A Consent-Driven Model for Reducing Echo Chambers in Social Media

Naomi Korem[0000−0003−0415−0287], Tammar Shrot[0000−0002−9611−2765], and Hadassa Daltrophe[0000−0003−2305−125X]

Shamoon College of Engineering, 84 Jabotinsky St., Ashdod 77245, Israel
{naomiko,tammash,hadasda1}@sce.ac.il

Abstract. Echo chambers in social media pose a growing threat to democratic discourse. Unlike other approaches that address this challenge by imposing obligations on the platforms or through regulatory measures that violate users' rights, our proposal preserves user autonomy by allowing them to decide whether to explore opinions outside their echo chamber. This paper investigates whether a regulation applied only to users who have granted *explicit consent* can promote balanced information flow across polarized communities throughout the network. We propose a general model that captures how information spreads on social media. The model accounts for users' friends, opinions, and message virality. We conduct simulations with varying parameters over real-world data to analyze the model's behavior. Our results show that even targeting a small fraction of consenting users can significantly enhance cross-group message reach (without overstepping user entitlements). This work underscores the potential for a consent-driven regulation to foster healthier public dialogue.

Keywords: Social Network · Regulation · Personalization Algorithms · Polarization · Privacy Preserving

1 Introduction

Polarization in contemporary societies has become increasingly pronounced, with stark divisions emerging along political, social, and cultural lines [9]. A key contributor to this trend is the role played by social media platforms, particularly through their personalization (a.k.a. recommendation) algorithms [10]. These algorithms are designed to maximize users' engagement by presenting content that aligns with their individual preferences. This often results in the creation of "echo chambers", where users are predominantly exposed to views that mirror their own or discredit opposing views. In such settings, beliefs are amplified [4] through repeated communication within a closed network, insulated from outside perspectives or rebuttal [3]. Such environments can undermine informed democratic discourse. Moreover, members of an echo chamber tend to develop feelings like hate, distrust and contempt towards people who don't share their views, those that exist outside their echo chamber [6]. In recent years, echo

chambers have been observed across various social networks [4]. While their precise role remains debated [5], various studies suggest that the existence of echo chambers contributes to social polarization [8].

These concerns point to the need for some form of regulation. One possible solution is to manipulate personalization algorithms to recommend potentially relevant and diverse friends from outside the users' echo chamber [12]. However, this solution requires the consent of social media platforms such as *Meta* or *X*, a consent that would be difficult to achieve or enforce. Alternatively, regulation could occur externally, without access to platform internals, but this too poses challenges: even if such a regulation has good reasons, there is also an important reason against it. After all, When the state decides for us that we can no longer consume only opinions similar to our own, it is compromising our *autonomy* to decide for ourselves. Autonomy is generally taken to be a central value of liberalism, and we all tend to get quite nervous when our autonomy is threatened.

In this paper, we propose a regulation that takes seriously the notion of autonomy. We take inspiration from, and base our research upon, a recent study [2] that explores the question of whether regulation can reduce the echo chamber effect, given a general information-spreading model that captures the essence of a social media, friends-based, information-spreading process. In this study, a crucial criterion for any such regulation is a commitment to respecting user privacy. Specifically, regulatory functions cannot access or use user opinions. We investigate whether improved exposure to diverse opinions can be achieved by allowing users to voluntarily renounce their privacy and grant access to their opinions. Such users reflect people in the real world who wish to escape their echo chamber, and therefore give the regulator access to their opinion. Their autonomy is, therefore, respected. We study how this voluntary disclosure affects exposure across the network.

Main contribution We propose a regulatory framework for social media platforms that utilizes a consent-based mechanism to respect user autonomy. To support this approach, we develop a network-based spreading model that integrates user consent into the regulatory intervention process. Additionally, we define an objective function for echo chamber mitigation that explicitly considers the network's structural topology. Finally, we validate the effectiveness of our proposed model through extensive simulations on real-world social network data. Specifically, we simulate message spreading over several models of network spreading, while also examining the effects of the parameter values.

2 Social-media spreading model

We present a simplified model for information propagation in a social-media network. The framework captures user interactions, platform behavior, and regulatory mechanisms. Our presentation is based on [2], where full technical details and extensions can be found.

We model a social network as a tuple $N = (G, c, s)$, where $G = (V, E)$ is an undirected graph, V denotes the set of users, and $E \subseteq V \times V$ represents social connections. Each user $v \in V$ is assigned an opinion via a coloring function $c : V \to \{\texttt{red}, \texttt{blue}\}$. Additionally, a response function $s : V \to \{\texttt{inactive}, \texttt{react}, \texttt{ignore}\}$ specifies each user's reaction to a message. All users begin in the neutral state `inactive`, and upon receiving a message, may transition to either `react` or `ignore`. The transition is monotonic: once a user changes state, the response cannot be reverted. A message originates from a `red` user v with $s(v) = \texttt{react}$ at time $t = 0$. The process unfolds in discrete time steps.

Let $\mathcal{A}_t = L_t \cup I_t$ be the set of users who have reacted by time t, with L_t and I_t denoting users who have `react`-ed or `ignore`-ed the message, respectively. The evolution $\mathcal{P}_t = \langle \mathcal{A}_0, \ldots, \mathcal{A}_t \rangle$ defines the *spreading sequence*. To model the spreading behavior, we define the *social media spreading function*, $\mathcal{F}_M$, the *user response function*, $\mathcal{F}_U$, and the *regulation spreading function*, $\mathcal{F}_R$. The specific implementations of $\mathcal{F}_M$, $\mathcal{F}_U$, and $\mathcal{F}_R$ are left abstract to allow for various theoretical and empirical investigations. We can now formally describe the evolution of a generic social media spreading process.

Definition 1 (Social Media Spreading Process). *A social media spreading process is defined by a 5-tuple $\langle v, N, \mathcal{F}_M, \mathcal{F}_R, \mathcal{F}_U \rangle$ where the initial message originates from user v in network N. The process unfolds in rounds indexed by t, generating the spreading sequence $\mathcal{P}_t$ using the following steps:*

1. ***Message Sharing:*** *At the end of time t, any user who responded with* `react` *shares the message. At $t = 0$, the originator v sets $s(v) =$* `react` *and shares the message.*
2. ***Social Media Choice:*** *The platform selects a candidate inbox set M_{t+1} of inactive users to receive the message at time $t + 1$, using $\mathcal{F}_M(N, \mathcal{P}_t, v)$.*
3. ***Regulation Choice:*** *A regulatory mechanism selects the final inbox set $Q_{t+1} = \mathcal{F}_R(N, M_{t+1})$, potentially extending the platform's selection. In the case of passive regulation, $\mathcal{F}_R = \emptyset_{\mathrm{R}}$ and $Q_{t+1} = M_{t+1}$.*
4. ***User Response:*** *Each user $u \in Q_{t+1}$ receives the message in their inbox and updates their status via $s(u) = \mathcal{F}_U(N, \mathcal{P}_t, v)$. Then:*

   $$L_{t+1} = L_t \cup \{u \mid u \in Q_{t+1}, s(u) = \texttt{react}\}, I_{t+1} = I_t \cup \{u \mid u \in Q_{t+1}, s(u) = \texttt{ignore}\}.$$

   Hence, $\mathcal{A}_{t+1} = L_{t+1} \cup I_{t+1}$.
5. ***Repeat or Stop:*** *If new users* `react` *the message ($L_{t+1} \neq L_t$), the process continues to the next round. Otherwise, it terminates.*

This framework outlines a general and extensible model for opinion spreading in social networks. In the next sections, we present concrete instantiations of $\mathcal{F}_M$, $\mathcal{F}_U$, and $\mathcal{F}_R$ to explore fundamental behaviors under different policies.

2.1 Modeling Social Media Spreading function, $\mathcal{F}_M$

Social media platforms employ complex proprietary algorithms to determine which content is displayed in the user feed. These algorithms may consider message content, user interaction history, and other contextual factors. However, in

this work, we focus specifically on the echo chamber and thus adopt a simplified yet principled model of message propagation. We use the following assumptions:

- A message is only propagated by users who previously chose to `react`.
- A message shared by a user v is eligible to be seen only by v's neighbors.
- A message shared at time t may appear only in the feed of users who are `inactive` at time t, and only for the next round $(t+1)$.
- The decision to deliver the message to each neighbor is made independently of other neighbors.

Let us define $p, q \in [0,1]$ as two probabilities controlling how likely a user is to receive the message based on opinion similarity. The probability that an `inactive` node $u \in \bar{\mathcal{A}}_t$, who is a neighbor of a newly active user v, is chosen to receive the message in round $t+1$ is given by:

$$\delta(v,u) = \begin{cases} 0 & (v,u) \neq E \\ p & (v,u) \in E \text{ AND } c(u) = c(v) \\ q & (v,u) \in E \text{ AND } c(u) \neq c(v) \end{cases}$$

This defines the **spreading probability** from v to u based on their connection and opinion similarity.

Definition 2 (social media spreading function). *The function $\mathcal{F}_M(N, \mathcal{P}_t, v)$ defines the candidate inbox set for round $t+1$, denoted by M_{t+1}. This set is generated as follows: for every edge $(v,u) \in E$ such that $v \in L_t \setminus L_{t-1}$ and $u \in \bar{\mathcal{A}}_t$, node u is included in M_{t+1} independently with probability $\delta(v,u)$.*

To explore the impact of social media algorithms on echo chambers, we consider the following specific configurations of the parameters p and q:

1. **Uniform spreading ($p = 1, q = 1$):** The message is forwarded to *all* inactive neighbors of newly active users, regardless of color.
2. **p-homophily ($p \geq \frac{1}{2},\ q = 1-p$):** The message is more likely to be forwarded to neighbors with the same opinion as the sender. This models homophilic behavior, where users with similar views are preferentially exposed to shared content.

2.2 Modeling User Response Function, $\mathcal{F}_U$

When a user receives a new message in their inbox (i.e., feed), how will they react to it? What will cause them to `react` to the message or `ignore` it? Obviously, this is a non-transparent, complex process that is hard to model exactly.

There are several factors that are known to influence the number of *likes* a post receives. These include the identity of the people doing the liking [11], and its existing *like* count [7]. First, when a `react` comes from friends, familiar people, or people with similar opinions to ours, they carry greater social proof and elicit more subsequent `reacts` from others, strengthening relational bonds and encouraging further engagement. Second, posts that already have more `reacts`

tend to attract disproportionately more new `reacts`— a "rich-get-richer" effect — whereby early popularity begets further popularity.

Hence, for the sake of tractability, we assume that a user's reaction is determined by three factors: (i) the user u opinion (modeled by its color $c(u)$), (ii) the identity of the sender, v, specifically whether the sender is a neighbor and their associated opinion (i.e., $c(v), c(u)$), and (iii) the cumulative number of users who have `react`-ed to the message up to time t.

Recall that the color of the message is fixed to `red`, therefore, the color-dependent function $g(u)$ provides high probability for a `red` user to `react` to a message: $g(u) = \begin{cases} 1 & c(u) = \texttt{red} \\ -1 & \text{otherwise} \end{cases}$

The friendship function $f(v, u)$ reflects the influence of the sender v identity and adjacency with user u; $f(v,u) = \begin{cases} -1 & (v,u) \neq E \\ 0 & (v,u) \in E \text{ AND } c(u) \neq c(v) \\ 1 & (v,u) \in E \text{ AND } c(u) = c(v) \end{cases}$

Finally, recall that n is the number of users in the network N, and $|L_t|$ is the number of users who response with `react`, the virality-function $h(L_t, n)$ models the impact of the social proof: $h(L_t, n) = |L_t|/n$. These three factors underlie the modeling of the *user reaction function*, denoted by $\mathcal{F}_U(N, \mathcal{P}_t, v)$.

Definition 3 (User reaction function, $\mathcal{F}_U$). *Let $|L_t|$ denote the number of users who have responded with* `react` *up to time t. Given a message recipient u and a sender v, the* user reaction function $\mathcal{F}_U(N, \mathcal{P}_t, v)$ *determines whether user u responded with* `react` *or* `ignore` *to the message, according to the following probabilistic rule:*

$$s(u) = \mathcal{F}_U(N, \mathcal{P}_t, v) = \begin{cases} \texttt{react} & \textit{with probability } \sigma\left(h(L_t, n) \cdot (\alpha g(u) + \beta f(v,u))\right) \\ \texttt{ignore} & \textit{otherwise} \end{cases}$$

where $\sigma(x) = \frac{1}{1+e^{-x-1/2}}$ is the sigmoid function, modeling the stochastic nature of the response, and α, β are weight parameters that modulate the relative influence of g and f.

2.3 Modeling the Consent-Driven Regulation Function, $\mathcal{F}_R$

The proposed regulatory mechanism is based on an opt-in principle: Only users who have explicitly consented to regulation are subject to intervention. This group of consenting users is referred to as the *opt-in set*, denoted by Q.

Consent may arise from various motivations, for instance, a willingness to be exposed to diverse perspectives or an awareness (enabled through transparency mechanisms) that the user is currently embedded within an ideological echo chamber. For the purposes of this model, all members of Q must hold the `blue` opinion, as the message being distributed in the system is of opinion `red`.

The regulation mechanism selectively intervenes for users in $Q \subseteq V_{\texttt{blue}}$ (the set of `blue` users) by potentially exposing them to content that challenges their existing opinion. This is achieved by injecting additional `red`-content into their feed, beyond what they would receive through the normal spreading function.

Definition 4 (Consent-Driven Regulation Function $\mathcal{F}_R$). *Given a set of opt-in users $Q \subseteq V_{\texttt{blue}}$, a candidate message recipient set $M_t \subseteq V$, and a regulation parameter $\rho \in [0,1]$, The regulation function at time t, denoted by $\mathcal{F}_R(M_t, Q, \rho)$, is computed as: $\mathcal{F}_R = M_t \cup Q_t$, where $Q_t \subseteq Q$ is a set of* `inactive` *users selected uniformly at random from the set Q, and $|Q_t| = \lceil \rho \cdot |M_t| \rceil$.*

3 Results

Experimental Setup. To evaluate the influence of the Consent-Driven Regulation mechanism on the spreading dynamics, a series of simulations were carried out on the *Bloggers*[52,48] graph. This network comprises 1222 nodes and 16,717 edges, representing the largest connected component of a weblog network centered on U.S. political discourse in 2005 [1]. Each node corresponds to a user whose political affiliation is classified as either conservative or liberal, with 636 users (52%) labeled as `red` and 586 users (48%) as `blue`. The graph structure includes 7841 `red` edges and 7301 `blue` edges, which denote intra-group connections, as well as 1575 inter-group (cross-party) edges.

The spreading process is initiated by selecting a `red` user uniformly at random and updating their status to `react`. The spreading then proceeds according to the dynamics defined in Definition 1, using the following parameters:

(i) **Spreading function:** $\mathcal{F}_M$ with parameter settings $(p = 1, q = 0)$ and $(p = 0.9, q = 0.1)$.
(ii) **User Response function:** $\mathcal{F}_U$ with parameter setting $(\alpha = 0.7, \beta = 0.3)$.
(iii) **Consent-Driven Regulation Function:** $\mathcal{F}_R$ parameterized by ρ, which denotes the relative fraction of additional users from the opt-in set who received the message. The values considered are $\rho \in \{0, 0.001, 0.01, 0.1\}$.

Once the spreading process terminates, we record the distribution of user responses (`react` or `ignore`) among those who were exposed to the message. Each configuration is simulated 1000 times, and the results are averaged to obtain the expected number of reactions.

Reference Point for Echo Chamber Reduction Evaluation. In addressing the challenge of mitigating the echo chamber phenomenon in social media, a fundamental question arises: what form of message distribution would constitute a meaningful step toward reducing ideological segregation in the network? We propose that the design of the objective function should take into account the *original topology* of the network. Specifically, we suggest a reference point in which a message originating in the `red` community is spreading uniformly across both `red` and `blue` neighbors of each node, without regulatory intervention. This objective, captured by the "Uniform spreading" model (see Section 2.1), serves as a natural reference point for evaluating the effects of any regulatory mechanism.

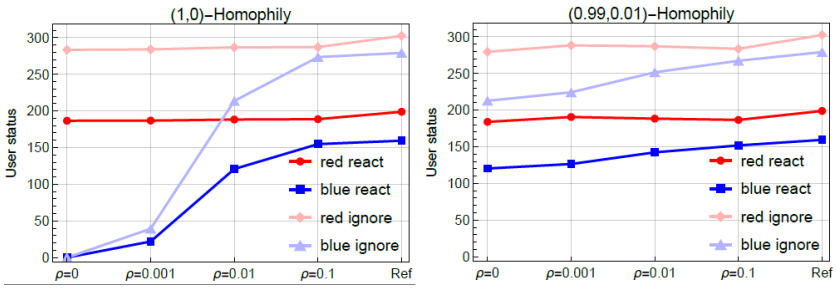

Fig. 1. Consent-Driven Regulation intervention (ρ) under two homophily settings: pure homophily ($p = 1, q = 0$) (left) and relaxed homophily ($p = 0.99, q = 0.01$) (right).

Experimental Output. Figure 1 shows the aggregated results. The rightmost value ('Ref') represents the reference scenario, where each user shares the message uniformly with all neighbors, without homophily or regulation.

The left plot in Figure 1 shows a pure homophily spreading function. When $\rho = 0$ (i.e., passive regulation), a `red` user spreads the message *only* to other `red` users. Hence, we can see that only `red` users got the message, which led to only `red` response in the population. As expected, increasing the regulation intervention by ρ, allows `blue` users to receive the message. However, even a minimal regulatory adjustment ($\rho = 0.001$) results in a substantial increase in both blue exposure and reactivity, demonstrating the high sensitivity of the network to small interventions. Similar results are observed even when the social media spreading mechanism is less strictly homophilic, for instance, under parameter settings such as ($p = 0.999, q = 0.001$).

The right plot in Figure 1 illustrates a spreading mechanism with a 1% tolerance for differing opinions, resulting in a less homogeneous echoing effect even in the absence of regulation (i.e., $\rho = 0$). Introducing the Consent-Driven Regulation mechanism progressively guides the system toward the desired distribution, as indicated by the reference point.

Notably, in both scenarios, our results show that even targeting a small fraction of consenting users can significantly enhance cross-group message reach. Furthermore, in both scenarios, the red group's reactivity remains largely unchanged across all levels of regulation.

4 Discussion

The results demonstrate that regulatory intervention mitigates the echo chamber effect by expanding message reach within the `blue` bubble. Notably, even regulation with only 1% additional users per phase, yields a substantial improvement in cross-group exposure. Moreover, the number of `blue` recipients far exceeds those chosen by the regulator, who are a negligible subset. These findings show that lightweight regulatory mechanisms can reduce polarization, even under strong homophily (e.g., $p = 0.999$, $q = 0.001$). Small interventions

($\sim 1\%$) achieve significant diversity in exposure. Thus, effective platform regulation need not involve sweeping algorithmic changes. Importantly, our results offer practical guidance for social media platforms and policymakers. Injecting messages into a small fraction of unreached consenting users can significantly improve exposure diversity. The regulation can be implemented flexibly, allowing users to opt-out and thus preserving user consent while improving balance. From a policy standpoint, our finding support digital regulation frameworks like the EU Digital Services Act Regulation improves exposure diversity without affecting `red` group exposure, suggesting that public–interest interventions can be both effective and minimally invasive. While the model captures key dynamics of polarization, its simplicity limits real-world applicability. Future work could explore alternative spreading and user reaction functions.

References

1. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 us election: divided they blog. In: Proceedings of the 3rd international workshop on Link discovery. pp. 36–43 (2005)
2. Avin, C., Daltrophe, H., Lotker, Z.: On the impossibility of breaking the echo chamber effect in social media using regulation. Scientific Reports **14**(1), 1107 (2024)
3. Baumann, F., Lorenz-Spreen, P., Sokolov, I.M., Starnini, M.: Modeling echo chambers and polarization dynamics in social networks. Physical Review Letters **124**(4), 048301 (2020)
4. Cinelli, M., Morales, G.D.F., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proceedings of the National Academy of Sciences **118**(9) (2021)
5. Haidt, J., Bail, C.: (ongoing) social media and political dysfunction: A collaborative review. new york university (2022), https://tinyurl.com/PoliticalDysfunctionReview, unpublished manuscript
6. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., Westwood, S.J.: The origins and consequences of affective polarization in the united states. Annual review of political science **22**(1), 129–146 (2019)
7. Muchnik, L., Aral, S., Taylor, S.J.: Social influence bias: A randomized experiment. Science **341**(6146), 647–651 (2013)
8. Settle, J.E.: Frenemies: How social media polarizes America. Cambridge University Press (2018)
9. Silver, L.: Most across 19 countries see strong partisan conflicts in their society, especially in south korea and the us. Pew Research Center (2022)
10. Stinson, C.: Algorithms are not neutral: Bias in collaborative filtering. AI and Ethics **2**(4), 763–770 (2022)
11. Stsiampkouskaya, K., Joinson, A., Piwek, L.: To like or not to like? an experimental study on relational closeness, social grooming, reciprocity, and emotions in social media liking. Journal of Computer-Mediated Communication **28**(2), zmac036 (2023)
12. Tommasel, A., Rodriguez, J.M., Godoy, D.: I want to break free! recommending friends from outside the echo chamber. In: Proceedings of the 15th ACM Conference on Recommender Systems. pp. 23–33 (2021)