforms (e.g., messaging apps, live updates).

FIFO (First-In, First-Out). FIFO implementation prioritizes older messages. While rarely used in mainstream platforms due to poor engagement, FIFO provides an additional mechanism for validating the simulations.

Random. A message is selected uniformly at random from the available set. This algorithm models environments with no prioritization or learning. While impractical in real-world interfaces, random selection is a control condition for our experiments to assess the influence of other algorithm designs.
Hot. The "hot" algorithm ranks and selects the message with the highest number of reshares, approximating a dynamic popularity score, making popular messages dominate the feed.
Like-Based. This algorithm selects the message with the most likes, reflecting the crowd's cumulative approval rather than engagement velocity.

3.1 Agent Specifications

Each agent is defined by the `agent_id`, a unique identifier assigned to the agent and the `network layer` which the agent belongs. Agents possess the ability to produce new messages, if they are a `producer` (e.g., a user can produce new messages but a subreddit cannot) and they can be a `resharer`, i.e., indicating whether the agent can reshare content received from others. Additionally, agents manage content flow through two main queues, an `input_queue` storing messages received from other agents and an `output_queue` containing messages the agent has prepared to distribute.
Users have a decision-making process that is governed by the following traits:

- `quality_preference` $\in [-1, 1]$: A scalar reflecting the agent's orientation toward factual content. This value is sampled from a specified random distribution.
- `reward_sensitivity` and `meaning_seeking`: These are independently drawn from triangular distributions with range $[0, 1]$, reflecting an agent's motivation to seek engagement (reward) and interpretive depth (meaning).
- `share_tendency`: Defined as the norm of the inverted motivational vector: $A = [1 - a_1,\ 1 - a_2]^\top$, where a_1 = `reward_sensitivity` and a_2 = `meaning_seeking`.
  It is calculated as `share_tendency` $= \|A\|_2 = \sqrt{(1 - a_1)^2 + (1 - a_2)^2}$. This norm captures the agent's overall inclination to share content, where lower motivational alignment (i.e., values of a_1 and a_2 closer to one) leads to a higher tendency to share.
- `liking_behavior`: In each cycle, agents like a message based on a `like_score` = (1 − `quality_preference`) × `clickbaitiness` + `quality_preference` × `informativeness`. The like score balances attention between clickbait and informative content. Agents with high fact checking (closer to 1) prioritize informativeness, while those with low fact checking (closer to 0) are more influenced by clickbait.

Message Representation. Each message is represented as a two-dimensional vector $\mathbf{Mk} = (\mathrm{mk}_1, \mathrm{mk}_2)$, where $\mathrm{mk}_1 \in [0, 1]$ denotes the clickbaitiness or motivational valence, indicating how attention-grabbing the message is, and $\mathrm{mk}_2 \in [0, 1]$ captures the illuminating quality, reflecting how meaningful or insightful the message is.

Additionally, each message is associated with a scalar misleading/informative score $\in \{-1, 0, 1\}$ where -1 is misleading, 0 is neutral and 1 informative (and factual).

The pair $(\mathrm{mk}_1, \mathrm{mk}_2)$ is sampled from a triangular distribution to model skewed preferences, such as most messages being moderately engaging but not highly insightful and is used to calculate a `message magnitude`, which is defined as $\|\mathbf{Mk}\| = \sqrt{\mathrm{mk}_1^2 + \mathrm{mk}_2^2}$, providing a combined measure of the message's overall strength.

Simulation. We implement a message-passing simulator over a graph G of agents (nodes) and connections (edges). At each discrete timestep, a random agent is selected to act. An agent may produce a new message with probability p_{post}, or attempt to reshare with probability $p_{reshare}$. For reshares, candidate messages are drawn from the agent's input queue and filtered by credibility (informativeness above the agent's quality preference) and share tendency (minimum message magnitude). If viable messages remain, one is chosen at random and forwarded. Message delivery is determined by edge type: broadcast edges deliver to all neighbors, while unicast edges deliver to a single randomly chosen neighbor. Additionally, global broadcasting can occur at fixed intervals, where all messages from designated broadcasters are pushed to their neighbors. The simulation records network-level outcomes such as message reach, popularity distributions, and persistence, enabling analysis of diffusion under varying algorithmic and architectural conditions.

4 Results

We selected key simulation parameters based on prior works to ensure realistic and representative dynamics. The *post probability* was set to 0.45, reflecting observed activity rates in Twitter diffusion networks [29]. The *reshare probability* was set to 0.25, consistent with reshare behaviors reported in [23]. The *queue size*, representing an agent's screen or attention span, was set to 10, aligning with the default screen size used in [23]. For network structures, we adopted the scale used to simulate Twitter user-user graphs, consisting of approximately 100,000 users and 3 million edges [29], and applied a similar structure to Facebook's user-user network. For group interactions, we followed the scale of Reddit's user-subreddit networks reported in [14], which include over 1.5 million users and 416 subreddits, and used comparable settings for Facebook's user-group network. Facebook's user-page edges were based on data from [5], involving approximately 1.4 million users and 303 pages, and this structure also informed the modeling of page-group connections. For TikTok, we modeled the network as a fully connected graph, where each user is connected to every other user, resulting in $(n \times (n-1))$ edges for a network of n users.

We ran our experiments with 10,000 users over 30,000 timesteps. As shown in Figure 2, for each combination of network architecture and algorithm, we plot the Complementary Cumulative Distribution Function (CCDF) of message

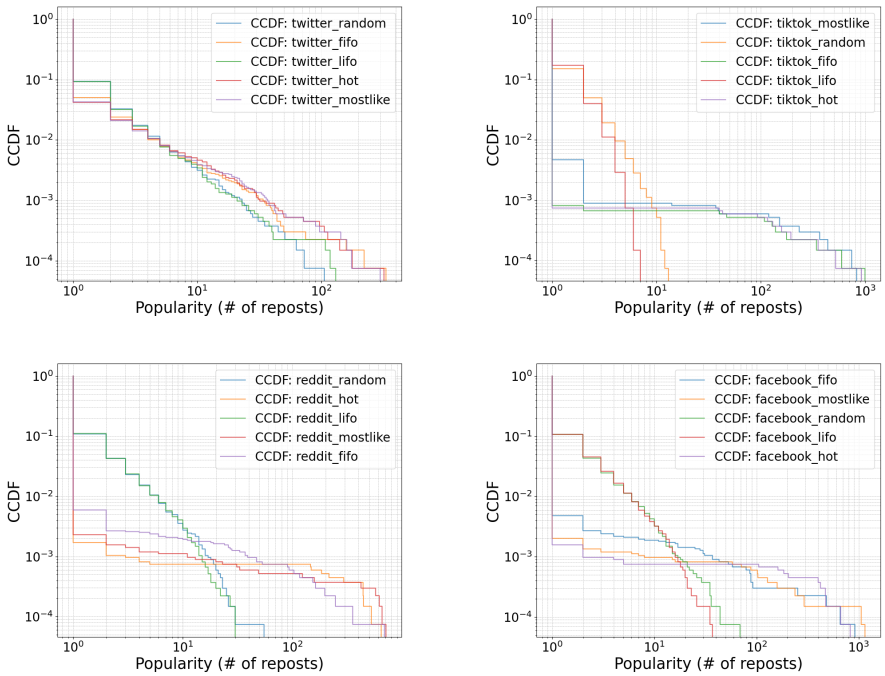

Fig. 2. CCDF vs Message Popularity for Architecture and Algorithms

popularity. The CCDF shows the probability that a message achieves a popularity greater than or equal to a given value, making it useful for analyzing the presence and extent of heavy-tailed distributions, where few messages go viral while most remain relatively unseen. Across platforms, we observe that Twitter was the least affected by algorithmic changes, while TikTok was the most sensitive, with algorithmic selection strongly dominating message spread. Reddit and Facebook show moderate sensitivity to algorithmic variation. The algorithmic conditions we tested include: *LIFO* (Last-In-First-Out) exhibited a relatively standard flow of information, while *Random* served as a control condition. *FIFO* (First-In-First-Out) created forced virality, where early messages continued to circulate through user queues, resulting in a pronounced long tail in the CCDF of message popularity. In contrast, *Hot* and *Most-Liked* algorithms, which are engagement-driven, tended to reinforce the spread of already popular messages, also producing longer tails as certain messages gained disproportionate reach and became viral.

Information Spread. Next, we examine the mean number of reshares over time to assess the dynamics of information spread across platforms. As shown in Figure 3, we find that Twitter, Reddit, and Facebook exhibit comparable levels of information diffusion, with similar average reshare rates throughout the sim-

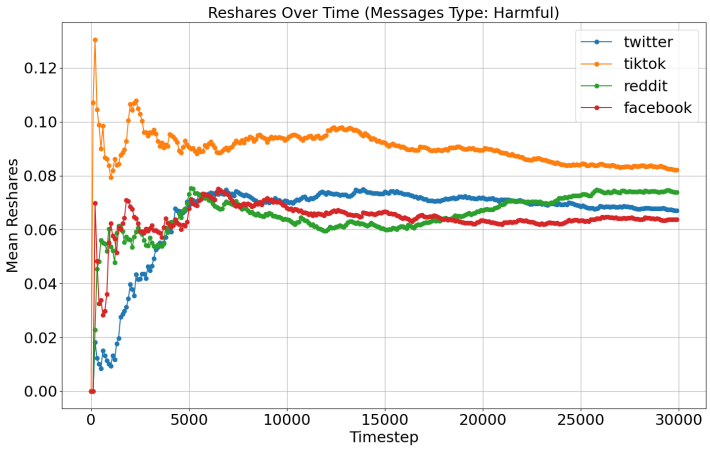

Fig. 3. Mean Reshares of Messages over Time

ulation. In contrast, TikTok consistently demonstrates the highest information spread, indicating a platform dynamic where algorithmic amplification and dense connectivity significantly accelerate the propagation of messages. This suggests that on TikTok, content can more easily achieve broad visibility, likely due to its fully connected structure and stronger algorithmic influence.

5 Discussion

Our findings collectively highlight that both the structural architecture of a platform and the algorithms it employs are critical in shaping how information flows, persists, and evolves. Engagement-optimized algorithms (e.g., "hot" ranking) inherently prioritize content that garners rapid information diffusion. This leads to the selective virality of messages that spreads disproportionately fast, irrespective of its factual grounding. Consequently, engagement algorithms function not merely as amplifiers of interest, but as filters that concentrate attention on a narrow subset of the available discourse, and could potentially reinforce confirmation biases or novelty-seeking behaviors. Platforms with minimal community-bound structure, such as TikTok, enable information to traverse rapidly. The lack of bounded communities contributes to ephemeral virality, where content goes viral quickly but can also disappear from relevance just as fast, reducing opportunities for deeper deliberation or sustained information verification.

6 Conclusion

Our work demonstrates that the flow of information in social media platforms is closely tied to the interplay between their architectural design and algorithmic

choices. We find that engagement-driven algorithms, coupled with loosely structured architectures can lead to rapid information diffusion often at the expense of reliability and critical evaluation. By contrast, platform designs that encourage more structured interactions may better support sustained deliberation and verification. These insights underscore the need for deliberate design interventions that balance engagement with reliability, ensuring that social media ecosystems promote trustworthy information exchange.

References

1. Bak-Coleman, J.B., Kennedy, I., Wack, M., Beers, A., Schafer, J.S., Spiro, E.S., Starbird, K., West, J.D.: Combining interventions to reduce the spread of viral misinformation. Nature Human Behaviour **6**(10), 1372–1380 (2022)
2. Baumann, F., Halpern, D., Procaccia, A.D., Rahwan, I., Shapira, I., Wüthrich, M.: Optimal engagement-diversity tradeoffs in social media. In: Proceedings of the ACM Web Conference 2024. pp. 288–299 (2024)
3. Broniatowski, D.A.: Does systems architecture drive risk perception? In: IISE Annual Conference. Proceedings. p. 1543. Institute of Industrial and Systems Engineers (IISE) (2015)
4. Broniatowski, D.A.: Flexibility due to abstraction and decomposition. Systems Engineering **20**(2), 98–117 (2017)
5. Broniatowski, D.A., Jamison, A.M., Johnson, N.F., Velasquez, N., Leahy, R., Restrepo, N.J., Dredze, M., Quinn, S.C.: Facebook pages, the "disneyland" measles outbreak, and promotion of vaccine refusal as a civil right, 2009–2019. American journal of public health **110**(S3), S312–S318 (2020)
6. Broniatowski, D.A., Moses, J.: Flexibility, complexity, and controllability in large scale systems (2014)
7. Broniatowski, D.A., Moses, J.: Measuring flexibility, descriptive complexity, and rework potential in generic system architectures. Systems Engineering **19**(3), 207–221 (2016)
8. Broniatowski, D.A., Simons, J.R., Gu, J., Jamison, A.M., Abroms, L.C.: The efficacy of facebook's vaccine misinformation policies and architecture during the covid-19 pandemic. Science Advances **9**(37), eadh2132 (2023)
9. Broniatowski, D.A., Zhong, W., Simons, J.R., Jamison, A.M., Dredze, M., Abroms, L.C.: Explaining twitter's inability to effectively moderate content during the covid-19 pandemic. Scientific Reports (2025in press)
10. Ciampaglia, G.L., Nematzadeh, A., Menczer, F., Flammini, A.: How algorithmic popularity bias hinders or promotes quality. Scientific reports **8**(1), 15951 (2018)
11. Cima, L., Tessa, B., Cresci, S., Trujillo, A., Avvenuti, M.: Taming toxicity or fueling it? the great ban role in shifting toxic user behavior and engagement. arXiv preprint arXiv:2411.04037 (2024)
12. Cinus, F., Minici, M., Monti, C., Bonchi, F.: The effect of people recommenders on echo chambers and polarization. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 16, pp. 90–101 (2022)
13. Forestal, J.: Social media, social control, and the politics of public shaming. American Political Science Review **118**(4), 1704–1718 (2024)
14. Hofmann, V., Schütze, H., Pierrehumbert, J.B.: The reddit politosphere: a large-scale text and network resource of online political discourse. In: Proceedings of the international AAAI conference on web and social media. vol. 16, pp. 1259–1267 (2022)

15. Jhaver, S., Frey, S., Zhang, A.X.: Decentralizing platform power: A design space of multi-level governance in online social platforms. Social Media+ Society **9**(4), 20563051231207857 (2023)
16. Li, M., Suk, J., Zhang, Y., Pevehouse, J.C., Sun, Y., Kwon, H., Lian, R., Wang, R., Dong, X., Shah, D.V.: Platform affordances, discursive opportunities, and social media activism: A cross-platform analysis of# metoo on twitter, facebook, and reddit, 2017–2020. new media & society p. 14614448241285562 (2024)
17. van der Linden, S.: How influencers and algorithms mobilize propaganda—and distort reality. Nature **633**(8029) (2024)
18. Moses, J.: The anatomy of large scale systems revisited. In: Second International Symposium on Engineering Systems. MIT, Cambridge, Massachusetts. Citeseer (2009)
19. Moses, J.: The anatomy of large scale systems (2012)
20. Perra, N., Rocha, L.E.: Modelling opinion dynamics in the age of algorithmic personalisation. Scientific reports **9**(1), 7261 (2019)
21. Piccardi, T., Saveski, M., Jia, C., Hancock, J.T., Tsai, J.L., Bernstein, M.: Social media algorithms can shape affective polarization via exposure to antidemocratic attitudes and partisan animosity. arXiv preprint arXiv:2411.14652 (2024)
22. Santos, F.P., Lelkes, Y., Levin, S.A.: Link recommendation algorithms and dynamics of polarization in online social networks. Proceedings of the National Academy of Sciences **118**(50), e2102141118 (2021)
23. Sasahara, K., Chen, W., Peng, H., Ciampaglia, G.L., Flammini, A., Menczer, F.: Social influence and unfollowing accelerate the emergence of echo chambers. Journal of Computational Social Science **4**(1), 381–402 (2021)
24. Smaldino, P.E., Russell, A., Zefferman, M.R., Donath, J., Foster, J.G., Guilbeault, D., Hilbert, M., Hobson, E.A., Lerman, K., Miton, H., et al.: Information architectures: a framework for understanding socio-technical systems. npj Complexity **2**(1), 13 (2025)
25. Tokita, C.K., Guess, A.M., Tarnita, C.E.: Polarized information ecosystems can reorganize social networks via information cascades. Proceedings of the National Academy of Sciences **118**(50), e2102147118 (2021)
26. Wang, S., Huang, S., Zhou, A., Metaxa, D.: Lower quantity, higher quality: Auditing news content and user perceptions on twitter/x algorithmic versus chronological timelines. Proceedings of the ACM on Human-Computer Interaction **8**(CSCW2), 1–25 (2024)
27. Wang, Y., Kleinberg, J.: On the relationship between relevance and conflict in online social link recommendations. Advances in Neural Information Processing Systems **36**, 36708–36725 (2023)
28. Wei, Z., Broniatowski, D.A.: Characterizing system architectures using network data. Procedia Computer Science **153**, 301–308 (2019)
29. Weng, L., Flammini, A., Vespignani, A., Menczer, F.: Competition among memes in a world with limited attention. Scientific reports **2**(1), 335 (2012)
30. Zhang, A.X., Bernstein, M.S., Karger, D.R., Ackerman, M.S.: Form-from: A design space of social media systems. Proceedings of the ACM on Human-Computer Interaction **8**(CSCW1), 1–47 (2024)
31. Zhou, B., Pei, S., Muchnik, L., Meng, X., Xu, X., Sela, A., Havlin, S., Stanley, H.E.: Realistic modelling of information spread using peer-to-peer diffusion patterns. Nature Human Behaviour **4**(11), 1198–1207 (2020)