Submitted to *Management Science*
manuscript ( )

On the Mediating Role of Two-Sided Platforms

Abhishek Ray
Donald G. Costello College of Business, George Mason University, aray8@gmu.edu

Cesar Martinelli
Department of Economics, George Mason University, cmarti33@gmu.edu

Are welfare motives compatible with profit maximization objectives? We explore this fundamental question in light of two-sided platforms and growing affective polarization. Multi-sided platforms increasingly need more efficient and equitable participation with rising affective polarization. However, platforms need engagement, polarized or otherwise, for sustained shareholder value generation. Thus, we explore the compatibility of welfare motives with profit maximization in the context of two-sided platforms. We propose a novel framework – Dolbear Axelrod Model (DAM) – that facilitates analyzing the balance between profit-maximizing platforms and welfare-maximizing planners in addressing participation trade-offs. Our findings indicate that while platforms attempt to engage users across the political spectrum, their strategies can often backfire, increasing polarization and decreasing user welfare. We explore various modeling approaches, including incentive-based mechanisms (algorithmic tweaks akin to subsidies and taxes) and control mechanisms (e.g., participation restrictions and interaction obfuscation), revealing that neither approach perfectly mediates polarization. Our analysis highlights platforms' challenges in reconciling profit motives with the need for equitable participation, suggesting that user-driven negotiations may yield more favorable outcomes than platform-driven regulations. Ultimately, we call for reevaluating platform policies to address the complexities of affective polarization better and promote healthier digital interactions.

Key words: game theory, polarization, platforms, externalities

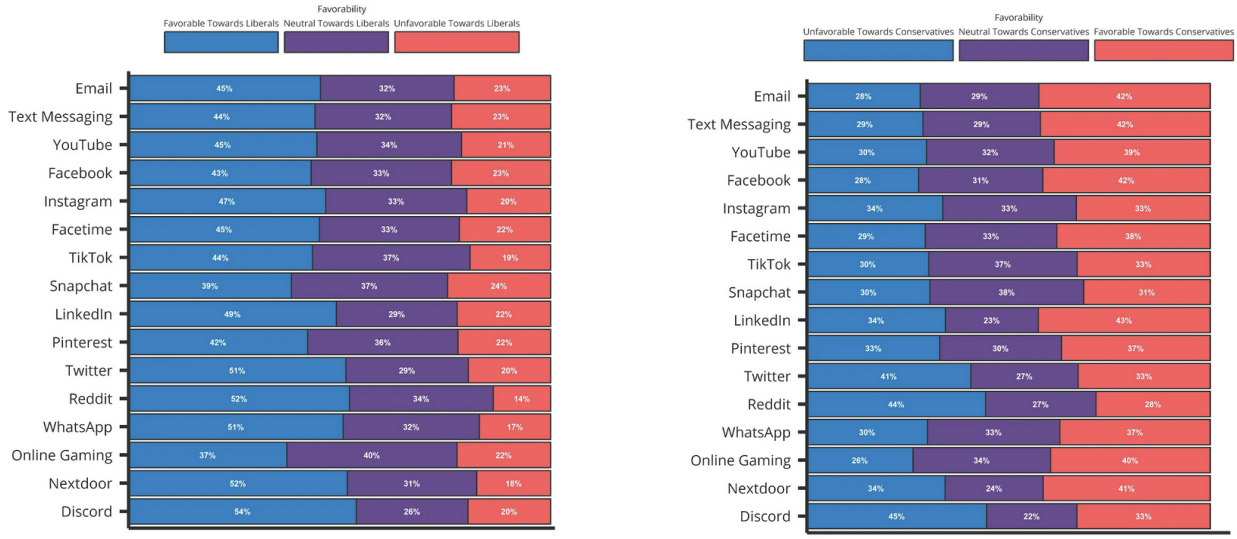
1. Introduction

How can digital platforms address the rise in affective polarization? We use two-sided platform theory to study this fundamental question, focusing on the need for more efficient and equitable platform participation (Baqir et al. 2025). Social media users are content consumers and producers, so affective polarization and 'us vs. them' attitudes aggravate partisan content consumption and exacerbate participation inequality and inefficiency. Empirical research suggests that fears of a broader 'sectarianism' fueled by affective polarization have replaced concerns due to socio-economic polarization (Garzia et al. 2023). Further, social identity-driven prominent political positions (e.g., conservative vs. liberal) promote 'othering,' hatred, suspicion of the opponent, and moral superiority for one's side (Schedler 2023). As political campaigns and discussions proliferate online, digital platforms have faced heated polarized engagement.

Notably, polarization has led to substantial numbers of users of one party leaving major (e.g., X, Facebook) platforms (Nazarro 2024, Nossel 2025). While polarized partisan groups differ in sizes widely across

platforms (Figure 1), calls for these platforms to address affective polarization have increased in recent years (Rathje et al. 2024). Policy researchers (e.g., Brookings) have noted that platforms can improve inter-group engagement, prioritize agreeable information across user groups, develop better user interfaces, and correct misconceptions to reduce friction (Whitaker 2022). Still, in polarized populations, these initiatives rarely reach the majority. For example, Facebook tested the news-feed filters to show rival partisan posts. Yet, only polarization-agnostic users benefited from these filters (Lydia 2020), while others remained in their partisan bubbles. Thus, typical solutions have enjoyed limited to no success in enhancing equitable participation.

Even so, platforms have attempted to woo users from across the political spectrum. For instance, Substack Notes and GETTR have made 'aggressive outreach efforts to Democrats' by hiring campaign staff and reaching out to left-leaning influencers (Cheadle 2021, Krishan 2021a). Further, Parler, GETTR, and MeWe have intentionally rebranded themselves as mainstream and objective, not conservative echo chambers (Krishan 2021b, Ingram and Brooks 2023, Kaufhold 2024). Still, conservative content performs marginally better on Facebook, making it appealing to both sides; X has historically been liberal-dominant but has become increasingly conservative under Elon Musk's ownership; YouTube and TikTok see increased use from both liberals and conservatives, but younger liberals use them more (González-Bailón et al. 2022, Vogels et al. 2024). Moreover, newcomers like BlueSky, Damus, and RTRO offer a balanced, non-partisan experience to keep all sides interested (Chillingworth 2025).



(a) Across platforms favorability towards liberals

(b) Across platforms favorability towards conservatives

Figure 1 Partisan favorability on major social media platforms (Motyl 2023, de Zúñiga et al. 2025)

But, evidence suggests that creating balanced political spaces to attract users and maximize profitability has backfired. For example, participation on X can lead to decreased well-being, heightened political division, outrage, and a diminished sense of belonging (Oldemburgo de Mello et al. 2024), even after controlling for demographics and personality. Hyperlocal platforms like Nextdoor are losing members due to

their reputation for being 'woke' and promoting 'cancel culture' by eliminating political posts (Alexander 2022, Stephens 2024). More broadly, recent online dating platform surveys show that 60.5% of respondents say a partner's different political views are a dealbreaker (Anderson 2024). Payment platforms (e.g., Paypal, Venmo) have seen dwindling participation after being accused of banning conservative-leaning users citing reasons such as 'hate speech' (Parke 2019, Samuelson 2024). Thus, participation incentives and profit maximization seem irreconcilable given polarization frictions.

We offer a foundational approach to this seemingly irreconcilable problem. Specifically, we propose a novel modeling framework, the Dolbear Axelrod Model (DAM), which combines the Dolbear Triangle (Dolbear 1967) with the Axelrod attraction-repulsion notions (Axelrod et al. 2021). DAM relaxes several crucial assumptions in previous platform models (see Section 2, Table 2) and considers the platform an intermediary between two affectively polarized user populations. Affective polarization is modeled as a user type, an identity-driven negative externality (West and Iyengar 2022, Versteegen 2024) that interacts with the profit-maximizing platform's pricing strategies or the welfare-maximizing social planner's optimal subsidies or taxes (Baumol 1972, Nault 1996). Additionally, DAM enables analysis of participation controlling strategies (e.g., banning) that the platform can use to tackle polarization. Finally, DAM helps consider a *quasi*-Coasian scenario where polarized groups are left to negotiate away their polarization-driven externality without platforms or social planners.

Analysis of these scenarios reveals several important insights (see Table 1) outlined along various dimensions of platform characteristics. We structure the analysis along four modeling approaches. The first is the benchmark, where the platform sets access fees to maximize payoff, and users decide on participation. The second approach analyzes whether subsidies or taxes reduce affective polarization-driven externalities. The platform offers blanket or type-driven (Pigouvian) taxes and subsidies, adjusting prices to maximize payoff while incentivizing users to participate. The third approach considers the social planner setting *optimal* subsidies and taxes, after which the platform sets access fees, and users decide to join. Finally, users tackle affective polarization in the Coasian scenario without a platform or social planner while splitting a common valuation. Table 1 provides insights for the first approach.

Considering the social planner, our analysis reveals a dilemma. Specifically, a *benevolent* (maximizes overall welfare) social planner setting optimal subsidies or taxes that mitigate polarization can only set the minimum subsidies or the maximum taxes such that the overall welfare (i.e., platform's payoff and user surplus) is maximized. Our results hold even when considering a *discriminating* social planner that aims to maximize user welfare while minimizing platform payoff. Finally, in a Coasian setting where two polarized groups negotiate the split of a shared resource without platforms or social planners, we show that affective polarization externality favors the less polarized population. Thus, our analysis presents a mixed picture: two-sided platforms are ill-equipped to handle polarized participation since participation pricing can capture the surplus from subsidy/tax interventions. Moreover, social planner-driven regulations threaten

Table 1 Overall insights using DAM Framework

Category	*Benchmark Model Insight*	*Incentive-based Mechanisms*	*Control-based Mechanisms*
Participation Size	Despite being more polarized, either side may have greater participation due to out-group averse participants.	Subsidies increase participation whereas taxes decrease it. User type-based (Pigouvian) subsidy may maximize participation.	Restricted participation limits participation size. All others do not affect participation.
Participation Diversity	Participants are mostly in-group aversion leaning.	Only Pigouvian tax-based mechanism allows more out-group averse participants and diversify participation more than benchmark.	None of the control-based mechanisms has any impact on improving diversity.
Pricing Asymmetry	Even when more polarized, the platform may charge higher access fee to that side.	Pricing asymmetry persists even under subsidy or tax regimes. Compared to benchmark, the platform is guaranteed to charge lower access fee to either side.	Pricing asymmetry persists under various control mechanisms. Compared to benchmark, platform may charge one or both sides lower.
Same-side Demand Elasticity	Same side demand elasticities are always negative.	Except blanket subsidies, same side demand elasticities are negative for all other incentive-based mechanisms.	Except valuation manipulation, all control-based procedures offer negative or positive same- side elasticities.
Cross-platform Demand Elasticity	Cross-platform demand elasticities are always positive.	Except blanket subsidies, cross-platform demand elasticities are positive for all other incentive-based mechanisms.	Restricted participation and interaction obfuscation can offer negative cross-platform elasticity. Filtering have no cross-platform externality.

'extremes' at the expense of the optimal, but user-driven negotiations can yield the a biased solution since certain parties may benefit from polarization.

For the remainder of the paper, Section 2 explores the literature and highlights the paper's contribution, including academic and policy-specific contributions given recent research. Then, in Section 3, we provide an overview of DAM through the framework preliminaries. Section 4 puts the analysis into perspective using derived insights and relevant empirical/anecdotal evidence to establish relevant intuition regarding the results. Section 5 explores three relevant extensions to DAM - one considering competition among two platforms (Subsection 5.4), then considering the platform as a profit and welfare maximizer (Subsection 5.2), and finally, considering a Rawlsian social planner (Subsection 5.3). Further, Section 6 dives deeper into the policy implications of our analysis, with special emphasis on Section 230 and the platform's responsibilities towards participants. Finally, in Section 7, we conclude with limitations and future planned work based on our analysis.

2. Literature Review

In this section, we provide a pointed and objective overview of platform literature, and where this work can be situated. We use the following Table 2 for this overview.

In addition to Table 2, we now compare and contrast recent works spanning economics, information systems, and political science that have attempted to capture negative externalities from platform participation. Recent economic and political science research has improved our understanding of how political biases affect economic results through novel models. Studies spanning 2023-2025 quantified how partisan forces affect economic decision-making. Jelveh et al. (2024) created a novel "political slant" index using machine learning to show that economists' ideological orientations change policy estimates by 12-15%, with right-leaning economists recommending 60% lower tax rates than left-leaning economists (77%). The Federal Reserve impression survey found that 66% of Democrat-leaning participants thought the Fed

Table 2 Two-Sided Platform features comparison with the DAM Framework features

Two Side Platform Feature	Established Insights	Related Literature	DAM Feature
Direct within group network effects	Positive externality due to presence of more participants from same side	Rochet and Tirole (2003), Armstrong (2006)	Non-uniform positive externality due to user-specific affective polarization tendencies
Cross group network effects	Positive externality due to presence of more participants from other side	Belleflamme and Peitz (2018, 2021)	Non-uniform *negative* externality due to user-specific affective polarization tendencies
Uniform platform valuation	Each side has a uniform platform valuation, regardless of type	Peitz (2024)	Platform valuation depends on how affectively polarized users are, i.e., type
Consumers can consider outside option	Consumers coordinating on outside option incentivizes the platform to design asymmetric pricing structure	Belleflamme and Peitz (2019), Peitz (2024)	No outside option - consumers are assumed to coordinate on participation equilibrium
		Rysman (2009), McIntyre and Srinivasan (2017)	No coordination allows platform charging access fee only, without incorporating transaction fee
Explicit and easily distinguishable group of users is always required	Platform adjusts prices to internalize network effects even when user groups are not easily distinguishable	Böhme (2016), Jeon et al. (2022), Peitz (2024)	Platform is able to distinguish users based on their partisan activity on platform in the past and thus set access fee
Social planner always maximizes total surplus	Monopoly platform analysis often considers social planners problem as maximizing overall surplus	Peitz (2024)	Social planner can be discriminating where the objective is to maximize participant surplus and minimize platform payoffs.

favored Republicans (Kuang et al. 2024). Meanwhile, 60% of Republican-leaning participants think it favors Democrats. Republican-leaning participants forecast higher inflation (7.35% vs. 5.80%) and unemployment (9.90% vs. 8.60%) than Democrats. These models show how partisan identity-driven activity impacts economic thinking beyond economic conditions.

These advancements in modeling political bias have helped us understand economic processes in our increasingly interconnected yet politically fractured globe. The Richmond Federal Reserve's 2023 analysis shows how partisan legislative disagreement creates policy uncertainty that affects business investment and hiring decisions, demonstrating a strong link between fiscal policy uncertainty and real economic activity (Azzimonti 2023). Allianz (2023) noted that 60% of global GDP faced elections in 2024, causing widespread "wait-and-see" attitudes that postponed important economic choices. Their research showed that populist governance, which is becoming more widespread worldwide, has lowered GDP per capita by 10% over 15 years. These studies show how political polarization causes economic inefficiencies through policy execution, erroneous economic expectations among professionals and consumers, and increased uncertainty during political transitions. These ideas are vital as economies manage technology integration and political fragmentation.

DAM contributes to this literature by providing a framework that compares various mechanisms that help a platform mediate affective polarization-driven interactions. Our two-dimensional model that incorporates aversion vs. affinity among users analyzes how should platforms set access fees and manage user interactions in an environment marked by affective polarization. Given the importance of mediating polarized frictions, DAM's analysis and insights, as outlined in Table 1, emphasize the crucial role of platforms in the overall digital ecosystem.

3. DAM Framework

This section addresses the DAM-specific modeling of affective polarization, utility representation, and platform pricing across the three extensive game setups (see Fig. 2): (1) Benchmark model, where the platform sets access fees and consumers participate to maximize profits in a 2-stage game; (2) Platform-Provision (PP) model, where the platform sets access fees given incentive or control mechanism; (3) Social-Planner-Platform (SPP) model, where a Social Planner sets subsidies/taxes to maximize platform and consumer welfare. Finally, without an intermediary platform, the two polarized consumer populations negotiate their externalities (*quasi*-Coasian bargaining) in the final No-Platform (NP) model. We begin by illustrating the representation of affective polarization and its relation to participation utility.

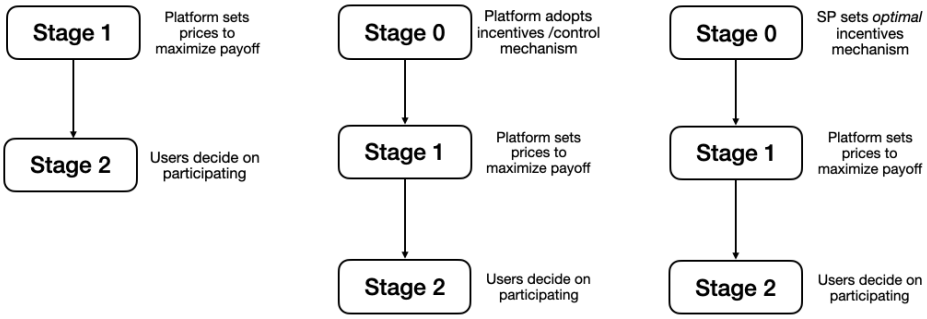


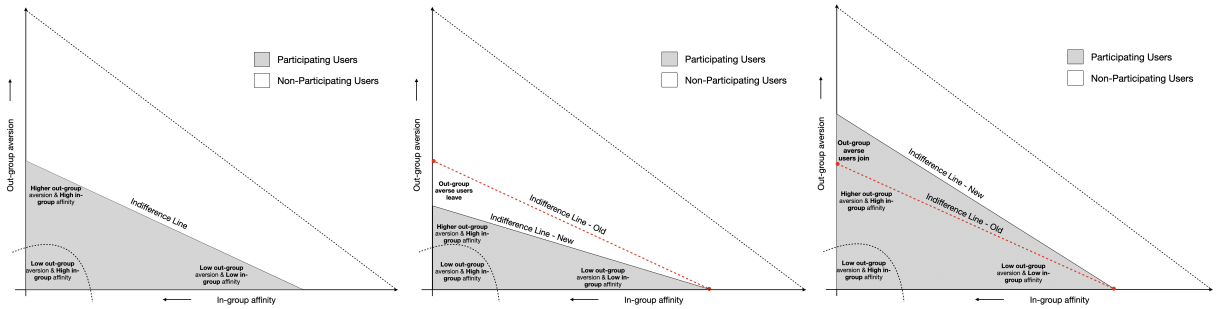
Figure 2 Three extensive game setups used for analysis.

3.1. Polarized Participation Representation

Affective polarization refers to the growing emotional divide between partisan in-groups and out-groups, characterized by varying degrees of bias toward one's group and hostility toward the opposing group (Rudolph and Hetherington 2021, Versteegen 2024). The Axelrod Attraction–Repulsion Model (ARM) (Axelrod et al. 2021) offers a valuable foundation for modeling polarized behavior. ARM begins with ideological polarization (distribution of political disagreements) and extends to affective polarization. Specifically, Axelrod et al. (2021) illustrates that the degree of affective polarization is the likelihood that a member of one party will tolerate a member of the same party (in-group) and feel resentful toward another (out-group), regardless of ideology. ARM suggests numerically assigning agent points in an attraction-repulsion space to represent affective polarization. We extend this feature to motivate a bi-dimensional representation of affectively polarized participants.

We enhance ARM's bi-dimensional representation using the 'Dolbear Triangle' approach (Dolbear 1967). Specifically, this approach examines the trade-off between utility and externalities by emphasizing individual preferences, market mechanisms, and the limitations of standard policy tools. Dolbear (1967) demonstrates the economic interactions of two individuals using a triangular Edgeworth Box. Here, one

individual (X) produces heat, which is related to smoke, while the other individual (Y) consumes bread and experiences the effects of smoke. The model employs indifference curves for these two commodities to analyze trade-offs and welfare outcomes. We apply this framework to polarized participants, taking into account in-group affinity and out-group aversion and the overall benefits of the platform.



(a) DAM representation of one side's participation

(b) Degree of polarization increases (out-group averse users leave)

(c) Degree of polarization decreases (out-group averse users join)

Figure 3 DAM representation of user participation on a given side. In Figs. 3b & 3c, the dotted line is the indifference user line in Fig. 3a.

Given the platform, DAM considers the trade-off between in-group affinity and out-group aversion. Each X and Y dimension represents a measure of increasing aversion (or decreasing affinity) in DAM. Without loss of generality, the X-axis represents in-group affinity, while the Y-axis represents out-group aversion. Higher values on X represent decreasing affinity, while higher values on Y represent increasing aversion. The Edgeworth Box is still triangular, but the upper bound on X and Y coordinates is 1, making the total population of users in the box $\frac{1}{2}$. Thus, each point in the triangle represents a user's type characterized by in-group affinity vs. out-group aversion. Given type, each user, when deciding to join the platform, considers the platform valuation v, access fee $p_i, i \in \{L, R\}$ (e.g., premium vs. free memberships, cost of advertising), and mass of users anticipated on the same n_i and the other side n_{-i}. Each factor's impact varies, and overall, the user joins only when the utility u_i from joining is non-negative (i.e., $u_i \geq 0$). The following equation gives the structure of utility.

$$u_i = \underbrace{(v - \alpha_i(n_{-i} - n_i) - p_i)}_{\text{Net Benefits}} - \overbrace{(y + a_i x)}^{\text{User type}} \tag{1}$$

Refer to Figure 3 for an illustration of user participation using Equation 1. Note that access fees and user type affect how users benefit from joining the platform. For instance, conservatives and liberals differ in content consumption and platform perceptions, according to research. Conservatives utilize social media less than liberals (Motyl 2023) while liberals consume 56% more problematic in-group content and retweet 19.4% more diverse policy problems than conservatives (Chang et al. 2023). Conservative accounts that promote low-quality or misinformation-heavy content have greater suspension rates (Zaman 2024). Still,

X users report greater balance, whereas Instagram, Facebook, and TikTok users report more liberal content (McClain et al. 2024). Therefore, users' preference considering in-group affinity vs. out-group aversion ($a_i \in (0,1)$ termed as *degree of polarization* following Stewart et al. (2020), Norman and Green (2025)) affects social media utility.

Note that Equation 1 omits the following. The model does not consider transaction-level utility and therefore transaction costs. DAM focuses on how platform participation is affected when users on one side know of the other and must choose between participation and non-participation. Moreover, since users on both sides can be producers and consumers of platform content, one side's participation size indicates their polarizing capability and is thus a disutility for the other. Lastly, user types are *static*, meaning each side's users maintain a fixed in-group affinity vs. out-group aversion towards the opponent side throughout the game. The following subsections describe platform and social planners' polarization friction mitigation strategies.

3.2. Incentive Mechanisms: Subsidies vs. Taxes

Digital platforms often provide algorithmic *tweaks* that function as subsidies or taxes, impacting user engagement (Bucher 2018). Essentially, algorithmic recommendations reallocate user engagement, akin to subsidies and taxes (Gillespie 2014). Platforms may subsidize engagement by boosting exposure to aligned partisan content. Similarly, platforms can 'tax' engagement by exposing users ('perspective-taking') to conflicting content (Tufekci 2014, Benson 2024). These digital mechanisms, whether meant to enhance engagement or resulting from data bias, can profoundly impact market dynamics by rewarding certain (user-generated) content and penalizing others. Recent research (see, e.g., González-Bailón et al. 2023, Guess et al. 2023) explores how such subsidies or taxes impact algorithm-driven content consumption, affecting user utility and partisan alignment. Thus, in the PP and SPP models, subsidies and taxes are incentive mechanisms to mediate polarized interactions.

Blanket subsidies (or taxes) are algorithm *tweaks* that users on both sides enjoy (or dislike), irrespective of type. Conversely, Pigouvian subsidies or taxes target user types instead of net utility. Typically, Pigouvian subsidy (or tax) is intended to target specific 'behaviors' that improve (or worsen) utility (Baumol 1972). Since DAM's user types are represented using behavioral biases – in-group affinity vs. out-group aversion – that determines platform participation (Howard and Rothbart 1980), Pigouvian subsidies reward while taxes penalize such biases, given the user's type. Both these incentive mechanisms can be represented using the following equations.

$$\hat{u}_i = \begin{cases} u_i + \Theta_I(a_i, s_i), & \text{Blanket Subsidy/Tax} \\ (v - \alpha_i(n_{-i} - n_i) - p_i) - (y + \Theta_I(a_i, s_i)x), & \text{Pigouvian Subsidy/Tax} \end{cases} \tag{2}$$

where $\Theta_I = \begin{cases} a_i s_i, & \text{Subsidy} \\ \frac{a_i}{s_i}, & \text{Tax} \end{cases}$, represent the structure of subsidies or taxes whether blanket or Pigouvian. Practically, for instance, taxes may boost conflicting content visibility that can discourage liberal users from exploring diverse opinions (e.g., perspective taking). At the same time, subsidies may boost aligned content, reinforcing existing beliefs (Tufekci 2014), and potentially deepening the echo chambers (Van Dijck 2013).

3.3. Control Mechanisms: Obfuscations vs. Restrictions

Under control mechanisms, we consider all other ways in which platforms can impact user participation, thus impacting polarization frictions. Four broad mechanisms are considered: restricting participation, obfuscating interactions, valuation manipulation and platform versioning.

Platforms may restrict or prohibit users by combining automated algorithms and human moderation to curb hate speech, misinformation, and harassment (Gillespie 2018, Gerrard 2020). These mechanisms can de-platform or suspend users whose content violates rules, even when the line between harmful and lawful expression is unclear (Balkin 2017). Due to policy formulations, the opacity and inconsistent application of these mechanisms often raise questions about accountability and ideological bias (Gorwa 2019). Thus, ongoing conflicts between protecting users and open online debate highlights platforms' struggles to regulate content while valuing free expression.

Instead of restricting participation, platforms might minimize (neutralize) polarizing user interactions. For instance, to reduce extremely heated user interactions, platforms may restrict comment display, limit controversial post reach, or prohibit responses from accounts users do not follow (Cinelli et al. 2021). Moreover, some platforms use friction measures like warning prompts when users try submitting unsafe content or sharing articles without reading them (Pennycook and Rand 2022). These interventions can encourage more thoughtful engagement and discourage impulsive, hostile interactions. Alternatively, reducing reliance on virality metrics, such as likes and shares, can slow the spread of polarizing content. For example, X removed engagement counts on replies to reduce frictions (Brady et al. 2017).

Platforms have also used psychological design, regulating algorithms, and social reinforcement to affect user utility evaluations. For example, ephemeral content (Snapchat Stories), engagement streaks, and Twitter Blue benefits create a false sense of scarcity, enhancing the platform's perceived worth (Hwang 2020). Further, engagement streaks provide dopamine-driven feedback, reinforcing that social media is crucial for self-worth and social belonging (Alter 2017, Ichihashi and Kim 2023). In some cases, platforms showcase tailored lifestyles and influencer content, altering users' ideas of what is desirable and driving the need to participate (Fardouly et al. 2015). Moreover, constant notifications, trending topics, and real-time updates create a sense of urgency, making users feel they must stay active or risk missing important events or conversations (Przybylski et al. 2013, Brailovskaia and Margraf 2024).

Finally, numerous platforms offer versions by area, regulatory compliance, or audience (e.g., Facebook Lite is for users in low-internet areas). Features, accessibility, and moderation policies change each version. Additionally, it provides different content and moderation policies depending on local laws (e.g., restricted content in certain countries) (Gillespie 2018). Due to China's strict internet regulations, TikTok, the global version, and Douyin, the Chinese version, operate as separate platforms with different algorithms, content moderation, and features (Fuchs 2021). LinkedIn offers a localized version in China that complies with

Chinese internet regulations, which has led to content and feature restrictions compared to its global counterpart (Napoli 2019).

Each of these mechanisms is represented, similar to Equation 2, by modifying Equation 1, as follows.

$$\breve{u}_i = \begin{cases} u_i(\eta), & \text{Participation restriction} \\ (v - \alpha_{ic} n_{-i} + \alpha_{is} n_i - p_i) - (y + a_i x), & \text{Interaction obfuscation} \\ (v_i - \alpha_i (n_{-i} - n_i) - p_i) - (y + a_i x), & \text{Valuation manipulation} \\ (v + \alpha_i (n_i - \epsilon) - p_i) - (y + a_i x), & \text{Versioning} \end{cases}, \tag{3}$$

where $\eta \in (0,1)$ is the fraction of users restricted from participating; $\alpha_{ic} \neq \alpha_i, \alpha_{ic} \in (0,1)$ is marginal cost from rival partisan users while $\alpha_{is} \neq \alpha_i, \alpha_{is} \in (0,1)$ is the marginal benefit from same side users; $v_i > 1, v_i \neq v$ is the partisan side-relevant valuation; ϵ is the 'lite' version disutility that users incur, which is significantly lesser than $\alpha_i n_{-i}$ in Equation 1.

3.4. Quasi-Coasian Bargaining

Coase (2013) defined Coasian bargaining as private negotiations between parties impacted by externalities to produce efficient results without government interference. Coasian bargaining works under specific conditions. First, transaction costs for negotiating, enforcing, and monitoring agreements should be minimal. Second, property rights must be clear and enforceable to define parties' obligations and entitlements. Third, parties need symmetric and enough externality cost-benefit information to reach fair agreements. Besides, increased stakeholder numbers can hinder negotiating by causing coordination and collective action issues (Demsetz 1967). Coasian bargaining reduces externalities under all three conditions.

However, two-sided digital platforms rarely meet all three conditions. First, assessing affective polarization externalities is difficult due to a broad and diverse user base, and bargaining among millions of people costs more than private agreement gains. Second, user-generated content and interactions have unclear or nonexistent property rights, making it difficult to assign responsibility and rewards for effective bargaining. Finally, online interactions change constantly, making it challenging for bilateral or multilateral agreements to create solid, mutually agreed-upon solutions that decrease externalities.

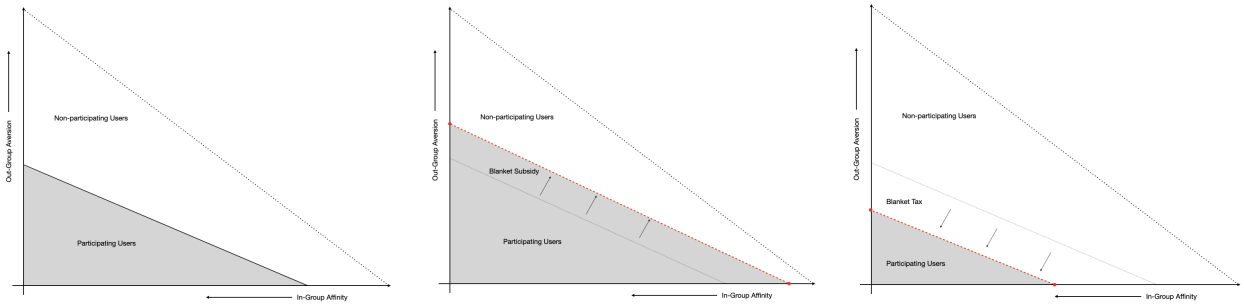
Consequently, we consider *quasi*-Coasian scenario where, without a platform or a social planner, the polarized populations are left to negotiate a fraction $\gamma \in (0,1)$ over a common valuation $v > 1$ of interacting with the other group. Each user is now represented using in-group affinity, out-group aversion, and valuation of negotiating a fraction of v. Without a mediating platform, direct negotiations are not uniformly valued; therefore, negotiating over that fraction also determines group sizes. More succinctly, in this case, the following equations, extending Equation 1 represent the utility from the interaction.

$$\tilde{u}_i = \begin{cases} z(v(1-\gamma)) - (y + a_i x), & \text{Side } i, \\ z(v\gamma) - (y + a_{-i} x), & \text{Side } -i \end{cases} \tag{4}$$

where $i \in \{L, R\}$. Note that while the valuation may be common v, there is heterogeneity in each individual's affinity towards direct interaction $z \in (0,1)$ distributed uniformly. Hence, even the fraction

split of the valuation depends on how each individual, given their 'type,' considers it worth interacting with the rival group.

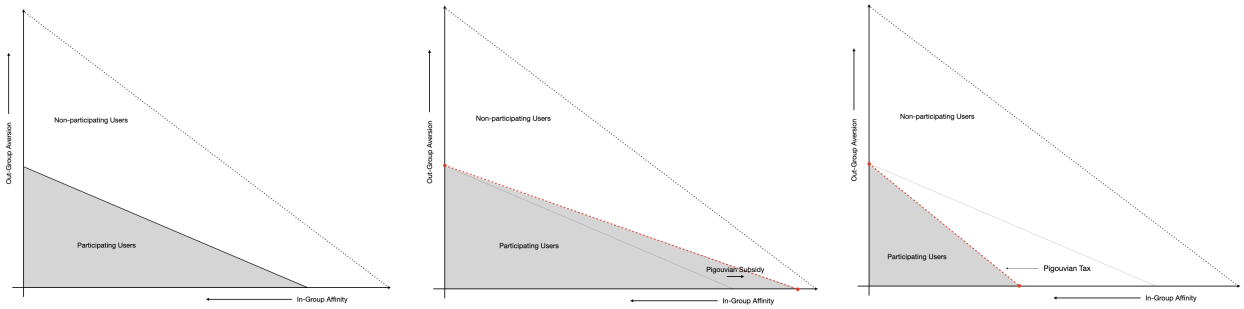
Example: Suppose $v = 2, \alpha = \frac{1}{2}, \alpha_{-i} = \frac{1}{4}, a_L = \frac{1}{2}, a_R = \frac{3}{4}, s_L = \frac{1}{4}, s_R = \frac{1}{4}$. Further, suppose the



(a) Base Model **(b)** Blanket Subsidy **(c)** Blanket Tax

Figure 4 DAM Illustration with Blanket Subsidy/Tax

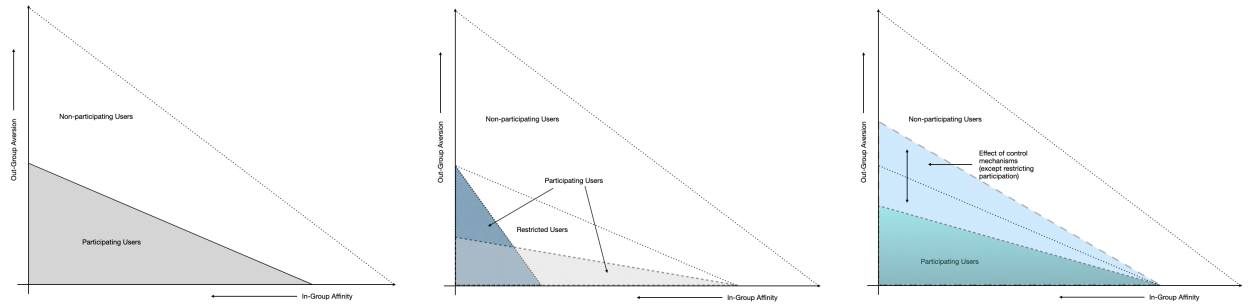
$x-$intercept for side L is $x_L = \frac{3}{4}$ and for side R is $x_R = \frac{1}{2}$. Then, we can construct the following participation size and illustrate how subsidies, taxes or control mechanisms may change participation on one side (and symmetrically the other side). First, consider the blanket incentives in Figs. 4a - 4c. Blanket subsidies



(a) Base Model **(b)** Pigouvian Subsidy **(c)** Pigouvian Tax

Figure 5 DAM Illustration with Pigouvian Subsidy/Tax

shift the indifference line out, while blanket taxes shift the line in. The degree (a_i) of polarization does not change, but everyone gets the same benefit. In contrast, Pigouvian subsidies and taxes change the slope a_i as shown in the following Figures 5a - 5c. Changing a_i implies changing the trade-off between in-group affinity vs. out-group aversion, thus impacting each user differently given their innate type. Finally, when considering control mechanisms, note that each such mechanism can have a different impact on the indifference line and thus, the participating masses, as shown in the following Figs 6a - 6c. Specifically, Figure 6b illustrates that restricting participants can lead to one of the shaded areas representing the final participation mass. All other control mechanisms have the effect of changing the slope of the indifference curve that essentially invites or repels users that lean more towards out-group aversion.



(a) Base Model **(b)** Restricted Participation **(c)** All other control mechanisms

Figure 6 DAM Illustration with Control Mechanisms

4. Equilibrium Analysis

The equilibrium analysis of DAM for the Benchmark, PP, SPP, and NP models, is along the usual dimensions of analysis of two-sided platforms (see Table 1). Further, DAM-specific analysis dimensions are explained below using Definitions 1 and 2. Finally, Definition 3 establishes the meaning of mediating polarized participation through the DAM framework lens.

Definition 1 ***Participation Diversity****: Participation diversity compares in-group affinity to out-group aversion for each incentive or control mechanism (indexed by k). The ratio ($\Lambda_k \coloneqq \frac{n_{ik,y\leq x}}{n_{ik,y\geq x}} in \mathbb{R}_+, i \in \{L,R\}, k \in \{1,2,\cdots\}$) divides the participating population (n_{ik}) into in-group affinity ($n_{ik,y<x}$) vs. out-group aversion leaning participants. If $\Lambda_k \to 1$, then participation diversity is maintained; otherwise, for $\Lambda_k \in (0,1)$ in-group affinity dominates or for $\Lambda_k > 1$, out-group aversion dominates.*

We consider participation diversity as representative of diversity of thought that breaks echo chambers and promotes critical discourse, vital for a healthy democratic process, reducing polarization. Diverse perspectives can challenge confirmation bias and promote reflective reasoning (Sunstein 2002). Research indicates that heterogeneous groups can produce more imaginative and reliable solutions to complicated challenges (Page 2008). Promoting diverse opinions in polarized environments can reduce ideological extremism and promote balanced public debate (Iyengar and Westwood 2015). By embracing multiple viewpoints, platforms can attempt to reduce polarization, promote inclusive communication, and better decision-making (Fiorina et al. 2004).

Definition 2 ***Pricing Asymmetry****: Pricing Asymmetry measures the change in access fee under incentive or control mechanisms (indexed by k), compared to the benchmark (indexed by B), using the ratio $\xi_k \coloneqq \frac{(p^*_{ik} - p^*_{iB})}{(p^*_{-ik} - p^*_{-iB})}, i \in \{L,R\}$.*

Asymmetry in access fees can explain how digital platforms use differential pricing to manage polarizing participation and user composition. Platforms often adjust prices to balance network effects, conceivably

discouraging extreme users and encouraging moderate participation (Rochet and Tirole 2003, Eisenmann et al. 2006). Platforms can filter participation and reduce the dissemination of polarizing information by charging greater fees to polarized users (Tucker et al. 2018). Our analysis focuses on targeted pricing methods as a key policy instrument to reduce platform participant polarization. These asymmetry measures also provide a natural experiment to explore economic incentives and ideological expression. Next, we define the aspects of platforms' mediating role in the context of DAM. Specifically, the following definition clarifies what it means when platforms mediate affectively polarized user interactions.

Definition 3 ***Mediating Polarized Participation:*** *Perfect mediation of polarized participation occurs when a platform's incentive or control mechanisms result in increase in participation size with participation diversity close to* 1*, positive pricing asymmetry so neither side is charged lesser than benchmark, and negative same- and cross-side demand elasticities. If some of these conditions are met, the platform mediates polarized participation imperfectly. Otherwise, the platform fails to mediate polarization.*

Each factor mentioned in Definition 3 draws from recent empirical and qualitative research. First, increased in-group affinity reinforces shared identities and ideals while portraying out-group individuals as fundamentally different or possibly toxic, promoting group cohesion (Tajfel et al. 1982). Conversely, increased out-group aversion reinforces existing stereotypes and negative perceptions, deepening rival group hostility (Iyengar et al. 2019). This dual process amplifies the emotional distance between polarized groups through a self-reinforcing cycle where extreme loyalty to one's group leads to the vilification of the other, ultimately exacerbating affective polarization (Mason 2018). Thus, we consider a balanced participation diversity as a measure of mediating polarized participation.

Second, economic constraints on digital platforms may increase affective polarization and political inequality. One side's higher access fee may exclude some members, reinforcing bias and reducing cross-group interaction. Social identity theory indicates exclusion increases in-group cohesion and worsens out-group perceptions (Tajfel et al. 1982). Recent research also suggests that economic gate-keeping might lead to partisan splits by limiting varied opinions and promoting echo chamber effects (Benkler 2007, Wagner and Praprotnik 2024). Reduced group contact can worsen biases and increase hostility, leading to increased affective polarization (Iyengar and Westwood 2015, Iyengar et al. 2019). Minimizing pricing asymmetry is another way to mediate polarized participation.

Finally, DAM considers participation from two political echo chambers on each side. Typical two-sided platform theory indicates negative same-side and cross-side price elasticity of demand, indicating that lowering prices on one side encourages more participants from both sides (Rochet and Tirole 2003, Belleflamme and Peitz 2018). Affective polarization and adverse cross-platform externalities may change elasticities (e.g., cross-side elasticity might be positive). Thus, a platform mediates polarized participation if its demand elasticities match the regular two-sided platform participation.

Given these definitions, we use four theorems to examine and present our findings. Theorems 1 and 2 present insights regarding platform-provisioned incentive and control mechanisms, respectively. Theorem 3 summarizes insights from quasi-Coasian bargaining. Theorem 4 summarizes insights from a social planner's optimal subsidies or taxes. Each theorem builds on equilibrium prices and participation size analysis. However, we begin by establishing the benchmark model equilibrium in Lemma 1, with associated insights in Proposition 1.

Lemma 1 *In the benchmark model, the following is the participation sizes, access fees and demand elasticities for sides $i, -i \in \{L, R\}, i \neq -i$:*

Benchmark Model Equilibrium

1	*Participation Size*	$n_i^* = \frac{a_i x_i^2}{2}$
2	*Access fees*	$p_i^* = v - a_i x_i + \alpha_i (n_i^* - n_{-i}^*)$
3	*Same Side Elasticity*	$\eta_{ii}^* \coloneqq \frac{\partial n_i^*}{\partial p_i^*} \frac{p_i^*}{n_i^*} = -\frac{x_i(2-\alpha_{-i}x_{-i})}{2(2-\alpha_i x_i - \alpha_i x_{-i})} \frac{p_i^*}{n_i^*} \leq 0$
4	*Cross Side Elasticity*	$\eta_{i,-i}^* \coloneqq \frac{\partial n_i^*}{\partial p_{-i}^*} \frac{p_{-i}^*}{n_i^*} = \frac{\alpha x_i x_{-i}}{2(2-\alpha_i x_i - \alpha_{-i} x_{-i})} \frac{p_{-i}^*}{n_i^*} \geq 0$

(5)

where x_i, x_{-i} are x-intercepts of the indifferent user curve $u_i = u_{-i} = 0$.

Each equilibrium outcome indicates a key aspect of DAM and how DAM-specific features illustrate polarized interactions. First, note that a change in the degree of polarization (a_i) directly impacts the participation size. Second, since the platform sets access fees to maximize profit, such fees depend on the net benefits (or, costs) one side gets from participation sizes on both sides. Finally, affective polarization-driven participation implies negative same-side elasticity – lower demand if the access fee is raised on one side – and positive cross-side elasticity – higher demand if the access fee is raised on the other. Importantly, this directly contradicts the typical demand elasticities on two-sided platforms where users have positive externalities from increased participation on either side of the platform.

Still, it is unclear how polarization-driven equilibrium outcomes indicate the unpredictable and, therefore, risky nature of mediating polarization on platforms. The following proposition sheds further light on this aspect of the benchmark model equilibrium.

Proposition 1 *The side with a higher degree of polarization is not always charged a higher access fee or has a smaller participation size at the benchmark equilibrium.*

Proposition 1 provides the first indication of how mediating between two groups affectively polarized is not straightforward. Consider the participation size. From Lemma 1, participation size for each side is characterized by its degree of polarization a_i and the in-group affinity of the extreme indifferent user x_i that has no out-group aversion. The value of x_i determines in-group cohesion while a_i determines the relative out-group aversion. Thus, given $a_L < a_R$ (i.e., L more polarized than R), $n_L^* > n_R^*$ is possible when

$\frac{a_R x_R^2}{x_L^2} < a_L < a_R$ and $x_R < x_L$. That is, greater participation from those with higher in-group affinity can result in greater participation from the group with a higher degree of polarization.

Now consider the access fee. Lemma 1 shows that each side is charged a fee that is proportional to the net benefit (i.e., $\alpha_i(n_i^* - n_{-i}^*)$) from the participation sizes of either side, degree of polarization a_i, and the in-group affinity indifference point x_i. Thus, again, there can be multiple scenarios where the price for the side with a higher degree of polarization (i.e., in this case, L) can be higher. For instance, when $x_L < x_R$, $x_L + x_R > \frac{2}{\alpha_L + \alpha_R}$, and $\frac{a_R x_R (2 - x_R(\alpha_L + \alpha_R))}{x_L(2 - x_L(\alpha_L + \alpha_R))} < a_L < a_R$, then $p_L^* > p_R^*$. That is, for the side with a higher degree of polarization when participation size is lesser in comparison, then the platform charges a higher access fee to maximize payoffs.

Based on Proposition 1, two effects influence participation: (1) in-group cohesion-driven interest to boost polarized participation (*tribe effect*); (2) exploitation of participation sizes to optimize profit, where lesser participation may increase access fees (*value capture effect*). Recent studies confirm these effects. High-in-group affinity participants can increase polarized group size and cohesiveness through homophilic interactions and echo chamber effects. Engaging with like-minded peers strengthens beliefs and encourages participation, leading to fast group growth and ideological consolidation (Bakshy et al. 2015). Further, computational models indicate that in-group preference attachment leads to network fragmentation, protecting clusters from criticism and boosting political polarization (Cinelli et al. 2021). Even so, affectively polarized individuals may reduce online interactions due to emotional weariness and animosity, lowering user engagement and revenue for social media advertising (Iyengar et al. 2012). Thus, social media companies may increase access fees (e.g., targeted advertising) to recover revenue if participation decreases and advertisers lose interest (Kox et al. 2017, Votta et al. 2024).

Importantly, Proposition 1 highlights the challenges platforms face when dealing with affectively polarized participants. When targeting degree of polarization (i.e., using incentive mechanisms) or regulating participation, the platform must balance minimizing in-group cohesion-driven size growth with maximizing revenues by charging for lower participation. Thus, the platform must balance maximizing profit and welfare using mediation mechanisms. The following subsections show equilibrium insights along this fundamental trade-off.

4.1. Incentive Mechanisms Equilibria

Theorem 1 *None of the incentive mechanisms perfectly mediates polarization. While subsidies generally increase participation, and taxes reduce it, blanket subsidies (or taxes) do not necessarily guarantee larger participation than Pigouvian. Moreover, only Pigouvian taxes improve participation diversity. Additionally, pricing asymmetry is consistently negative regardless of the mechanism employed. Finally, except for blanket subsidies, same-side demand elasticities remain negative while cross-side demand elasticities are positive across all other incentive mechanisms.*

Theorem 1 is established through analysis of the benchmark equilibrium (Lemma 1); equilibrium analysis of participation size and diversity (Proposition 1), pricing asymmetry and same-side and cross-side elasticities (Lemma 2 and Proposition 2). We begin by illustrating the equilibrium participation size and diversity using the following lemma.

Lemma 2 *Under each incentive mechanism* $k \in \{\text{Blanket tax}, \text{Blanket subsidy}, \cdots\}$*, the equilibrium participation size* n_{ik}*, access fee* p^*_{ik}*, and demand elasticities are shown in Appendix A Table 3.*

Detailed expressions are in Appendix A. Lemma 2 establishes the structure of the outcomes, in comparison with Lemma 1. Observe that while participation size is impacted by the subsidy or tax levels s_i, potentially dampening the *tribe effect*, the access fee still displays the *value capture effect*, where the platform ends up extracting some of the value generated by the incentive through modified participation sizes. Additionally, note that even under incentives, the degree of polarization impacts participation sizes. Finally, affective polarization-driven participation suggests negative same-side elasticity and positive cross-side elasticity even under Pigouvian subsidy and taxes and blanket tax mechanisms. As explained below, the platform can partially mediate polarization only under blanket subsidy, as the elasticities indicate.

Proposition 2 *Participation diversity* Λ_k *can be close to* 1 *only under Pigouvian tax. Further, pricing asymmetry* ξ_k *is consistently negative across all mechanisms while demand elasticities only under blanket subsidy imperfectly mediate polarized participation.*

The overall takeaway from Proposition 2 points to how incentive mechanisms are imperfect at mediating polarized participation. First, consider participation diversity. Among the mechanisms used, Pigouvian subsidy or tax changes the degree of polarization. Effectively, changing the degree of polarization can increase (subsidy) or decrease (tax) the participation size. However, for diversity, it is important to note who joins or leaves the platform on each side. If the degree of polarization increases, more individuals with greater in-group affinity can join; conversely, a decrease in the degree of polarization can result in more individuals with greater in-group affinity leaving. Definition 1 says that balanced participation of out-group averse and in-group affinitive individuals improves diversity, dampening the *tribe effect*. Thus, Pigouvian tax can help achieve diversity close to 1 when $a_i \sim s_i$. Besides, recent studies suggest that algorithmic tweaks purposely exposing polarized partisan users to opposing views can breach echo chambers and reduce polarization. For instance, adding cross-cutting content to recommendation algorithms increases political diversity and reduces homophilic reinforcement (Saveski et al. 2022, Gillissen et al. 2023). Following earlier research (see, e.g., Cinelli et al. 2021), even small increases in exposure to opposing views can reduce extreme partisan attitudes and affective polarization. These findings suggest that *targeted* algorithmic interventions could improve public debate and diminish social media's profound ideological differences.

Second, consider pricing asymmetry. As Table 3 shows, the pricing structure is similar to benchmark equilibria, which implies that the platform's pricing strategy demonstrates the *value capture effect*. That is, when the platform adopts such mechanisms, the users pay for the value generated and the platform can end up charging one side higher and the other lower than the benchmark. Moreover, asymmetric access fee (e.g., advertising costs) often reinforces existing group boundaries rather than bridging them (Iyengar et al. 2012). Instead of incentivizing cross-partisan engagement, asymmetric access fees tend to marginalize users vulnerable to affective polarization, further entrenching echo chambers. More broadly, recent studies have indicated that asymmetric targeted advertising on social media—where partisan users receive disproportionately tailored political messages—decreases partisan involvement relative to their base. For instance, asymmetric ad-targeting may cause feelings of manipulation and inaccuracies, especially among liberals who are wary of data-driven campaigns (Benkler et al. 2018). Research indicates that liberals may become disengaged and less active online due to overexposure to partisan messages since they regard them as intrusive and biased (Bakshy et al. 2015, Tucker et al. 2018).

Finally, consider demand elasticities. Typically, blanket subsidies result in negative demand elasticities, depending on consumer responsiveness and market dynamics (Mas-Colell et al. 1995). By reducing the cost of goods or services, subsidies can increase consumption among price-sensitive customers, leading to negative demand elasticity (Kreps 2020). Still, the nature of the good (e.g., Veblen or Giffen) can result in positive price elasticity, irrespective of subsidy. Considering DAM, a typical blanket subsidy incentivizes all polarized individuals to reduce disutility from rival partisan participation. Yet, the two-sided nature of platforms implies that blanket subsidies can have mixed outcomes. When side i's increase in price decreases the demand from side $-i$, side i can see an increase in demand. Conversely, when such a price increase results in increased demand from side $-i$, side i can see a decrease in demand. Unlike other mechanisms, subsidies draw everyone to the platform, so the side that benefits most from participation does so even when prices rise. Our findings mirror the 'confrontation effect,' where more polarized individuals use social media more, even with targeted partisan ads (Mochon and Schwartz 2024).

Nevertheless, platforms have other tools at their disposal to control polarized participation. While incentive mechanisms explore algorithmic tweaks to content consumption, control mechanisms, in the following subsection, explore explicit changes to other aspects of platform use experience that might help mediate polarization.

4.2. Control Mechanisms Equilibria

Theorem 2 *No control mechanism perfectly mediates polarized participation. While participation sizes are weakly lower than the benchmark, in-group affinity dominates participation diversity. Still, pricing asymmetry can improve depending on the mechanism. Further, only under platform versioning, same side elasticities are negative while cross-side elasticity is perfectly inelastic.*

Theorem 2 is established through similar background analysis of the benchmark equilibrium (Lemma 1); equilibrium analysis of participation size and diversity (Proposition 1). Lemma 3 and Proposition 3 establish pricing asymmetry and same-side and cross-side elasticities under control mechanisms. We begin by illustrating the equilibrium participation size and diversity using the following lemma.

Lemma 3 *Under each control mechanism* $k \in \{\text{Restricting participation}, \text{Obfuscating interactions}, \cdots\}$*, the equilibrium participation size* n_{ik}*, access fee* p^*_{ik}*, and demand elasticities are shown in Appendix B Table 4.*

Detailed expressions are in Appendix B. Lemma 3 establishes the structure of the outcomes, in comparison with Lemma 1. In this scenario, participation sizes are almost the same as the benchmark, reinforcing the *tribe effect*, unlike incentive mechanisms. At the same time, access fees demonstrate the *value capture effect*. In particular, the platform accounts for the lost fraction of users through access fees when participation is restricted. Additionally, note that even under control mechanisms, affective polarization-driven participation suggests negative same-side elasticity and positive cross-side elasticity under all mechanisms except platform versioning. Since different versions of platforms rule out interactions with the other side, there is no impact of changing the access fee of one side on the other. The following proposition illustrates how control mechanisms might help mediate polarization, albeit imperfectly, given these outcomes.

Proposition 3 *Participation diversity* $\Lambda_k < 1$ *for all mechanisms. Further, pricing asymmetry* ξ_k *across all mechanisms, as shown below, is not always negative. Finally, demand elasticities across mechanisms are similar to benchmark but under platform versioning, cross-side elasticity is* 0*.*

	Pricing Asymmetry
1 *Restricting Participation*	$\xi^*_{RP} = -\frac{\alpha_i}{\alpha_{-i}} < 0$
2 *Interaction Obfuscation*	$\xi^*_{IO} = \frac{a_i x_i^2(\alpha_{is}-\alpha_i)+\alpha_{-i}x_{-i}^2(\alpha_i-\alpha_{ic})}{a_i x_i^2(\alpha_{-i}-\alpha_{ic})+\alpha_{-i}x_{-i}^2(\alpha_{-is}-\alpha_{-i})}$
3 *Manipulating Valuation*	$\xi^*_{MV} = \frac{v-v_i}{v-v_{-i}}$
4 *Platform Versioning*	$\xi^*_{PV} = \frac{\alpha_i a_{-i} x_{-i}^2 - 2\alpha_i \epsilon}{a_i \alpha_{-i} x_i^2 - 2\alpha_{-i}\epsilon} \geq 0$

(6)

Proposition 3's overall takeaway, in conjunction with Proposition 2, points to the fundamental difference between using incentives vs. control when attempting to mitigate externalities. Incentives offer utility- or user-type specific benefits or costs, whereas control approaches aim to alter the total utility derived from participation, which may be ineffective. To begin with, consider participation diversity and size. Participation sizes for all control measures except restricting participation are similar to the benchmark since the platform uses access fees to counteract diversity-related gains. Thus, none of the mechanisms deliver diversity close to one. More generally, participation controls do not address selective exposure and network homophily. Therefore, they do not diversify partisan voices. For instance, users flock toward

like-minded communities even after platforms restrict participation, sustaining ideological echo chambers (Garrett 2009). Additionally, such restrictions can be counterproductive since excluded individuals can migrate to alternative platforms where they may further radicalize, ultimately undermining efforts to cultivate a more inclusive and deliberative digital public sphere (Mutz 2006).

Second, consider pricing asymmetry (Equation 6). Only under participation restrictions and platform versioning do platforms establish access prices to benefit one side or retain the *value capture effect* for comparable charges. Obfuscating interaction determines whether the platform raises or lowers the access fees compared to the benchmark. For instance, $\alpha_{is} < \alpha_i < \alpha_{ic}$ but $\alpha_{-is} > \alpha_{-i} > \alpha_{-ic}$ ensures the platform advantages side i with lower access fee while charges side $-i$ higher. That is compared to benchmark, lesser (greater) obfuscation of side i's (side $-i$'s) cross-side externality coupled with greater (lesser) obfuscation of side i's (side $-i$'s) same-side externality acts as an added cost (benefit) that the platform delivers, for which the platform incentivizes participation from side i but increases access fee to capture value from side $-i$. More generally, obfuscating social media interactions between polarized partisans may encourage targeted partisan advertising. Platforms can group users into homogeneous groups by purposely hiding the complete network of cross-ideological connections, making it easier for advertisers to target political leanings (Kobayashi et al. 2024). Online ad micro-targeting research suggests that obscuring inter-group interactions gives advertisers a clearer signal of user preferences and allows them to use more aggressive partisan ad strategies, amplifying polarization and undermining balanced public discourse (Benkler et al. 2018, Roozenbeek and Van der Linden 2024).

Finally, consider demand elasticities. Participation controls, unlike blanket subsidies, have a definitive impact on demand. Except for platform versioning, all other control mechanisms do not change demand elasticities compared to the benchmark. Controlling market participation through regulatory hurdles or selective entry rules changes the competitive environment and substitutes availability, affecting demand elasticity (Mas-Colell et al. 1995). Typically, restricting market participants or product types may make demand for existing products more inelastic since buyers have limited choices. Since consumers have few options, firms with market strength can raise prices without losing sales (Kreps 2013). More options in open, competitive markets increase price sensitivity and demand elasticity. Yet, two-sided platforms present a different picture when considering the relationship between elasticity and participation controls. Specifically, since platforms are constrained by their need for both sides participating to guarantee revenue, participation controls that fail to remove dependency of access fees on participation sizes on either side display similar elasticity as the benchmark – negative for same side and positive for cross-side.

However, under platform versioning, the platform successfully mitigates the impact of cross-side negative externality on demand. In general, platform versioning presents an opportunity to redesign social media platforms that can provide polarization mediation. Combs et al. (2023) provides experimental evidence where a platform version with anonymized user handles mitigates polarized interactions. Further, Greene

et al. (2024) suggests a scalable design where participants of opposing partisan views but equal social status are put in contact and thus show reduced polarized interactions. In effect, participation controls have a different albeit imperfect way of mediating polarized participation compared to incentive mechanisms.

4.3. Quasi-Coasian Equilibrium

Theorem 3 *In quasi-Coasian bargaining without a platform, when polarized groups negotiate over a common valuation, the side that enjoys a larger share has a lower degree of polarization.*

Theorem 3 is established using Lemma 4 and Proposition 4. Essentially, since this is the case where there is no platform, we have two three-dimensional planes intersecting in a feasible region defined by a cube of volume 1. Lemma 4 establishes the intersection line for two intersecting planes that defines this feasible region. Further, Proposition 4 establishes the conditions for which a valid bargaining solution exists.

Lemma 4 *Assuming a scaling factor $t \in \mathbb{R}_+$ and an outside option utility of w, the line that defines the indifference curve given $\tilde{u}_i$ (Equation 4) is given by* $\left\{x = \frac{(1-2\gamma)t}{a_i - a_{-i}}, y = \frac{a_i(\gamma t - w) + a_{-i}(w-(1-\gamma)t)}{a_i - a_{-i}}, z = \frac{t}{v}\right\}$.

Parametric equations can neatly express a line that connects two non-parallel planes in three dimensions. The parametric representation defines the intersection and simplifies multidimensional geometric and analytic research. Drawing on this fundamental method, we represent the indifferent users locus using Lemma 4. Observe that in this case, without access fee or platform valuation, the indifferent user(s) considers the valuation split γ, the outside option w and the degree of polarization a_i, a_{-i}. Thus, the equilibrium split is decided not just on the split level γ but who gains from such a split, accounting for the valuation of such splits and the affectively polarized interaction externalities. The following proposition further elucidates the equilibrium split.

Proposition 4 *Any equilibrium split $\gamma \in (0,1)$ is possible only when $\gamma > \frac{1}{2}$. Additionally, a possible equilibrium split holds when $a_i < \frac{1}{2} < a_{-i}$ and $2\gamma - 1 \leq a_{-i} - a_i$ along with $t < v$ and $w < \frac{t(a_i\gamma - a_{-i}(1-\gamma))}{a_i - a_{-i}}$.*

Proposition 4 provides the conditions that must hold for an equilibrium with a finite split $\gamma \in (0,1)$. Observe that across all equilibria, the split disproportionately favors side $-i$ with degree of polarization $a_{-i} > a_i$. That is, the side that has lower degree of polarization tends to enjoy a larger share of the bargain. Another angle to understanding this result is through the effect the bargain splits have on interaction utilities. Specifically, Equation 4 shows that benefit from interacting on a common valuation is not constant implying that individuals that have similar preference for out-group aversion vs. in-group affinity (e.g., larger a_{-i}) need a larger share to offset their type-specific affective polarization disutility. Thus, sticking to in-group or strong aversion to out-groups can end up hurting one's chances at a better bargain share.

Proposition 4 suggests opportunity for better bargains without platform-driven mediation. Recent studies have shown that structured offline, in-person contacts can reduce affective polarization. Grounded in

Allport et al. (1954) seminal contact hypothesis and supported by meta-analytic evidence from Pettigrew and Tropp (2006), face-to-face encounters have been shown to reduce inter-group hostility and foster empathy by humanizing individuals beyond partisan labels. These interactions provide an environment where rival partisans can engage in genuine dialogue, challenge stereotypes, and share personal narratives, thereby softening the rigid boundaries that fuel affective divisions (Mutz 2006). Recent studies in political psychology further underscore that deliberative, interpersonal exchanges reduce hostility and enhance mutual understanding, paving the way for more constructive political discourse.

Recent practical attempts by charitable groups to bridge affective polarization have increasingly focused on building structured, offline exchanges that bring opposing partisans together in regulated settings. For example, programs like Living Room Conversations and Braver Angels enable face-to-face discussions that provide safe spaces for dialogue, lowering inter-group fear and neutralizing negative feelings toward political opponents (Oliver-Blackburn and Chatham-Carpenter 2023). Preliminary evaluations of these interventions indicate that moderated, in-person exchanges significantly decrease affective polarization, supporting the notion that structured interpersonal interactions can effectively supplement broader efforts to bridge partisan divides (Mason 2018). More recently, NACo's Bridging Divides Toolkit (McGuire et al. 2024) provides local leaders with intergroup strategies, such as organizing stereotype-challenging discussions, building inclusive public places, and identifying shared community challenges to facilitate collaborative problem-solving.

However, due to the recent wave of lawsuits filed by the public against social media, the involvement of the government and lawmakers cannot be overlooked. The following section considers this scenario, where, given the platform, a social planner sets subsidies or taxes to maximize overall social welfare.

4.4. Social Planner Equilibrium

Theorem 4 *Whether the social planner's maximizes the overall welfare of both consumers and the platform or just the consumers , the only feasible subsidy or tax strategy is offering: (1) maximum blanket or minimum Pigouvian subsidy; (2) minimum blanket or maximum Pigouvian tax to platform participants.*

Theorem 4 is established using Lemma 5 and Proposition 5. Specifically, the SPP model adds a stage before the platform sets prices, where a social planner sets the optimal subsidy or tax. Thus, the social planner's observed subsidy and tax allow the platform to strategize and set access fee that maximizes revenue from participating individuals. Lemma 5 establishes the equilibrium outcomes for the SPP game, while Proposition 5 establishes the social planner's strategy. We consider two social planner approaches. A discriminating approach enhances participant welfare but minimizes platform revenue; a non-discriminating approach maximizes both. The final incentive scheme depends on which approach dominates. Given the platform's access fee setting strategy, *ex-ante* conclusions about which approach dominates are non-trivial.

Lemma 5 *Assuming $x_i = \frac{1}{2}$. Then, under each incentive mechanism $k \in \{$*Blanket tax, Pigouvian tax, $\cdots\}$*, the equilibrium participation size, access fees, and optimal incentives are shown in Appendix C, where $i \in \{L, R\}$.*

SPP model analysis extends the PP model by exploring the optimal social planner-driven incentives scenario. Note that assuming $x_i = \frac{1}{2}$ allows for mathematical tractability and ease of comparison across different incentive schemes. Further, in each incentive setting game, the Stage 0 problem for the social planner (discriminating or not) becomes an optimization over a non-concave function. The interior solution turns out to be a saddle point, which ensures that corners are optimal. Stage 1 platform's revenue maximization problem follows certain constraints, which form the basis for the corner solutions for the social planner. Hence, the social planner sets only the maximum or minimum possible subsidies or taxes, as the scenario dictates. The underlying reasons behind setting these subsidy levels is explored in the following proposition.

Proposition 5 *When considering blanket incentives, the social planner sets maximum subsidy or minimum tax. However, when considering Pigouvian incentives, the social planner sets the minimum subsidy or maximum tax.*

Proposition 5's major contribution to the overall analysis is highlighting the impact incentive mechanisms have on driving participation vs. the platform's revenue maximization strategy. First, consider blanket incentives. In the subsidy scenario, Proposition 2 shows that blanket subsidies attract everyone to the platform, due to which demand elasticities are not guaranteed to be positive. Thus, the social planner faces a trade-off: set the minimum possible subsidy and let the platform set prices to attract participation or set the maximum subsidy that boosts demand for the platform. When discriminating against the platform, setting the minimum subsidy does hurt the platform's revenues by undercutting demand. However, it does nothing to maximize participant welfare, so it is ruled out as a possible solution. Consequently, setting the maximum subsidy remains the only solution for the social planner, irrespective of discriminating or non-discriminating objectives.

Similarly, consider the tax scenario. Again, Proposition 2's central insight on blanket tax indicates that such taxes uniformly dampen the demand for the platform. Thus, here, too, the social planner faces a trade-off: setting the minimum possible taxes boosts demand, helping the platform *capture value* while setting the maximum taxes decreases demand, hurts participant welfare, but reduces the platform's revenue too. When discriminating against the platform, the social planner cannot afford to hurt participant welfare. So, the platform ends up selecting minimum taxes. More generally, the major insight from blanket incentive analysis reflects some recent trends in the lawsuits and legislative efforts in the US (Prothero 2023). Lawmakers have increasingly argued that all-embracing benefits for all social media users could reduce affective polarization by ensuring diversified content consumption and decreasing algorithmic biases that support partisan echo

chambers (Mills 2023). This approach also proposes transparent algorithms, standardized content moderation, and universal digital literacy to level the playing field and create a healthier digital public sphere where no group disproportionately benefits.

Now consider Pigouvian incentives. In the subsidy scenario, recall that Proposition 2 shows that Pigouvian subsidies can help more in-group affinitive participants join the platform, thus intensifying the in-group cohesion, entrenching echo chambers and overall affective polarization. The social planner faces the following trade-off: either set maximum subsidy and allow more in-group affinitive users to join, boosting echo-chamber but also allowing the platform to *capture value*; or set the minimum subsidy, ensure echo-chambers are discouraged while also ensuring the platform does not over-exploit the resulting demand. Again, a discriminating social planner will not set a maximum subsidy since that will allow the platform to reap rewards from boosted demand. Moreover, comparing welfare under discrimination and non-discrimination shows that maximum subsidy results in higher access fees but lower participation, thus reducing the platform's revenue. Thus, the social planner will set the minimum possible Pigouvian subsidy that achieves both objectives, discriminating or otherwise.

Similarly, consider the tax scenario. Pigouvian taxes are the only incentive to improve participation diversity (Proposition 2). The social planner can intuitively set the minimum possible taxes to ensure diversified participation, while the platform also sets prices to maximize revenue. However, here, too, there is a potential dilemma: minimum taxes can potentially attract a relatively higher number of out-group-averse participants, degrade participant diversity, and thus decrease welfare while the platform exploits participation. Consequently, the platform sets the maximum possible taxes, ensuring relatively greater participation diversity, while the platform enjoys platform revenue from the diversified participation. Given rising polarization frictions, some lawmakers have called for targeted penalties for all parties, including users, marketers, and content providers. By fining or sanctioning provocative or divisive activity, this regulatory strategy reduces the incentives for extreme partisan language and promotes peaceful online conversation (Napoli 2019). This technique focuses on digital accountability, demonstrating that fair and persistent punishments can potentially disrupt the animosity that deepens polarization (Gorwa et al. 2020).

Overall, setting optimal incentives depends fundamentally on how the incentive impacts participation, tackles polarization, and thus, influences the platform's revenue maximization strategy. It is non-trivial to set such optimal incentives and Proposition 5 demonstrates typical trade-offs the social planner faces: profit maximization vs. consumer welfare. Recent efforts from lawmakers have oscillated between imposing public law norms through regulatory involvement or asking platforms to accept them voluntarily (Prothero 2023). However, bipartisan backing is insufficient to guarantee success for any legislative measures. This seeming deadlock informs the exploration of the further extensions to DAM and the related analysis in the following section.

5. Extensions

We extend DAM to explore the scenarios where the platform may 'enrage' one side to 'engage' the other (Section 5.1), is faced with trolls that participate to enrage the other side (Section 5.2), maximizes revenue when faced with a social planner that cares about the participation of the least polarized and therefore most 'vulnerable' side, i.e., Rawlsian planner (Section 5.3), and finally, competes with another platform for attracting users across two echo chambers (Section 5.4).

5.1. Enrage to Engage Strategy

Using 'Enrage to Engage' strategy, the platform tries to mediate between two sides by taxing the more polarized side (L WLOG), while subsidizing the less polarized side (R).

Proposition 6 *The 'Enrage to Engage' strategy has the following impact: only a Pigouvian tax on L improves participation diversity when $a_L > s_L$, while a subsidy on R worsens participation diversity. Additionally, only a Pigouvian subsidy to R guarantees consistently negative same-side and positive cross-platform demand elasticity, regardless of subsidies or taxes on L.*

Proposition 6 shows further evidence of the differential impact of Pigouvian vs. blanket incentives when mediating polarized participation. The primary strategy the platform considers is 'taxing' the more polarized side (L) while subsidizing the less polarized one (R). In doing so, the platform is *enraging* L side, while *engaging* R side. There are four cases to consider: (1) blanket overall, (2) blanket subsidy vs. Pigouvian tax, (3) blanket tax vs. Pigouvian subsidy, and (4) Pigouvian overall. Consider the first case – blanket overall. Intuitively, blanket subsidy to R boosts while blanket tax to L dampens demand. Yet, the two-sided nature of consumption does not guarantee such an outcome. Similar to Proposition 2, demand elasticities are not guaranteed negative for any side, implying that, for instance, a higher access fee to the side that is taxed might increase demand from that side if the other side decreases its demand. Thus, blanket incentives overall have similar impact as the symmetric scenario considered in Proposition 2.

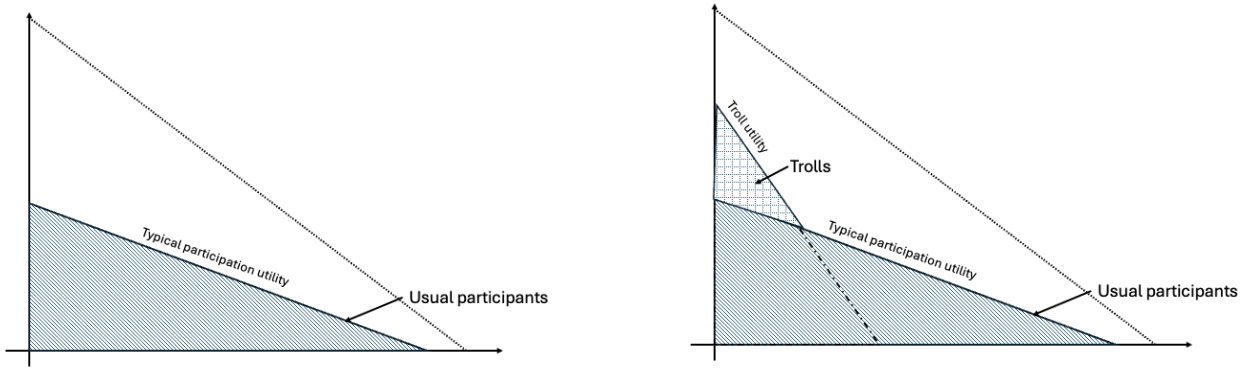
Next, consider the asymmetric cases (i.e., the second and third cases) – blanket subsidy (tax) vs. Pigouvian tax (subsidy). Observe that while blanket incentives impacts demand for all, Pigouvian incentives impacts participation from some. In effect, since participation sizes determine utility (Equation 1), if the size of those participating on R is comparable to the size discouraged on L, demand elasticities on both sides will not be straightforward. Thus, here too we have non-trivial elasticity outcomes (e.g., positive same side demand and negative cross-side elasticity under certain conditions). Moreover, subsidy on one side does not help improve participation diversity, which has an associated impact on the side being taxed. Consequently, in this case too, the platform fails to mediate polarized participation.

Finally, consider the fourth case – Pigouvian tax vs. Pigouvian subsidy. A Pigouvian incentive overall is a targeted incentive – with some targeted for penalties and some for rewards. Since the less polarized side

R is provided with a Pigouvian subsidy, their improved participation already hurts L. Further, a Pigouvian tax on L further dampens demand for the platform, resulting in a uniform overall demand elasticity – negative on the same and positive cross sides. Consequently, this scenario, like the blanket overall case, has similar outcomes to the symmetric case in Proposition 2. Recent research on social media platforms shows preferential content curation, escalating political division. These platforms use algorithmic tweaks to prioritize emotionally charged or contentious content to maximize user engagement, inflaming one partisan party while reinforcing the other (Avalle et al. 2024). This technique creates an asymmetry in engagement and emotional response by amplifying outrage-inducing information among reactive users and selectively delivering affirming content to moderate users (Bonini and Treré 2024). Such tailored content streams also create echo chambers, entrenching partisan divides and reducing cross-ideological dialogue (Lazer et al. 2018). Thus, Proposition 6 reflects these findings and further highlights social media algorithm design's influence on shaping political discourse.

5.2. Impact of Trolls

In this section, we explore demand elasticity and pricing asymmetry under the presence of 'trolls' on each side. Recent research shows that social media trolling intensifies hostile emotions and ideological divisions, increasing affective polarization. Trolling—provocative, aggressive, and sometimes anonymous online behavior—can increase in-group affinity and out-group aversion (Buckels et al. 2014). This phenomenon aligns with broader findings on affective polarization, where exposure to aggressive online content heightens negative perceptions of political opponents and diminishes the willingness for constructive dialogue (Iyengar et al. 2012). Experimental evidence further suggests that the incivility in trolling behaviors can amplify partisan animosity, thereby deepening societal divisions and challenging the integrity of digital discourse (Avalle et al. 2024).



(a) Typical equilibrium without trolls **(b)** Equilibrium with trolls

Figure 7 DAM Illustration with Trolls

Given this impact of trolls on polarized engagement, it is unclear ex-ante the impact trolls have on the demand for platform participation and associated incentives. We therefore extend DAM to model trolling activity on each side of the platform, by introducing troll participants that enjoy the following utility

$$u_{iT} = \underbrace{(v + \alpha_i(n_{-i} - n_{-iT}) - p_i)}_{\text{Net Benefits}} - \overbrace{(y + a_i x)}^{\text{User type}},\ i \in \{L, R\}. \tag{7}$$

Note the differences between Equation 7 and Equation 1. Further, the participation on each side is now composed of usual participants (indexed G) and those trolls (indexed T) that find it beneficial ($u_{iG} \geq 0, u_{iT} \geq 0$ and $u_{iT} \geq u_{iG}$) to participate. Assuming $l \in \{G, T\}$ as the index that denotes usual vs. troll users, the following lemma characterizes the equilibrium outcome for this case. At the same time, the Proposition 7 outlines the impact of trolls on pricing asymmetry and demand elasticities. Note that the troll and usual participants have individual participation sizes, but face the same access fee. Thus, their demand elasticities are different, too, as shown below.

Lemma 6 *When considering trolls, the following is the participation sizes, access fees and demand elasticities for usual participants and trolls ($l \in \{G, T\}$) across sides $i \in \{L, R\}$:*

Equilibrium under Impact of Trolls

1 *Participation Size*	$\hat{n}^*_{il} = \begin{cases} n^*_i, & l = G \\ \alpha_i x_{iT}\left(n^*_{-i} - \frac{n^*_i}{2}\right), & l = T \end{cases}$
2 *Access fees*	$\hat{p}^*_i = v - a_i x_i + \alpha_i n^*_i(1 - \alpha_{-i}x_{-iT}) - \alpha_i n^*_{-i}(1 - \frac{\alpha_{-i}x_{-iT}}{2})$
3 *Same Side Elasticity*	$\hat{\eta}^*_{iil} \coloneqq \begin{cases} \frac{\hat{p}^*_i}{n^*_i}\left(\frac{4x_i(\alpha_{-i}x_{-i}(\alpha_i x_{iT}-1)+2)}{\alpha_i x_i(\alpha_{-i}x_{-iT}(\alpha_{-i}x_{-i}(3\alpha x_{iT}-2)+8)-2(\alpha_i\alpha_{-i}x_{iT}x_{-i}+4))+8\alpha_{-i}x_{-i}(\alpha_i x_{iT}-1)+16}\right) \leq 0, & l = G \\ \frac{\hat{p}^*_i}{n^*_{iT}}\left(\frac{2\alpha_i x_i x_{iT}(\alpha_{-i}x_{-i}+2)}{\alpha_i x_i(\alpha_{-i}x_{-iT}(\alpha_{-i}x_{-i}(3\alpha_i x_{iT}-2)+8)-2(\alpha_i\alpha_{-i}x_{iT}x_{-i}+4))+8\alpha_{-i}x_{-i}(\alpha_i x_{iT}-1)+16}\right) \geq 0, & l = T \end{cases}$
4 *Cross Side Elasticity*	$\hat{\eta}^*_{i,-il} \coloneqq \begin{cases} \frac{\hat{p}^*_{-i}}{n^*_i}\left(-\frac{2\alpha_i x_i x_{-i}(\alpha_{-i}x_{-iT}-2)}{\alpha_i x_i(\alpha_{-i}x_{-iT}(\alpha_{-i}x_{-i}(3\alpha_i x_{iT}-2)+8)-2(\alpha_i\alpha_{-i}x_{iT}x_{-i}+4))+8\alpha_{-i}x_{-i}(\alpha_i x_{iT}-1)+16}\right) \geq 0, & l = G \\ \frac{\hat{p}^*_{-i}}{n^*_{iT}}\left(\frac{\alpha_i x_{iT}x_{-i}(\alpha_i x_i(2-3\alpha_{-i}x_{-iT})-8)}{\alpha_i x_i(\alpha_{-i}x_{-iT}(\alpha_{-i}x_{-i}(3\alpha_i x_{iT}-2)+8)-2(\alpha_i\alpha_{-i}x_{iT}x_{-i}+4))+8\alpha_{-i}x_{-i}(\alpha_i x_{iT}-1)+16}\right) \leq 0, & l = T \end{cases}$

where the usual participants, indexed with $l = G$, are shown without any subscript, while trolls are indexed with $l = T$.

Lemma 6 has several important implications for an equilibrium with trolls. First, observe that trolls equilibrium participation is determined by the number of usual participants on the other side n^*_{-i} but moderated by the number of usual participants on the same side n^*_i. That is, the more *net* rival participants, the greater the troll participation. Second, since the platform does not distinguish or discriminate between trolls and usual participants, the access fee now accounts for not just the net benefit externality from usual participants on the same side but also how such externality is moderated by the presence of trolls (i.e., $1 - \alpha_{-i}x_{-iT}$). That is, a greater troll population participating on the other side greatly diminishes the net benefits from participation on the same side. Finally, the demand elasticities are different since platform demands from usual and troll participants depend on a common access fee. Specifically, the usual participants demonstrate the same elasticity nature as in Lemma 1. However, trolls display the opposite nature of elasticity, thus establishing

the inherent difference between trolls and usual participants. The following proposition outlines more ways in which this equilibrium displays a different scenario compared to the benchmark.

Proposition 7 *In the presence of trolls, negative pricing asymmetry is not guaranteed. Moreover, the demand elasticity of trolls is exactly the inverse of that of usual participants.*

Proposition 7 essentially lays out the case of why a platform can tolerate and, in some cases, benefit from the presence of trolls. First, consider pricing asymmetry. By Definition 2, negative pricing asymmetry implies the platform charges a lower access fee, compared to the benchmark, to one side only. Intuitively, the presence of trolls should be treated as a threat to civil discourse on the platform, thus allowing a higher access fee charge to the side that hosts such trolls. However, the platform depends on participation revenue. Higher access fees can discourage not just trolls but also usual participants. Thus, the platform can lower access fees at the expense of the other side's disutility to maintain and optimize revenue.

Second, consider demand elasticity. Usual participants suffer affectively polarized interactions and thus display negative same-side and positive cross-side demand elasticities. Indeed, Lemma 1 illustrates this nature of demand sensitivity. However, observe that trolls, who derive utility from provoking the other side, exhibit the opposite demand elasticity. That is, trolls have a positive same-side demand elasticity and a negative cross-side demand elasticity. Again, the two-sided nature of platform demand and utility can explain this result. Trolls enjoy higher participation from the other side, and a higher access fee on the same side incentivizes greater participation from the other side – thus, positive same-side demand elasticity. Conversely, any changes in access fees that lower participation from the other side discourage troll participation. So, higher access fees on the other side dampen demand on the other side, thus lowering trolls' demand – negative cross-side demand elasticity.

The major takeaway from Proposition 7 points to a worrying aspect of troll-ridden online social media interactions. Recent studies show that online trolls may create perverse incentives for social media companies, which increasingly rely on engagement-based business models. Boffey (2025) argues that Facebook and Instagram algorithmically magnify controversial or inflammatory content— "rage-farming"—to improve user engagement, which boosts ad revenue. Moreover, a 2024 study indicated that undergraduate business students exposed to trolling spend 53% more time on social media (increasing to 4.7 hours from 2021), with Instagram and Snapchat witnessing 66% increases in trolling and user participation (Case et al. 2024). This feedback loop occurs when trolls use anonymity and platform structures to incite anger, and algorithms boost such content due to its 'virality' potential (Montez and Kim 2025). Proposition 7 breaks down these insights using elasticity and access fee trends that encourage the platform to tolerate trolls, thus impacting revenue but deteriorating participation utility.

These insights might raise the question - *does a social planner benefit from protecting the most vulnerable, rather than everyone*? The next extension essentially answers this question.

5.3. Rawlsian Social Planner

In the Rawlsian Social Planner scenario, the SP sets subsidies or taxes to the less polarized side R since they suffer the most from interacting with the more polarized side L. Essentially, the first stage objective function is now only the welfare (i.e., participation size) of side R.

Proposition 8 *The Rawlsian Social Planner sets minimum (maximum) taxes and maximum (minimum) subsidies under blanket (Pigouvian) incentives.*

Proposition 8 echoes the same insight as Proposition 5. Essentially, the social planner's strategy does not change even when planning to protect the most vulnerable. More broadly, lawyers and lawmakers have increasingly sued social media platforms, claiming their algorithms use polarized engagement to hurt vulnerable users, particularly minors. The 42 state attorneys general in a 2024 federal lawsuit accused Meta of using Instagram and Facebook features to "addict children and teens" by amplifying harmful, polarizing content and compulsive engagement, violating consumer protection laws and COPPA by secretly collecting data on under-13 users (James 2023). Courts have permitted novel legal strategies to bypass Section 230 immunity, such as Massachusetts' public nuisance claim asserting Meta's platform architecture—including infinite scroll and algorithmic recommendations—directly caused mental health harms like anxiety and eating disorders (Miller 2024).

Meanwhile, bipartisan litigation has targeted platforms' role in exacerbating political extremism, as seen in Texas and Florida laws (HB 20 and SB 7072) that sought to bar viewpoint-based moderation (Henshall 2024). However, courts have questioned their constitutionality while acknowledging platforms' systemic amplification of polarizing content. These instances emphasize reducing engagement-maximizing algorithm-fueled partisan hatred and protecting vulnerable populations like kids from addiction and grooming. More recently, in 2023–2024, federal judges allowed negligence claims against Meta and Snapchat, finding that platforms must prevent obsessive use and mental health crises in young users regardless of political alignment (Miller 2025). Additionally, Massachusetts' 2024 lawsuit against Meta emphasized how algorithmic curation of toxic content—regardless of partisan slant—violates consumer protection statutes by exploiting developmental vulnerabilities in minors (Miller 2024). Thus, when protecting the most vulnerable or everyone involved, the social planner's strategy remains the same - targeted incentives should not follow the provision modalities of blanket incentives.

5.4. Competing Platforms

In this section, we consider a duopoly – two competing platforms that attract participation from the two polarized sides. Each platform induces some polarization degree a_{L1}, a_{L2} and b_{R1}, b_{R2} given the net benefits and in- vs. out-group aversion of the users. Comparing a_{L1} with a_{L2} (or b_{R1} with b_{R2}) determines relative polarization degree.

Proposition 9 *When platforms compete for polarized users, relative polarization degrees determine participation size. Further, although one side may be relatively less polarized than the other, neither platform is guaranteed fewer participants from the less polarized side.*

Competition between platforms, given polarized participants can lead to the intuition that whichever platform faces participation from the side that is more polarized can drive away participation from the less polarized side. Indeed, asymmetrically polarized social media platforms risk alienating less polarized users through algorithmic amplification and unpleasant discourse. For instance, a 2025 analysis of 4.2 million Facebook postings from 31 election-year countries found that right-leaning accounts engaged 53% more on politically hot themes like migration than left-leaning ones on education and health (Pecile et al. 2025). This imbalance creates self-reinforcing cycles where platforms' engagement-based algorithms prioritize content from polarized groups. A 2024 X (Twitter) simulation showed verified ideologues with entrenched views increased network polarization by 38% through prioritized content distribution (de Arruda et al. 2024). Moreover, a 2025 study of Ukraine-related disputes showed that 72% of users who encountered polarized information passively consumed it, with multimedia and inaccurate news prevalence inversely associated (Baqir et al. 2025). Further, experimental research shows that partisan users share out-group insulting materials 67% more than neutral users, producing unfriendly environments that hinder moderate engagement (Smith et al. 2025). Thus, it can be natural to conclude that the more polarized dominate participation everywhere.

Yet, Proposition 9 shows such participation dominance is not always guaranteed. Specifically, given a platform, each side trades-off their in-group affinity with out-group aversion. Such a trade-off can result in heterogeneity in proportion of in-group affinitive vs. out-group averse participants on both sides. Consequently, less polarized side may have more out-group averse users join, thus increasing their size over the more polarized side. For instance, Facebook research shows high levels of polarized interactions and sustained involvement from less polarized individuals. A 2024 cross-platform investigation indicated that poisonous language generally arises from frequent partisan arguments, yet polarization does not prevent user engagement (Avalle et al. 2024). Facebook's balance of political and non-political content, interest groups, and local community posts creates 'attenuated polarization' where less partisan users can compartmentalize their relationships. Meta's 2024 data showed that 68% of centrist users avoided politically contentious forums and regularly participated in family or fitness networks (Zimmerman et al. 2024). Design choices allow users to self-segregate into interest-based subgroups, even as controversial content proliferates in other platform corners.

A 2025 longitudinal study found that moderate political Facebook users increased their posting frequency by 22% in non-political spaces compared to 2019 baselines, despite a 53% increase in polarized debate within political groups during election cycles (Durrheim and Schuld 2025). Researchers ascribe this

resilience to the platform's focus on strong-tie networks (friends and family) rather than weak-tie ideological communities, which decreases temptation to engage with divisive information. Facebook's default privacy settings and tailored feeds allow less polarized users to curate experiences that reflect local social norms rather than national political storylines, unlike Twitter (Stray et al. 2023). Critics say compartmentalization risks generating hyper-partisan and apolitical information ecosystems without cross-group communication. DAM's utility representation, especially related to user type-dependent trade-off between in-group affinity vs. out-group aversion helps model this aspect of polarized participation.

Still, across the fundamental and extensions analysis of DAM, a common theme emerges. The platform, far from being a 'town-square' where people can gather to argue and celebrate is increasingly becoming a focal point in the fight against vitriolic and divisive content consumption and production. While Section 230 and parts of Communication Decency Act in the US (Digital Services Act in EU) ensure limited liability of the platform as an entity in this scenario, not all stakeholders are convinced. Evidence of harms due to inaction of platforms is piling up and there are growing calls for platforms to 'do something' about this issue. The following section outlines some implications of DAM-driven insights generated.

6. Policy Implications

DAM presents a novel way to study affectively polarized interactions on social media platforms. Effectively, DAM analysis insights point to platforms' inability to mediate polarization frictions perfectly. However, there are opportunities for platforms to partially mediate polarization by either modifying access fees, diversifying participation, or altering demand elasticities that mirror typical two-sided platform characteristics. Our analysis points to incentive and control-driven mechanisms that can achieve at most two objectives. Thus, platforms have a significant role to play in mediating polarization.

Concerns are growing about the legal immunities granted to social media platforms and how these protections have worsened affective polarization in today's digital landscape. Immunity provisions, originally established under Section 230 of the Communications Decency Act and similar laws in other areas, were created during the early stages of the Internet when platforms were viewed simply as passive conduits for user-generated content. However, as social media firms have transformed into active curators and amplifiers of content through sophisticated, engagement-maximizing algorithms, these legal doctrines now seem outdated and ill-equipped to address the realities of modern content ecosystems. DAM emphasizes that platforms are no longer neutral intermediaries but rather influential actors that shape online discourse by promoting divisive and polarizing content.

Evidence suggests that recent judicial opinions have grown skeptical toward expansive interpretations of these immunities. For example, in August 2024, Judge Paul B. Matey of the Third Circuit notably criticized the current application of Section 230, arguing that it effectively shields corporations from any claim that may be even loosely related to content posted by third parties, regardless of the underlying cause of action

or the platform's involvement (Chrostowski 2024). Similar concerns have been voiced by judges in other circuits, indicating a potential shift in judicial attitudes that might eventually lead to a Supreme Court reexamination of these longstanding legal protections. This judicial discomfort is part of a broader critique questioning whether platforms should continue to benefit from an immunity framework that no longer reflects their active role in content distribution.

Still, analysis using DAM contrasts the traditional legal view of platforms as passive conduits and the emerging regulatory approaches in other jurisdictions. Underlying DAM's analysis is the observation that while the existing immunity doctrines were founded on the assumption of platforms being neutral infrastructure providers, new regulatory frameworks, such as the EU Digital Services Act and the UK Online Safety Act, explicitly acknowledge that these companies are active shapers of the digital content landscape (Kohl 2024). This disconnect between old legal doctrines and new regulatory realities highlights a fundamental problem: *platforms continue to enjoy broad immunity even as they utilize algorithms that actively amplify polarizing and harmful content.* As a result, legal scholars are advocating for a reformed approach in which only genuinely neutral platforms—those that do not invest in content curation or moderation—retain immunity (Botero Arcila and Griffin 2023). In contrast, platforms that leverage algorithmic tools to drive engagement should be held accountable for the adverse societal impacts of their practices – DAM's central premise.

In addition to the legal critique, various empirical studies reveal how algorithmic content amplification contributes to affective polarization. DAM's normative recommendations argue that the design of these algorithms, which prioritize sensational and emotionally charged material to maximize user engagement, inadvertently fosters an environment where divisive content thrives. DAM's analysis explains that this process intensifies political and social polarization, creating echo chambers where extreme viewpoints are continuously reinforced and deepening societal divisions. Therefore, DAM emphasizes that the responsibility for this harmful amplification should not be overlooked simply because outdated legal protections currently shield the platforms. It suggests that holding platforms accountable could incentivize them to adopt more balanced and responsible content distribution practices. The analysis further elaborates on several potential technical interventions to mitigate the impact of affective polarization. Among these, our study highlights control mechanisms (e.g., polarizing content warnings), which serve as a preemptive measure to alert users about the nature of the content they are about to encounter.

Overall, our analysis indicates that a synergistic approach, combining legal reform with technical interventions, represents the most promising pathway to address the challenges posed by algorithm-driven polarization. If legal frameworks were updated to limit immunity only to platforms that function as truly passive intermediaries, companies would be compelled to take greater responsibility for the negative externalities created by their algorithms. This legal accountability, paired with practical measures like content warnings and inoculation strategies, could fundamentally shift the incentives for social media companies. Instead of

prioritizing engagement at any cost, platforms might be encouraged to recalibrate their algorithms to promote a healthier, more balanced public discourse. Thus, DAM and its extensions are presented as a means to reduce the amplification of divisive content and foster a digital environment where free speech and open dialogue can coexist alongside a commitment to minimizing societal harm.

Further, empirical studies cited in our analysis have shown that such warnings can effectively reduce the perception of societal polarization and subsequently lower emotional responses to divisive content. Our analysis also indicates promising outcomes from inoculation interventions and educational tools designed to equip users with the skills to identify and resist hyper-partisan and manipulative content. These interventions, often delivered through video-based formats, have been demonstrated in experiments conducted in both the United States and the United Kingdom to significantly reduce user engagement with polarizing stimuli by making the underlying techniques more transparent (Smith et al. 2025). The research suggests that if platforms were required to integrate these interventions into their content delivery systems, it could help break the cyclical polarization pattern currently exacerbated by algorithmic amplification. Thus, a key premise of DAM's analysis stands vindicated - that platforms can and should mediate interactions between polarized populations.

7. Conclusion & Future Directions

Our research examines how digital platforms can address affective polarization among users, highlighting the complex interplay of user behavior, platform policies, and economic incentives. It discusses various strategies platforms might employ—such as bans, suspensions, or modifications to user interfaces—to mitigate polarization, yet finds that current attempts are mainly ineffective. The study introduces the Dolbear Axelrod Model, analyzing how subsidies and taxes can influence user participation based on their aversions. It argues that platforms often exploit polarization for profit, presenting a dilemma between maximizing participation and promoting welfare. Ultimately, it suggests that consumer-level strategies may offer better solutions than platform-based regulations, questioning whether traditional economic theories adequately address the challenges of polarization. Our main contention is that, unlike other types of markets, two-sided marketplaces are not intended to reduce affective polarization. A surplus that results from tax or subsidy initiatives can be captured by pricing. Thus, regulations that social planners lead pose a threat to "extremes," which come at the expense of the optimal. Our fervent hope is that our research informs calls for regulation- and business-model redesign of digital platforms policies.

References

Alexander, D. (2022). Banned Users Decry Nextdoor's Lack of Transparency - The Silicon Valley Voice.

Allianz (2023). 2024: A year of political turmoil and economic uncertainty. https://www.allianz.com/en/mediacenter/news/studies/231215-allianz-2024-a-year-of-political-turmoil-and-economic-uncertainty.html. Accessed: 2025-03-15.

Allport, G. W., Clark, K., and Pettigrew, T. (1954). *The nature of prejudice*. Addison-wesley publishing company Cambridge, MA.

Alter, A. (2017). *Irresistible: The rise of addictive technology and the business of keeping us hooked*. Penguin.

Anderson, L. (2024). How Politics Influence Dating Habits: Study.

Armstrong, M. (2006). Competition in two-sided markets. *The RAND journal of economics*, 37(3):668–691.

Avalle, M., Di Marco, N., Etta, G., Sangiorgio, E., Alipour, S., Bonetti, A., Alvisi, L., Scala, A., Baronchelli, A., Cinelli, M., et al. (2024). Persistent interaction patterns across social media platforms and over time. *Nature*, 628(8008):582–589.

Axelrod, R., Daymude, J. J., and Forrest, S. (2021). Preventing extreme polarization of political attitudes. *Proceedings of the National Academy of Sciences*, 118(50):e2102139118.

Azzimonti, M. (2023). Partisan conflict in the us and potential impacts on the economy. *Richmond Fed Economic Briefs*, 23(20):1–5.

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

Balkin, J. M. (2017). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UCDL rev.*, 51:1149.

Baqir, A., Chen, Y., Diaz-Diaz, F., Kiyak, S., Louf, T., Morini, V., Pansanella, V., Torricelli, M., and Galeazzi, A. (2025). Unveiling the drivers of active participation in social media discourse. *Scientific Reports*, 15(1):4906.

Baumol, W. J. (1972). On taxation and the control of externalities. *The American Economic Review*, 62(3):307–322.

Belleflamme, P. and Peitz, M. (2018). *Platforms and network effects*. Edward Elgar Publishing.

Belleflamme, P. and Peitz, M. (2019). Price disclosure by two-sided platforms. *International Journal of Industrial Organization*, 67:102529.

Belleflamme, P. and Peitz, M. (2021). *The economics of platforms*. Cambridge University Press.

Benkler, Y. (2007). The wealth of networks: How social production transforms markets and freedom.

Benkler, Y., Faris, R., and Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Benson, J. (2024). Democracy and the epistemic problems of political polarization. *American Political Science Review*, 118(4):1719–1732.

Boffey, M. (2025). Rage farming, OnlyFans and brain rot. *Campaign US*.

Böhme, E. (2016). Second-degree price discrimination on two-sided markets. *Review of Network Economics*, 15(2):91–115.

Bonini, T. and Treré, E. (2024). *Algorithms of resistance: The everyday fight against platform power*. MIT Press.

Botero Arcila, B. and Griffin, R. (2023). Social media platforms and challenges for democracy, rule of law and fundamental rights. Technical Report PE 743.400, Policy Department for Citizens' Rights and Constitutional Affairs, Directorate-General for Internal Policies, European Parliament. Study requested by the LIBE Committee. Available at: `http://www.europarl.europa.eu/supporting-analyses`.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., and Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Brailovskaia, J. and Margraf, J. (2024). From fear of missing out (fomo) to addictive social media use: The role of social media flow and mindfulness. *Computers in Human Behavior*, 150:107984.

Bucher, T. (2018). *If... then: Algorithmic power and politics*. Oxford University Press.

Buckels, E. E., Trapnell, P. D., and Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, 67:97–102. The Dark Triad of Personality.

Case, C. J., King, D. L., and Case, J. A. (2024). Undergraduate business students social media and online trolling behavior: Trends during and after covid-19. In *Proceedings of International Academic Conferences*, number 14616334 in 00. International Institute of Social and Economic Sciences.

Chang, H.-C. H., Druckman, J., Ferrara, E., and Willer, R. (2023). Liberals engage with more diverse policy topics and toxic content than conservatives on social media. *Institute of Policy Research*.

Cheadle, H. (2021). Why everyone is so mad at Substack all the time.

Chillingworth, A. (2025). New social media platforms [2025] — Epidemic Sound.

Chrostowski, G. (2024). ANALYSIS: Social Media Immunity Inches Closer to Supreme Court.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Coase, R. H. (2013). The problem of social cost. *The Journal of Law and Economics*, 56(4):837–877.

Combs, A., Tierney, G., Guay, B., Merhout, F., Bail, C. A., Hillygus, D. S., and Volfovsky, A. (2023). Reducing political polarization in the united states with a mobile chat platform. *Nature human behaviour*, 7(9):1454–1461.

de Arruda, H. F., Oliveira, K. A., and Moreno, Y. (2024). Echo chamber formation sharpened by priority users. *iScience*, 27(11).

de Zúñiga, H. G., González-González, P., and Goyanes, M. (2025). Pathways to political persuasion: Linking online, social media, and fake news with political attitude change through political discussion. *American Behavioral Scientist*, 69(2):240–261.

Demsetz, H. (1967). Towards a theory of property rights. *The American Economic Review*, 57(2):347–359.

Dolbear, F. T. (1967). On the theory of optimum externality. *The American Economic Review*, 57(1):90–103.

Durrheim, K. and Schuld, M. (2025). Polarization on social media: Comparing the dynamics of interaction networks and language-based opinion distributions. *Political Psychology*, n/a(n/a):00.

Eisenmann, T., Parker, G., and Van Alstyne, M. (2006). Strategies for Two-Sided Markets. *Harvard Business Review*, 1463:https://hbr.org/2006/10/strategies–for–two–sided–markets.

Fardouly, J., Diedrichs, P. C., Vartanian, L. R., and Halliwell, E. (2015). Social comparisons on social media: The impact of facebook on young women's body image concerns and mood. *Body image*, 13:38–45.

Fiorina, M. P., Abrams, S. J., and Pope, J. (2004). *Culture War? The Myth of a Polarized America*. Longman.

Fuchs, C. (2021). *Social media: A critical introduction*. SAGE publications Ltd.

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of computer-mediated communication*, 14(2):265–285.

Garzia, D., Ferreira da Silva, F., and Maye, S. (2023). Affective polarization in comparative and longitudinal perspective. *Public Opinion Quarterly*, 87(1):219–231.

Gerrard, Y. (2020). Behind the screen: Content moderation in the shadows of social media. *New Media & Society*, 22(3):579–582.

Gillespie, T. (2014). 9 the relevance of algorithms. *Media Technologies: Essays on Communication, Materiality, and Society*, page 167.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moder ation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Gillissen, M., Rooduijn, M., and Schumacher, G. (2023). Empathic concern and perspective-taking have opposite effects on affective polarization. *Journal of Experimental Political Science*, pages 1–19.

González-Bailón, S., d'Andrea, V., Freelon, D., and De-Domenico, M. (2022). The advantage of the right in social media news sharing. *PNAS Nexus*, 1(3):pgac137.

González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A. M., Iyengar, S., Kim, Y. M., Malhotra, N., Moehler, D., Nyhan, B., Pan, J., Rivera, C. V., Settle, J., Thorson, E., Tromble, R., Wilkins, A., Wojcieszak, M., de Jonge, C. K., Franco, A., Mason, W., Stroud, N. J., and Tucker, J. A. (2023). Asymmetric ideological segregation in exposure to political news on facebook. *Science*, 381(6656):392–398.

Gorwa, R. (2019). The platform governance triangle: Conceptualising the informal regulation of online content. *Internet Policy Review*, 8(2):1–22.

Gorwa, R., Binns, R., and Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Greene, K. F., Rossiter, E. L., Seira, E., and Simpser, A. (2024). Interacting as equals reduces partisan polarization in mexico. *Nature Human Behaviour*, pages 1–9.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., Thorson, E., Tromble, R., Wilkins, A., Wojcieszak, M., Xiong, B., de Jonge, C. K.,

Franco, A., Mason, W., Stroud, N. J., and Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404.

Henshall, W. (2024). The Supreme Court could determine the future of social media content moderation. *TIME*.

Howard, J. W. and Rothbart, M. (1980). Social categorization and memory for in-group and out-group behavior. *Journal of personality and social psychology*, 38(2):301.

Hwang, T. (2020). *Subprime attention crisis: Advertising and the time bomb at the heart of the Internet*. FSG originals.

Ichihashi, S. and Kim, B.-C. (2023). Addictive platforms. *Management Science*, 69(2):1127–1145.

Ingram, D. and Brooks, A. (2023). Conservative social media app Parler planning to relaunch ahead of 2024 election.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual review of political science*, 22(1):129–146.

Iyengar, S., Sood, G., and Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public opinion quarterly*, 76(3):405–431.

Iyengar, S. and Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American journal of political science*, 59(3):690–707.

James, L. (2023). Attorney General James and multistate coalition sue Meta for harming youth.

Jelveh, Z., Kogut, B., and Naidu, S. (2024). Political language in economics. *The Economic Journal*, 134(662):2439–2469.

Jeon, D.-S., Kim, B.-C., and Menicucci, D. (2022). Second-degree price discrimination by a two-sided monopoly platform. *American Economic Journal: Microeconomics*, 14(2):322–369.

Kaufhold, K. (2024). Does tribe trump facts? novel measures of hostile media. *Electronic News*, 0(0):19312431241277960.

Kobayashi, T., Zhang, Z., and Liu, L. (2024). Is partisan selective exposure an american peculiarity? a comparative study of news browsing behaviors in the united states, japan, and hong kong. *Communication Research*, 0(0):00936502241289109.

Kohl, U. (2024). Toxic recommender algorithms: immunities, liabilities and the regulated self-regulation of the digital services act and the online safety act. *Journal of Media Law*, pages 1–35.

Kox, H., Straathof, B., and Zwart, G. (2017). Targeted advertising, platform competition, and privacy. *Journal of economics & management strategy*, 26(3):557–570.

Kreps, D. M. (2013). *Microeconomic foundations*, volume 1. Princeton university press.

Kreps, D. M. (2020). *A course in microeconomic theory*. Princeton university press.

Krishan, N. (2021a). Challenge for conservative social media platforms: attracting 'libs' to own - Washington Examiner. *Washington Examiner*.

Krishan, N. (2021b). Parler attempts to rebrand itself as nonpartisan and mainstream - Washington Examiner. *Washington Examiner*.

Kuang, P., Weber, M., and Xie, S. (2024). Perceived political bias of the federal reserve. Technical report, National Bureau of Economic Research.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lydia (2020). How do we combat polarization online? - the new modality.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. (1995). *Microeconomic theory*, volume 1. Oxford university press New York.

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. University of Chicago Press.

McClain, C., Anderson, M., and Gelles-Watnick, R. (2024). How Americans navigate politics on TikTok, X, Facebook and Instagram.

McGuire, M., Ring, J., and Bartholomay, D. (2024). Bridging Divides: Tools for county leaders to overcome division and foster collaboration. Technical report, National Association of Counties (NACo).

McIntyre, D. P. and Srinivasan, A. (2017). Networks, platforms, and strategy: Emerging views and next steps. *Strategic management journal*, 38(1):141–160.

Miller, Jr, B. R. V. (2025). Social media addiction lawsuit.

Miller, G. (2024). Social media lawsuits by state Attorneys General Surmount Section 230, other challenges.

Mills, G. (2023). A contractual approach to social media governance. *Yale L. & Pol'y Rev.*, 42:522.

Mochon, D. and Schwartz, J. (2024). The confrontation effect: When users engage more with ideology-inconsistent content online. *Organizational Behavior and Human Decision Processes*, 185:104366.

Montez, D. and Kim, D. H. (2025). How do silent trolls become overt trolls? fear of punishment and online disinhibition moderate the trolling path. *Social Media+ Society*, 11(1):20563051251320437.

Motyl, M. (2023). The Political Attitudes of Social Media Users. *Designing Tomorrow*.

Mutz, D. C. (2006). *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.

Napoli, P. (2019). *Social media and the public interest: Media regulation in the disinformation age*. Columbia university press.

Nault, B. R. (1996). Equivalence of taxes and subsidies in the control of production externalities. *Management Science*, 42(3):307–320.

Nazarro, M. (2024). Democrats flee X for Bluesky amid Musk-Trump alliance .

Norman, J. M. and Green, B. (2025). Why can't we be friends? untangling conjoined polarization in america. *Political Psychology*, 0(0):1–22.

Nossel, S. (2025). What do Meta's changes mean for Facebook users around the world? - Los Angeles Times. *LA Times*.

Oldemburgo de Mello, V., Cheung, F., and Inzlicht, M. (2024). Twitter (x) use predicts substantial changes in well-being, polarization, sense of belonging, and outrage. *Communications Psychology*, 2(1):15.

Oliver-Blackburn, B. M. and Chatham-Carpenter, A. (2023). 'but i don't know if i want to talk to you': strategies to foster conversational receptiveness across the united states' political divide. *Journal of Applied Communication Research*, 51(1):55–71.

Page, S. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition*. Princeton University Press.

Parke, C. (2019). Conservatives call for PayPal boycott after CEO says Southern Poverty Law Center helps ban users.

Pecile, G., Di Marco, N., Cinelli, M., and Quattrociocchi, W. (2025). Mapping the global election landscape on social media in 2024. *PLOS ONE*, 20(2):1–16.

Peitz, M. (2024). The economic theory of two-sided platforms. In Sokol, D., Dukes, A., Sun, T., and Zhu, F., editors, *Cambridge Handbook on Digital Platforms*, chapter 8, pages 1–22. Cambridge University Press, Cambridge.

Pennycook, G. and Rand, D. G. (2022). Nudging social media toward accuracy. *The Annals of the American Academy of Political and Social Science*, 700(1):152–164.

Pettigrew, T. F. and Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology*, 90(5):751.

Prothero, A. (2023). Lawmakers want to regulate young social media users. Will it work? *Education Week*.

Przybylski, A. K., Murayama, K., DeHaan, C. R., and Gladwell, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in Human Behavior*, 29(4):1841–1848.

Rathje, S., Robertson, C., Brady, W. J., and Van Bavel, J. J. (2024). People think that social media platforms do (but should not) amplify divisive content. *Perspectives on Psychological Science*, 19(5):781–795.

Rochet, J.-C. and Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the european economic association*, 1(4):990–1029.

Roozenbeek, J. and Van der Linden, S. (2024). *The psychology of misinformation*. Cambridge University Press.

Rudolph, T. J. and Hetherington, M. J. (2021). Affective polarization in political and nonpolitical settings. *International Journal of Public Opinion Research*, 33(3):591–606.

Rysman, M. (2009). The economics of two-sided markets. *Journal of Economic Perspectives*, 23(3):125–143.

Samuelson, P. (2024). Do social media platforms have free speech rights to 'censor'conservatives? *Communications of the ACM*, 67(11):24–26.

Saveski, M., Gillani, N., Yuan, A., Vijayaraghavan, P., and Roy, D. (2022). Perspective-taking to reduce affective polarization on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 885–895.

Schedler, A. (2023). Rethinking political polarization. *Political science quarterly*, 138(3):335–359.

Smith, F., Simchon, A., Holford, D., and Lewandowsky, S. (2025). Inoculation reduces social media engagement with affectively polarized content in the uk and us. *Communications Psychology*, 3(1):11.

Stephens, N. (2024). Nextdoor neighbors and cancel culture.

Stewart, A. J., McCarty, N., and Bryson, J. J. (2020). Polarization under rising inequality and economic decline. *Science Advances*, 6(50):eabd4201.

Stray, J., Iyer, R., and Puig Larrauri, H. (2023). The algorithmic management of polarization and violence on social media.

Sunstein, C. R. (2002). The law of group polarization. *The Journal of Political Philosophy*, 10(2):175–195.

Tajfel, H., Turner, J. C., Austin, W. G., et al. (1982). Social psychology of intergroup relations. *Annual review of psychology*, 33(1):1–39.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., and Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Tufekci, Z. (2014). Engineering the public: Big data, surveillance and computational politics. *First Monday*.

Van Dijck, J. (2013). *The Culture of Connectivity: A Critical History of Social Media*. Oxford University Press.

Versteegen, P. L. (2024). We love, they hate: Emotions in affective polarization and how partisans may use them. *Political Psychology*.

Vogels, E., Auxier, B., and Anderson, M. (2024). Partisan differences in social media use show up for some platforms, but not Facebook.

Votta, F., Kruschinski, S., Hove, M., Helberger, N., Dobber, T., and de Vreese, C. (2024). Who does (n't) target you? mapping the worldwide usage of online political microtargeting. *Journal of Quantitative Description: Digital Media*, 4.

Wagner, M. and Praprotnik, K. (2024). Affective polarization and coalition signals. *Political Science Research and Methods*, 12(2):336–353.

West, E. A. and Iyengar, S. (2022). Partisanship as a social identity: Implications for polarization. *Political Behavior*, 44(2):807–838.

Whitaker, B. (2022). Social media's role in America's polarized political climate - 60 Minutes.

Zaman, T. (2024). Do social media platforms suspend conservatives more?

Zimmerman, F., Bailey, D. D., Muric, G., Ferrara, E., Schöne, J., Willer, R., Halperin, E., Navajas, J., Gross, J. J., and Goldenberg, A. (2024). Attraction to politically extreme users on social media. *PNAS Nexus*, 3(10):pgae395.

Appendix A: Incentive Mechanism Equilibria

Incentive	Equilibrium Outcome	Value
Blanket Subsidy (BS)	Participation Size $n^*_{i,BS}$	$\frac{1}{2}a_i(x_i+s_i)^2$
	Access Fee $p^*_{i,BS}$	$v-a_ix_i+\alpha_i\left(n^*_{i,BS}-n^*_{-i,BS}\right)$
	Same Side Elasticity $\eta^*_{ii,BS}$	$\left(\frac{(x_i+s_i)^2\left(2x_{-i}-\alpha_{-i}(x_{-i}+s_{-i})^2\right)}{2\alpha_ix_{-i}(x_i+s_i)^2+2\alpha_{-i}x_i(x_{-i}+s_{-i})^2-4x_ix_{-i}}\right)\frac{p^*_{i,BS}}{n^*_{i,BS}}$
	Cross Side Elasticity $\eta^*_{i,-i,BS}$	$\left(-\frac{\alpha_i(x_i+s_i)^2(x_{-i}+s_{-i})^2}{2\alpha_ix_{-i}(x_i+s_i)^2+2\alpha_{-i}x_i(x_{-i}+s_{-i})^2-4x_ix_{-i}}\right)\frac{p^*_{-i,BS}}{n^*_{i,BS}}$
Blanket Tax (BT)	Participation Size $n^*_{i,BT}$	$\frac{1}{2}a_i(x_i-s_i)^2$
	Access Fee $p^*_{i,BT}$	$v-a_ix_i+\alpha_i\left(n^*_{i,BT}-n^*_{-i,BT}\right)$
	Same Side Elasticity $\eta^*_{ii,BT}$	$\left(-\frac{(x_i-s_i)^2\left(\alpha_{-i}(x_{-i}-s_{-i})^2-2x_{-i}\right)}{2\alpha_ix_{-i}(x_i-s_i)^2+2\alpha_{-i}x_i(x_{-i}-s_{-i})^2-4x_ix_{-i}}\right)\frac{p^*_{i,BT}}{n^*_{i,BT}}\leq 0$
	Cross Side Elasticity $\eta^*_{i,-i,BT}$	$\left(-\frac{\alpha_i(x_i-s_i)^2(x_{-i}-s_{-i})^2}{2\alpha_ix_{-i}(x_i-s_i)^2+2\alpha_{-i}x_i(x_{-i}-s_{-i})^2-4x_ix_{-i}}\right)\frac{p^*_{-i,BT}}{n^*_{i,BT}}\geq 0$
Pigouvian Subsidy (PS)	Participation Size $n^*_{i,PS}$	$\frac{a_ix_i^2}{2s_i}$
	Access fee $p^*_{i,PS}$	$v-a_ix_i+\alpha_i\left(n^*_{i,PS}-n^*_{-i,PS}\right)$
	Same Side Elasticity $\eta^*_{ii,PS}$	$\left(-\frac{x_i(\alpha_{-i}x_{-i}-2s_{-i})}{2(\alpha_ix_is_{-i}+\alpha_{-i}x_{-i}s_i-2s_is_{-i})}\right)\frac{p^*_{i,PS}}{n^*_{i,PS}}\leq 0$
	Cross Side Elasticity $\eta^*_{i,-i,PS}$	$\left(\frac{\alpha_ix_ix_{-i}}{-2\alpha_ix_is_{-i}-2\alpha_{-i}x_{-i}s_i+4s_is_{-i}}\right)\frac{p^*_{-i,PS}}{n^*_{i,PS}}\geq 0$
Pigouvian Tax (PT)	Participation Size $n^*_{i,PT}$	$\frac{1}{2}a_ix_i^2s_i$
	Access fee $p^*_{i,PT}$	$v-a_ix_i+\alpha_i\left(n^*_{i,PT}-n^*_{-i,PT}\right)$
	Same Side Elasticity $\eta^*_{ii,PT}$	$\left(\frac{x_is_i(2-\alpha_{-i}x_{-i}s_{-i})}{2(\alpha_ix_is_i+\alpha_{-i}x_{-i}s_{-i}-2)}\right)\frac{p^*_{i,PT}}{n^*_{i,PT}}\leq 0$
	Cross Side Elasticity $\eta^*_{i,-i,PT}$	$\left(\frac{\alpha_ix_ix_{-i}s_is_{-i}}{-2\alpha_ix_is_i-2\alpha_{-i}x_{-i}s_{-i}+4}\right)\frac{p^*_{-i,PT}}{n^*_{i,PT}}\geq 0$

Table 3 Equilibria outcomes along the four parameters analyzed for Benchmark Model in Lemma 1

Appendix B: Control Mechanism Equilibria

Control Mechanism	Equilibrium Outcome	Value
Restricting Participation (RP)	Participation Size $n^*_{i,PS}$	$\frac{1}{2}a_i x_i^2 \eta,\ \eta \in (0,1)$ (less than Benchmark)
	Access Fee $p^*_{i,RP}$	$v - a_i x_i + \alpha_i \left(n^*_{i,RP} - n^*_{-i,RP}\right)$
	Same Side Elasticity $\eta^*_{ii,RP}$	$\left(\frac{\eta x_i(\alpha_{-i}x_{-i}-2)}{2(2-\alpha_i\eta x_i-\alpha_{-i}x_{-i})}\right)\frac{p^*_{i,RP}}{n^*_{i,RP}} \leq 0$
	Cross Side Elasticity $\eta^*_{i,-i,RP}$	$\left(\frac{\alpha_i\eta x_i x_{-i}}{4-2\alpha\eta x_i-2\alpha_{-i}x_{-i}}\right)\frac{p^*_{-i,RP}}{n^*_{i,RP}} \geq 0$
Interaction Obfuscation (IO)	Participation Size $n^*_{i,IO}$	n^*_i (Benchmark)
	Access Fee $p^*_{i,IO}$	$v - a_i x_i + \alpha_{is} n^*_i - \alpha_{ic} n^*_{-i}$
	Same Side Elasticity $\eta^*_{ii,IO}$	$\left(\frac{x_i(x_{-i}\alpha_{-is}-2)}{x_i\alpha_{is}(x_{-i}\alpha_{-is}-2)-x_i x_{-i}\alpha_{ic}\alpha_{-ic}-2x_{-i}\alpha_{-is}+4}\right)\frac{p^*_{i,IO}}{n^*_{i,IO}} \leq 0$
	Cross Side Elasticity $\eta^*_{i,-i,IO}$	$\left(\frac{x_i x_{-i}\alpha_{ic}}{x_i\alpha_{is}(x_{-i}\alpha_{-is}-2)-x_i x_{-i}\alpha_{ic}\alpha_{-ic}-2x_{-i}\alpha_{-is}+4}\right)\frac{p^*_{-i,IO}}{n^*_{i,IO}} \geq 0$
Manipulating Valuation (MV)	Participation Size $n^*_{i,MV}$	n^*_i (Benchmark)
	Access fee $p^*_{i,MV}$	$v_i - a_i x_i + \alpha_i \left(n^*_{i,MV} - n^*_{-i,MV}\right)$
	Same Side Elasticity $\eta^*_{ii,MV}$	η^*_{ii} (Benchmark)
	Cross Side Elasticity $\eta^*_{i,-i,MV}$	$\eta^*_{i,-i}$ (Benchmark)
Platform Versioning (PV)	Participation Size $n^*_{i,PV}$	n^*_i (Benchmark)
	Access fee $p^*_{i,PV}$	$v - a_i x_i + \alpha_i \left(n^*_{i,PV} - \frac{1}{2}a_i x_i \epsilon\right)$
	Same Side Elasticity $\eta^*_{ii,PV}$	$\left(\frac{x_i}{2-\alpha_i x_i}\right)\frac{p^*_{i,PV}}{n^*_{i,PV}}$
	Cross Side Elasticity $\eta^*_{i,-i,PV}$	0

Table 4 Equilibria outcomes along the four parameters analyzed for Benchmark Model in Lemma 1

Appendix C: Social Planner Equilibria

Incentive	Equilibrium Outcome	Value
Social Planner's Pigouvian Subsidy ($SPPS$)	Participation Size $n^*_{i,SPPS}$	$\frac{1}{4}a_i$
	Access Fee $p^*_{i,SPPS}$	$v - \frac{a_i}{2} + \alpha_i \left(n^*_{i,SPPS} - n^*_{-i,SPPS}\right)$
	Optimal Subsidy $s^*_{i,SPPS}$	$\frac{1}{2}$
Social Planner's Blanket Subsidy ($SPBS$)	Participation Size $n^*_{i,SPBS}$	$\frac{1}{2}a_i$
	Access Fee $p^*_{i,SPBS}$	$v - \frac{a_i}{2} + \alpha_i \left(n^*_{i,SPBS} - n^*_{-i,SPBS}\right)$
	Optimal Subsidy $s^*_{i,SPBS}$	$\frac{1}{2}$
Social Planner's Pigouvian Tax ($SPPT$)	Participation Size $n^*_{i,SPPT}$	$\frac{1}{8}a^*_i$
	Access fee $p^*_{i,SPPT}$	$v - \frac{a_i}{2} + \alpha_i \left(n^*_{i,SPPT} - n^*_{-i,SPPT}\right)$
	Optimal Tax $s^*_{i,SPPT}$	1
Social Planner's Blanket Tax ($SPBT$)	Participation Size $n^*_{i,SPBT}$	$\frac{1}{8}a_i$
	Access fee $p^*_{i,SPBT}$	$v - \frac{a_i}{2} + \alpha_i \left(n^*_{i,SPBT} - n^*_{-i,SPBT}\right)$
	Optimal Tax $s^*_{i,SPBT}$	0

Table 5 Equilibria outcomes along the four incentives analyzed in Lemma 2