Online Platforms and the Fair Exposure Problem under Homophily

Jakob Schoeffer[1*], Alexander Ritchie[2*], Keziah Naggita[3*], Faidra Monachou[4*], Jessie Finocchiaro[4,5*], Marc Juarez[6]

[1]Karlsruhe Institute of Technology (KIT)
[2]University of Michigan
[3]Toyota Technological Institute at Chicago
[4]Harvard University
[5]Center for Research on Computation and Society (CRCS)
[6]University of Edinburgh
jakob.schoeffer@kit.edu, aritch@umich.edu, knaggita@ttic.edu, monachou@stanford.edu, jessie@seas.harvard.edu, marc.juarez@ed.ac.uk

Abstract

In the wake of increasing political extremism, online platforms have been criticized for contributing to polarization. One line of criticism has focused on echo chambers and the recommended content served to users by these platforms. In this work, we introduce the *fair exposure problem*: given limited intervention power of the platform, the goal is to enforce balance in the spread of content (e.g., news articles) among two groups of users through constraints similar to those imposed by the *Fairness Doctrine* in the United States in the past. Groups are characterized by different affiliations (e.g., political views) and have different preferences for content. We develop a stylized framework that models intra- and intergroup content propagation under homophily, and we formulate the platform's decision as an optimization problem that aims at maximizing user engagement, potentially under fairness constraints. Our main notion of fairness requires that each group see a mixture of their preferred and non-preferred content, encouraging information diversity. Promoting such information diversity is often viewed as desirable and a potential means for breaking out of harmful echo chambers. We study the solutions to both the fairness-agnostic and fairness-aware problems. We prove that a fairness-agnostic approach inevitably leads to group-homogeneous targeting by the platform. This is only partially mitigated by imposing fairness constraints: we show that there exist optimal fairness-aware solutions which target one group with different types of content and the other group with only one type that is not necessarily the group's most preferred. Finally, using simulations with real-world data, we study the system dynamics and quantify the price of fairness.

1 Introduction

In the wake of increasing political extremism (US Department of Justice 2021), online platforms (e.g., social media networks) have been extensively criticized for exacerbating political polarization in the United States (Boxell, Gentzkow, and Shapiro 2017; Bail et al. 2018; Hawdon et al. 2020; Yarchi, Baden, and Kligler-Vilenchik 2020) and elsewhere.[1] This phenomenon is often attributed to platform designs that aim to generate revenue by maximizing user engagement with promoted or shared content (e.g., news articles, opinions, ads). Motivated by the need to promote pluralism online, this paper focuses on understanding the spread of information under a limited platform intervention scheme, where the platform exposes (a subset of) users of the same affiliation to content of contrasting views. We introduce this problem and its study as the *fair exposure problem*.

From a historical perspective, parallels can be drawn between the fair exposure problem and the *Fairness Doctrine* (Ashford 2021; Pickard 2021), a past media policy which required that news media cover issues of public importance by presenting diverse, opposing perspectives in an attempt to ensure media diversity. Over the decades, the effectiveness and ethical use of this policy was questioned (Pickard 2021): for example, the doctrine enabled activists to help combat racist broadcasting, but it also helped promote the Anti-Equal Rights Amendment campaign (Pickard 2021). As history has shown, interventions aimed at balancing the exposure of the public to opposing views might have ambiguous results. Thus, the goal of this paper is to shed light on the trade-offs that the adoption of such policies may introduce for online platforms.

Towards this goal, we develop a stylized model to understand the impact of platform interventions on the propagation of different articles over time to understand the effects of *positional polarization* (cf. Yarchi, Baden, and Kligler-Vilenchik (2020)). Our model considers two groups of users with different affiliations and different preferences for articles. Among two opposing articles, we assume that each group tends to like more the article that aligns with the group's views. Moreover, due to homophily in social net-

*These authors contributed equally.

[1]See https://arxiv.org/abs/2202.09727 for the full paper including the appendix and https://github.com/jfinocchiaro/fair-exposure for all code.

works, users in a given group see mostly articles shared by other users in the same group. In this framework, the platform wishes to maximize user exposure (measured through the aggregate number of clicks and likes), potentially subject to fairness constraints. We only consider interventions where the platform chooses the articles that an *initial set* of users in each group sees. Our main fairness notion aims at approximately equalizing the relative exposure to a mixture of preferred and non-preferred articles across groups, by imposing certain lower and upper bounds. We analyze the platform's optimization problem and compare the solutions for its unconstrained (fairness-agnostic) version to the solutions for its constrained (fairness-aware) version. We prove that the fairness-agnostic solution always targets each group with one article. When the platform must abide by the fairness constraints, we show that at least one group will be targeted with a mixture of articles. However, depending on the model parameters, it may be optimal that the other group is targeted with only one article type; interestingly, the selected article may not be the group's preferred article. Thus, one group incurs the "cost of fairness," whereas the other one the "cost of maximizing engagement." When the content refers to high-stakes procedures (e.g., referendums, elections), such an outcome can be problematic.

We supplement our theoretical results with empirical results to gain additional insights by estimating our model parameters from real-world datasets collected from Twitter and Facebook (Garimella et al. 2017; Bakshy, Messing, and Adamic 2015). Moreover, we measure the *price of fairness*, i.e., the difference in the platform's utility between the fairness-aware and the fairness-agnostic settings. Using parameters estimated from Bakshy, Messing, and Adamic (2015), we observe an optimal fairness-aware solution that heavily favors one group.

2 Related Work

The spread of information in social networks is well-studied; the structure of these social networks tends to be homophilous (McPherson, Smith-Lovin, and Cook 2001; Lazarsfeld and Merton 1948). Yarchi, Baden, and Kligler-Vilenchik (2020) formalize three notions of polarization that emerge from social networks: *interactional, positional, and affective* polarization. We study a model most suited to study positional polarization through information exposure. Balancing information exposure has also been studied through several different technical methods; however, to our knowledge, the impact of platform interventions to ensure balanced exposure via fairness constraints has not been studied before. Celis et al. (2019) study a similar problem of controlling polarization in bandit settings, though our model differs by assuming that intervention is only possible at the first time step; their constrained problem is similar to our approximately fair average exposure constraint in (3). Our model is sequential like the social learning models of Banerjee (1992); Bikhchandani, Hirshleifer, and Welch (1992), which also study information spread, but without balancing content exposure. Papanastasiou (2020); Candogan and Drakopoulos (2020) study stylized models for fact-checking news articles in social networks when the platform can intervene to inspect the content or incentivize fact-checking by users through information design; Cisternas and Vásquez (2020) take a market design approach. Allon, Drakopoulos, and Manshadi (2021) further show that polarization arises due to uncertainty in content accuracy.

Starbird et al. (2018) demonstrate the emergence of echo chambers by a mixed methods analysis of perceptions of the White Helmets, particularly enabled by content sharing platforms such as Twitter, and Jeon et al. (2021) gamify the balance of seeking influence and reputability simultaneously on Twitter. Our setting is also similar to influence maximization literature (Kempe, Kleinberg, and Tardos 2003) in the sense that platform interventions are limited. However, our model is sequential and aims for balance in article exposure, while the influence maximization literature seeks to maximize information diffusion (Fish et al. 2019; Stoica, Han, and Chaintreau 2020; Ali et al. 2019). Finally, balancing information propagation is well-studied in literature on recommender systems (Zoetekouw 2019; Hu et al. 2012; Farajtabar et al. 2016) and the emergence of echo chambers (Barberá et al. 2015; Mukerjee, Jaidka, and Lelkes 2020; Dubois and Blank 2018; Hosseinmardi et al. 2020). Bakshy, Messing, and Adamic (2015) and Garimella et al. (2017) empirically study the extent of disparity in intragroup exposure of ideas and do not aim to balance it. In general, although previous works (Bakshy, Messing, and Adamic 2015; Garimella et al. 2017) investigate the empirics of information flow in similar models, they do not study the mechanisms that lead to (imbalanced) exposure; our model addresses this.

Many of the standard metrics of group fairness are not applicable in our setting as we work with heterogeneous preferences of outcomes: members of one group prefer seeing content that aligns with their group identity. Graph-based models of opportunity flow have considered similar, yet inherently different, fairness constraints and problems. For example, Liu et al. (2021) consider fair equality of opportunity in settings where flow of opportunity proceeds along an acyclic graph and everyone is striving for the same desired outcome. Similarly, Arunachaleswaran et al. (2021) approximately optimize social welfare in settings where opportunity flows along an acyclic graph. Recently, Chen et al. (2022) apply fairness constraints to other online platform operations, specifically assortment planning. However, neither of the approaches in this paper are directly applicable to our setting. The definitions of *fair exposure* presented in § 3 are stylized for this particular setting.

3 Model

3.1 General Setup

We consider a platform with a finite mass M of users with affiliation group $g \in \{A, B\}$. Let $\pi_g \in (0, 1)$ denote the fraction of users from group g (at any time). We assume that $\pi_A = 1 - \pi_B = \pi$. Time is discrete with $t = 1, 2, \ldots, T$, $T \leq M$. All notation is summarized in Table 1.

Before time $t = 1$, the platform receives two articles representing different views a, b that are aligned with groups A, B, respectively (e.g., sponsored posts on Facebook or Twitter). For simplicity, we refer to the two articles as a

Symbol	Definition
M	Finite mass of users
$g \in \{A, B\}$	Affiliation group
$\pi_g \in (0, 1)$	Fraction of users in group g
$t \in \{1, \ldots, T\}$	Time step (discrete) with horizon $T \leq M$
$s \in \{a, b\}$	Article sources affiliated with groups A, B
$\theta_{g,s} \in [0, 1]$	Fraction of group g users who are shown article s by the platform at $t = 1$
$p_{g,s} \sim F_{g,s}$	Probability for users of group g to like an article of source s
$F_{g,s}$	Distribution with support $[0, 1]$
$c_{g,s} > 0$	Cost users in g occur when clicking on an article s
$v_{g,s} > 0$	Valuation of users in g when liking an article s
$q_g \in (1/2, 1)$	Probability of intragroup propagation
$l_{g,s}(t, \theta)$	Mass of users in g born at time t who have clicked and liked an article s
$e_{g,s}(t)$	Exposure of users in g to article s at time t
$\underline{\delta} < 1 < \overline{\delta}$	Fairness lower and upper bound parameters

Table 1: Overview of notation.

and b, where a (resp. b) is the *in-group/preferred* (resp. *out-group/non-preferred*) article type of group A, and similarly for group B. At time $t = 1$, the platform decides how many users in group g to show an article s to. Let $\theta_{g,s}$ denote the fraction of users in group g who are shown article s by the platform at time $t = 1$.

Each user observes the source $s \in \{a, b\}$ of the article they are shown. Users of group g have a probability $p_{g,s} \sim F_{g,s}$ of "liking" an article of source s, where $F_{g,s}$ is a known distribution with support $[0, 1]$. Each user in group g knows their own realized probabilities $p_{g,s}$ for $s \in \{a, b\}$. Users from group A have a higher preference for articles of source a; the same holds for users of group B and articles of source b. To model this (stochastically) biased behavior of users in each group g, we assume that $p_{A,a} \succ_{FSD} p_{A,b}$ and $p_{B,b} \succ_{FSD} p_{B,a}$.[2]

At every time period $t > 0$, a unit mass of users arrives. At time t, each user in group g sees one article s and decides whether to click with probability dependent on $p_{g,s} \sim F_{g,s}$. If the user clicks on the article, they incur a constant cost $c_{g,s} > 0$ for reading the article. If they like it, they get valuation $v_{g,s} > 0$ (minus the cost $c_{g,s}$), so their final payoff is $v_{g,s} - c_{g,s} > 0$. If they do not like it, their final payoff is $-c_{g,s} < 0$.

At the next period $t + 1$, an equal mass of users arrives. This modeling choice reflects the general format of content sharing on social platforms, in which at different time steps, there are different batches of people on the online platform. We assume synchronicity in individual arrivals rather than time-step measures so we can reduce to a discretized time analysis without loss of generality. Specifically, each user from group g gives their position to a user[3] from the same group g with probability $q_g \in (1/2, 1)$, where the lower bound comes from homophily assumptions; we refer to this event as *intragroup propagation*. With probability $1 - q_g$, this user is replaced by a user in $g' \neq g$ (*intergroup propagation*). In order to ensure consistency with the fraction π_g of each group g over time, we require the parameters q_A, q_B, and π satisfy $q_A \pi_A + (1 - q_B)\pi_B = \pi_A$.[4] If a user i arriving at time t liked the article, then the new user i', replacing user i at time $t+1$, sees the same article as i. If user i did not like an article, then user i' is not shown any article at time $t + 1$.

For $t \geq 1$, let $l_{g,s}(t, \theta)$ denote the mass of users born at time t who belong to group g and have clicked and liked an article s. The objective of the platform is to maximize user exposure over time, i.e.,

$$\max_{\theta_{A,a}, \theta_{B,a} \in [0,1]} \sum_{t=1}^{T} \sum_{g \in \{A,B\}} \sum_{s \in \{a,b\}} l_{g,s}(t, \theta), \tag{1}$$

potentially subject to fair exposure constraints. We measure *user exposure* in the number of users who click and like an article. The strengths of this metric are two-fold: first, because the platform has to plan for T time steps, ensuring an article is liked means it will continue to propagate in the next time step. Second, we assume that liking an article is a proxy for more *meaningful* engagement than simply clicking on it.

While our model makes many simplifying assumptions, this strengthens our negative results (e.g., Lemma 3) as they do not hold even in an oversimplified model. Moreover, while our model is not graph-based, it is an abstraction of the Erdős-Rényi random graph in expectation with different attachment parameters for each group. As many social networks generally closely resemble preferential attachment models rather than Erdős-Rényi graphs (Clauset 2021), we compare our model's performance to graph-based simulations in § G, and observe similar results.

[2]Recall that a random variable X with CDF F_X *first-order stochastically dominates* Y with CDF F_Y, that is $X \succ_{FSD} Y$, if $F_X(z) < F_Y(z)$ for all z.

[3]Our model and analysis can be directly extended to the case where a user in period t is replaced by $n_{t+1} > 1$ users in period $t + 1$. The current assumption is made for clarity of exposition.

[4]This is necessary for theoretical results, but it does not hold for the parameters used in § 5, and does not affect results there.

3.2 Notions of Fair Exposure

Broadly speaking, we define *fair exposure* as a situation where users of different affiliation are similarly exposed to non-preferred content. Promoting such information diversity—as opposed to *selective exposure* (Freedman and Sears 1965)—is often viewed as desirable and a potential means for breaking out of harmful echo chambers that are detrimental to "the quality, safety, and diversity of discourse online," as Gillani et al. (2018) put it. Garrett and Resnick (2011), among others, likewise suggest that "software designers ought to create tools that encourage and facilitate consumption of diverse news streams, making users, and society, better off." Diversity of perspectives might also help users to see things from novel perspectives or become aware that they might be already stuck in an echo chamber. We operationalize fair exposure through two types of constraints: first, we ask that exposure rates for both types of content be *equal at each point in time* ("constant fair exposure"). Acknowledging that this is a rather restrictive constraint, we also examine fair *average* exposure, where we further allow a certain deviation from equality ("approximately fair average exposure").

Constant Fair Exposure The rate of exposure of users to their preferred article s is constant at level $e \in [0, 1]$ at each time step and equal across groups, i.e.,

$$\frac{l_{A,s}(t,\theta)}{\pi_A} = \frac{l_{B,s'}(t,\theta)}{\pi_B} = e \quad \forall t \leq T,\ \forall s, s' \in \{a, b\}, s \neq s'. \tag{2}$$

Approximately Fair Average Exposure The total exposure of users to their preferred article s (resp. non-preferred article s') is approximately equal across groups, i.e., for given parameters $\underline{\delta} < 1 < \overline{\delta}$,

$$\underline{\delta} \leq \frac{\sum_{t=1}^{T} l_{A,a}(t,\theta)}{\sum_{t=1}^{T} l_{B,b}(t,\theta)} \leq \overline{\delta} \quad \text{and} \quad \underline{\delta} \leq \frac{\sum_{t=1}^{T} l_{A,b}(t,\theta)}{\sum_{t=1}^{T} l_{B,a}(t,\theta)} \leq \overline{\delta}. \tag{3}$$

4 Theoretical Analysis

4.1 Preliminaries

We begin with preliminaries. We define the users' decision problem, analytically describe the system dynamics, and finally transform them to a tractable non-recursive form.

Users' Decision Problem A user in group g with realized probability $p_{g,s}$ of liking an article shown to them clicks on the article if and only if their expected utility is non-negative, that is

$$v_{g,s} p_{g,s} \geq c_{g,s}. \tag{4}$$

Therefore, the fraction of users in g who click on article s shown to them is $1 - F_{g,s}(\frac{c_{g,s}}{v_{g,s}})$. Since $p_{A,a} \succ_{FSD} p_{A,b}$ and $p_{B,b} \succ_{FSD} p_{B,a}$, users tend to click more on their in-group articles.

Understanding System Dynamics As a warm-up, we show how the different masses of users evolve in the first time period. We then generalize to any $t > 1$.

Time $t = 1$. Fix the fractions $\theta_{A,a}$ and $\theta_{B,a}$ of users in groups A and B, respectively, who are shown article a at time $t = 1$; recall that $\theta_{A,a}$ and $\theta_{B,a}$ are the platform's decision. Let L denote the Bernoulli random variable that a user likes the article after clicking on it. Then, the mass of users in g who clicked on the article from source a and liked it during period $t = 1$ is

$$\begin{aligned} l_{g,s}(1,\theta) =& \pi_g \theta_{g,s} \int_{c_{g,s}/v_{g,s}}^{1} \Pr[L = 1 \mid p_{g,s}] dF_{g,s}(p_{g,s}) \\ =& \pi_g \theta_{g,s} \int_{c_{g,s}/v_{g,s}}^{1} p dF_{g,s}(p). \end{aligned}$$

Symmetrically, the mass of users in g who clicked on article s but did *not* like it equals $\pi_g \theta_{g,s} \int_{c_{g,s}/v_{g,s}}^{1} (1-p) dF_{g,s}(p)$. The rest of users in group g who were shown article s did not click on it; their mass equals $\pi_g \theta_{g,s} F_{g,s}\left(\frac{c_{g,s}}{v_{g,s}}\right)$.

Time $t > 1$. For general $t > 1$, recall that a user in group g who was shown article s is replaced by a user also in g in the next time period with probability q_g (and by a user in the opposite group $g' \neq g$ with probability $1 - q_g$). For brevity, we refer to the new user as the *replacing user*. Figure 1 illustrates how an article "travels" throughout the network via intra- and intergroup propagation.

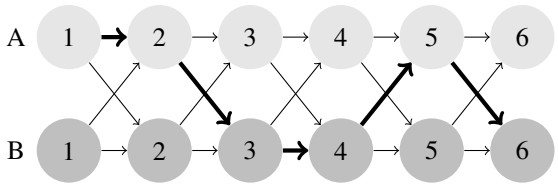

Figure 1: *Article sharing over time with* $T = 6$*:* The horizontal and diagonal edges represent intragroup and intergroup propagation, respectively. With thicker edges, we give an example of how an article s initially read by a user in A is propagated through the network.

Generalizing the system dynamics for $t > 1$, we obtain the following recursive formula:

$$l_{g,s}(t+1,\theta) = \psi_{g,s}(q_g l_{g,s}(t,\theta) + (1 - q_{g'}) l_{g',s}(t,\theta)), \tag{5}$$

where we used[5]

$$\psi_{g,s} := \int_{c_{g,s}/v_{g,s}}^{1} p dF_{g,s}(p).$$

For a visual demonstration, see Figure 4 in § A. The next lemma follows from (5). All proofs can be found in § B.

Lemma 1. *The mass function can be written as* $l_{g,s}(t,\theta) = \theta_{g,s} w_{g,s}(t) + \theta_{g',s} u_{g,s}(t)$*, where* $u_{g,s}(1) = 0$*,* $u_{g,s}(t) > 0$ *for* $t \geq 2$*, and* $w_{g,s}(t) > 0$ *for* $t \geq 1$*.*

Lemma 1 says that $l_{g,s}$ is a strictly increasing linear function of $\theta_{g,s}$ and $\theta_{g',s}$ except at time $t = 1$, when $l_{g,s}$ is not a function of $\theta_{g',s}$. We note that $w_{g,s}(t)$ corresponds to the mass of intragroup propagation and $u_{g,s}(t)$ to that of intergroup propagation.

[5]We assume $\psi_{g,s} > 0$, i.e., we do not consider the trivial case of $\psi_{g,s} = 0$.

Unfortunately, the recursive expression for the mass function given in (5) is intractable. Thus, in Theorem 1 we derive an equivalent non-recursive expression using the one-sided $\mathcal{Z}$-transform.

Theorem 1. *For all $t \geq 1$, regardless of group g and article s, we have*

$$w_{g,s}(t) = A^w_{1,g,s} a^{t-1}_{1,s} + A^w_{2,g,s} a^{t-1}_{2,s}, \quad t \geq 1, \tag{6}$$

$$u_{g,s}(t) = A^u_{g,s}(a^{t-1}_{1,s} - a^{t-1}_{2,s}), \quad t \geq 1, \tag{7}$$

with $w_{g,s}(t)$ and $u_{g,s}(t)$ as introduced in Lemma 1, and

$$a_{1,s} := \frac{1}{2}\Big(\psi_{g,s}q_g + \psi_{g',s}q_{g'} + \big[(\psi_{g,s}q_g + \psi_{g',s}q_{g'})^2 - 4\psi_{g,s}\psi_{g',s}(q_g + q_{g'} - 1)\big]^{\frac{1}{2}}\Big)$$

$$a_{2,s} := \psi_{g,s}q_g + \psi_{g',s}q_{g'} - a_{1,s}$$

$$A^w_{1,g,s} := \frac{\pi_g\psi_{g,s} + \psi_{g',s}\left(\pi_{g'}\psi_{g',s}(1-q_{g'}) - \pi_g\psi_{g,s}q_{g'}\right)a^{-1}_{1,s}}{1 - a_{2,s}a^{-1}_{1,s}}$$

$$A^w_{2,g,s} := \frac{\pi_g\psi_{g,s} + \psi_{g',s}\left(\pi_{g'}\psi_{g',s}(1-q_{g'}) - \pi_g\psi_{g,s}q_{g'}\right)a^{-1}_{2,s}}{1 - a^{-1}_{2,s}a_{1,s}}$$

$$A^u_{g,s} := \frac{\psi_{g',s}\left(\pi_{g'}\psi_{g',s}(1-q_{g'}) - \pi_g\psi_{g,s}q_{g'}\right)a^{-1}_{1,s}}{1 - a_{2,s}a^{-1}_{1,s}}.$$

We note that all quantities in Theorem 1 are real numbers, which is shown in Lemma 5 in § C. An intuitive interpretation of these quantities is as follows: the terms $a_{1,s}$ and $a_{2,s}$ are the roots of a quadratic in $\mathcal{Z}$-space that roughly corresponds to a kinematic equation describing the homophilic sharing process. $A^u_{g,s}$ corresponds roughly to the difference between contributions to the mass that would have been realized if intergroup propagation had not occurred and those that would have been realized if intragroup propagation had not occurred. The quantities $A^w_{1,g,s}$ and $A^w_{2,g,s}$ correspond roughly to $A^u_{g,s}$ and $-A^u_{g,s}$, respectively, plus an additional term relating to the mass generated by propagation *within* group g and propagation *from* group g' to g.

4.2 Platform's Optimization Problem

Building upon our previous results, in this section we proceed to formulate the platform's problem, i.e., the maximization of user exposure, as a linear program subject to *approximately fair average exposure* constraints. More specifically, at time $t = 1$ the platform needs to decide the fraction of users in each group to show articles a and b. Recall that we denote the proportion of users in g that are shown article s by $\theta_{g,s}$. The platform wants to maximize the total number of users across all groups that click on and like the two articles, but also faces a fair exposure constraint (see (3)). Thus, the platform's optimization problem becomes:

$$\max_{\theta_{A,a},\theta_{B,a}\in[0,1]} \sum_{t=1}^{T} \sum_{g\in\{A,B\}} \sum_{s\in\{a,b\}} l_{g,s}(t,\theta) \tag{P}$$

$$\text{s.t. } \underline{\delta} \leq \frac{\sum_{t=1}^{T} l_{A,a}(t,\theta)}{\sum_{t=1}^{T} l_{B,b}(t,\theta)} \leq \overline{\delta} \tag{C1}$$

$$\underline{\delta} \leq \frac{\sum_{t=1}^{T} l_{A,b}(t,\theta)}{\sum_{t=1}^{T} l_{B,a}(t,\theta)} \leq \overline{\delta}. \tag{C2}$$

Intuitively, to avoid the extreme, but feasible, case where each group g is only shown their preferred article, definition (3) introduces constraints (C1) and (C2). These constraints require that each group is exposed to their preferred article and their non-preferred article in a balanced way, i.e., the exposure ratio is similar for both articles within a group (within bounds $\underline{\delta} < 1 < \overline{\delta}$).

From Lemma 1, we know that, given $t \in \{1, \ldots, T\}$, $l_{g,s}(t,\theta)$ is a linear and strictly increasing function in $\theta_{g,s}$, $\theta_{g',s}$. Thus, the objective function of (P) is linear in two dimensions; similarly, the exposure constraints can also be transformed to linear inequalities. Consequently, (P) is a linear program.

Fairness-Agnostic Optimization Problem As a natural benchmark, we first consider the optimization problem (P) *without* exposure constraints (C1) and (C2), while retaining the constraint $\theta_{g,s} \in [0,1]$ for all g, s. We refer to this as the *fairness-agnostic* problem. We show that the exclusion of fairness constraints *always* results in all members of the same group being shown the same article by the platform at time $t = 1$. Specifically, the solution to the fairness-agnostic exposure problem is given in the following proposition.

Proposition 1. *The solution to the fairness-agnostic optimization problem is*

$$\theta^*_{A,a} = \mathbf{1}\Big\{\sum_{t=1}^{T}\left(w_{A,a}(t) - w_{A,b}(t) + u_{B,a}(t) - u_{B,b}(t)\right) > 0\Big\},$$

$$\theta^*_{B,a} = \mathbf{1}\Big\{\sum_{t=1}^{T}\left(w_{B,a}(t) - w_{B,b}(t) + u_{A,a}(t) - u_{A,b}(t)\right) > 0\Big\}.$$

From a theoretical perspective, this result follows from the linearity of (P). From a practical perspective, Proposition 1 suggests that targeting a group with their preferred article is not necessarily optimal for maximizing user engagement. Albeit counter-intuitive, it might be optimal for the platform to ignore group preferences and target the whole user network with a single article. Two additional implications of Proposition 1 are given in Corollary 1 and Lemma 2 below.

Corollary 1. *The feasible solution $\theta_{A,b} = 1$, $\theta_{B,a} = 1$ is never optimal for* (P).

Lemma 2. *Assume $\theta_{A,a} = 1$, $\theta_{B,b} = 1$. If*

$$\frac{q_A\pi_A}{(1-q_B)\pi_B} < 1, \quad \frac{\psi_{A,a}\psi_{B,a}}{\pi_B\psi_{A,b}\psi_{B,b}} < 1, \tag{8}$$

then group A is exposed more to article b than a over time, i.e., $e_{A,b}(T) = \frac{l_{A,b}(T,\theta)}{\pi_A} > e_{A,a}(T) = \frac{l_{A,a}(T,\theta)}{\pi_A}$ for any $T > 2$.

Under homophily, one might expect a group to be preferentially exposed to in-group articles. However, as Lemma 2 shows, this may not be the case if group sizes are radically different or if one group displays much lower levels of homophily than the other. Lemma 2 can shed light on several counter-intuitive possibilities for article exposure over time. More specifically, it suggests that, due to the network structure and the dynamics of propagation, targeting each group with their preferred article might *not always* bring

the intended targeting and thus potentially lead to suboptimal outcomes for the platform. Even if the platform targets each group only with their preferred (in-group) article, one group may—after several rounds—be exposed to their non-preferred (out-group) article. For example, given a significantly larger group B, weak homophily for both groups ($q_A \simeq q_B \simeq 1/2$) and similar preferences for compatible articles ($\psi_{A,a} \simeq \psi_{B,b}$), group A is exposed more to article b even if $\theta_{A,a} = 1$ and $\theta_{B,b} = 1$. A similar property holds when there is an extreme preference for article b in group B compared to moderate preference in group A, i.e., $\psi_{B,b} \gg \psi_{A,a}$.

In contrast to Lemma 2, Corollary 1 offers a quite intuitive insight, showing that the opposite strategy (i.e., targeting both groups with their out-group article) is *never* optimal. Indeed, depending on the model parameters, either the network will eventually favor the article with the largest sharing rate in total or the users in each group will start clicking more on their in-group article. In both cases, the platform's initial targeting $\theta_{A,b} = 1$, $\theta_{B,a} = 1$ would only manage to delay any of these events thus leading to a suboptimal number of clicks and likes at the initial stages of propagation.

Fairness-Aware Optimization Problem: Constant Fair Exposure In this section, we explore the feasibility of a natural but stricter fairness notion, i.e., constant fair exposure, as defined in (2). As detailed in Lemma 3 below, we show that it is generally not possible to achieve equal and constant exposure at every time step unless certain restrictive conditions hold.

Lemma 3. *Let $e \in (0,1)$ be the platform's targeted fair exposure level. Achieving constant fair exposure is possible if and only if for both $s \in \{a,b\}$,*

$$\begin{aligned} &\psi_{A,s}\left(q_A + \frac{1-\pi_A}{\pi_A}(1-q_B)\right) \\ &= \psi_{B,s'}\left(\frac{1-\pi_A}{\pi_A}q_B + (1-q_A)\right) \end{aligned} \tag{9}$$

and the platform sets $\theta_{A,a} = 1 - \theta_{B,a} = e$ at time $t = 1$.

The conditions of Lemma 3 guarantee that the mass of users clicking on a given article will be the same across all groups and time steps. However, this will almost certainly never occur in practice due to differing preferences in content across groups. Therefore, we ask if average exposure over time can be equalized across groups, i.e., if $\frac{1}{T}\sum_{t=1}^{T}\frac{l_{A,s}(t,\theta)}{\pi_A} = \frac{1}{T}\sum_{t=1}^{T}\frac{l_{B,s'}(t,\theta)}{\pi_B} = e$ is possible. Lemma 4 shows that it is very difficult to achieve any desired average exposure rate:

Lemma 4. *For any $\pi_g \in (0,1)$, average exposure levels for group g to article s are achievable only in the range $0 \leq e \leq \frac{1}{T\pi_g}\sum_{t=1}^{T}(w_{g,s}(t) + u_{g,s}(t))$.*

Fairness-Aware Optimization Problem: Approximately Fair Average Exposure Given the restrictive nature of constant fair exposure, we turn to a relaxed notion. Specifically, we explore the feasibility of the optimization problem (P) *with* fairness constraints (C1) and (C2), and analytically describe the solution by deriving expressions for the extreme points of the constraint polytope. Let

$$\begin{aligned} \overline{m}_{g,s} &\coloneqq \sum_{t=1}^{T} u_{g,s}(t) + \overline{\delta}\sum_{t=1}^{T} w_{g',s'}(t), \\ \underline{m}_{g,s} &\coloneqq \sum_{t=1}^{T} u_{g,s}(t) + \underline{\delta}\sum_{t=1}^{T} w_{g',s'}(t), \\ \overline{n}_{g,s} &\coloneqq \sum_{t=1}^{T} w_{g,s}(t) + \overline{\delta}\sum_{t=1}^{T} u_{g',s'}(t), \\ \underline{n}_{g,s} &\coloneqq \sum_{t=1}^{T} w_{g,s}(t) + \underline{\delta}\sum_{t=1}^{T} u_{g',s'}(t), \\ m_{g,s} &\coloneqq \sum_{t=1}^{T} u_{g,s}(t) + \sum_{t=1}^{T} w_{g,s}(t). \end{aligned}$$

From constraints (C1) and (C2) and using Theorem 1, we can infer the feasible bounds on $\theta_{B,a}$ (dependent on $\theta_{A,a}$), in addition to $\theta_{A,a}, \theta_{B,a} \in [0,1]$. We state these bounds as well as the axes intersects of the hyperplanes that induce the half-spaces containing the feasible region in § D. Evaluating the relative positions of these hyperplanes, we can then infer when the fairness-aware optimization problem is infeasible:

Theorem 2. *The fairness-aware optimization problem is infeasible if and only if one of the following holds:*

$$\begin{aligned} &\frac{\underline{\delta}m_{B,b}}{\underline{m}_{A,a}} > \frac{m_{A,b}}{\underline{m}_{A,b}} \quad \textit{and} \quad \frac{\underline{\delta}m_{B,b}}{\underline{n}_{A,a}} > \frac{m_{A,b}}{\underline{n}_{A,b}}; \\ &\frac{m_{A,b}}{\overline{m}_{A,b}} > \frac{\overline{\delta}m_{B,b}}{\overline{m}_{A,a}} \quad \textit{and} \quad \frac{m_{A,b}}{\overline{n}_{A,b}} > \frac{\overline{\delta}m_{B,b}}{\overline{n}_{A,a}}; \\ &\underline{\delta}\sum_{t=1}^{T} w_{B,b}(t) > \sum_{t=1}^{T} w_{A,a}(t) \quad \textit{and} \quad \frac{\underline{\delta}m_{B,b}}{\underline{m}_{A,a}} > \frac{\underline{\delta}m_{B,b}}{\underline{\delta}m_{B,b} - \underline{n}_{A,a}}; \\ &\sum_{t=1}^{T} u_{A,b}(t) > \overline{\delta}\sum_{t=1}^{T} u_{B,a}(t) \quad \textit{and} \quad \frac{m_{A,b}}{\overline{m}_{A,b}} > \frac{m_{A,b}}{m_{A,b} - \overline{n}_{A,b}}. \end{aligned}$$

It follows that we can always make the problem feasible by setting $\underline{\delta}$ and $\overline{\delta}$ accordingly. As noted in § D, letting $\underline{\delta} \to 0$ and $\overline{\delta} \to \infty$, the fairness-agnostic problem is recovered. By the intermediate value theorem, there exist infinitely many values of $\underline{\delta}, \overline{\delta}$ that define a non-empty feasible region strictly contained in the unit box. Otherwise, if the problem is feasible, the fundamental theorem of linear programming states that an optimal solution will occur at a corner point of the feasible region, or on a line segment between two corner points. Theorem 3 (deferred to § C) states the collection of possible solutions $\theta^i_{g,s}$ to the *fairness-aware* optimization problem. In particular, note that all of these solutions may not be feasible for a particular problem instance. Which of these solutions is feasible and optimal will depend on the true problem parameters. In particular, define $c_{g,s} \coloneqq \sum_{t=1}^{T}(w_{g,s}(t) - w_{g,s'}(t) + u_{g',s}(t) - u_{g',s'}(t))$ and write the objective as $\theta_{A,a}c_{A,a} + \theta_{B,a}c_{B,a}$. As in Proposition 1, the particular solution then depends on the signs and relative magnitudes of $c_{A,a}$ and $c_{B,a}$. For example, if $c_{A,a} \gg c_{B,a} > 0$, then the largest feasible value of $\theta^i_{A,a}$ and the corresponding $\theta^i_{B,a}$ will be the optimal solution; see Figure 5 in § D for an illustration.

The main difference to Proposition 1 is that, due to the imposed fairness constraints, some of the optimal unconstrained solutions might be out of the feasible region. At a higher level, the more restrictive the bounds $\underline{\delta}, \overline{\delta}$ get, the further we move from the optimal binary solution of the fairness-agnostic problem. Thus, some solutions correspond to a mixture of articles shown to each group, and no group is targeted with one article type. However, others may correspond to cases where exactly one group is targeted with only one article, while the other sees both articles at unequal rates. Observe that it is still possible that a group is only shown their out-group article.

Our results offer novel insights for platform design. Even though satisfying (C1) and (C2) imposes a significant restriction on the platform and ostensibly seems to ensure a balanced exposure up to some extent, extreme solutions may still arise. Introducing fairness constraints does not automatically imply that the final outcome is *truly* fair—or even balanced. Furthermore, in any solution i where $\theta^i_{g,s} \in \{0, 1\}$ while $0 < \theta^i_{g',s} < 1$, only one group incurs the "price of fairness" whereas the other group, which is targeted with only one article, serves the platform's major goal of maximizing clicks. (Note that one can verify that half of the solutions in Theorem 3 have this property.) Thus, when the content is related to sensitive or high-stakes procedures (e.g., a referendum), ensuring fair exposure is not just a technical challenge; if the interventions are not carefully designed (e.g., choosing $\underline{\delta}, \overline{\delta}$ thoughtfully), they can lead to unintended outcomes, potentially with legal consequences.

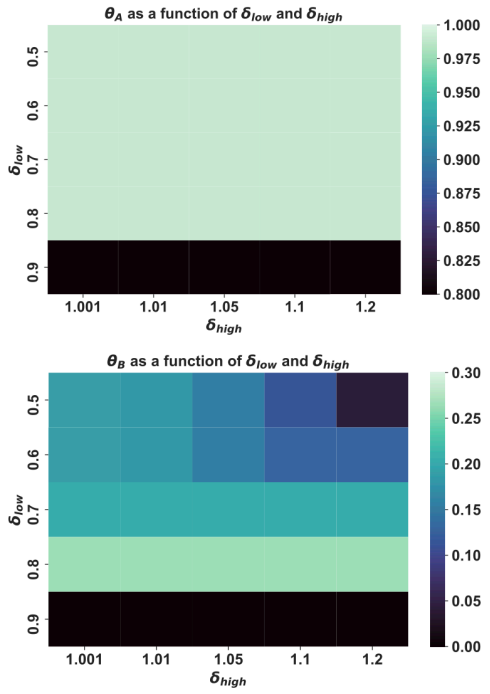

Figure 2: Calculating $\theta_{A,a}$ (top) and $\theta_{B,a}$ (bottom) as a function of $\underline{\delta}$ and $\overline{\delta}$ with parameters estimated from Bakshy, Messing, and Adamic (2015). Black cells at the bottom indicate no feasible solution to the fairness-constrained problem.

5 Simulations

We use our model to empirically study the effects of different model parameters from real-world click data. Tables 2 and 3 in § E describe the parameters used, such as number of runs, proportional representation of each group, among others. We use maximum likelihood estimation to fit parameter values from Bakshy, Messing, and Adamic (2015) in this section, and study three datasets from Garimella et al. (2017) deferred to § F. For p, we fit a beta distribution and present the parameters α and β in the appendix. See § E for additional experiments evaluating the effects of population-based parameters.

Effect of Fairness Bounds $\underline{\delta}$ and $\overline{\delta}$ on θ We start by studying the effect of different model parameters on the platform's optimization and outcomes. In particular, we focus on the change of $\underline{\delta}$ and $\overline{\delta}$, and its impacts on exposure and click rates, both en masse and across groups. Figure 2 illustrates that the optimal solution is to almost always show article a to members of group A, and fair exposure is then enforced by restricting how group B is shown articles. In this case, setting $\overline{\delta}$ closer to 1 (making the constraint more restrictive) generally increases the proportion of members of group B who are shown article a. It is helpful to understand when the fairness-aware problem is (i) feasible and (ii) restrictive; if $\underline{\delta}$ and $\overline{\delta}$ are too close to 1, the feasible region may be empty (as in the bottom row of Figure 2), but if they are too far from 1, they may not constrain the fairness-agnostic problem around the agnostic optimum. For intuition on how these parameters may affect the feasible region, see Figure 5 in § D.

Exposure Disparity We are also interested in understanding how imposing balanced exposure constraints might affect disparity in expected exposure and clicks. Figure 6 in § E highlights the disparity in exposure for different optimization policies θ. There, we observe that a uniformly randomized policy (*random*) and proportional policy (*proportional*; $\theta_{g,s} = \pi_g$) yield a large disparity in article exposure between article a and article b, while this disparity is lower in the fairness-agnostic (*unconstrained*) and fairness-aware (*fair*) settings, though there is no significant difference between the two. When evaluating differences in how often the articles get liked, this gap closes across all four policies.

Engagement Disparity Perhaps unsurprisingly, we can see in Figure 3 (left) that intergroup exposure is significantly higher when randomizing exposure than when optimizing exposure as in the fairness-agnostic (*unconstrained*) and fairness-aware (*fair*) settings. When evaluating the number of *likes* across groups in Figure 3 (middle), this arises as an artifact of the model more generally, though the gap significantly decreases. Optimizing in fairness-aware and -agnostic settings yield relatively similar distributions of intergroup likes on articles.

Price of Fairness We consider the price of fairness similar to that of Bertsimas, Farias, and Trichakis (2011), given

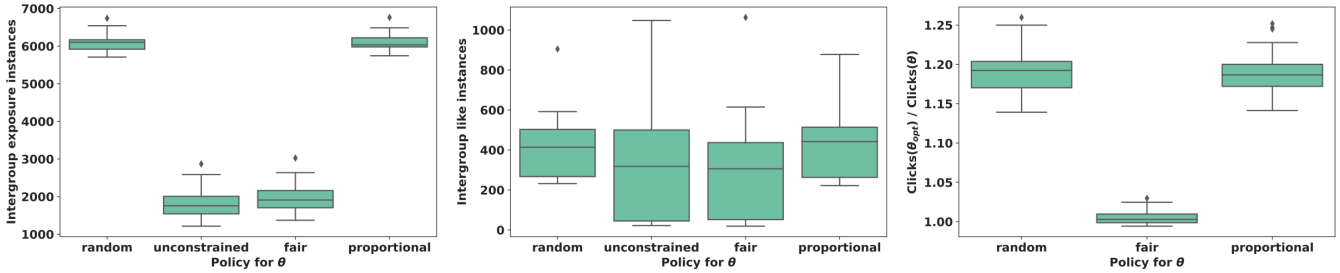

Figure 3: Intergroup exposure (left) and liking (middle), as well as the price of fairness (right), using model parameters from Bakshy, Messing, and Adamic (2015).

in (10). Here, a lower price of fairness for a given policy is better, as it indicates being closer to the fairness-agnostic optimization problem.

$$POF(\theta) = \frac{\#clicks(\theta_{opt})}{\#clicks(\theta)} \tag{10}$$

We can see in Figure 3 (right) that the price of fairness for the fairness-aware optimization problem is close to 1 in most trials, which is observationally lower than the price of fairness for a uniformly randomized or proportional policy. This suggests that our fairness-aware optimization problem yields approximately the same number of clicks as the fairness-agnostic solution. Figures 9, 11, and 13 in § F show the price of fairness for adding constraints compared to a uniformly random policy using the parameters estimated from Garimella et al. (2017).

6 Discussion and Conclusion

Motivated by the concerning increase in polarization in social media platforms, this paper introduces the fair exposure problem and develops a theoretical dynamic model to study its implications. Albeit simple and intuitive, our model is highly stylized (as other models in the literature (Papanastasiou 2020; Allon, Drakopoulos, and Manshadi 2021)). One simplification is the propagation scheme which aims at approximating article sharing and user exposure in a computationally tractable way. Thus, our framework offers novel insights about the propagation in expectation across groups (instead of propagation from individual to individual). Nevertheless, a theoretical analysis using an underlying graph structure would be a natural extension; we study this more realistic scenario through simulations in § G. Another assumption of our model is that each user can see only one article. We make this modeling choice merely for technical simplicity that offers tractability and clearer insights. However, a partial interpretation of this assumption would be that the platform has limited slots for promoted content or that users most likely click on the first article they see (Robertson and Belkin 1978; Wang et al. 2013; Craswell et al. 2008).

As the Fairness Doctrine was introduced to ensure opinion diversity, a modern version of this policy could be similarly introduced in online platforms (Pickard 2021). Although both the Fairness Doctrine and our model expose users of different groups to a diverse set of news articles at the same time, our model might expose each individual to only one article. Such platform interventions have the great potential to ensure diversity of viewpoints; however, the design of such policies entails the careful examination of any ethical concerns. A question that naturally arises is whether it is ethical for the platform to algorithmically control and potentially randomize the content that a user sees and, ultimately, who—if anyone—has the responsibility to ensure fair exposure in online spaces. This question has been under close scrutiny in interpreting Section 230 of the Communications Decency Act in the United States (Chintalapoodi 2021). For example, given that news sharing and discussions in social media can determine important political outcomes and thus the passive or more restrictive role that the platform chooses to undertake matters (see, e.g., British Broadcasting Corporation (BBC) (2020); Isaac and Frenkel (2020)), it is unclear how a fair representation of content should be defined. Thus, we acknowledge that balancing exposure to different ideologies of content might not actually be fair in a given context. For instance, Bail et al. (2018) suggest that showing people opposing viewpoints makes them more polarized, whereas Becker, Porter, and Centola (2019) show that echo chambers do not necessarily increase polarization. Furthermore, considering the amount of disinformation and the technical challenges in identifying problematic content (e.g., fake news, hate speech) in platforms, the fair exposure constraints should not be applied to all content. Implementing fair exposure can thus become particularly challenging, and more interdisciplinary research is needed to understand where to draw the boundary. Our work is an initial step towards this broader goal.

Finally, our framework highlights how the introduction of fairness constraints can only partially mitigate group-homogeneous targeting and points to problematic outcomes, as sometimes only one group incurs the "price of fairness" while the other pays the "cost of user engagement." It also gives rise to a series of emerging, challenging directions for future research related to platforms and algorithmic fairness. These include the study of fair exposure notions, the design of dynamic interventions and more sophisticated targeting, ad pricing and revenue maximization under fair exposure constraints, and their implications on the competition among different platforms.

Acknowledgements

This project has been part of the MD4SG working group on Bias, Discrimination, and Fairness. The material is based on work supported by the National Science Foundation under Graduate Research Fellowship No. DGE-1650115 and National Science Foundation Award No. 2202898 (Jessie Finocchiaro). Keziah Naggita was supported in part by the National Science Foundation under Grant No. CCF-1815011 and by the Simons Foundation under the Simons Collaboration on the Theory of Algorithmic Fairness. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding sources.

References

Ali, J.; Babaei, M.; Chakraborty, A.; Mirzasoleiman, B.; Gummadi, K. P.; and Singla, A. 2019. On the fairness of time-critical influence maximization in social networks. *arXiv preprint arXiv:1905.06618*.

Allon, G.; Drakopoulos, K.; and Manshadi, V. 2021. Information inundation on platforms and implications. *Operations Research*, 69(6): 1784–1804.

Arunachaleswaran, E. R.; Kannan, S.; Roth, A.; and Ziani, J. 2021. Pipeline interventions. In *12th Innovations in Theoretical Computer Science Conference*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Ashford, N. A. 2021. Not on Facebook? You're still likely being fed misinformation. The New York Times, https://www.nytimes.com/2021/03/29/opinion/misinformation-television-radio.html. Accessed: 2023-03-07.

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. B. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239): 1130–1132.

Banerjee, A. V. 1992. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3): 797–817.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10): 1531–1542.

Becker, J.; Porter, E.; and Centola, D. 2019. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, 116(22): 10717–10722.

Bertsimas, D.; Farias, V. F.; and Trichakis, N. 2011. The price of fairness. *Operations Research*, 59(1): 17–31.

Bikhchandani, S.; Hirshleifer, D.; and Welch, I. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5): 992–1026.

Boxell, L.; Gentzkow, M.; and Shapiro, J. M. 2017. Is the internet causing political polarization? Evidence from demographics. Technical report, National Bureau of Economic Research.

British Broadcasting Corporation (BBC). 2020. Facebook ad campaign helped Donald Trump win election, claims executive. BBC News, https://www.bbc.com/news/technology-51034641. Accessed: 2023-03-07.

Candogan, O.; and Drakopoulos, K. 2020. Optimal signaling of content accuracy: Engagement vs. misinformation. *Operations Research*, 68(2): 497–515.

Celis, L. E.; Kapoor, S.; Salehi, F.; and Vishnoi, N. 2019. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 160–169.

Chen, Q.; Golrezaei, N.; Susan, F.; and Baskoro, E. 2022. Fair assortment planning. *arXiv preprint arXiv:2208.07341*.

Chintalapoodi, P. 2021. Understanding the controversy over Section 230. Chip Law Group, https://www.lexology.com/library/detail.aspx?g=38784375-b9ba-4c31-8f06-5c30c37534dd. Accessed: 2023-03-07.

Cisternas, G.; and Vásquez, J. 2020. Fake news in social media: A supply and demand approach. *Available at SSRN 3698788*.

Clauset, A. 2021. Random graph models. https://aaronclauset.github.io/courses/5352/. Accessed: 2023-03-07.

Craswell, N.; Zoeter, O.; Taylor, M.; and Ramsey, B. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 87–94.

Dubois, E.; and Blank, G. 2018. The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, 21(5): 729–745.

Farajtabar, M.; Ye, X.; Harati, S.; Song, L.; and Zha, H. 2016. Multistage campaigning in social networks. In *Advances in Neural Information Processing Systems*, volume 29, 4725–4733.

Fish, B.; Bashardoust, A.; Boyd, D.; Friedler, S.; Scheidegger, C.; and Venkatasubramanian, S. 2019. Gaps in information access in social networks? In *The World Wide Web Conference*, 480–490.

Freedman, J. L.; and Sears, D. O. 1965. Selective exposure. In *Advances in Experimental Social Psychology*, volume 2, 57–97. Elsevier.

Garimella, K.; Gionis, A.; Parotsidis, N.; and Tatti, N. 2017. Balancing information exposure in social networks. In *Advances in Neural Information Processing Systems*, volume 30, 4663–4671.

Garrett, R. K.; and Resnick, P. 2011. Resisting political fragmentation on the Internet. *Daedalus*, 140(4): 108–120.

Gillani, N.; Yuan, A.; Saveski, M.; Vosoughi, S.; and Roy, D. 2018. Me, my echo chamber, and I: Introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, 823–831.

Hawdon, J.; Ranganathan, S.; Bookhultz, S.; and Mitra, T. 2020. Social media use, political polarization, and social capital: Is social media tearing the US apart? In *International Conference on Human-Computer Interaction*, 243–260. Springer.

Hosseinmardi, H.; Ghasemian, A.; Clauset, A.; Rothschild, D. M.; Mobius, M.; and Watts, D. J. 2020. Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube. *arXiv preprint arXiv:2011.12843*.

Hu, C.; Zhang, C.; Wang, T.; and Li, Q. 2012. An adaptive recommendation system in social media. *2012 45th Hawaii International Conference on System Sciences*, 1759–1767.

Isaac, M.; and Frenkel, S. 2020. Facebook braces itself for Trump to cast doubt on election results. The New York Times, https://www.nytimes.com/2020/08/21/technology/facebook-trump-election.html. Accessed: 2023-03-07.

Jeon, Y.; Kim, B.; Xiong, A.; Lee, D.; and Han, K. 2021. ChamberBreaker: Mitigating the echo chamber effect and supporting information hygiene through a gamified inoculation system. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–26.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137–146.

Lazarsfeld, P. F.; and Merton, R. K. 1948. *Mass communication, popular taste and organized social action*. Bobbs-Merrill, College Division.

Liu, D. M.; Shafi, Z.; Fleisher, W.; Eliassi-Rad, T.; and Alfeld, S. 2021. RAWLSNET: Altering Bayesian networks to encode Rawlsian fair equality of opportunity. *Available at SSRN 3816196*.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1): 415–444.

Mukerjee, S.; Jaidka, K.; and Lelkes, Y. 2020. The political landscape of the U.S. Twitterverse. *OSF Preprints*.

Papanastasiou, Y. 2020. Fake news propagation and detection: A sequential model. *Management Science*, 66(5): 1826–1846.

Pickard, V. 2021. The Fairness Doctrine won't solve our problems – but it can foster needed debate. The Washington Post, https://www.washingtonpost.com/outlook/2021/02/04/fairness-doctrine-wont-solve-our-problems-it-can-foster-needed-debate/. Accessed: 2023-03-07.

Robertson, S. E.; and Belkin, N. J. 1978. Ranking in principle. *Journal of Documentation*.

Starbird, K.; Arif, A.; Wilson, T.; Van Koevering, K.; Yefimova, K.; and Scarnecchia, D. 2018. Ecosystem or echosystem? Exploring content sharing across alternative media domains. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Stoica, A.-A.; Han, J. X.; and Chaintreau, A. 2020. Seeding network influence in biased networks and the benefits of diversity. In *Proceedings of The Web Conference 2020*, 2089–2098.

US Department of Justice. 2021. Thirteen charged in federal court following riot at the United States Capitol. https://www.justice.gov/opa/pr/thirteen-charged-federal-court-following-riot-united-states-capitol. Accessed: 2023-03-07.

Wang, C.; Liu, Y.; Zhang, M.; Ma, S.; Zheng, M.; Qian, J.; and Zhang, K. 2013. Incorporating vertical results into search click models. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 503–512.

Yarchi, M.; Baden, C.; and Kligler-Vilenchik, N. 2020. Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 38: 98 – 139.

Zoetekouw, K. 2019. A critical analysis of the negative consequences caused by recommender systems used on social media platforms. University of Twente, http://essay.utwente.nl/78500/. Accessed: 2023-03-07.