Reaching a Consensus in Predictive Loops

Jiduan Wu[1,2], Rediet Abebe[*1,3], and Celestine Mendler-Dünner[*1,3]

[1]*Max Planck Institute for Intelligent Systems, Tübingen, and Tübingen AI Center*
[2]*ETH Zurich*
[3]*ELLIS Institute Tübingen*

Abstract

Predictions in digital platforms must adapt over time as individuals update their beliefs through social interactions. At the same time, changing predictions alter the content people are exposed to and, consequently, the very beliefs they aim to forecast. This recursive coupling between predictions and outcomes complicates the analysis of the long-term societal impact of predictive systems. In this work, we propose a minimal model combining opinion dynamics with concepts from performative prediction to highlight the qualitative implications of predictive loops on equilibrium outcomes in social networks. We assume a platform's predictions influence each individual's opinion, which then evolve through peer interactions and form the training data for future platform model updates. In this model, we show how standard predictive objectives can drive networks toward consensus even under conditions where classical opinion-dynamics models lead to disagreement. This emerges because predictive systems dynamically adapt to changing opinions, and learning objectives create spillover effects among individuals beyond the topology of the network. We complement our non-parametric theory with simulations of popular parametric learners on real social network data. Together, our results illustrate performativity as an important, yet so far neglected, qualifying factor in social networks.

1 Introduction

The formation of opinions has traditionally been modeled as a social process unfolding through interpersonal influence: individuals exchange views with their peers, update their beliefs, and gradually cluster or polarize depending on the structure of the network and the strength of social ties. From classical models of diffusion (DeGroot, 1974; Holley and Liggett, 1975; Friedkin and Johnsen, 1990) to contemporary theories of polarization (Bail et al., 2018), opinion dynamics have focused on how local interactions between peers aggregate into collective outcomes. However, this assumption no longer accurately reflects how opinions form in today's society where digital platforms play an increasingly dominant role as a source of information.

The effect of digital platforms on beliefs and opinions has been documented in various settings, ranging from preferences for restaurants (Anderson and Magruder, 2012) and travel providers (Ursu, 2018) to political leaning (Gauthier et al., 2026; Huszár et al., 2022) and cultural tastes (Salganik et al., 2006). To account for such platform influence on the formation of opinions, several extensions of classical opinion dynamics models have been proposed. Among others, they discuss how platforms change beliefs, modify social ties, and mediate trust. However, these approaches predominantly treat platforms as static or exogenous to the opinion formation process (Out et al., 2024). As a result, they miss how digital platforms respond to changing opinions and repeatedly distill aggregate information into future decisions.

[*]Joint supervision.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

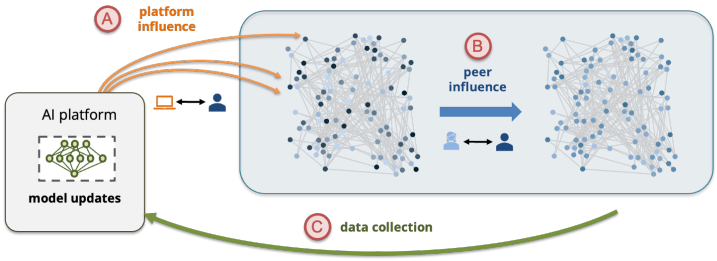


Figure 1: *Peer-platform co-influence in opinion dynamics*: (A) Individuals interact with a digital platform that deploys a predictive system and update their opinions based on the content they are exposed to. (B) Individuals' opinions evolve through peer interaction. (C) Expressed opinions inform future platform predictions.

To address this, the core focus of this work lies in the interaction of individuals with machine-learning-powered systems. These are algorithms that are continuously fed with behavioral data about users, make predictions about user preferences and determine what content to show next. As argued above, even when framed as purely predictive, such algorithmic decisions shape downstream user behavior and future data observed by the algorithm—a phenomenon known as *performativity*. As a consequence, learning systems are not merely descriptive or static; rather, they co-evolve with the data they create. This dynamic is conceptualized in the machine learning literature under the umbrella of performative prediction (Perdomo et al., 2020; Hardt and Mendler-Dünner, 2025). Applied to opinion dynamics, performativity implies that any platform applying predictive algorithms, necessarily becomes an active participant in the opinion formation process. This socio-technical feedback-loop is illustrated in Figure 1. Predictions and opinions are recursively coupled: observations of opinions determine predictions and the performative effects of predictions determine opinions. We refer to this feedback mechanism, characteristic of machine learning powered platforms, as the predictive loop.

1.1 Our work

We blend models of social influence (Friedkin and Johnsen, 1990) with concepts from performative prediction (Perdomo et al., 2020) and model the co-evolution of opinions and predictions as a nested feedback loop. The inner-loop describes how opinions evolve through peer interactions, and the outer-loop describes how platforms influence the opinion formation process and subsequently learn from observations of expressed opinions to determine future predictions. An equilibrium state emerges where the platform has no reason to change its predictions based on observed data. This new equilibrium state is not only determined by network characteristics and individuals' susceptibility to peer influence, as in traditional network models, but also the platform's choice of learning algorithm and individuals' susceptibility to platform influence. With technical results and simulations, we perform investigations into the characteristics of this new equilibrium and illustrate interesting deviations from the classical network equilibria absent performativity.

More concretely, we instantiate peer dynamics with the classical Friedkin-Johnsen (FJ) model (Friedkin and Johnsen, 1990), and focus on self-fulfilling predictions on the side of the platform. Through a theoretical analysis of non-parametric prediction and extensive simulations using parametric models, our work offers the following key take-aways:

- *Predictive loops as an explanation for consensus.* We show that by moving with expressed opinions, performative predictions can act as a homogenizing force in networks. For our model we prove that under natural assumptions on the network, the variance of opinions at equilibrium is negatively correlated with the susceptibility of individuals to platform influence, or, in other words, the strength of performativity. Notably, in the limit of high platform susceptibility, a consensus is reached. This consensus emerges beyond primitive graphs and despite stubbornness during peer interactions—standard conditions under which classical results in networks would suggest disagreement.
- *Performativity as a pathway for peer influence.* We illustrate how extrapolating predictions to unobserved individuals can induce consensus under even weaker conditions on the network. In particular, we prove that it is sufficient for information to flow in the observed subgroup for a

consensus to emerge across the entire population, again, in the limit of high platform susceptibility. The reason is that individuals who are not susceptible to peer influence can indirectly be impacted by the opinions of their peers through model predictions. We offer a lower bound on this performativity-mediated spillover effect unique to predictive loops. Similar in vein, we show how interventions targeted at a single individual impact equilibrium opinions across the entire population in a way that is correlated with the strength of performativity.

Both of these take-aways constitutes an important conceptual deviation from standard theories of networks and showcase performativity as an important qualifying factor in social dynamics.

2 Preliminaries and related work

Throughout this work, graphs serve as a mathematical object to characterize interpersonal relationships in a population of individuals. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected simple graph where $\mathcal{V} = [n]$ is the node set and $\mathcal{E} = \{(i, j) \mid i \text{ and } j \text{ are connected}\}$ represents the set of edges. The adjacency matrix of $\mathcal{G}$ is denoted as A where $a_{ij} = 1$ if $(i, j) \in \mathcal{E}$, and $a_{ij} = 0$ otherwise. We denote the degree of each node as d_i and use $\deg := \mathrm{diag}(d_1, \cdots, d_n)$. Each individual i on the network holds a time-dependent opinion x_i. Let $\boldsymbol{x} = [x_1, ..., x_n]$ denote opinions of all n individuals. These opinions are our subject of study and they can be influenced by both peer interactions and platform predictions.

2.1 Peer influence

An individual's opinion can be influenced by social interactions. This is the basis for the long-standing field of opinion dynamics that seeks to model how beliefs evolve on a social network. DeGroot (1974) explains how repeated linear averaging mechanisms among neighbors in a graph can lead to consensus, and, later, the FJ model (Friedkin and Johnsen, 1990, 1999) suggests to anchor individuals' opinions in their initial beliefs, allowing persistent disagreement even in connected networks.

Consider a graph with n nodes and let the vector $\boldsymbol{x}^{(k)} \in [0, 1]^n$ denote individuals' opinions at time step k, with $\boldsymbol{x}^{(0)} = \boldsymbol{x}_{\text{init}} \in [0, 1]^n$ referred to as initial opinions. Then, the FJ model defines

$$\boldsymbol{x}^{(k+1)} = (I_n - \Lambda_\alpha)\, \boldsymbol{x}_{\text{init}} + \Lambda_\alpha W \boldsymbol{x}^{(k)}. \tag{1}$$

The influence matrix W defines how individuals weight the opinions of their peers during social interaction. For the purpose of this work, the influence matrix is defined by the graph structure as $W := \deg^{-1} A$, hence, it is constant and row-stochastic. The peer influence matrix $\Lambda_\alpha = \mathrm{diag}(\alpha_1, \cdots, \alpha_n)$ with $\alpha_i \in [0, 1]$ for $i \in [n]$ parameterizes individual i's susceptibility to peer influence. As α_i increases, individual i becomes more susceptible to peer influence. For $\alpha_i \equiv 1, \forall i \in [n]$, the dynamics model in Equation 1 recovers the linear averaging mechanism of the DeGroot model as a special case. The complement $1 - \alpha_i$ is commonly referred to as the stubbornness parameter.

Akin to Jia et al. (2015, 2019) and Friedkin et al. (2016), we consider the evolution of opinions across multiple stages of peer interactions. But instead of reasoning about changes in the weighting of individuals through social dynamics, we leave the network untouched and study the opinion evolution when the platform updates the prediction for each individual before each round of peer interactions.

2.2 Platform influence

It has been widely documented that beyond peer interaction, exposure to online content on digital platforms shapes individuals' opinions (Anderson and Magruder, 2012; Ursu, 2018; Mendler-Dünner et al., 2024). Prior work in opinion dynamics accounts for platform influence by modeling how platforms intervene in the network structure. For example, Zhou et al. (2024) and Chitra and Musco (2020) model the influence of platform recommendations as changes to the edge weights. Santos et al. (2021) and Wang and Kleinberg (2023) study how link recommendations can influence network structure and equilibria. In our model, platforms directly influence individuals' opinions through self-fulfilling predictions. Further, the platform is external to the graph and offers personalized predictions to every individual. Lanzetti et al. (2023) show the macro-level opinions can remain unchanged while individual opinions are deeply influenced by the platform. They focus on a homogeneous population with no peer interactions and the platform tries to maximize the instant reward via exploration-exploitation trade-off. Subsequent to our work, Tomašev et al. (2026) study how the long-term use of personalized AI assistants in pursuit of value alignment can lead to normative mode collapse, with the aligned AI assistants modelled as perfectly matching users' historical values.

Algorithm 1 Algorithmically mediated opinion dynamics

Parameters: $\boldsymbol{\alpha}, \boldsymbol{\beta}, K$, a connected simple graph $\mathcal{G}$
Input: Innate opinion $\boldsymbol{x}^*$, platform algorithm $\mathrm{Algo}(\cdot)$.

Initial platform predictions $\boldsymbol{f}^{(1)} = \mathrm{Algo}(\boldsymbol{x}^*)$.
for $t > 0$ **do**
  $\boldsymbol{x}_{\mathrm{init}}^{(t)} = (I_n - \Lambda_\beta)\boldsymbol{x}^* + \Lambda_\beta \boldsymbol{f}^{(t)}$ ▷ (A) platform influence
  $\boldsymbol{x}_0 = \boldsymbol{x}_{\mathrm{init}}^{(t)}$.
  for $k = 0, 1, 2, \ldots, K-1$ **do**
    $\boldsymbol{x}_{k+1} = (I_n - \Lambda_\alpha)\boldsymbol{x}_0 + \Lambda_\alpha W \boldsymbol{x}_k$. ▷ (B) peer influence
  end for
  Platform observes expressed opinions $\boldsymbol{x}_{\mathrm{ex}}^{(t)} = \boldsymbol{x}_K$.
  $\boldsymbol{f}^{(t+1)} = \mathrm{Algo}(\boldsymbol{x}_{\mathrm{ex}}^{(t)})$. ▷ (C) platform model update
end for

2.3 Predictive loops

Predictive models are rarely static but repeatedly updated based on fresh data. If this data is influenced by the model itself it creates a dynamic in which predictions and data co-evolve. In the machine learning literature such dynamics have been formalized under the umbrella of performative prediction (Perdomo et al., 2020). Convergence guarantees and equilibria are derived under different learning settings (e.g., Perdomo et al., 2020; Mendler-Dünner et al., 2020; Drusvyatskiy and Xiao, 2023; Wang et al., 2025; Mofakhami et al., 2023; Farina and Perdomo, 2026). While these results characterize the asymptotic behavior of the platform, they model the population as a distribution without specifying the internal structure. Instead, we leverage network structure in the population and focus on equilibrium outcomes for the population, rather than properties of the learned model.

Wang et al. (2023); Narang et al. (2023); Saig and Rosenfeld (2025) consider network effects among decision makers. Eilat and Rosenfeld (2023) discuss a bipartite graph between strategic users and recommended items. But none of these existing models aims to characterize equilibrium outcomes for individuals in networked populations. More recently, Góis et al. (2025) provides a framework studying how trust in predictions mediates performativity over a networked population, highlighting key trade-offs between predictive accuracy and social welfare. Unlike our focus on opinion dynamics, their work investigates best-response dynamics and the cooperative subpopulation over the network. We refer to Hardt and Mendler-Dünner (2025) for a more comprehensive overview.

Mathematical conventions. We use $\sigma(A)$ to denote the spectrum of A and $\rho(A)$ to denote the spectral radius of A. Let $\boldsymbol{e}_i$ be the i-th standard basis vector whose i-th entry is 1 and remaining entries are 0. The notation $\mathbf{0}, \mathbf{1}$ denotes a matrix or vector of appropriate size with all 0 or 1 entries respectively. An identity matrix of dimension d is denoted by I_d. We write $A = \mathrm{diag}(a_1, \ldots, a_d)$ for a diagonal matrix with entries $(A)_{ij} = a_i$ for $i = j$ and 0 otherwise. We say that a real matrix is non-negative (positive) if all entries $(A)_{ij}$ are non-negative (positive). Moreover, A is primitive if $\exists m \in \mathbb{N}$ such that $(A^m)_{ij} > 0, \forall i, j$. A non-negative matrix A is row-stochastic if $\sum_j A_{i,j} = 1, \forall i$. We write $a = \mathcal{O}(b)$ if there exist a constant $C > 0$ such that $a \leq Cb$.

3 A model of peer-platform co-influence

We consider a network of n individuals repeatedly interacting with a platform. During each interaction the platform's predictions influence individuals' opinions. Subsequently, individuals share opinions with their peers. After a fixed time period of peer interaction, the platform observes expressed opinions and adjusts future predictions accordingly. We use the FJ model to describe individual instances of peer interaction. We are interested in the long-term equilibrium of this coupled dynamic.

Our model is summarized in Algorithm 1. We use t to index platform updates and k to index peer interactions within each platform round. The vector $\boldsymbol{x}^*$ denotes individuals' innate opinions. We encode active predictions at time step t as $\boldsymbol{f}^{(t)} \in [0, 1]^n$. To model platform influence, we assume

initial opinions are a biased version of individuals' innate opinions. More formally,

$$\boldsymbol{x}_{\text{init}}^{(t)} = (I_n - \Lambda_\beta)\boldsymbol{x}^* + \Lambda_\beta \boldsymbol{f}^{(t)}, \tag{2}$$

where the diagonal matrix $\Lambda_\beta = \text{diag}(\beta_1, \ldots, \beta_n)$ with $\beta_i \in [0, 1]$ denotes the susceptibility of each individual $i \in [n]$ to platform influence. The larger β_i, the more x_i is confounded by the predictions.

Then, given the initial opinion $\boldsymbol{x}_{\text{init}}^{(t)}$, we assume peer dynamics follow the FJ model, as detailed in Equation 1. This means, individuals repeatedly adjust their opinion based on the opinion expressed by their neighbors. The number of peer interaction steps K performed in each instance is a parameter of our model. For $K = \infty$, we let the FJ dynamics evolve until an equilibrium is reached. These modeling assumptions mirror that, in each round individuals evaluate a new exposure to predictions $\boldsymbol{f}^{(t+1)}$, against their core underlying identity, $\boldsymbol{x}^*$, to form a new private initial stance $\boldsymbol{x}_{\text{init}}^{(t+1)}$. In the language of performative prediction this process of how the expressed opinion forms defines the distribution map.

The digital platform then observes the resulting expressed opinions. For each instance t we denote this observed state of opinions after K interaction steps as

$$\boldsymbol{x}_{\text{ex}}^{(t)} := \text{FJ}_K(\boldsymbol{x}_{\text{init}}^{(t)}).$$

Observations of $\boldsymbol{x}_{\text{ex}}^{(t)}$ are then collected into a dataset and form the basis for the platform to retrain their predictive model. Thus, observations from time step t determine predictions deployed by the platform at time step $t + 1$. To indicate that a learning algorithm is involved in this step, we write

$$\boldsymbol{f}^{(t+1)} = \text{Algo}\big(\boldsymbol{x}_{\text{ex}}^{(t)}\big). \tag{3}$$

The algorithm $\text{Algo}(\cdot)$ maps observations to predictions. It serves as a flexible placeholder for the platform's learning algorithm and we discuss specific instantiations in the following sections. Think of the argument $\boldsymbol{x}$ as the variable that the platform observes, which is simultaneously the target of prediction. What is implicit in our notation is that the platform might have access to additional meta-data about individuals to train a parametric model to predict opinions. For our theoretical analysis we first assume platforms can fully observe opinions, which constitute the target variable in our prediction task. We relax this assumption in Section 5. Challenges of estimation and measurement are left for future work. Below we provide a running example to motivate our framework.

Example 1 (Content recommendation)**.** Envision a setting in which individuals interact with a recommendation platform while forming opinions on a personal lifestyle choice such as smoking or a consequential public issue. Each individual's opinion is represented as a continuous value on a spectrum between two opposing positions. The platform recommends content, including videos, podcasts, and news articles, with varying positions on the issue based on a predictive algorithm and users' previously expressed opinions. On a daily basis, individuals consume the recommended content, update their opinions, and exchange views with peers through online sharing or in-person conversations. As this process repeats, the platform continuously adapts its recommendations using newly observed behavioral data.

3.1 Performative stability

We are interested in the equilibrium opinions under our new model. We say the dynamics in Algorithm 1 have converged if there is no reason for the platform to update their predictive model based on the data they observe. This means, the algorithm continues to recover the model that is already deployed—and hence it repeatedly recovers the same outcome after peer interactions. Following the terminology of Perdomo et al. (2020), we call this equilibrium state *performative stability*. In our model, predictions $\boldsymbol{f}_{\text{PS}}$ are performatively stable iff they satisfy

$$\boldsymbol{f}_{\text{PS}} = \text{Algo}\big(\text{FJ}_K((I - \Lambda_\beta)\boldsymbol{x}^* + \Lambda_\beta \boldsymbol{f}_{\text{PS}})\big).$$

We use $\boldsymbol{x}_{\text{PS}}$ to denote the equilibrium opinions implied by $\boldsymbol{f}_{\text{PS}}$, i.e., $\boldsymbol{x}_{\text{PS}} = \text{FJ}_K((I - \Lambda_\beta)\boldsymbol{x}^* + \Lambda_\beta \boldsymbol{f}_{\text{PS}})$. Throughout the paper, we discuss instantiations of our model where Algorithm 1 converges to a unique performatively stable point. Accordingly, the equilibrium of interest is

$$\lim_{t \to \infty} \text{FJ}_K\big(\boldsymbol{x}_{\text{init}}^{(t)}\big) \to \boldsymbol{x}_{\text{PS}}. \tag{4}$$

Going forward, we characterize how $\boldsymbol{x}_{\mathrm{PS}}$ depends on β, α and K for different algorithmic strategies, and analyze how it differs from platform-free FJ equilibrium $\boldsymbol{x}_{\mathrm{FJ}} := \mathrm{FJ}(\boldsymbol{x}^*)$. A particularly interesting case of our model is when $K \to \infty$. Here the platform observes opinions at equilibrium and $\boldsymbol{x}_{\mathrm{PS}}$ corresponds to a state where both, learning dynamics and peer dynamics have converged.

4 Peer dynamics when perfectly predicting the past

Predictions rely on observations of expressed opinions from the past. At the same time they influence opinions the moment they are used to make recommendations. This can couple instances of peer interaction happening at different time steps. To analyze the effects, we first consider an idealized policy that accurately predict each individual's most recently expressed opinion. We refer to this policy as perfect prediction. Hence, the learning algorithm simplifies to

$$\mathrm{Algo}(\boldsymbol{x}_{\mathrm{ex}}^{(t)}) := \boldsymbol{x}_{\mathrm{ex}}^{(t)}, \quad t \geq 1. \tag{5}$$

This baseline instantiation exposes important qualitative effects of performative learning systems without the complexities of data collection, model training and uncertainty quantification. The effect it captures is how predictions introduce memory to peer dynamics and opinions from time step t impact next-step dynamics through the information distilled into the machine learning model.

For perfect prediction, the first instance of peer interaction ($t = 1$) recovers the classical FJ dynamics anchored at $\boldsymbol{x}^{(0)} = \boldsymbol{x}^*$. When individuals are not susceptible to platform influence and $K \to \infty$ it thus holds that $\boldsymbol{x}_{\mathrm{PS}} = \boldsymbol{x}_{\mathrm{FJ}}$. When $\beta_i > 0$ and t grows our model deviates from this classical treatment. The resulting equilibrium will differ qualitatively, as we will show.

4.1 Performative stability

We characterize the learning dynamics under perfect prediction. Instantiating our model with the platform policy in Equation 5, we observe that the predictions $\boldsymbol{f}$ admit the following recursion

$$\boldsymbol{f}^{(t+1)} = \Psi_K \left[(I_n - \Lambda_\beta)\boldsymbol{x}^* + \Lambda_\beta \boldsymbol{f}^{(t)} \right] \text{ with } \Psi_K := \sum_{i=0}^{K-1} (\Lambda_\alpha W)^i (I_n - \Lambda_\alpha) + (\Lambda_\alpha W)^K. \tag{6}$$

The initial opinions are a linear combination of the current predictions $\boldsymbol{f}^{(t)}$ and the innate opinion $\boldsymbol{x}^*$. The matrix Ψ_K characterizes peer influence during K consecutive steps of peer interaction, analogous to classical characterizations of opinion dynamics (e.g., Friedkin and Johnsen, 1990). What is new to our work is an outer loop implied by repeated updates to the platform's predictions. Interestingly, convergence of our model does not depend on individuals' platform susceptibility, instead, it is sufficient for peer dynamics to be stable.

Proposition 1 (Existence of a unique performatively stable equilibrium)**.** Assume the population is characterized by a simple connected graph $\mathcal{G}$, and $\exists i : \alpha_i < 1$. Consider a platform that performs perfect prediction of past data. Then, the equilibrium state $\boldsymbol{x}_{\mathrm{PS}}$ exists and is unique. Furthermore, the opinions $\boldsymbol{x}_{\mathrm{ex}}^{(t)}$ converge towards $\boldsymbol{x}_{\mathrm{PS}}$ at a linear rate

$$\|\boldsymbol{x}_{\mathrm{ex}}^{(T)} - \boldsymbol{x}_{\mathrm{PS}}\|_2 = \mathcal{O}(c^T)$$

where c can be characterized as follows:

- If $\beta_i < 1 \; \forall i \in [n]$ we have $c = \mathcal{O}(\max_i \beta_i)$.
- If $\exists i : \beta_i = 1$ and $\rho(\Psi_K \Lambda_\beta) < 1$ we have $c = \rho(\Psi_K \Lambda_\beta)$.
- If $\exists i : \beta_i = 1$ and $\rho(\Psi_K \Lambda_\beta) = 1$ we have $c = \max\{|\lambda| : \lambda \in \sigma(\Psi_K \Lambda_\beta), |\lambda| < 1\}$.

Note that the assumptions coincide with the conditions that guarantee the existence of equilibrium under the FJ model (Parsegov et al., 2016). Notably, the result in Proposition 1 holds for any value of $\beta_i \in [0, 1]$. This implies that our coupled dynamics (6) converge whenever the inner-loop dynamics converge. We provide a general closed-form characterization of $\boldsymbol{x}_{\mathrm{PS}}$ in Appendix C.2. The convergence of the outer loop is governed by the individual who is most susceptible to platform influence. The case distinction arises for $\beta_i = 1$ when dynamics are no longer anchored in the innate opinion of individual i. But in general, it holds that higher platform susceptibility slows down convergence which is in line with classical results in performative prediction that high sensitivity ε of the data distribution to changes in the predictive model implies slower convergence (Perdomo et al., 2020). We show the distribution shift is $\varepsilon \propto \|\Lambda_\beta\|$-sensitive in the predictions in Appendix C.3.

4.2 Homogenizing force of performativity

Next, we look at how platform influence impacts opinions at equilibrium. Note that absent performativity, the FJ model induces a heterogeneous equilibrium $\boldsymbol{x}_{\mathrm{FJ}}$, whenever innate opinions $\boldsymbol{x}^*$ are non-uniform. The following result explains how performativity acts as a homogenizing force.

Proposition 2 (Homogenizing force of performativity)**.** Consider a network characterized by a simple connected regular graph $\mathcal{G}$, its nodes with peer susceptibilities $\alpha_i \equiv \alpha \in (0,1)$ and platform susceptibilities $\beta_i \equiv \beta \in (0,1)$. The platform performs perfect prediction of past data. Then, the mean opinion at equilibrium is given by $\mathrm{Mean}(\boldsymbol{x}_{\mathrm{PS}}) = \frac{1}{n}\mathbf{1}^\top \boldsymbol{x}^*$ and the variance satisfies

$$\frac{\partial \mathrm{Var}(\boldsymbol{x}_{\mathrm{PS}})}{\partial \beta} < 0.$$

The result is a direct consequence of the equilibrium characterization $\boldsymbol{x}_{\mathrm{PS}} = (1-\beta)(I_n - \beta\Psi_K)^{-1}\Psi_K \boldsymbol{x}^*$ when $\alpha \in (0,1)$, $\beta < 1$, together with the observation that Ψ_K is row-stochastic.

Thus, as platform susceptibility β grows the variance of equilibrium opinions decreases. Remember that for $\beta = 0$ the stable point corresponds to $\boldsymbol{x}_{\mathrm{PS}} = \mathrm{FJ}_K(\boldsymbol{x}^*)$ and hence the variance reduction can be as large as $\mathrm{Var}(\boldsymbol{x}^*)$ for networks with small α. The mean, in turn, remains unaffected by platform interactions. We defer the discussion on the dependence of $\mathrm{Var}(\boldsymbol{x}_{\mathrm{PS}})$ over α to Appendix C.4.

4.3 Consensus in the limit

As the heterogeneity of equilibrium opinions gets reduced when β grows, a particularly interesting case is $\beta \to 1$. In this case platform influence is maximized. We show how, in combination with perfect prediction, this induces consensus at equilibrium.

Theorem 1 (Consensus under strong performativity)**.** Consider a network characterized by a simple connected graph $\mathcal{G}$, individuals with peer susceptibilities $\alpha_i \in (0,1)\ \forall i$, and a platform that performs perfect prediction of past data. Then, in the limit $\beta_i = \beta \to 1$ for all i,[1] a consensus is reached at equilibrium. Namely, for any $\boldsymbol{x}^*$ there exists a constant $c^* = c^*(\boldsymbol{x}^*) \in [0,1]$ such that:

$$\lim_{\beta \to 1} (\boldsymbol{x}_{\mathrm{PS}})_i = c^* \quad \forall i \in [n]. \tag{7}$$

In particular, if $K = \infty$, the consensus value $c^* = \sum_{i=1}^n \frac{(1-\alpha_i)d_i x_i^*}{\alpha_i} / \sum_{j=1}^n \frac{(1-\alpha_j)d_j}{\alpha_j}$.

Let us first contrast this result with the classical equilibrium under the FJ dynamics. The conventional understanding is that as long as $\boldsymbol{x}^*$ is heterogeneous and $\alpha_i < 1$ for some i, the platform-free equilibrium leads to disagreement (Friedkin and Johnsen, 1990). However, with performativity this is different. When $\beta_i = 1$ for all i this results in a consensus at equilibrium *even if* the dynamics are initialized at a heterogeneous $\boldsymbol{x}^*$. The reason is that, under strong performativity, the equilibrium at each step is carried over to initialize the next step, which progressively reduces the influence of the innate opinions $\boldsymbol{x}^*$ from the first instance. At a high level, this process can be viewed as a gradual transition from FJ dynamics to DeGroot dynamics, as the effect of innate opinions vanishes.

Now, let us contrast the result with the equilibrium in the DeGroot model (DeGroot, 1974) which explains consensus when $\alpha_i = 1$ for all i and primitive $\mathcal{G}$. Theorem 1 can explain convergence to consensus for a broader class of graphs without primitivity. The reason is that anchoring of initial opinions, with $\alpha_i \in (0,1)$, makes the random walk induced by Ψ_K aperiodic.

5 Indirect peer influence under partial observation

Standard opinion-dynamics models assume that peer influence travels along the edges of a graph and the network structure governs how opinions flow. Our next result shows that platforms can add an overlaying structure to the network. In doing so, they can create new pathways for peer influence, linking people who are not otherwise connected in the graph.

We demonstrate how these alternate pathways for peer influence arise under a simple model of prediction with limited information. Formally, suppose the platform observes opinions for individuals

[1] we can relax the homogeneity assumption on β with an additional technical assumption on the speed at which the limit is attained for individual β_i, see Appendix C.5.

in a subgraph of $\mathcal{G}$, denoted as $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$. To make prediction for unobserved individuals, the platform then extrapolated these observations, as detailed in the predictive policy below:

$$(\boldsymbol{f}^{(t+1)})_i = \begin{cases} (\boldsymbol{x}_{\text{ex}}^{(t)})_i, & \text{for} \quad i \in \mathcal{V}', \\ \sum_{j \in \mathcal{V}'} r_j (\boldsymbol{x}_{\text{ex}}^{(t)})_j, & \text{otherwise,} \end{cases} \tag{8}$$

for fixed scalars $r_i \in (0,1), \forall i \in \mathcal{V}'$ with $\sum_{i \in \mathcal{V}'} r_i = 1$. The algorithm deviates from perfect prediction for unobserved individuals $i \notin \mathcal{V}'$. The following result shows how such deviations from perfect prediction couple individuals, even if they are not susceptible to peer influence.

Proposition 3. Consider a network characterized by a simple connected graph $\mathcal{G}$. Assume the platform has access to expressed opinions of individuals in a subgraph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ of $\mathcal{G}$ and $\alpha_i \in (0,1)$, $\beta_i \in (0,1)$ for all $i \in \mathcal{V}'$. Consider innate opinions $\boldsymbol{x}^*$ and $\tilde{\boldsymbol{x}}^*$ that differ in a single coordinate $i_0 \in \mathcal{V}'$, and let $\boldsymbol{x}_{\text{PS}}$ and $\tilde{\boldsymbol{x}}_{\text{PS}}$ be the corresponding stable points when platform predictions follow Equation 8. Fix $q \notin \mathcal{V}'$. Then, even if $\alpha_q = 0$, as long as $\beta_q > 0$, it holds that

$$|(\boldsymbol{x}_{\text{PS}})_q - (\tilde{\boldsymbol{x}}_{\text{PS}})_q| > c\beta_q |x^*_{i_0} - \tilde{x}^*_{i_0}|$$

where constant c is determined by Λ_α, $\{r_i\}_{i \in \mathcal{V}'}$, $\mathcal{G}$, and $\mathcal{G}'$.

The result explains that changes to the innate opinions of individual i impact the opinion of individual q even if $\alpha_q = 0$. Thus, there can be indirect spillover effects through predictive models even if individuals are not susceptible to peer influence. This is an important deviation from classical network theory. Moreover, $\mathcal{G}'$ is not necessarily connected, while the platform adds another layer of connectivity among individuals in $\mathcal{V}'$ via the platform-susceptible individuals in $\mathcal{V} \setminus \mathcal{V}'$.

A consequence of these indirect spillover effects is that a consensus can be achieved at equilibrium even in the presence of stubborn individuals.

Theorem 2 (Consensus despite stubbornness)**.** Consider a network characterized by a simple connected graph $\mathcal{G}$. Assume the platform has access to expressed opinions of individuals in a subgraph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ of $\mathcal{G}$, $\alpha_i \in (0,1)$ for all $i \in \mathcal{V}'$, and applies estimator (8) to $\mathcal{G}$. Then, if $\beta_i = 1$ for all i, the equilibrium $\boldsymbol{x}_{\text{PS}}$ corresponds to a consensus. Namely, for any $\boldsymbol{x}^*$ there exists a constant $c^*(\boldsymbol{x}^*) \in [0,1]$ such that:

$$\lim_{\beta \to 1} (\boldsymbol{x}_{\text{PS}})_i = c^*(\boldsymbol{x}^*) \quad \forall i \in [n]. \tag{9}$$

The proof of Theorem 2 goes through by observing the platform acts as a virtual node that connects nodes in $\mathcal{G}'$ via the weighted mean estimation. The difference to the result in Theorem 1 is that we do not require $\alpha_i > 0$ for every i. Instead, individuals for whom the platform derives predictions from alternate observations can have peer susceptibility $\alpha_i = 0$. Because rather than through direct peer influence, they get pulled towards the consensus opinion through predictions instead.

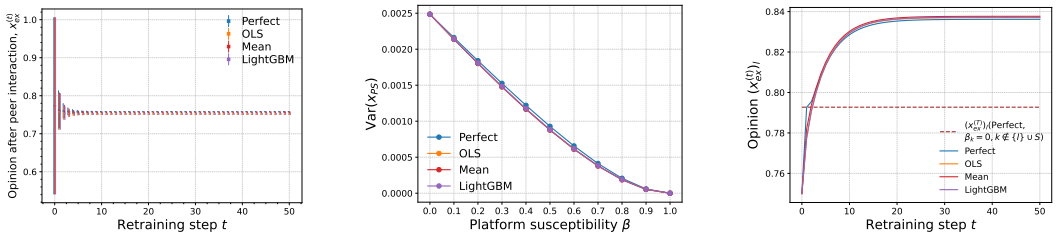
Another consequence of the spillover effect is the consensus in the absence of connectivity, which is a standard assumption in network science literature. The consensus is guaranteed via the connections created by the shared predictions and strong platform influences.

Corollary 1 (Consensus in disconnected social graphs)**.** Consider a population characterized by a simple connected network $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ with $\alpha_i \in (0,1)$, $\forall i \in \mathcal{V}_1$, and another population characterized by a simple network $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$. We assume $\mathcal{G}_2$ is disconnected from $\mathcal{G}_1$ while $\mathcal{G}_2$ itself is not necessarily connected. Suppose the platform has access to expressed opinions of individuals of $\mathcal{G}_1$ and apply estimator (8) to $\mathcal{G}_1 \cup \mathcal{G}_2$. Then if $\beta_i = 1 \; \forall i$ on $\mathcal{G}_1 \cup \mathcal{G}_2$, a consensus is reached in equilibrium $\boldsymbol{x}_{\text{PS}}$ as in Theorem 2.

6 Equilibrium response to platform intervention

The qualitative differences in how opinions spread across the population through platform and peer interactions affects the equilibrium response to platform interventions. In the following, we focus on platform interventions that systematically deviate from perfect prediction for a single individual.

$$\text{For a fixed } j \in [n], \text{ let} \quad (\boldsymbol{f}^{(t+1)})_i = \begin{cases} s & \text{for } i = j \\ (\boldsymbol{x}_{\text{ex}}^{(t)})_i & \text{otherwise.} \end{cases} \tag{10}$$



(a) Opinions across t (Yelp) (b) Opinions at equilibrium (Yelp) (c) Indirect effect (Yelp)

Figure 2: In (a), we show how performativity homogenizes opinions across retraining steps. In (b), we show how platform susceptibility decreases variance of opinions at equilibrium. In (c), we show how indirect platform influence increases over retraining steps.

In words, the platform intervenes by repeatedly predicting $s \in [0, 1]$ for individual j and remains faithful to perfect prediction for all other individuals. We assume that individual j has $\beta_j > 0$ and we study the indirect platform influence on individual $l \neq j$. To attribute the observed effect to spillovers, we let $\beta_l = 0$. Further, we let $\boldsymbol{x}^* = \mathbf{0}$ so that any deviation from $\mathbf{0}$ can be attributed to the platform intervention s. We set $\beta_i = \gamma$ to control the platform influence on individuals $i \notin \{j, l\}$ exposed to perfect prediction. The next result explains how the effect of the intervention on equilibrium opinions grows with the platform susceptibility of peers.

Theorem 3 (Peer-mediated platform influence)**.** Consider a network characterized by a simple fully-connected graph $\mathcal{G}$, and peer susceptibilities $\alpha_i \equiv \alpha \in (0, 1)$. Let $\boldsymbol{x}^* = \mathbf{0}$ and consider a platform that always predicts $s > 0$ for randomly sampled individual j according to Equation (10). Consider $K = \infty$. Let platform susceptibilities satisfy $\beta_i > 0$ for $i = j$, $\beta_i = 0$ for randomly sampled individual $i = l$, and $\beta_i = \gamma \in (0, 1]$ for $i \notin \{j, l\}$. Then, the influence of intervention s strictly grows with the platform susceptibility γ of their peers. In particular, it holds that

$$\frac{\partial \Delta_l^\gamma}{\partial \gamma} > 0 \quad \text{for} \quad \Delta_l^\gamma := (\boldsymbol{x}_{\mathrm{PS}})_l - (\boldsymbol{x}_{\mathrm{PS}})_l^{\gamma=0} \quad \text{and} \quad \frac{\partial \mathrm{Mean}(\boldsymbol{x}_{\mathrm{PS}})}{\partial \gamma > 0}.$$

To appreciate the result, note that $(\boldsymbol{x}_{\mathrm{PS}})_l^{\gamma=0}$ equals the opinion for individual l under the FJ dynamics initialized with $\boldsymbol{x}_{\mathrm{init}} = s\beta_j \boldsymbol{e}_j$. This corresponds to the classical network equilibrium under platform intervention. Now, what Theorem 3 tells us is that the presence of a perfect predictor except for individual j amplifies the indirect influence of the intervention s on the equilibrium opinions. For the population, the influence also strictly grows with the platform susceptibility γ of peers. Consequently, classical models of peer interaction that treat platforms as static may miss the amplification of the intervention effect because of this repeated coupling of platform and peer influence.

7 Experimental results

To complement our theoretical study, we conduct semi-synthetic experiments on real social networks using popular machine learning methods to derive predictions. We extract features for training the predictive model from public data about individual's profile information. We then simulate our model and investigate how practical deviations from perfect prediction impact dynamics and equilibria.

Simulation environment. We build on the Pokec and Yelp datasets.[2] The Pokec dataset contains anonymized profile data of more than 1.6 million individuals in different modalities together with an underlying social network graph. For our simulations, we sample a connected subgraph of 2,163 nodes which we treat as an undirected graph. The Yelp dataset contains numerous reviews with user profiles describing more than 150,000 businesses. In our simulations, we build an undirected connected network of 1,789 users who left reviews for a randomly chosen business.

We present results on the Yelp dataset and defer the ones using Pokec dataset and further details to Appendix G. For the Yelp dataset, we normalize the feature "stars" of the reviews to $[0, 1]$ and use them as the opinions of users towards the business. We simulate retraining and the dynamics

[2]The dataset are available from `https://snap.stanford.edu/data/soc-Pokec.html` and `https://business.yelp.com/data/resources/open-dataset/`

over the network (Algorithm 1), treating the opinion extracted from the data as the individuals' innate opinions. For the platform's learning algorithm we contrast perfect prediction (Perfect) with a) mean estimation (Mean), b) ordinary least square (OLS) with predictions clipped to $[0, 1]$, and c) Light Gradient-Boosting Machine (Ke et al., 2017) (LightGBM). We approximate equilibrium using $K = 100$ peer interactions and $T = 50$ platform updates. All the simulations are conducted on commodity hardware using Python 3.13. The code is available on github.

Retraining and equilibria. We visualize how opinions evolve across retraining steps for the different learning settings. In Figure 2a, we plot the mean and variance of opinions $\boldsymbol{x}_{\mathrm{ex}}^{(t)}$ across retraining steps for $\beta_i = 1$ and heterogeneous $\{\alpha_i\}$. We see that for all four predictive strategies, opinions converge, and the variance approximately converges to 0 as t increases. This suggests that our intuition of performativity as an explanation of consensus also extends to parametric predictors. In Figure 2b we again hold $\{\alpha_i\}$ fixed, but set $\beta_i \equiv \beta$, and vary β between 0 and 1. We can see that the variance of the equilibrium opinion decreases as β grows. Also note that the Yelp network is not regular suggests that the take-away for Proposition 2 may extend to a broader class of networks.

Indirect platform influence. Next, we show how peer interactions amplify platform influence, as analyzed in Theorem 3. We randomly choose an individual l and set $\beta_l = 0$. Then, instead of one individual, we randomly sample a set S that contains 10% of individuals of the network and steer them towards $s = 1$. In Figure 2c, we observe the opinion of individual l increasingly deviates from x_l^* as retraining step t increases. The red dashed line shows the setting where $\beta_i = 0$ for individuals who are not steered, i.e., $i \notin S$. This corresponds to the outcomes after a single instance of peer interactions. We see how opinions deviate from this state through repeated peer-platform interactions.

8 Discussion

With theoretical results and simulations we demonstrated how predictive loops can qualitatively alter equilibria in social networks. We identify two key mechanisms that lead to divergent behavior. First, learning systems introduce memory into peer dynamics by deriving future predictions from past observations, thereby coupling opinions across time. Second, learning systems create new pathways for peer influence beyond the topology of social networks by repeatedly aggregating data across a population and distilling it into personalized predictions. Both mechanisms operate for a platform whose only objective is predictive accuracy. Our model is deliberately minimal, and we close by making the reasoning behind them explicit; a list of future directions is deferred to Appendix B.

Perfect prediction. We establish our core results under perfect prediction in our theoretical analysis. This idealized setting isolates the mechanics of the predictive loop from estimation error and model misspecification. It is the most favorable case for the platform and hence a stress test for our claims: when homogenization and consensus emerge already here they must be intrinsic structural properties of the loop rather than artifacts of noise. Our simulations with parametric learners on real social graphs show that the qualitative findings persist under realistic estimation errors. Extending our theory and investigations of indirect spillovers to parametric learning algorithms is left for future work.

Anchoring at innate opinions. Re-anchoring individuals on their innate opinion $\boldsymbol{x}^*$ rather than the temporary peer compromise $\boldsymbol{x}_{\mathrm{ex}}^{(t)}$ is grounded in sociological and mathematical principles. Sociologically, in line with the FJ model, individuals maintain persistent baseline values when consuming new platform content rather than rewriting their core identity after every peer debate. The memory that the identity policy (5) introduces across deployment rounds is held entirely within the platform's deployed predictive state $\boldsymbol{f}^{(t)}$ and enters opinions only through β_i, rather than in unobservable, path-dependent internal user drift, and setting $\beta = 0$ recovers $\boldsymbol{x}_{\mathrm{FJ}}$. By contrast, using $\boldsymbol{x}_{\mathrm{ex}}^{(t)}$ would cause the dynamics to degenerate to the DeGroot model, and conflict with the FJ framework's identity persistence assumption during social interactions. That said, investigating predictive loops in combination with other models of peer interaction would be an interesting extension.

Network assumptions. Our focus on self-fulfilling performativity ($\beta_i \in [0, 1]$) and our network assumptions, such as a fixed undirected topology and the complete graph in Theorem 3, are standard in the foundational papers on peer and platform influence (Friedkin and Johnsen, 1990; Perdomo et al., 2020). Adopting them permits direct comparisons with established results such as the original FJ model, which is essential to isolate the qualitative shifts emerging from the peer-platform co-influence. Our experiments on real-world social networks with heterogeneous susceptibilities suggest that homogenization and intervention amplification persist beyond these assumptions.

Acknowledgements

Rediet Abebe and Celestine Mendler-Dünner acknowledge the financial support of the Hector Foundation. Abebe was supported by the Andrew Carnegie Fellowship Program. Jiduan Wu would like to thank for the financial support from the Max Planck ETH Center for Learning Systems (CLS). We would like to thank Yatong Chen, Ana-Andreea Stoica, and anonymous reviewers for helpful feedback on the manuscript.

References

Rediet Abebe, Jon Kleinberg, David Parkes, and Charalampos E. Tsourakakis. 2018. Opinion Dynamics with Varying Susceptibility to Persuasion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 1089–1098.

Daron Acemoglu, Tianyi Lin, Asuman Ozdaglar, and James Siderius. 2026. *How AI aggregation affects knowledge*. Technical Report. National Bureau of Economic Research.

Gediminas Adomavicius, Jesse C Bockstedt, Shawn P Curley, and Jingjing Zhang. 2013. Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research* 24, 4 (2013), 956–975.

Michael Anderson and Jeremy Magruder. 2012. Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. *The Economic Journal* 122, 563 (2012), 957–989.

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

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

Abraham Berman and Robert J Plemmons. 1994. *Nonnegative matrices in the mathematical sciences*. SIAM.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. 2022. Performative prediction in a stateful world. In *International conference on artificial intelligence and statistics*. PMLR, 6045–6061.

Kartik Chandra, Max Kleiman-Weiner, Jonathan Ragan-Kelley, and Joshua B Tenenbaum. 2026. Sycophantic chatbots cause delusional spiraling, even in ideal Bayesians. *arXiv preprint arXiv:2602.19141* (2026).

Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*. 224–232.

Gary Cheng, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Causal Inference out of Control: Estimating Performativity without Treatment Randomization. In *International Conference on Machine Learning*. 8077–8103.

Uthsav Chitra and Christopher Musco. 2020. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th international conference on web search and data mining*. 115–123.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the national academy of sciences* 118, 9 (2021), e2023301118.

Pranav Dandekar, Ashish Goel, and David T Lee. 2013. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5791–5796.

Sarah Dean, Mihaela Curmei, Lillian Ratliff, Jamie Morgenstern, and Maryam Fazel. 2024a. Emergent specialization from participation dynamics and multi-learner retraining. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 343–351.

Sarah Dean, Evan Dong, Meena Jagadeesan, and Liu Leqi. 2024b. Accounting for ai and users shaping one another: The role of mathematical models. *arXiv preprint arXiv:2404.12366* (2024).

Morris H DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical association* 69, 345 (1974), 118–121.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the national academy of Sciences* 113, 3 (2016), 554–559.

Dmitriy Drusvyatskiy and Lin Xiao. 2023. Stochastic Optimization with Decision-Dependent Distributions. *Mathematics of Operations Research* 48, 2 (2023), 954–998.

Cynthia Dwork, Chris Hays, Lunjia Hu, Nicole Immorlica, and Juan Perdomo. 2025. Inducing Efficient and Equitable Professional Networks Through Link Recommendations. *arXiv preprint arXiv:2503.04542* (2025).

Dean Eckles, Brian Karrer, and Johan Ugander. 2017. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference* 5, 1 (2017), 20150021.

Itay Eilat and Nir Rosenfeld. 2023. Performative recommendation: diversifying content via strategic incentives. In *International Conference on Machine Learning*. PMLR, 9082–9103.

Gabriele Farina and Juan Carlos Perdomo. 2026. The Stability of Online Algorithms in Performative Prediction. *arXiv preprint arXiv:2602.24207* (2026).

Noah E Friedkin, Peng Jia, and Francesco Bullo. 2016. A theory of the evolution of social power: Natural trajectories of interpersonal influence systems along issue sequences. *Sociological Science* 3 (2016), 444–472.

Noah E Friedkin and Eugene C Johnsen. 1990. Social influence and opinions. *Journal of mathematical sociology* 15, 3-4 (1990), 193–206.

Noah E. Friedkin and Eugene C. Johnsen. 1999. Social Influence Networks and Opinion Change. In *Advances in Group Processes*. Vol. 16. Emerald, 1–29.

Jason Gaitonde, Jon Kleinberg, and Éva Tardos. 2021. Polarization in geometric opinion dynamics. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 499–519.

Germain Gauthier, Roland Hodler, Philine Widmer, and Ekaterina Zhuravskaya. 2026. The political effects of X's feed algorithm. *Nature* (2026), 1–8.

António Góis, Mehrnaz Mofakhami, Fernando P Santos, Simon Lacoste-Julien, and Gauthier Gidel. 2025. Performative Prediction on Games and Mechanism Design. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1855–1863.

Andrew M Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. 2023. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* 381, 6656 (2023), 398–404.

Moritz Hardt. 2026. Retraining Seeks Stable Signals. *arXiv preprint arXiv:2607.15623* (2026).

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. 2022. Performative power. In *Advances in Neural Information Processing Systems*, Vol. 35. 22969–22981.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic Classification. In *ACM Conference on Innovations in Theoretical Computer Science*. 111–122.

Moritz Hardt and Celestine Mendler-Dünner. 2025. Performative Prediction: Past and Future. *Statist. Sci.* 40, 3 (2025), 417–436.

Jan Hązła, Yan Jin, Elchanan Mossel, and Govind Ramnarayan. 2024. A geometric model of opinion polarization. *Mathematics of Operations Research* 49, 1 (2024), 251–277.

Luzie Helfmann, Nataša Djurdjevac Conrad, Philipp Lorenz-Spreen, and Christof Schütte. 2023. Modelling opinion dynamics under the impact of influencer and media strategies. *Scientific Reports* 13, 1 (2023), 19375.

Richard A Holley and Thomas M Liggett. 1975. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability* (1975), 643–663.

Petter Holme and Jari Saramäki. 2012. Temporal networks. *Physics reports* 519, 3 (2012), 97–125.

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. Algorithmic amplification of politics on Twitter. *Proceedings of the national academy of sciences* 119, 1 (2022), e2025334119.

Zachary Izzo, Lexing Ying, and James Zou. 2021. How to Learn when Data Reacts to Your Model: Performative Gradient Descent. In *International Conference on Machine Learning*, Vol. 139. PMLR, 4641–4650.

Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. 2021. Alternative Microfoundations for Strategic Classification. In *International Conference on Machine Learning*, Vol. 139. PMLR, 4687–4697.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–15.

Peng Jia, Noah E Friedkin, and Francesco Bullo. 2019. Opinion dynamics and social power evolution: A single–timescale model. *IEEE Transactions on Control of Network Systems* 7, 2 (2019), 899–911.

Peng Jia, Anahita MirTabatabaei, Noah E Friedkin, and Francesco Bullo. 2015. Opinion dynamics and the evolution of social power in influence networks. *SIAM review* 57, 3 (2015), 367–397.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

Nicolas Lanzetti, Florian Dörfler, and Nicolò Pagan. 2023. The impact of recommendation systems on opinion dynamics: Microscopic versus macroscopic effects. In *2023 62nd IEEE conference on decision and control (CDC)*. IEEE, 4824–4829.

Qiang Li, Chung-Yiu Yau, and Hoi-To Wai. 2022. Multi-agent performative prediction with greedy deployment and consensus seeking agents. *Advances in Neural Information Processing Systems* 35 (2022), 38449–38460.

Wenjun Mei, Francesco Bullo, Ge Chen, Julien M. Hendrickx, and Florian Dörfler. 2022. Micro-foundation of opinion dynamics: Rich consequences of the weighted-median mechanism. *Phys. Rev. Res.* 4 (2022), 023213. Issue 2.

Celestine Mendler-Dünner, Gabriele Carovano, and Moritz Hardt. 2024. An engine not a camera: Measuring performative power of online search. In *Advances in Neural Information Processing Systems*.

Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. 2022. Anticipating Performativity by Predicting from Predictions. In *Advances in Neural Information Processing Systems*. Article 2260.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. 2020. Stochastic Optimization for Performative Prediction. In *Advances in Neural Information Processing Systems*, Vol. 33. 4929–4939.

John P Miller, Juan C Perdomo, and Tijana Zrnic. 2021. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*. PMLR, 7710–7720.

Atsushi Miyauchi, Yuko Kuroki, Federico Cinus, Stefan Neumann, and Francesco Bonchi. 2026. A survey on algorithmic interventions in opinion dynamics. *arXiv preprint arXiv:2603.10756* (2026).

Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. 2023. Performative Prediction with Neural Networks. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 206)*. PMLR, 11079–11093.

Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J. Ratliff. 2023. Multiplayer Performative Prediction: Learning in Decision-Dependent Games. *Journal of Machine Learning Research* 24, 202 (2023), 1–56.

Sridhar Narayanan and Kirthi Kalyanam. 2015. Position Effects in Search Advertising and their Moderators: A Regression Discontinuity Approach. *Marketing Science* 34, 3 (May 2015), 388–407.

Buddhika Nettasinghe and Kang Zhao. 2026. Dynamics of Human-AI Collective Knowledge on the Web: A Scalable Model and Insights for Sustainable Growth. In *Proceedings of the ACM Web Conference 2026*. 1797–1808.

Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. 2023. Like-minded sources on Facebook are prevalent but not polarizing. *Nature* 620, 7972 (2023), 137–144.

OECD. 2025. Social Media Governance Policy Brief: How the DSA can enable a public science of digital platform social impacts. `https://oecd.ai/en/wonk/documents/social-media-governance-policy-brief-how-the-dsa-can-enable-a-public-science-of-digital-platform-social-impacts-policy-brief`.

Charlotte Out, Sijing Tu, Stefan Neumann, and Ahad N Zehmakan. 2024. The impact of external sources on the friedkin–johnsen model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1815–1824.

Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Sergey E Parsegov, Anton V Proskurnikov, Roberto Tempo, and Noah E Friedkin. 2016. Novel multidimensional models of opinion dynamics in social networks. *IEEE Trans. Automat. Control* 62, 5 (2016), 2270–2285.

Antonio F Peralta, János Kertész, and Gerardo Iñiguez. 2025. Opinion dynamics in social networks: From models to data. In *Handbook of Computational Social Science*. Edward Elgar Publishing Limited, 384–406.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. In *International Conference on Machine Learning*, Vol. 119. PMLR, 7599–7609.

Jacopo Perego and Sevgi Yuksel. 2022. Media competition and social disagreement. *Econometrica* 90, 1 (2022), 223–265.

Nicola Perra and Luis EC Rocha. 2019. Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific reports* 9, 1 (2019), 7261.

Niccolo Pescetelli, Daniel Barkoczi, and Manuel Cebrian. 2022. Bots influence opinion dynamics without direct human-bot interaction: the mediating role of recommender systems. *Applied Network Science* 7, 1 (2022), 46.

Jinghua Piao, Jiazhen Liu, Fang Zhang, Jun Su, and Yong Li. 2023. Human–AI adaptive dynamics drives the emergence of information cocoons. *Nature Machine Intelligence* 5, 11 (2023), 1214–1224.

Georgios Piliouras and Fang-Yi Yu. 2023. Multi-Agent Performative Prediction: From Global Stability and Optimality to Chaos. In *ACM Conference on Economics and Computation*. 1047–1074.

Anton V Proskurnikov and Roberto Tempo. 2017. A tutorial on modeling and analysis of dynamic social networks. Part I. *Annual Reviews in Control* 43 (2017), 65–79.

Walter Quattrociocchi, Guido Caldarelli, and Antonio Scala. 2014. Opinion dynamics on interacting networks: media competition and social influence. *Scientific reports* 4, 1 (2014), 4938.

Hegselmann Rainer and Ulrich Krause. 2002. Opinion Dynamics and Bounded Confidence: Models, Analysis and Simulation. *Journal of Artificial Societies and Social Simulation* 5, 3 (2002).

Eden Saig and Nir Rosenfeld. 2025. Evolutionary Prediction Games. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.

Fernando P. Santos, Yphtach Lelkes, and Simon A. Levin. 2021. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences* 118, 50 (2021), e2102141118.

Lukas Schüepp, Carmen Amo Alonso, Florian Dörfler, and Giulia De Pasquale. 2026. Socially-Aware Recommender Systems Mitigate Opinion Clusterization. *arXiv preprint arXiv:2601.02412* (2026).

Eugene Seneta. 2006. *Non-negative matrices and Markov chains*. Springer Science & Business Media.

Alina Sîrbu, Dino Pedreschi, Fosca Giannotti, and János Kertész. 2019. Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. *PloS one* 14, 3 (2019), e0213246.

Stigler Committee. 2019. Final Report: Stigler Committee on Digital Platforms.

Haoxin Sun and Zhongzhi Zhang. 2023. Opinion optimization in directed social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4623–4632.

Nenad Tomašev, Matija Franklin, and Simon Osindero. 2026. AI Value Alignment for Evolving Social Norms. *arXiv preprint arXiv:2607.18506* (2026).

Stratis Tsirtsis, Kai Rawal, Chris Russell, Brent Mittelstadt, and Sandra Wachter. 2026. AI-Mediated Communication Can Steer Collective Opinion. *arXiv preprint arXiv:2605.16245* (2026).

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. 2013. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 329–337.

Raluca M. Ursu. 2018. The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions. *Marketing Science* 37, 4 (August 2018), 530–552.

Xiaolu Wang, Chung-Yiu Yau, and Hoi To Wai. 2023. Network effects in performative prediction games. In *International Conference on Machine Learning*. PMLR, 36514–36540.

Yanbang Wang and Jon Kleinberg. 2023. On the relationship between relevance and conflict in online social link recommendations. *Advances in Neural Information Processing Systems* 36 (2023), 36708–36725.

Zifan Wang, Changxin Liu, Thomas Parisini, Michael M Zavlanos, and Karl H Johansson. 2025. Constrained optimization with decision-dependent distributions. *IEEE Trans. Automat. Control* (2025).

Ercan Yildiz, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione. 2013. Binary opinion dynamics with stubborn agents. *ACM Transactions on Economics and Computation (TEAC)* 1, 4 (2013), 1–30.

Tianyi Zhou, Stefan Neumann, Kiran Garimella, and Aristides Gionis. 2024. Modeling the impact of timeline algorithms on opinion dynamics using low-rank updates. In *Proceedings of the ACM Web Conference 2024*. 2694–2702.

Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. 2021. Who leads and who follows in strategic classification? *Advances in neural information processing systems* 34 (2021), 15257–15269.

A Extended related work

Peer influence. There are numerous extensions and variations of DeGroot and FJ models that consider stubbornness of individuals (Yildiz et al., 2013), bounded confidence models (Rainer and Krause, 2002), as well as alternative mechanisms for the aggregation of opinions of peers (Mei et al., 2022). We refer interested readers to recent reviews of opinion dynamics in social networks (Proskurnikov and Tempo, 2017; Peralta et al., 2025). In our work, we build on the Friedkin-Johnsen model (Friedkin and Johnsen, 1990) to describe how opinions spread along the edges of the graph.

Standard network equilibria. We use $\mathrm{FJ}_k(\boldsymbol{x}_{\mathrm{init}})$ to denote the state of opinions $\boldsymbol{x}^{(k)}$ after k discrete time steps of FJ dynamics when initialized at $\boldsymbol{x}_{\mathrm{init}}$. Similarly, we denote the equilibrium state for $k \to \infty$ by $\mathrm{FJ}(\boldsymbol{x}_{\mathrm{init}})$. When $\alpha_i < 1$ for at least one individual i, the dynamics are guaranteed to converge when $\mathcal{G}$ is connected. The equilibrium can be characterized in closed form: $\mathrm{FJ}(\boldsymbol{x}_{\mathrm{init}}) = (I_n - \Lambda_\alpha W)^{-1}(I_n - \Lambda_\alpha)\boldsymbol{x}_{\mathrm{init}}$ (Parsegov et al., 2016). For the special case of DeGroot, the equilibrium corresponds to $\mathrm{DG}(\boldsymbol{x}_{\mathrm{init}}) = \lim_{t\to\infty} W^t\, \boldsymbol{x}_{\mathrm{inift}}$ while additional assumptions are required to guarantee convergence for $\alpha_i \equiv 1$, typically expressed as a primitivity assumption on $\mathcal{G}$ (see, e.g., DeGroot, 1974). We use these standard network equilibria as a baseline for our analysis.

Platform influence. It has been widely documented that beyond peer interaction, the exposure to online content on digital platforms shapes individuals' opinions. For example, Yelp recommendations shape restaurant choices (Anderson and Magruder, 2012), TV advertisements influence consumer preferences (Adomavicius et al., 2013), and ranking algorithms shape clicks (Narayanan and Kalyanam, 2015; Ursu, 2018; Mendler-Dünner et al., 2024). More recently, LLMs have been observed to exert latent influence on users' opinions (Jakesch et al., 2023), mediate opinion exchange (Tsirtsis et al., 2026), and affect knowledge aggregation (Nettasinghe and Zhao, 2026). Phenomena such as the spread of misinformation (Del Vicario et al., 2016), filter bubble (Pariser, 2011), polarisation (Perra and Rocha, 2019), echo chamber (Cinelli et al., 2021; Chaney et al., 2018), information cocoons (Piao et al., 2023), and delusional spiraling (Chandra et al., 2026) can be unwanted outcomes from it. However, evidence against online platforms' role in causing polarization has also been provided (Guess et al., 2023; Nyhan et al., 2023). Besides empirical studies, researchers have also developed theoretical models to study AI-user interactions (Dean et al., 2024b).

A line of research models the platform as a shared source of influence. Out et al. (2024) study the influence of biased media by modeling the platform as a node with its own expressed opinion, connected to every individual in the network. Gaitonde et al. (2021) extend a geometric model of opinion dynamics and study how a random and shared platform intervention can induce polarization. Acemoglu et al. (2026) study AI systems that aggregate mass beliefs and in turn, influence users. Their work identifies a threshold of the updating frequency of the aggregator for robust improvement.

Other works model the platform as a mediator between the content and users. Pescetelli et al. (2022) treat platforms as a filter that ranks the opinions of neighbors and study the equilibrium effect of bots using agent-based simulation. Schüepp et al. (2026) model the recommender system as a distribution over content creators and study how exploiting social network structure can help with designing the recommender system that mitigates filter bubble effects. Sîrbu et al. (2019) capture how platforms promote interactions among similar individuals and extend the bounded confidence model, finding that platform influence can lead to fragmentation in settings that would otherwise yield consensus.

Beyond our single predictor setting, many researchers have explored the multi-platform scenario. Perego and Yuksel (2022) show that competition among multiple platforms can lead to informational specialization where they tend to provide more information on topics of which users' opinions are more heterogeneous. Quattrociocchi et al. (2014) empirically show that different strategies: copying or opposing the most successful media can lead to consensus or heterogeneous opinions in equilibrium. They model medias and users as two networks that can interact with each other. Helfmann et al. (2023) model the feedback loop between influencers, medias, and users using a system of stochastic differential equations. Users are influenced by the influencers they follow, while each influencer is also drawn to the average opinion of their followers.

Performative prediction. Other works in performative prediction that study explicit models for the data generating process, such as rational agent models (Hardt et al., 2016; Jagadeesan et al., 2021), causal models (Mendler-Dünner et al., 2022; Cheng et al., 2024), or parametric models (Miller et al., 2021; Izzo et al., 2021), have not considered network effects among individuals. Li et al.

(2022) study a network of performative learning agents that learn a common model from disjoint populations. They seek consensus via updates that consider weighted parameters of their neighbors. Subsequent work (Hardt, 2026) studies retraining dynamics in the presence of stable signals, which can be interpreted as individuals' innate opinions in our framework.

B Future directions

Our model is deliberately minimal, and the number of unexplored questions at the intersection of performative prediction and network science is vast. We collect here the directions that our results point to most directly.

Beyond perfect prediction and consensus. In our model we mainly focus on the case with access to perfect measurements of opinions. Measurement uncertainty, stochasticity and limited information are natural extensions, and our simulations suggest that many qualitative findings hold beyond perfect prediction. A simple additive noise model with M samples per instance already shows the direction: the equilibrium is unchanged in expectation and noise adds a variance term of order $\sigma^2\beta^2/(M(1-\beta^2))$, see Appendix F. A theory of parametric predictors is also of interest.

Our results are descriptive rather than normative, and one may question whether consensus is desirable. Consensus may reflect agreement on important proposals, but it may also entail an undesirable collapse of diversity (Tomašev et al., 2026). A further direction is to study platforms that pursue objectives beyond predictive accuracy, such as maximizing their influence over the population (Sun and Zhang, 2023; Hązła et al., 2024) or preserving opinion diversity.

Multiple platforms. Everything in this paper concerns a single platform. When alternatives are available in the market, individuals may switch to a better service and hence influence the training data for the subsequent round, which connects to multiplayer performative prediction (Dean et al., 2024a; Narang et al., 2023; Piliouras and Yu, 2023). Another interesting extension of our model is a setting where platforms adopt more active competition strategies (Perego and Yuksel, 2022; Quattrociocchi et al., 2014).

Extensions to network and interaction dynamics. We leave the network topology and the susceptibilities untouched, whereas platforms can also change these characteristics (e.g., Zhou et al., 2024; Chitra and Musco, 2020; Abebe et al., 2018; Dwork et al., 2025). We refer to Miyauchi et al. (2026) for a survey of more algorithmic interventions. Studying how directed (Sun and Zhang, 2023) and temporal (Holme and Saramäki, 2012) networks will influence the co-evolution of peer and platform influence are also natural follow-ups.

We assume that performative effects are self-fulfilling, with $\beta_i \in [0, 1]$, whereas the users may also be pushed away from recommended content. Together with the linear averaging dynamics, this makes our consensus result differ from the polarization literature (Sîrbu et al., 2019; Dandekar et al., 2013; Hązła et al., 2024), which incorporates mechanisms such as biased assimilation, homophily, and bounded confidence. Nonlinear individual responses to predictions provide a future direction for studying polarization in predictive loops. This is of particular interest with the emergence and wide adoption of LLMs (Tsirtsis et al., 2026; Jakesch et al., 2023).

The number K of peer-interaction steps between two updates of platform policies is also connected to stateful performative prediction, where the population responds gradually to a deployed model (Brown et al., 2022; Li et al., 2022), and to the question of who leads and who follows in performative settings (Zrnic et al., 2021). Theorem 1 establishes that a peer-stubborn and well-connected individual dominates the consensus when $K = \infty$. Characterizing the influence of individuals at equilibrium for heterogeneous β_i and finite K would be an interesting next step.

Measurement, identification and regulation. Our results indicate that platform mediation amplifies peer influence. This also links our results to measures of performative power (Hardt et al., 2022; Mendler-Dünner et al., 2024), suggesting that standard causal designs that ignore platform-mediated interference may misestimate population-level performative power. Developing estimation methods that account for both platform mediation and social interferences (Ugander et al., 2013; Eckles et al., 2017; Aral and Walker, 2012) is an interesting direction for future research.

More broadly, the need to understand how digital platforms reshape our society is not only scientific but also regulatory. Recognizing the important role that emerging dynamics involving platforms and users play, the Digital Services Act explicitly calls for greater transparency (OECD, 2025):

> "[...] to shed light on how the algorithms and processes deployed by these platforms influence the way information flows in our society, and influence individual platform users"

The effect of platforms over users is also a central concern in digital antitrust (Stigler Committee, 2019). Our model says what such transparency should measure: individual susceptibility β, the spillover onto individuals who are not themselves exposed, and the amplification of interventions through peers; the data-access route for vetted researchers under Article 40 of the Digital Services Act is the natural mechanism. To understand the true scale of platform influence, we must stop evaluating predictive models in isolation and start auditing how they co-evolve with society.

C Proofs

Throughout the appendix, we say $A \preceq B$ for two arbitrary symmetric matrices A, B, if $B - A$ is positive semi-definite. We let $(A)_{ij}$ denote the entry at i-th row j-th column of matrix A. The transpose of a matrix A or a vector $\boldsymbol{v}$ are denoted by $A^\top$, $\boldsymbol{v}^\top$ respectively. We denote $A \leq B$ if matrices A, B with same dimensions satisfying $(A)_{ij} \leq (B)_{ij}\ \forall i, j$. The same convention is used for vectors. A non-negative square matrix $A \in \mathbb{R}^{n\times n}$ is irreducible if $\forall i, j \in [n]$, there exists $m_{ij} \in \mathbb{N}$ such that $(A^{m_{ij}})_{ij} > 0$.

C.1 Auxiliary results

Lemma 1. (Primitive matrices) Let $M \in \mathbb{R}^{n\times n}$ be irreducible and non-negative. Further let $D \in \mathbb{R}^{n\times n}$ be a diagonal matrix with non-negative diagonal entries and at least one positive entry. Then, the sum of the two matrices $M + D$ is primitive.

Proof. Suppose for $i_0 \in [n]$, $D_{i_0,i_0} > 0$. The proof is immediate by noticing that the matrix $M + D$ is irreducible and aperiodic by adding a self-loop. □

Lemma 2. Let $M \in \mathbb{R}^{n\times n}$ be an arbitrary primitive and row-stochastic matrix, then the sequence of matrices $\{M^t\}_{t\in\mathbb{N}}$ converges as follows

$$\lim_{t\to+\infty} M^t = \mathbf{1}\boldsymbol{y}_M^\top.$$

where $\boldsymbol{y}_M$ is the eigenvector of $M^\top$ with respect to eigenvalue 1 and $\sum_{i=1}^n y_i = 1$.

Proof. The result is a corollary of the Perron-Frobenius theorem and Theorem 1.2 in Seneta (2006). From Seneta (2006) we have

$$M^t = \mathbf{1}\boldsymbol{y}_M^\top + \mathcal{O}(|\lambda_2|^t)$$

where $\{\lambda_i\}_{i=1}^e$ are eigenvalues of M such that $\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_e|$. Thus, the second term vanishes as $t \to \infty$. □

C.2 Proof of Proposition 1

Proof. To characterize the equilibrium under repeated perfect prediction, we note that

$$\begin{aligned}
\boldsymbol{x}_{\text{ex}}^{(t+1)} &= \Psi_K \boldsymbol{x}_{\text{init}}^{(t)} \\
&= \Psi_K((I_n - \Lambda_\beta)\boldsymbol{x}^* + \Lambda_\beta \boldsymbol{f}^{(t)}) \\
&= \Psi_K((I_n - \Lambda_\beta)\boldsymbol{x}^* + \Lambda_\beta \boldsymbol{x}_{\text{ex}}^{(t)}), \quad t \geq 0.
\end{aligned} \tag{11}$$

From this recursion we can see that for proving the existence of a stable point we need to be able to control $\lim_{t\to\infty}(\Psi_K\Lambda_\beta)^t$.

Applying Corollary 1 in Parsegov et al. (2016), we know Ψ_∞ is well-defined. Further, for any $K \in \{\infty\} \cup \mathbb{N}$ the matrix Ψ_K is row-stochastic by observing

$$\begin{aligned}
\Psi_K \mathbf{1} &= \left[\sum_{i=0}^{K-1} (\Lambda_\alpha W)^i (I_n - \Lambda_\alpha) + (\Lambda_\alpha W)^K \right] \mathbf{1} \\
&= (\Lambda_\alpha W)^{K-1} \Lambda_\alpha \mathbf{1} + (\Lambda_\alpha W)^{K-1} (I_n - \Lambda_\alpha) \mathbf{1} \\
&\quad + \sum_{j=0}^{K-2} (\Lambda_\alpha W)^j (I_n - \Lambda_\alpha) \mathbf{1} \\
&= (\Lambda_\alpha W)^{K-2} \Lambda_\alpha \mathbf{1} + \sum_{j=0}^{K-2} (\Lambda_\alpha W)^j (I_n - \Lambda_\alpha) \mathbf{1} \\
&= \cdots = \mathbf{1}, \quad K < \infty,
\end{aligned}$$

because W is row-stochastic. For $K = \infty$ we have $\Psi_\infty = \lim_{k\to\infty} \Psi_k$ and hence Ψ_∞ is also row-stochastic. Perron-Frobenius theorem implies $\rho(\Psi_K) \leq 1$.

Now, we take a detour and observe, under permutation of the node indices, Ψ_K can be decomposed as follows

$$\Psi_K = \begin{bmatrix} I_{n'} & \mathbf{0} \\ (\Psi_K)_{21} & (\Psi_K)_{22} \end{bmatrix}$$

where $n' = |\{i : i \in [n], \alpha_i = 0\}|$ denotes the number of individuals who are fully stubborn towards their peers. With this decomposition, we'll show the existence of $\boldsymbol{x}_{\mathrm{PS}}$ by proving $\lim_{t\to\infty} (\Psi_\infty)^t$ exists first.

- When $n' > 0$, if $\rho((\Psi_K)_{22}) < 1$ the proof is immediate, and it remains to discuss $\rho((\Psi_K)_{22}) = 1$. Here, we apply Lemma 2 in Parsegov et al. (2016) to $(\Psi_K)_{22}$, and under node permutation, we have,

  $$(\Psi_K)_{22} = \begin{bmatrix} ((\Psi_K)_{22})_{11} & ((\Psi_K)_{22})_{12} \\ \mathbf{0} & ((\Psi_K)_{22})_{22} \end{bmatrix}$$

  where block $((\Psi_K)_{22})_{11}$ contains indices $P := \{i : i \in [n]/[n'], \exists_j \in [n'] : (\Psi_K)_{ij} > 0\}$, i.e., nodes that are directly connected to completely stubborn nodes $[n']$. Since $\sum_{j=n'+1}^{n} (((\Psi_K)_{22})_{11})_{ij} < 1, \forall i \in P$ we know that $\rho(((\Psi_K)_{22})_{11}) < 1$ and $((\Psi_K)_{22})_{22}$ is row-stochastic. However, note $\mathcal{G}$ is connected and thus a node is either completely stubborn or connected to such a node. We conclude $(\Psi_K)_{22} = ((\Psi_K)_{22})_{11}$ and hence $\rho((\Psi_K)_{22}) = \rho(((\Psi_K)_{22})_{11}) < 1$. The limit $\lim_{t\to\infty} (\Psi_K)^t$ hence exists and corresponds to

  $$\lim_{t\to\infty} (\Psi_K)^t = \begin{bmatrix} I_{n'} & \mathbf{0} \\ (I_{n-n'} - (\Psi_K)_{22})^{-1} \Psi_{21} & \mathbf{0} \end{bmatrix}.$$

- When $n' = 0$, we know that Ψ_K is well-defined, as $\Lambda_\alpha \neq I_n$. For $K < \infty$, we expand the recursion and have $\Psi_K = I_n - \Lambda_\alpha + (\Lambda_\alpha W)^K + \ldots$, which shows it is connected and aperiodic, and hence primitive. When $K = \infty$, we can write $\Psi_\infty = \sum_{i=0}^{\infty} (\Lambda_\alpha W)^i (I_n - \Lambda_\alpha)$ in the following form under node permutation

  $$\Psi_\infty = \begin{bmatrix} (\Psi_\infty)_{11} & \mathbf{0} \\ (\Psi_\infty)_{21} & \mathbf{0} \end{bmatrix}$$

  where $(\Psi_\infty)_{11}$ is positive and row-stochastic, and this block corresponds to nodes with $\alpha_i < 1$. Observe

  $$\lim_{t\to\infty} (\Psi_\infty)^t = \begin{bmatrix} \lim_{t\to\infty} ((\Psi_\infty)_{11})^t & \mathbf{0} \\ (\Psi_\infty)_{21} \lim_{t\to\infty} ((\Psi_\infty)_{11})^t & \mathbf{0} \end{bmatrix}.$$

  where $\lim_{t\to\infty} ((\Psi_\infty)_{11})^t$ exists since $(\Psi_\infty)_{11}$ is primitive.

We conclude that $\lim_{t\to\infty}(\Psi_K)^t$ exists for any $\Lambda_\alpha \neq I_n$. Hence it follows that for any $\mathbf{0} \leq \Lambda_\beta \leq \boldsymbol{I}_n$, $\lim_{t\to\infty}(\Psi_K\Lambda_\beta)^t$ exists. Further, given the assumption that $\mathbf{0} \leq \Lambda_\beta \leq I_n$ with $\Lambda_\beta \neq I_n$, we have $\rho(\Psi_K\Lambda_\beta) \leq \rho(\Psi_K) = 1$ since $\Psi_K\Lambda_\beta \leq \Psi_K$ by applying a corollary of Perron-Frobenius theorem (Corollary 1.5.(a) in Berman and Plemmons (1994)).

Building on these results, we can now characterize the stable point.

If $\rho(\Psi_K\Lambda_\beta) < 1$, the dynamics in (11) are Schur stable and the convergence is immediate. We have

$$\boldsymbol{x}_{\mathrm{PS}} = (I_n - \Psi_K\Lambda_\beta)^{-1}\Psi_K(I_n - \Lambda_\beta)\boldsymbol{x}^*.$$

Now we discuss the case when $\rho(\Psi_K\Lambda_\beta) = 1$. Applying Lemma 2 in Parsegov et al. (2016), we know for row-substochastic matrix $\Psi_K\Lambda_\beta$, under node permutation, it can be decomposed into an upper triangular form

$$\Psi_K\Lambda_\beta = \begin{bmatrix} (\Psi_K\Lambda_\beta)_{11} & (\Psi_K\Lambda_\beta)_{12} \\ \mathbf{0} & (\Psi_K\Lambda_\beta)_{22} \end{bmatrix}$$

where $(\Psi_K\Lambda_\beta)_{11}$ is Schur stable and $(\Psi_K\Lambda_\beta)_{22}$ is row-stochastic. Note Ψ_K is row-stochastic and $0 \leq \beta_i \leq 1\ \forall i \in [n]$. Then for any node i in row-stochastic matrix $(\Psi_K\Lambda_\beta)_{22}$,

$$1 = \sum_j (\Psi_K)_{ij}\beta_j \leq 1.$$

The equality holds only when for any node j in row-stochastic matrix $(\Psi_K\Lambda_\beta)_{22}$, $\beta_j = 1$. Hence $(\Psi_K\Lambda_\beta)_{22} = (\Psi_K)_{22}$. Let $\Psi^*_{22} := \lim_{t\to\infty}(\Psi_{22})^t$ that exists since we've proved the existence of $\lim_{t\to\infty}(\Psi_K)^t$. The equilibrium is characterized as follows

$$\boldsymbol{x}_{\mathrm{PS}} = \begin{bmatrix} (I - (\Psi_K\Lambda_\beta)_{11})^{-1} & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} (\Psi_K)_{11} - (\Psi_K\Lambda_\beta)_{11} & (\Psi_K)_{12}\Psi^*_{22} \\ \mathbf{0} & \Psi^*_{22} \end{bmatrix} \boldsymbol{x}^*. \tag{12}$$

In particular, when $\Lambda_\beta = I_n$, the block $(\Psi_K\Lambda)_{11}$ is empty, and we have $\boldsymbol{x}_{\mathrm{PS}} = \lim_{t\to\infty}\Psi^t_K$. We proved earlier that this limit exists when $\Lambda_\alpha \neq I_n$.

Now, we prove that the inverse, $(I_n - \Psi_K\Lambda_\beta)^{-1}$, in the equilibrium characterization (12) is well-defined when $\alpha_i > 0\ \forall i \in [n]$, $\Lambda_\alpha \neq I_n$, and $\Lambda_\beta \neq I_n$. We apply a corollary of Perron-Frobenius theorem (Corollary 1.5.(b) in Berman and Plemmons (1994)) and notice that $\Psi_K\Lambda_\beta \leq \Psi_K$, $\Psi_K\Lambda_\beta \neq \Psi_K$, and $\Psi_K + \Psi_K\Lambda_\beta$ is irreducible since $\alpha_i > 0\ \forall i \in [n]$. Hence $\rho(\Psi\Lambda_\beta) < 1$ and we have

$$\boldsymbol{x}_{\mathrm{PS}} = (I_n - \Psi_K\Lambda_\beta)^{-1}\Psi_K(I_n - \Lambda_\beta)\boldsymbol{x}^*. \tag{13}$$

For non-negative matrices, we apply a corollary from Perron-Frobenius theorem (Corollary 1.5.(b) in Berman and Plemmons (1994)): If $\mathbf{0} \leq A \leq B$, then $\rho(A) \leq \rho(B)$. Observe

$$\begin{aligned} &\mathbf{0} \leq \Psi_K \leq \max_k \beta_k \Psi_K, \\ &\rho(\Psi_K\Lambda_\beta) \leq \rho(\max_k \beta_k \boldsymbol{\Psi}_K) = \max_k \beta_k \rho(\Psi_K) = \max_k \beta_k < 1. \end{aligned}$$

As for the case when some $\beta_i = 1$, the convergence rate and the choice of c is immediate from the proof above. □

C.3 Proof of distribution sensitivity

Proof. Set $\boldsymbol{f}^{(t)}$ to $\boldsymbol{f}_1^{(t)}, \boldsymbol{f}_2^{(t)}$ respectively, and let $\boldsymbol{f}_1^{(t+1)}, \boldsymbol{f}_2^{(t+1)}$ be the corresponding predictions at $t+1$. We have $\|\boldsymbol{f}_1^{(t+1)} - \boldsymbol{f}_2^{(t+1)}\| = \|\Psi_K\Lambda_\beta(\boldsymbol{f}_1^{(t)} - \boldsymbol{f}_2^{(t)})\| \leq \|\Lambda_\beta\|\|\Psi_K\|\|\boldsymbol{f}_1^{(t)} - \boldsymbol{f}_2^{(t)}\|$ where $\|\Lambda_\beta\|\|\Psi_K\|$ corresponds to the sensitivity parameter in Perdomo et al. (2020). □

C.4 Proof of Proposition 2

To study the relationship between variance and susceptibilities we build on the following intermediate result decomposing the equilibrium into a consensus component and a sum of heterogeneous components.

Proposition 4 (Decomposition of equilibrium)**.** Consider a network characterized by a regular, simple, and connected graph $\mathcal{G}$, its nodes with peer susceptibilities $\alpha_i \equiv \alpha \in (0,1)$, platform susceptibilities $\beta_i \equiv \beta \in [0,1]$, and a platform that enforces perfect predictions based on policy (5). Suppose $K = \infty$ and FJ dynamics converge. Then, the equilibrium $\boldsymbol{x}_{\mathrm{PS}}$ can be decomposed as follows

$$\boldsymbol{x}_{\mathrm{PS}} = \frac{\mathbf{1}^\top \boldsymbol{y}^*}{n(1-\beta)}\mathbf{1} + \sum_{i=2}^{n} \frac{\lambda_i \boldsymbol{v}_i^\top \boldsymbol{y}^*}{1-\beta\lambda_i}\boldsymbol{v}_i, \tag{14}$$

where $\boldsymbol{y}^* = (1-\beta)\boldsymbol{x}^*$ and $(\lambda_i, \boldsymbol{v}_i)$ are eigenvalues and corresponding eigenvectors of Ψ_∞ with $\boldsymbol{v}_i \perp \mathbf{1}\ \forall i \geq 2$.

Proof of Proposition 4. We work with the explicit expression of $\boldsymbol{x}_{\mathrm{PS}}$ in (13) where

$$\Psi_K \coloneqq \sum_{i=0}^{K-1} (\Lambda_\alpha W)^i (I_n - \Lambda_\alpha) + (\Lambda_\alpha W)^K.$$

Let us first focus on the matrix W. The matrix is symmetric and we apply the Schur decomposition $W = U\Lambda U^\top$ with $\Lambda = \mathrm{diag}(\mu_1, \cdots, \mu_n)$ and $U \coloneqq [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]$ where $\{\mu_i \boldsymbol{v}_i\}_{i\in[n]}$ denote the eigenvalues and eigenvectors of W. By definition we have $U^\top U = UU^\top = I_n$. Further, since W is row-stochastic, apply Gershgorin circle theorem, we have $\rho(W) = 1$. Apply Perron-Frobenius theorem on irreducible matrix W, w.l.o.g., let $\mu_1 = 1$ and $\boldsymbol{v}_1 = \frac{1}{\sqrt{n}}\mathbf{1}^\top$ and $\mu_i \neq 1$ and $|\mu_i| \leq 1\ \forall i \geq 2$. Using this expression in the characterization of stability, we obtain $\Psi_\infty = U\mathrm{diag}(\lambda_1, \cdots, \lambda_n)U^\top$ with $\lambda_i \coloneqq \frac{1-\alpha}{1-\alpha\mu_i}$, and hence

$$\boldsymbol{x}_{\mathrm{PS}} = U\mathrm{diag}\left(\frac{\lambda_1}{1-\beta\lambda_1}, \cdots, \frac{\lambda_n}{1-\beta\lambda_n}\right)U^\top.$$

We can easily verify that $\lambda_1 = 1$, and $0 < \lambda_i < 1\ \forall i \geq 2$ by noticing $\mu_i \neq 1$, $|\mu_i| \leq 1\ \forall i \geq 2$, and $1 - \alpha\mu_i - 1 + \alpha = \alpha(1-\mu_i) > 0$. This concludes the proof of Proposition 4. □

Before we move to the proof of Proposition 2 let us comment on this result. The characterization applies to regular networks and homogeneous parameters. It shows that the equilibrium has an invariant consensus component $\frac{\mathbf{1}}{\sqrt{n}}$ and its coefficient $\frac{1}{1-\beta}$ is invariant with respect to the choice of Λ_α and $\mathcal{G}$. The other components characterize the heterogeneity of opinion dynamics in different directions. The more the heterogeneous components $\boldsymbol{v}_i$, $i \geq 2$ are aligned with the invariant component $(1-\beta)\boldsymbol{x}^*$ and the larger the eigenvalue λ_i, the more $\boldsymbol{v}_i$ will contribute to the equilibrium. While we notice that the consensus component "dominates" in the sense that

$$0 < \frac{\lambda_i}{1-\beta\lambda_i} < \frac{1}{1-\beta},\ \forall i.$$

Next we build on this decomposition to prove Proposition 2.

Proof of Proposition 2. Note $\boldsymbol{v}_i \perp \mathbf{1}\ \forall i \geq 2$,

$$\begin{aligned}\bar{\boldsymbol{x}}_{\mathrm{PS}} &= \frac{(1-\beta)\mathbf{1}^\top \sum_{i=1}^{n} \frac{\lambda_i}{1-\beta\lambda_i}\boldsymbol{v}_i\boldsymbol{v}_i^\top \boldsymbol{x}^*}{n} \\ &= \frac{\mathbf{1}^\top \boldsymbol{x}^*}{n}.\end{aligned}$$

As for the variance of opinions, we have

$$\begin{aligned}\mathrm{Var}(\boldsymbol{x}_{\mathrm{PS}}) &= \left\| \sum_{i=1}^{n} \frac{\lambda_i}{1-\beta\lambda_i}\boldsymbol{v}_i^\top((1-\beta)\boldsymbol{x}^*)\boldsymbol{v}_i - \bar{\boldsymbol{x}}_{\mathrm{PS}} \right\|^2 \\ &= \left\| \sum_{i=2}^{n} \frac{\lambda_i}{1-\beta(1-\eta)\lambda_i}\boldsymbol{v}_i^\top((1-\beta)\boldsymbol{x}^*)\boldsymbol{v}_i \right\|^2 \\ &= \sum_{i=2}^{n} \left(\frac{\lambda_i}{1-\beta\lambda_i}\boldsymbol{v}_i^\top((1-\beta)\boldsymbol{x}^*) \right)^2.\end{aligned}$$

As for the relationship between platform susceptibility β and $\mathrm{Var}(\boldsymbol{x}_{\mathrm{PS}})$, we have

$$\frac{\partial \mathrm{Var}(\boldsymbol{x}_{\mathrm{PS}})}{\partial \beta} = \sum_{i=2}^{n} \frac{2\lambda_i^2(1-\beta)(\lambda_i - 1)(\boldsymbol{v}_i^\top \boldsymbol{x}^*)^2}{(1-\beta\lambda_i)^3} < 0.$$

□

Similarly we can also show that under the conditions of Proposition 2 the variance decreases as α grows. We omit this result from the main body as it follows from classical insights in peer dynamics. Namely, via basic calculations, we have

$$\frac{\partial \left(\frac{\lambda_i}{1-\beta\lambda_i}\right)^2}{\partial \alpha} = \frac{-\lambda_i(1-\mu_i)}{(1-\beta\lambda_i)^3(1-\alpha\mu_i)^2} < 0 \quad i \geq 2$$

and hence if there exists $i_0 \geq 2$ such that $\boldsymbol{v}_{i_0}^\top \boldsymbol{x}^* \neq 0$, the larger α is, the larger $\mathrm{Var}(\boldsymbol{x}_{\mathrm{PS}})$ is. Also, note $\mathrm{Var}(\boldsymbol{x}^*) = \sum_{i=2}^n (\boldsymbol{v}_i^\top \boldsymbol{x}^*)^2$, hence $\mathrm{Var}(\boldsymbol{x}_{\mathrm{PS}})$ is also implicitly determined by $\mathrm{Var}(\boldsymbol{x}^*)$. It follows that

$$\frac{\partial \mathrm{Var}(\boldsymbol{x}_{\mathrm{PS}})}{\partial \alpha} < 0.$$

C.5 Proof of Theorem 1

Next we show consensus in the limit $\beta_i \to 1 \ \forall i$. We prove the claim without the homogeneity assumption on β. Instead assume the weaker numerical condition on how the limit is attained for different individuals: All we assume is that the limit $\lim_{\beta_i,\beta_j \to 1} \frac{1-\beta_i}{1-\beta_j}$ exists.

Proof. To prove Theorem 1 we first show that Ψ_K is primitive under the conditions of the theorem. We have

$$\Psi_1 = (I_n - \Lambda_\alpha) + \Lambda_\alpha W$$

$$\Psi_K = \Lambda_\alpha W(I_n - \Lambda_\alpha) + (I_n - \Lambda_\alpha) + \sum_{i=2}^{K-1} (\Lambda_\alpha W)^i (I_n - \Lambda_\alpha) + (\Lambda_\alpha W)^K, \quad K \geq 2.$$

We note that $\Lambda_\alpha W(I_n - \Lambda_\alpha)$ is irreducible given that $\alpha_i \in (0,1) \ \forall i \in [n]$.[3] By Lemma 1, we know that Ψ_K is primitive for any $K \in \{\mathbb{N} \cup \infty\}$.

From the equilibrium characterization in Equation (13), we have that for $\Lambda_\beta = \mathbf{1}$ it holds that $\boldsymbol{x}_{\mathrm{PS}} = \lim_{t\to\infty} \Psi_K^t \boldsymbol{x}^*$. Hence, we get convergence to

$$c^* = \boldsymbol{y}\boldsymbol{f}^{(0)} = \boldsymbol{y}^\top \boldsymbol{x}^*$$

where $\boldsymbol{y}$ is the eigenvector of $(\Psi_K)^\top$ and satisfies $\mathbf{1}^\top \boldsymbol{y} = 1$.

It remains to show that in the limit $\Lambda_\beta \to \mathbf{1}$ this behavior is maintained. Thus, for any sequence $\Lambda_\beta \to \mathbf{1}$, we choose $\epsilon > 0$ and an arbitrary $c > 1$ such that $\frac{\epsilon}{c} \preceq I_n - \Lambda_\beta \preceq \epsilon I_n$

$$\begin{aligned}\left(\sum_{m=0}^{\infty} (\Psi_K(I_n - \epsilon I_n))^m \frac{\epsilon}{c}\right)_{ij} &\leq \left(\sum_{m=0}^{\infty} (\Psi_K \Lambda_\beta)^m (I_n - \Lambda_\beta)\right)_{ij} \\ &\leq \left(\sum_{m=0}^{\infty} (\Psi_K(I_n - \frac{\epsilon}{c} I_n))^m \epsilon\right)_{ij}\end{aligned}$$

where the inequality holds since Ψ is non-negative and $I_n - \Lambda_\beta$ is diagonal with positive entries. Under the assumption that $\lim_{i,j} \frac{1-\beta_i}{1-\beta_j}$ exists, let $\epsilon \to 0$ and apply the theorem from Friedkin and Johnsen (1990), we get

$$\begin{aligned}\left(\frac{1}{c} \lim_{m\to\infty} (\Psi_K)^m\right)_{ij} &\leq \lim_{\beta \to \mathbf{1}} \left(\sum_{m=0}^{\infty} (\Psi_K \Lambda_\beta)^m (I_n - \Lambda_\beta)\right)_{ij} \\ &\leq \left(c \lim_{m\to\infty} (\Psi_K)^m\right)_{ij}.\end{aligned}$$

[3] A side note is that for $K < \infty$, Ψ_K is primitive with a weaker assumption that $\Lambda_\alpha \neq I_n$.

Since $c > 1$ can be chosen arbitrarily, apply equilibrium (13) we have

$$\lim_{\Lambda_\beta \to \mathbf{1}} (I_n - \Psi_K \Lambda_\beta)^{-1}(I_n - \Lambda_\beta) = \lim_{j \to \infty} (\Psi_K)^j.$$

Apply Corollary 2, opinions reach consensus as follows

$$\lim_{\Lambda_\beta \to \mathbf{1}} \boldsymbol{x}_{\mathrm{PS}} = \mathbf{1}^\top \boldsymbol{y} \lim_{\Lambda_\beta \to \mathbf{1}} (I_n - \Lambda_\beta)^{-1} \Psi_K (I_n - \Lambda_\beta) \boldsymbol{x}^*.$$

In the end, we show that when $K = \infty$, consensus $c^* = \sum_{i=1} \frac{(1-\alpha_i) d_i x_i^*}{\alpha_i} / \sum_{j=1}^n \frac{(1-\alpha_j) d_j}{\alpha_j}$ by proving $\boldsymbol{y} \propto (I_n - \Lambda_\alpha) \cdot \Lambda_\alpha^{-1} \cdot \deg \cdot \mathbf{1}$. By definition of $\boldsymbol{y}$ and Ψ_∞, we have $\boldsymbol{y}^\top \Psi_\infty = \boldsymbol{y}^\top$ and hence

$$\begin{aligned}
\boldsymbol{y}^\top (I_n - \Lambda_\alpha W)^{-1}(I_n - \Lambda_\alpha) &= \boldsymbol{y}^\top \\
\boldsymbol{y}^\top (I_n - \Lambda_\alpha W)^{-1} - \boldsymbol{y}^\top (I_n - \Lambda_\alpha W)^{-1} \Lambda_\alpha &= \boldsymbol{y}^\top (I_n - \Lambda_\alpha W)^{-1}(I_n - \Lambda_\alpha W) \\
\boldsymbol{y}^\top (I_n - \Lambda_\alpha W)^{-1} \Lambda_\alpha &= \boldsymbol{y}^\top (I_n - \Lambda_\alpha W)^{-1} \Lambda_\alpha W.
\end{aligned}$$

Since W is irreducible and apply Perron-Frobenius theorem, 1 is a simple eigenvalue of W (e.g. Theorem 1.4 (b) of Chapter 2 in Berman and Plemmons (1994)). Hence we obtain

$$\boldsymbol{y}^\top (I_n - \Lambda_\alpha W)^{-1} \Lambda_\alpha \propto \deg \cdot \mathbf{1},$$

and conclude the proof. □

C.6 Proof of Proposition 3 and Theorem 2

Proof. First of all, we prove the convergence of the dynamics under Equation 8. Throughout this section, let $n' := |\mathcal{V}'|$ and define

$$\begin{aligned}
&\Lambda_{\beta,1} := (\Lambda_\beta)_{1:n',1:n'} \qquad \Lambda_{\beta,2} := (\Lambda_\beta)_{n'+1:n,n'+1:n}, \qquad R := \mathbf{1} \begin{bmatrix} r_1 & \cdots & r_{n'} \end{bmatrix}^\top \\
&\boldsymbol{x}'_{\mathrm{PS}} := (\boldsymbol{x}_{\mathrm{PS}})_{1:n'} \qquad \Psi'_K := (\Psi_K)_{1:n',1:n'}.
\end{aligned}$$

Without loss of generality we can rewrite the predictions as follows,

$$\boldsymbol{f}^{(t+1)} = M \boldsymbol{x}_{\mathrm{ex}}^{(t)}, \quad M := \begin{bmatrix} I_{n'} & \mathbf{0}_{n' \times (n-n')}, \\ R & \mathbf{0}_{(n-n') \times (n-n')} \end{bmatrix}.$$

Hence we have

$$\boldsymbol{x}_{\mathrm{ex}}^{(t+1)} = \Psi_K ((I_n - \Lambda_\beta) \boldsymbol{x}^* + \Lambda_\beta M \boldsymbol{x}_{\mathrm{ex}}^{(t)}).$$

We look at $\Psi_K \Lambda_\beta M$ in a block-wise way,

$$\Psi_K \Lambda_\beta M = \begin{bmatrix} (\Psi_K \Lambda_\beta M)_{1:n',1:n'} & \mathbf{0}_{n' \times (n-n')} \\ * & \mathbf{0}_{(n-n') \times (n-n')} \end{bmatrix},$$

- When $\Lambda_\beta = I_n$, we have

  $$\boldsymbol{x}_{\mathrm{ex}}^{(t)} = (\Psi_K M)^t \boldsymbol{x}_{\mathrm{ex}}^{(0)}.$$

  It is easy to verify that $\Psi_K M$ is still a row-stochastic matrix and

  $$\begin{aligned}
  \Psi_K M &= \begin{bmatrix} (\Psi_K M)_{1:n',1:n'} & \mathbf{0}_{n' \times (n-n')} \\ * & \mathbf{0}_{(n-n') \times (n-n')} \end{bmatrix}, \\
  (\Psi_K M)^t &= \begin{bmatrix} ((\Psi_K M)_{1:n',1:n'})^t & \mathbf{0}_{n' \times (n-n')} \\ * & \mathbf{0}_{(n-n') \times (n-n')} \end{bmatrix}
  \end{aligned}$$

  The primitivity of the block $(\Psi_K M)_{1:n',1:n'}$ comes from an observation: We assume a virtual node $n+1$ assigned with the weighted opinion of $\mathcal{V}'$, $\sum_{i \in \mathcal{V}'} r_i (\boldsymbol{x}_{\mathrm{ex}}^{(t)})_j$, which represents the union of $\mathcal{V} \backslash \mathcal{V}'$. From each node in $\mathcal{V}'$, there's a directed edge to node $n+1$. While the node $n+1$ is connected with some of nodes in $\mathcal{V}'$ with edges starting at $n+1$ because of the connectivity assumption. Hence for any pair of nodes in $\mathcal{V}'$, there is a path between them, the primitivity is immediate by noticing $(\Psi_K)_{ii} > 0$ for $i \in [n']$. Hence opinions in $\mathcal{V}'$ will converge to consensus using the same proof as Theorem 1. As the platform is using the weighted mean predictor, we conclude that the whole population will converge to consensus as $t \to \infty$, which concludes our proof of Theorem 2. We observe that the consensus opinion c' is influenced by nodes in $\mathcal{V}'$ and no assumption about $\alpha_j, j \in \mathcal{V}/\mathcal{V}'$, is needed.

- For general $\Lambda_{\boldsymbol{\beta}}$, since $\lim_{t\to\infty}((\Psi_K M)_{1:n',1:n'})^t$ exists, we know $\lim_{t\to\infty}((\Psi_K \Lambda_\beta M)_{1:n',1:n'})^t$ also exists and apply the proof of Proposition 1. The dynamics over $\mathcal{V}'$ converges. And the equilibrium can be characterized as follows

$$\boldsymbol{x}_{\mathrm{PS}} = \Psi_K \begin{bmatrix} \Lambda_{\boldsymbol{\beta},1}\boldsymbol{x}'_{\mathrm{PS}} + (I_{n'} - \Lambda_{\beta,1})(\boldsymbol{x}^*)_{1:n'} \\ \Lambda_{\beta,2}R\boldsymbol{x}'_{\mathrm{PS}} + (I_{n-n'} - \Lambda_{\beta,2})(\boldsymbol{x}^*)_{n'+1:n} \end{bmatrix}.$$

  Especially, when $\Lambda_{\beta,1} < I_{n'}$ and consider $\boldsymbol{x}'_{\mathrm{PS}}$,

$$\begin{aligned} \boldsymbol{x}'_{\mathrm{PS}} &= (\Psi_K \Lambda_\beta M)_{1:n',1:n'}\boldsymbol{x}'_{\mathrm{PS}} + (\Psi_K(I_n - \Lambda_\beta)\boldsymbol{x}^*)_{1:n'} \\ (\boldsymbol{x}'_{\mathrm{PS}} \mid_{\boldsymbol{x}^*} - \boldsymbol{x}'_{\mathrm{PS}} \mid_{\tilde{\boldsymbol{x}}^*}) &= (x^*_{i_0} - \tilde{x}^*_{i_0})(I_{n'} - (\Psi_K \Lambda_\beta M)_{1:n',1:n'})^{-1}(\Psi_K(I_n - \Lambda_\beta)\boldsymbol{e}_{i_0})_{1:n'}. \end{aligned}$$

  Note that since $(I_{n'} - (\Psi_K \Lambda_\beta M)_{1:n',1:n'})^{-1} = \sum_{t=0}^{\infty}((\Psi_K \Lambda_\beta M)_{1:n',1:n'})^t$ is positive. The variation of $\boldsymbol{x}'_{\mathrm{PS}}$ has the same sign as $x^*_{i_0} - \tilde{x}^*_{i_0}$.

Now we are ready to prove Proposition 3. Since we assume $\Lambda_{\beta,1} < I_{n'}$, $\boldsymbol{x}^*_{i_0} \neq \tilde{\boldsymbol{x}}^*_{i_0}$, for some $i_0 \in [n']$, $\boldsymbol{x}^*_i = \tilde{\boldsymbol{x}}^*_i, \forall i \in [n']/\{i_o\}$. Note $\alpha_q = 0$ and hence $(\Psi_K)_{qq} = 1$, we have

$$\begin{aligned}
&\boldsymbol{x}_{\mathrm{PS}} \mid_{\boldsymbol{x}^*} - \boldsymbol{x}_{\mathrm{PS}} \mid_{\tilde{\boldsymbol{x}}^*} = \Psi_K \begin{bmatrix} \Lambda_{\beta,1}(\boldsymbol{x}_{\mathrm{PS}} \mid_{\boldsymbol{x}^*} - \boldsymbol{x}_{\mathrm{PS}} \mid_{\tilde{\boldsymbol{x}}^*}) + (1-\beta_{i_0})(x^*_{i_0} - \tilde{x}^*_{i_0})\boldsymbol{e}_{i_0} \\ \sum_{i\in\mathcal{V}'} r_i((\boldsymbol{x}_{\mathrm{PS}} \mid_{\boldsymbol{x}^*})_i - (\boldsymbol{x}_{\mathrm{PS}} \mid_{\tilde{\boldsymbol{x}}^*})_i)\Lambda_{\beta,2}\mathbf{1} \end{bmatrix} \\
&(\boldsymbol{x}_{\mathrm{PS}} \mid_{\boldsymbol{x}^*})_q - (\boldsymbol{x}_{\mathrm{PS}} \mid_{\tilde{\boldsymbol{x}}^*})_q \\
&= \sum_{i=1}^{n'}(\Psi_K)_{q,i}\beta_i((\boldsymbol{x}_{\mathrm{PS}} \mid_{\boldsymbol{x}^*})_i - (\boldsymbol{x}_{\mathrm{PS}} \mid_{\tilde{\boldsymbol{x}}^*})_i) \\
&\quad + \sum_{i\in[n']} r_i((\boldsymbol{x}_{\mathrm{PS}} \mid_{\boldsymbol{x}^*})_i - (\boldsymbol{x}_{\mathrm{PS}} \mid_{\tilde{\boldsymbol{x}}^*})_i) \sum_{j=n'+1}^{n}(\Psi_K)_{q,j}\beta_j \\
&= \beta_q \sum_{i\in[n']} r_i((\boldsymbol{x}_{\mathrm{PS}} \mid_{\boldsymbol{x}^*})_i - (\boldsymbol{x}_{\mathrm{PS}} \mid_{\tilde{\boldsymbol{x}}^*})_i) \\
&= \beta_q(x^*_{i_0} - \tilde{x}^*_{i_0}) \begin{bmatrix} r_1 & \cdots & r_{n'} \end{bmatrix}^\top (I_{n'} - (\Psi_K \Lambda_\beta M)_{1:n',1:n'})^{-1}(\Psi_K(I_n - \Lambda_\beta)\boldsymbol{e}_{i_0})_{1:n'} \\
&> \beta_q(x^*_{i_0} - \tilde{x}^*_{i_0}) \begin{bmatrix} r_1 & \cdots & r_{n'} \end{bmatrix}^\top \begin{bmatrix} 0 \\ \vdots \\ (\Psi_K)_{i_0 1}(1-\beta_{i_0}) \\ \vdots \\ 0 \end{bmatrix} \\
&> \beta_q r_{i_0}(1-\beta_{i_0})(\Psi_K)_{i_0 i_0}(x^*_{i_0} - \tilde{x}^*_{i_0}) > 0
\end{aligned}$$

Since $\alpha_i \in (0,1)$ for $i \in [n']$, $(\Psi_K)_{i_0 i_0} > 0$. Let $c := r_{i_0}(\Psi_K)_{i_0 i_0}$ and we complete the proof for Proposition 3. □

C.7 Proof of Theorem 3

To characterize the indirect platform influence, we mainly prove and utilize the fact that the equilibrium under the assumptions of Theorem 3 can be written as

$$\boldsymbol{x}_{\mathrm{PS}} = \frac{\beta s(1-\gamma+\gamma\psi_2)\psi_1}{1-\gamma+2\gamma\psi_2-\gamma^2\psi_1\psi_2}e_j - \frac{\beta s\gamma\psi_1\psi_2}{1-\gamma+2\gamma\psi_2-\gamma^2\psi_1\psi_2}e_l + \frac{\beta s\psi_2}{1-\gamma+2\gamma\psi_2-\gamma^2\psi_1\psi_2}\mathbf{1} \tag{15}$$

where $\psi_1 := \frac{(1-\alpha)(n-1)}{n-1+\alpha} > 0$, $\psi_2 := \frac{\alpha}{n-1+\alpha} > 0$.

With the decomposition in (15), we have

$$\mathrm{Mean}(\boldsymbol{x}_{\mathrm{PS}}) = \frac{\mathbf{1}^\top \boldsymbol{x}_{\mathrm{PS}}}{n} = \frac{\beta s(1-\gamma\psi_1)}{n(1-\gamma+2\gamma\psi_2-\gamma^2\psi_1\psi_2)}.$$

In particular, $(\boldsymbol{x}_{\mathrm{PS}})_l = \frac{\beta s \psi_2(-\gamma\psi_1+1)}{1-\gamma+2\gamma\psi_2-\gamma^2\psi_1\psi_2}$. Take the derivative of $\mathrm{Mean}(\boldsymbol{x}_{\mathrm{PS}})$, $(\boldsymbol{x}_{\mathrm{PS}})_l$ with respect to γ and we have

$$\frac{\partial \mathrm{Mean}(\boldsymbol{x}_{\mathrm{PS}})}{\partial \gamma} = \frac{\beta s(1-\psi_1-2\psi_2+2\gamma\psi_1\psi_2-\gamma^2\psi_1^2\psi_2)}{(1-\gamma+2\gamma\psi_2-\gamma^2\psi_1\psi_2)^2} > 0.$$
$$\frac{\partial (\boldsymbol{x}_{\mathrm{PS}})_l}{\partial \gamma} = \frac{\beta s\psi_2(1-\psi_1-2\psi_2+2\psi_1\psi_2-\gamma^2\psi_1^2\psi_2)}{(1-\gamma+2\gamma\psi_2-\gamma^2\psi_1\psi_2)^2} > 0.$$

The above inequalities hold since

$$1-\psi_1-2\psi_2 \geq 1-\psi_1-n\psi_2 = 0, \text{ and}$$
$$2\psi_1\psi_2-\gamma^2\psi_1^2\psi_2 > 2\gamma\psi_1\psi_2-\gamma^2\psi_1^2\psi_2 > \gamma^2\psi_1\psi_2-\gamma^2\psi_1^2\psi_2 \geq 0.$$

Hence $\frac{\partial\left((\boldsymbol{x}_{\mathrm{PS}})_l-(\boldsymbol{x}_{\mathrm{PS}})_l^{\gamma=0}\right)}{\partial\gamma} > 0$.

Proof of Equation (15). Now we prove the decomposition of $\boldsymbol{x}_{\mathrm{PS}}$ step by step. With the dynamics described in Proposition 3, we have the equations

$$\boldsymbol{x}_{\mathrm{init}}^{(t+1)} = (I_n-\Lambda')\boldsymbol{x}^* + \tilde{\Lambda}\boldsymbol{x}_{\mathrm{ex}}^{(t)} + \beta s e_j,$$
$$\boldsymbol{x}_{\mathrm{ex}}^{(t+1)} = \Psi_\infty \boldsymbol{x}_{\mathrm{init}}^{(t+1)},$$

where $\Lambda' := \gamma I_n - \gamma e_l e_l^\top + (\beta-\gamma)e_j e_j^\top$ and $\tilde{\Lambda} := \gamma I_n - \gamma e_l e_l^\top - \gamma e_j e_j^\top$. Let $S := I_n - \Psi_\infty \Lambda''_\beta$, $\Lambda'' := \gamma I_n - \gamma e_l e_l^\top$. Since $\alpha, \gamma, \beta \in (0,1)$, the equilibrium is well-defined:

$$\begin{aligned}
\boldsymbol{x}_{\mathrm{PS}} &= (I_n - \Psi_\infty\tilde{\Lambda})^{-1}\Psi_\infty((I_n-\Lambda')\boldsymbol{x}^* + \beta s e_j) \\
&= \left(S^{-1} - \frac{\gamma S^{-1}\Psi_\infty e_j e_j^\top S^{-1}}{1+\gamma e_j^\top S^{-1}\Psi_\infty e_j}\right)\Psi_\infty\left((I_n-\Lambda''-(\beta-\gamma)e_j e_j^\top)\boldsymbol{x}^* + \beta s e_j\right) \\
&= S^{-1}\Psi(I_n-\Lambda'')\boldsymbol{x}^* - \frac{\gamma S^{-1}\Psi_\infty e_j e_j^\top S^{-1}\Psi_\infty(I_n-\Lambda'')\boldsymbol{x}^*}{1+\gamma e_j^\top S^{-1}\Psi_\infty e_j} \\
&\quad + \left(S^{-1} - \frac{\gamma S^{-1}\Psi_\infty e_j e_j^\top S^{-1}}{1+\gamma e_j^\top S^{-1}\Psi_\infty e_j}\right)\Psi_\infty(\beta s - (\beta-\gamma)e_j^\top \boldsymbol{x}^*)e_j \\
&= \beta s\left(S^{-1} - \frac{\gamma S^{-1}\Psi_\infty e_j e_j^\top S^{-1}}{1+\gamma e_j^\top S^{-1}\Psi_\infty e_j}\right)\Psi_\infty e_j
\end{aligned}$$

where we apply the Sherman–Morrison formula in the second equation. Furthermore, we apply Sherman–Morrison formula again and observe

$$S^{-1} = (I_n-\gamma\Psi_\infty)^{-1} - \frac{\gamma(I_n-\gamma\Psi_\infty)^{-1}\Psi_\infty e_l e_l^\top (I_n-\gamma\Psi_\infty)^{-1}}{1+\gamma e_l^\top (I_n-\gamma\Psi_\infty)^{-1}\Psi_\infty e_l}.$$

Note

$$\begin{aligned}
\Psi_\infty &= (1-\alpha)(I_n - \frac{\alpha}{n-1}(\mathbf{1}\mathbf{1}^\top - I_n))^{-1} \\
&= \frac{(1-\alpha)(n-1)}{n-1+\alpha} I_n + \frac{\alpha}{n-1+\alpha}\mathbf{1}\mathbf{1}^\top \\
&= \psi_1 I_n + \psi_2 \mathbf{1}\mathbf{1}^\top
\end{aligned}$$

where $\psi_1 := \frac{(1-\alpha)(n-1)}{n-1+\alpha}$, $\psi_2 := \frac{\alpha}{n-1+\alpha}$, and $\psi_1 + n\psi_2 = 1$. Hence

$$(I_n-\gamma\Psi_\infty)^{-1} = \frac{1}{1-\gamma\psi_1} I_n + \frac{\gamma\psi_2}{(1-\gamma\psi_1)(1-\gamma)}\mathbf{1}\mathbf{1}^\top = i_1 I_n + i_2 \mathbf{1}\mathbf{1}^\top$$

where $i_1 := \frac{1}{1-\gamma\psi_1}$, $i_2 := \frac{\gamma\psi_2}{(1-\gamma\psi_1)(1-\gamma)}$, and note $1+\gamma i_1\psi_1 = i_1$, $\gamma\psi_2 i_1 = (1-\gamma)i_2$. Then we have

$$\begin{aligned}
S^{-1} &= i_1 I_n + \left(i_2 - \frac{\gamma(i_2^2 + i_1 i_2\psi_2)}{1+\gamma(i_1\psi_1 + i_1\psi_2 + i_2)}\right)\mathbf{1}\mathbf{1}^\top - \frac{\gamma(\psi_1 i_1^2 e_l e_l^\top + i_1(i_2+i_1\psi_2)\mathbf{1}e_l^\top + i_1 i_2\psi_1 e_l \mathbf{1}^\top)}{1+\gamma(i_1\psi_1+i_1\psi_2+i_2)} \\
&= i_1 I_n + a_2\mathbf{1}\mathbf{1}^\top + a_3 e_l e_l^\top + a_4 \mathbf{1}e_l^\top + a_5 e_l \mathbf{1}^\top,
\end{aligned}$$

where

$$a_2 := i_2 - \frac{\gamma(i_2^2 + i_1 i_2 \psi_2)}{1+\gamma(i_1\psi_1 + i_1\psi_2 + i_2)} = \frac{i_2(1+\gamma i_1\psi_1)}{1+\gamma(i_1\psi_1 + i_1\psi_2 + i_2)} = \frac{i_2 i_1}{i_1 + i_2},$$
$$a_3 := \frac{-\gamma\psi_1 i_1^2}{1+\gamma(i_1\psi_1 + i_1\psi_2 + i_2)} = \frac{i_1 - i_1^2}{i_1 + i_2}, \quad a_4 := \frac{-\gamma i_1(i_2 + i_1\psi_2)}{1+\gamma(i_1\psi_1 + i_1\psi_2 + i_2)} = \frac{-i_1 i_2}{i_1 + i_2},$$
$$a_5 := \frac{-\gamma i_1 i_2 \psi_1}{i_1 + i_2} = \frac{i_2 - i_1 i_2}{i_1 + i_2}.$$

Note $j \neq l$ and then

$$\begin{aligned}
\boldsymbol{x}_{\mathrm{PS}} &= \frac{\beta s}{1+\gamma(i_1\psi_1 + a_2 + a_4\psi_1 + i_1\psi_2 + a_4\psi_2)}(i_1\psi_1 e_j + (a_3\psi_2 + a_5)e_l + (a_2 + i_1\psi_2 + a_4\psi_2)\mathbf{1}) \\
&= \frac{\beta s(i_1 + i_2)}{\gamma(i_1^2 + i_1 i_2 + i_2)}\left((i_1 - 1)e_j + \frac{i_2 - i_1 i_2}{(i_1 + i_2)}e_l + \frac{i_1 i_2}{(i_1 + i_2)}\mathbf{1}\right), \\
&= \frac{\beta s(1 - \gamma + \gamma\psi_2)\psi_1}{1 - \gamma + 2\gamma\psi_2 - \gamma^2\psi_1\psi_2}e_j - \frac{\beta s\gamma\psi_1\psi_2}{1 - \gamma + 2\gamma\psi_2 - \gamma^2\psi_1\psi_2}e_l + \frac{\beta s\psi_2}{1 - \gamma + 2\gamma\psi_2 - \gamma^2\psi_1\psi_2}\mathbf{1}.
\end{aligned}$$

which concludes the proof. □

C.8 Proof of Corollary 1

Apply Theorem 1 to network $\mathcal{G}_1$ and we know that a consensus is reached on $\mathcal{G}_1$. Since the nodes in $\mathcal{G}_2$ are strongly influenced by the predictions of the platform, which is a convex combination of opinions on $\mathcal{G}_1$, the opinions of $\mathcal{V}_2$ converge to the same consensus on $\mathcal{G}_1$ as well.

D Peer-platform co-influence under the DeGroot model

For the main body of this paper we have focused on the Friedkin-Johnsen model to characterize peer dynamics. However, our analysis can also be extended to the DeGroot model (DeGroot, 1974). In the following we show how the consensus result and how the intuition form the Friedkin-Johnsen model transfers. The technical challenge is resolved by the additional assumption that graph $\mathcal{G}$ is primitive. The other resultss can also be extended with minor modifications.

Proposition 5 (Consensus under DeGroot model)**.** Consider a network characterized by a simple primitive graph $\mathcal{G}$, its nodes with peer susceptibilities $\Lambda_\alpha = I_n$ and platform susceptibilities $\beta_i \in [0, 1]\ \forall i \in [n]$. Suppose $K = \infty$ and the platform performs perfect prediction of past data. Then, there exists a constant $d^* \in [0, 1]$ such that a consensus is reached at equilibrium, i.e.,

$$(\boldsymbol{x}_{\mathrm{PS}})_i = d^*, \quad \forall i \in [n].$$

Furthermore, if W is doubly-stochastic, we have $d^* = \frac{\sum_{i=1}^n (1-\beta_i)x_i^*}{n - \sum_{i=1}^n \beta_i}$ when $\Lambda_\beta \neq I_n$, and $d^* = \frac{\sum_{i=1}^n x_i^*}{n}$ when $\Lambda_\beta = I_n$.

Proof. Under supervised learning policy (5), apply (6) and set $\Lambda_\alpha = I_n$, we have

$$\boldsymbol{x}_{\mathrm{ex}}^{(t+1)} = W^\infty((I_n - \Lambda_\beta)\boldsymbol{x}^* + \Lambda_\beta \boldsymbol{x}_{\mathrm{ex}}^{(t)}).$$

- When $\Lambda_\beta \neq I_n$, note $W^\infty\Lambda_\beta \leq W^\infty$, $W^\infty\Lambda_\beta \neq W^\infty$ and $W^\infty\Lambda_\beta + W^\infty$ is irreducible since W is primitive, apply a corollary of Perron-Frobenius theorem (Corollary 1.5.(b) in Berman and Plemmons (1994)) and we have $\rho(W^\infty\Lambda_\beta) < \rho(W^\infty) = 1$. Hence we characterize the equilibrium as follows

  $$\boldsymbol{x}_{\mathrm{PS}} = (I_n - W^\infty\Lambda_\beta)^{-1}W^\infty(I_n - \Lambda_\beta)\boldsymbol{x}^* = \sum_{t=0}^{\infty}(W^\infty\Lambda_\beta)^t W^\infty(I_n - \Lambda_\beta)\boldsymbol{x}^*.$$

  Apply Perron-Frobenius theorem, we have $(I_n - W^\infty\Lambda_\beta)^{-1}$ is well-defined since $\beta_i \in (0, 1)$ and W^∞ is primitive. Moreover, we know $(W^\infty\Lambda_\beta)^t W^\infty(I_n - \Lambda_\beta)\boldsymbol{x}^* = c_t\mathbf{1}$ for some constant $c_t \in [0, 1], \forall t \geq 0$. Hence there exists $d^* \in [0, 1]$ such that $(\boldsymbol{x}_{\mathrm{PS}})_i = d^*, \forall i \in [n]$.

- When $\Lambda_\beta = I_n$, we have consensus again since W is primitive and

$$\boldsymbol{x}_{\mathrm{PS}} = W^\infty \boldsymbol{x}^*.$$

When W is doubly-stochastic, we know each node has the same long-run influence by applying Lemma 2. In other words, $\bar{\boldsymbol{x}}_{\mathrm{init}}^{(t)} = \bar{\boldsymbol{x}}_{\mathrm{ex}}^{(t)}$, i.e., when $\Lambda_\beta \neq I_n$,

$$nd^* = \sum_{i=1}^n (1-\beta_i)x_i^* + d^* \sum_{j=1}^n \beta_j.$$

When $\Lambda_\beta = I_n$, we have

$$d^* = \frac{\sum_{i=1}^n x_i^*}{n},$$

and we conclude our proof. □

E Results with finite K

We consider the following steering policy (3) where only finite K steps are observed for updating policies and the following equilibrium point performative stability

$$\boldsymbol{x}_{\mathrm{PS}} \coloneqq \lim_{t\to\infty} \boldsymbol{x}_{\mathrm{ex}}^{(t)}. \tag{16}$$

Theorem 4. (Decomposition of equilibrium under steering with finite K) For connected simple graph $\mathcal{G}$, let $d_i \equiv d \in \mathbb{N}^*$, $\alpha_i \equiv \alpha \in (0,1)$, $\beta_i \equiv \beta \in [0,1]$ $\forall i \in [n]$. Under (10), the equilibrium opinion, $\boldsymbol{x}_{\mathrm{PS}}$ can be decomposed into consensus and heterogeneous components as follows

$$\boldsymbol{x}_{\mathrm{PS}} = \sum_{i=1}^n \frac{\lambda_i}{1-\beta\lambda_i} \boldsymbol{v}_i \boldsymbol{v}_i^\top ((1-\beta)\boldsymbol{x}^*),$$

where $\{\lambda_i\}$ are eigenvalues of Ψ_K with $\{\boldsymbol{v}_i\}$ being the corresponding eigenvectors. W.l.o.g., let $\lambda_1 = 1, \boldsymbol{v}_1 = \frac{1}{\sqrt{n}}\mathbf{1}^\top$ and we have

$$\frac{\lambda_1}{1-\beta\lambda_1} = \frac{1}{1-\beta},$$
$$\left|\frac{\lambda_i}{1-\beta\lambda_i}\right| < \frac{1}{1-\beta}, \forall i \geq 2.$$

Proof. First of all, since W is symmetric, W can be decomposed as follows

$$W = U\mathrm{diag}(\mu_1, \cdots, \mu_n)U^\top,\ U \coloneqq [\boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_n],$$

where $U^\top U = UU^\top = I_n$ and $\{\mu_i\}_{i\in[n]}$ are eigenvalues of W and $\{\boldsymbol{v}_i\}_{i\in[n]}$ are corresponding eigenvectors. Since W is row-stochastic, apply Gershgorin circle theorem, we have $\rho(W) = 1$. Apply Perron-Frobenius theorem since W is also irreducible, we obtain $\mu_1 = 1$ and $\boldsymbol{v}_1 = \frac{1}{\sqrt{n}}\mathbf{1}^\top$ w.l.o.g. and $\mu_i \neq 1$ and $|\mu_i| \leq 1$ $\forall i \geq 2$. From (13), we can easily obtain

$$\Psi_K = U\mathrm{diag}(\lambda_1, \cdots, \lambda_n)U^\top,$$
$$\lambda_i \coloneqq \frac{1-\alpha}{1-\alpha\mu_i} + \left(1 - \frac{1-\alpha}{1-\alpha\mu_i}\right)\alpha^K \mu_i^K,$$
$$\boldsymbol{x}_{\mathrm{PS}} = U\mathrm{diag}\left(\frac{\lambda_1}{1-\beta\lambda_1}, \cdots, \frac{\lambda_n}{1-\beta\lambda_n}\right)U^\top.$$

We can easily verify that $\lambda_1 = 1$, $|\lambda_i| < 1$ $\forall i \in [n]$ by noticing $\mu_i \neq 1$, $|\mu_i| \leq 1$ $\forall i \geq 2$, and observe

$$\frac{1-\alpha}{1-\alpha\mu_i} < 1,\ |\lambda_i| < \max\{1, \alpha^K \mu_i^K\} = 1.$$

Hence we conclude the proof. □

Lemma 3. (Mean and variance of $\boldsymbol{x}_{\mathrm{PS}}$ with finite K) Under the same setting as Theorem 4, and consider the mean and the variance of $\boldsymbol{x}_{\mathrm{PS}}$ defined in (16), the following claims apply.

1. The invariant opinions $\boldsymbol{x}^*$ determine the mean
$$\bar{\boldsymbol{x}}_{\mathrm{PS}} = \frac{\mathbf{1}^\top \boldsymbol{x}^*}{n}.$$
2. The variance $\mathrm{Var}(\boldsymbol{x}_{\mathrm{PS}})$ increases as α increases if K is large enough. As for dependence over β, we have the variance decreases as β increases.

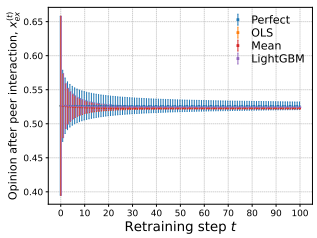
Proof. The structure of proof of Proposition 2 can be applied here. The only difference in the result is the dependence of variance over peer susceptibilities Λ_α. Notice λ_i can be negative. Hence by assuming K is large enough,
$$K \geq \max_i \left\{ \log \frac{|1-\alpha|}{\alpha(1-\mu_i)} / \log(\frac{1}{\alpha|\mu_i|}) \right\}$$
we ensure $\lambda_i > 0$ and the claims still apply. □

F Additive noise model

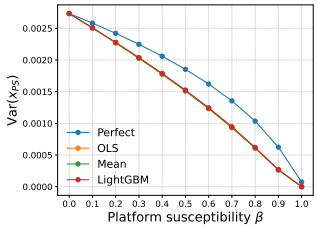
Consider an additive noise model where the platform observes perturbed opinions, and let $\beta_i \equiv \beta, \forall i \in [n]$, we have
$$\boldsymbol{x}_{ex}^{(t+1)} = \Psi_K((1-\beta)x^* + \beta \boldsymbol{f}^{(t)}), \quad f^{(t)} = \frac{1}{M}\sum_{j=1}^{M} \mathrm{clip}(\boldsymbol{x}_{ex}^{(t)} + \boldsymbol{\xi}_j^{(t)}, 0, 1).$$
Here M is the number of noisy samples the platform gathers for estimation and $\boldsymbol{\xi}_j \in \mathbb{R}^n, j = 1, \ldots, n$, are random noises. We drop the clipping step for brevity of the presentation and analyze:
$$x_{ex}^{(t+1)} = \Psi_K((1-\beta)x^* + \beta f^{(t)}), \quad f^{(t)} = \frac{1}{M}\sum_{j=1}^{M} (\boldsymbol{x}_{ex}^{(t)} + \boldsymbol{\xi}_j^{(t)}) \quad \text{where} \quad \boldsymbol{\xi}_j^{(t)} \sim N(0, \sigma^2 I_n).$$
The noise sensitivity can hence be formalized as
$$\text{sensitivity} = \frac{|\mathbb{E}[\mathrm{Var}(\boldsymbol{x}_{ex}^{(T)} \mid_{\sigma_1})] - \mathbb{E}[\mathrm{Var}(\boldsymbol{x}_{ex}^{(T)} \mid_{\sigma_2})]|}{|\sigma_1 - \sigma_2|}.$$
For the above noise model, we can directly observe that our core results (Theorem 1-3) hold in expectation. Assume the noises at each step t are i.i.d. Then the variance can be decomposed as follows:
$$\mathrm{Var}(\boldsymbol{x}_{ex}^{(T)} \mid_{\sigma>0}) = \mathrm{Var}(\boldsymbol{x}_{ex}^{(T)} \mid_{\sigma=0}) + T\sigma^2(\beta^2 - \beta^{2T+2})/((1-\beta^2)M).$$
Moreover, apply Hoeffding's inequality, we easily obtain that if $M \sim \mathcal{O}((\beta\sigma)^2 * \log(1/(\varepsilon\delta))/(\varepsilon^4(1-\beta^2)))$, then after $\mathcal{O}(1/\varepsilon)$ steps, we obtain an ε-approximation of exact equilibrium, $\boldsymbol{x}_{ex}^{(T)} \mid_{\sigma=0}$, with probability $1-\delta$. This guarantee carries over to the clipped dynamics, since clipping the predictions between 0 and 1 never increases the prediction errors.

G Additional simulations and details

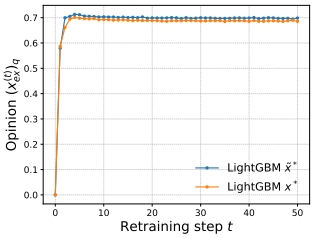
Settings If not stated otherwise, we use the following settings for simulations. We use $K = 100$ to approximate inner-loop convergence and $T = 50$ for the convergence in Algorithm 1. We use a heterogeneous setup where we sample $\{\beta_i\}_{i\in[n]}$ and $\{\alpha_i\}_{i\in[n]}$ from a normal distribution with mean 0.9 and standard deviation 0.1, and the resulting values are clipped within $[0.01, 0.99]$. If the predictions are not perfect, we assume opinions of 80% of the nodes are observed and use these observations as the training dataset to predict the opinions of the remaining 20% population, and we denote these two sets of individuals as O and U respectively.



(a) Opinions across retraining steps (Pokec)



(b) Opinions at equilibrium (Pokec)



(c) Evolving opinion of stubborn individual (Yelp)

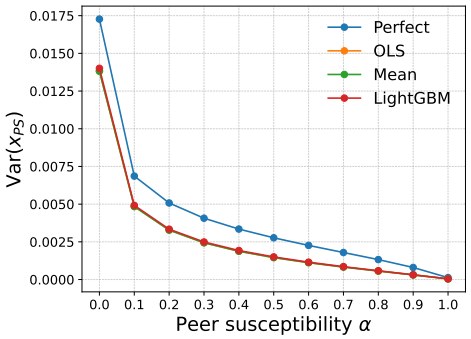
Figure 3: In (a), we show how performativity homogenizes opinions across retraining steps. The x-axis denotes the retraining step t, and the y-axis corresponds to individuals' expressed opinions in each time step, where the error bars indicate the variance. In (b), we show how platform susceptibility decreases variance of opinions at equilibrium. The x-axis denotes the varying homogeneous platform susceptibility while the y-axis denotes the variance of opinions in equilibrium. In (c), we show how a (peer)-stubborn individual $q \in U$ is influenced by their peers across retraining steps for two different scenarios.

In simulations using Pokec dataset, to mimic a machine learning setting we extract node features from the available profile information using a transformer model. For the Pokec network, we choose "relation_to_smoking" as the target of prediction. This field contains free form text from which we extract a sentiment score as a proxy for each individual's opinion towards smoking.

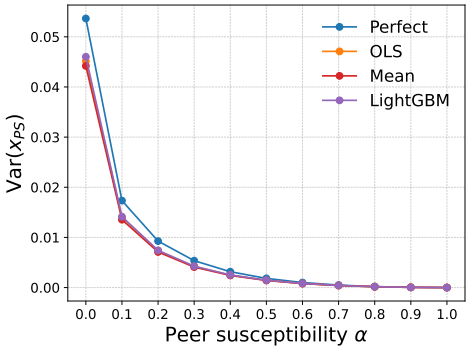
We also defer part of the simulation results to this section. In Figure 3 and Figure 6, the same procedure for generating plots with Yelp dataset is applied to Pokec dataset. Note that we use $T = 50$ for generating Figure 3b for computation efficiency. The finite approximation leads to the slight deviation of $\mathrm{Var}(\boldsymbol{x}_{\mathrm{PS}})$ from zero in the case when $\beta = 1$. Particularly, in Figure 3c, we show how platform interactions create indirect peer influence beyond mean estimation. In particular, we extrapolate predictions to unseen individuals using parametric models. In line with Theorem 3, we randomly sample an individual q in unobserved group U and set $\alpha_q = 0$, $x_q^* = 0$. We then randomly select 10% of individuals in observed group O and set their innate opinions to 1. We denote this new innate opinions as $\tilde{\boldsymbol{x}}^*$ and observe how the opinions $(\boldsymbol{x}_{\mathrm{ex}}^{(t)})_q$ converge to a value different from the equilibrium opinion under the original innate opinions $\boldsymbol{x}^*$.

In Figure 4, we use homogeneous peer susceptibilities, $\alpha_i \equiv \alpha$, and vary them while holding platform susceptibilities fixed. We observe the variance of opinion equilibrium decreases as peer susceptibility increases, which was proved for regular networks in Appendix C.4.

In Figure 5, under the mean estimation, the result illustrate Proposition 3 where the (peer)-stubborn individual $l \in U$ converges to a different equilibrium with different opinions in observed set O with different prediction methods/datasets from Figure 3c.

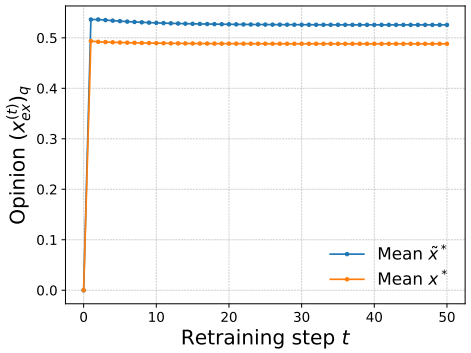


(a) Opinions at equilibrium (Pokec)

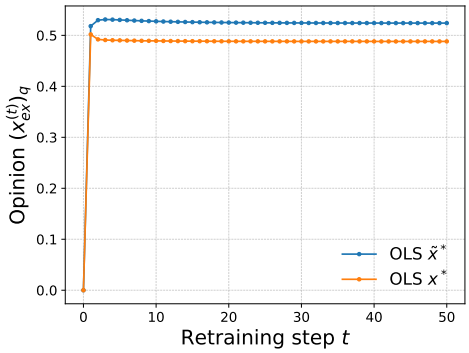


(b) Opinions at equilibrium (Yelp)

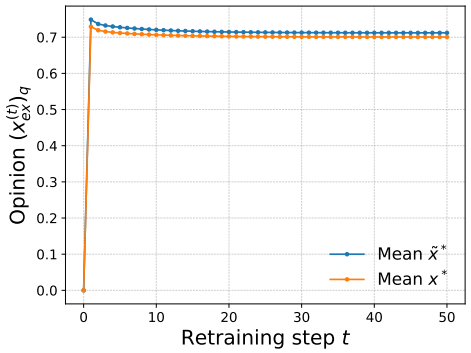
Figure 4: We show how peer susceptibility decreases variance at equilibrium. The x-axis denotes the varying homogeneous peer susceptibility while the y-axis denotes the variance of opinions in equilibrium. In (a), we use Pokec dataset and in (b), we use Yelp dataset.



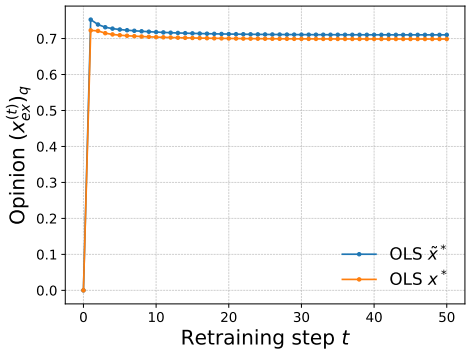
(a) Evolving opinion of stubborn individual (Pokec)



(b) Evolving opinion of stubborn individual (Pokec)

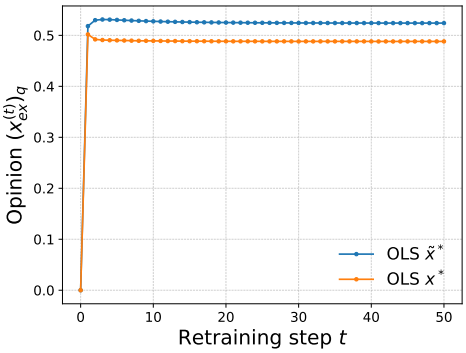


(c) Evolving opinion of stubborn individual (Yelp)

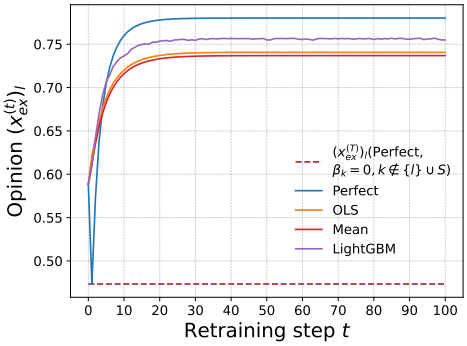


(d) Evolving opinion of stubborn individual (Yelp)

Figure 5: We show the individual $l \in U$ with $\alpha_l = 0$ converges to different equilibria under different innate opinions of individuals in set O. The x-axis denotes the retraining steps, and the y-axis denotes the opinions of individual l. In (a), the platform uses the mean estimation as in Proposition 3. We use Pokec dataset for (a) and (b), and Yelp for (c) and (d). In (b), the platform uses OLS for platform predictions.



(a) Evolving opinion of stubborn individual (Pokec)



(b) Indirect effect of platform intervention. (Pokec)

Figure 6: In (a), we show how a stubborn individual $q \in U$ with $\alpha_q = 0$ and $x_q^* = 0$ is influenced by their peers across retraining steps for two different scenarios. In (b), we show how indirect platform influence increases over retraining steps. We compare opinion trajectories of the stubborn individual l with $\beta_l = 0$. The purple dashed line denotes the opinion of the stubborn individual if we set $\beta_k = 0$, $k \notin \{l\} \cup S$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in abstract and introduction are accurate about the main contribution of this paper, with technical details omitted for the clear presentation of the result.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work in the section named "Discussion".

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We present the results with the full set of assumptions in the main text, and include the full proofs in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include all the details to reproduce our theoretical results and full codes to regenerate the plots in the paper with an anonymized repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the link to download the public datasets we used in the simulations. We also provide full access to our codes via an anonymized link.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We include important details in the main text for simulations and provide all details with full code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The error bars are computed with standard procedures and included in the plots.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the details about the commondity hardware used to run the simulations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The codes and the link are completely anonymized in our submission.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential impact of this work in discussion sections.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We use public datasets and our codes pose no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The public datasets are properly referred in the our work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code are well-documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work doesn't involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work doesn't involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM is only used for writing, editing and formatting purposes and is not used in the core method development in this paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.