Peer Filtering: Democratic Misinformation Control in Social Networks

KRISHNAMURTHY IYER, ANKUR MANI, CALVIN ROTH

Online social networking platforms suffer from the problem of misinformation spread that has serious societal implications in areas such as politics and public health, among others. Social networking platforms spend a lot of effort on curtailing misinformation. The techniques are often top down involving content moderation such as labeling, tagging, removing content and user centered such as nudging, debunking, and debunking. Research on misinformation in social networks has overlooked the ability of the network as a whole to moderate the content in a more democratic way. In this paper, we study the power of peers in filtering out false information (peer filtering) in social networks and how the platform can assist in peer filtering without direct content moderation and judging individuals and content. We present a tractable model of content spread in a social network of Bayesian users who derive utility from sharing a content depending upon its veracity and alignment with their opinions. User types differ along two axes, namely in their affinity to truth in misaligned content and in their aversion to misinformation in aligned content. After being exposed to a content, users receive a private signal that determines their posterior belief about the content's veracity. Based on their posterior beliefs, the users choose whether or not to further share the content. We study the resulting equilibrium in the network and find that depending upon the distribution of the types of users, and the fraction of true content introduced in the network, different types of equilibria emerge that differ in the spread and the virality of the content. We show that, under all types of equilibria, true content spreads more in the network than false content, but true content cannot go viral without false content also going viral. Using this model, we describe how key metrics of the network vary with the parameters of the model. These metrics include how much false content is successfully filtered out and how much true content remains as well as the volume and engagement in the network. We study the impact of population characteristics (proportions of types of users), average content veracity and platform control (number of peers who observe the content) on the power of peer filtering effect.

Contents

Manuscript submitted for review to the 25th ACM Conference on Economics & Computation (EC'24).

1 INTRODUCTION

Online social networking platforms suffer from the problem of misinformation spread that has serious societal implications in areas such as politics and public health, among others. In critical situations such as an election or during a public health crisis such as COVID-19, when the need for accurate information is essential, it has been observed that the level of misinformation in such platforms is very high. In particular, misinformation about masks, treatments, and vaccines led to adverse consequences in the US and the rest of the world [Caceres et al., 2022, Loomba et al., 2021, Rocha et al., 2021].

Social networking platforms spend a lot of effort on curtailing misinformation and have clear policies about removing harmful misinformation [Facebook, 2024, X.com, 2024]. Facebook spent over $13 billion between 2016 and 2021 on "safety and security", employs 40,000 employees to reduce misinformation [Robertson, 2021, Rosen, 2021] and its employees and contractors spent 3.2 million hours alone on searching, labelling and removing misinformation in 2020 [Justin Scheck and Horwitz, 2021]. There are four common techniques employed by platforms to counter misinformation, with varying levels of success and adverse side effects. These techniques include (i) *Detection and labeling* of content, sources and users [Facebook, 2020, Kennedy et al., 2022, Patwa et al., 2021, Pennycook and Rand, 2019, Zade et al., 2023], (ii) *Nudging* users to verify content before sharing [Fazio, 2020, Pennycook et al., 2021, 2020, Pennycook and Rand, 2021], (iii) *Debunking* content after it has been spreading in the network [Chan et al., 2017, Nyhan et al., 2014, Van Der Linden, 2022], and (iv) *Prebunking* or *inoculating* users before misinformation gets introduced in the network [Cook et al., 2017, Niederdeppe et al., 2015, Pfau and Burgoon, 1988, Van Der Linden, 2022, Van der Linden et al., 2017]. While some have shown small or moderate effectiveness [Brashier et al., 2021, Clayton et al., 2020, Mena, 2020, Walter and Murphy, 2018, Wood and Porter, 2019], in general they have been inadequate in combating ideological reasoning [Bhargava et al., 2023, Ecker et al., 2020, Lyons et al., 2020, Nyhan and Reifler, 2010, Paynter et al., 2019, Rich et al., 2020, Schwarz et al., 2016], only have short term benefits [Banas and Rains, 2010, Carey et al., 2022, Lewandowsky and van der Linden, 2021, Nyhan and Reifler, 2010, Swire-Thompson et al., 2020], and require readily available reliable information [Fazio, 2020, Pennycook et al., 2021, 2020, Pennycook and Rand, 2021].

Almost all of these methods employed by the social networking platforms are at the individual content or user level. The problem of misinformation still persists and the applications of these methods leads to other problems for the platforms. Social networking companies often face public anger for their failures. Facebook detects and removes less that 1% of the false information (mostly in English) and is often criticized for bias [Haugen, 2021]. Discouraged with the lack of success and high cost, the social networking platforms are questioning if they should be responsible for moderating content introduced by users on their platforms. For instance, Mark Zuckerberg [Rodriguez, 2020] recently said "I don't think that Facebook or Internet platforms in general should be arbiters of truth." Despite the need to moderate content on the social networking platforms, it is unclear how to do it while respecting the freedom of expression.

Research on misinformation in social networks has overlooked the ability of the network as a whole to moderate the content. The wisdom of the crowds, even if not perfect, has been the cornerstone of democratic society. Even in the absence of any explicit content moderation or censoring, if the wisdom of the crowds emerges then it has the potential to filter out most of the false information.

In this paper we investigate the power of peers in filtering out false information (peer filtering) in social networks and how the platform can assist in peer filtering without direct content moderation and judging individuals and content. For our study, we introduce a parsimonious model of users in

a social network, differing in their opinions, observing content of varying positions and veracity, forming beliefs about content veracity, and choosing to share or not given their alignment with the content and its veracity. We abstract away from the source of content that could be news sources, user generated content, or AI generated content. Thus the focus in our study is on the collective actions of users in the network and its impact on the outcomes. Even with all its simplicity, the model captures very rich dynamics of content spread in the social network and through the analysis of the network equilibrium arising from the model we are able to demonstrate the phenomena of *peer filtering*, that allows the network as a whole to filter out a large fraction of false content while spreading true content. We find that the peer filtering effect in equilibrium leads to higher odds of encountering a true content obtained from a peer than encountering a fresh content obtained outside the network. Yet, we find that the peer filtering effect cannot completely remove all false content in the network. In particular, in any equilibrium if a positive fraction of true content goes viral then a smaller, yet positive fraction of false content must also go viral.

We quantify the level of peer filtering in equilibrium using metrics including (i) *sensitivity*, (ii) specificity, and (iii) effectiveness, i.e.- the ratio between the fraction of true content in the network and the fraction of true content outside the network. We also quantify the user engagement level and content volume in the network in the equilibrium. Finally, through a numerical study of the equilibrium characteristics, we study the impact of the population composition, average veracity of the outside content, and the platform policies on the level of peer filtering effect, user engagement and volume of the content in the network. We find, surprisingly, that a population consisting of more polarized users that only prefer to share content aligned with them has a stronger peer filtering effect than a population consisting of unpolarized users who do not distinguish content based upon its alignment. This is an important finding because it decouples the problems of polarization and misinformation. Polarization is often blamed for the the spread of misinformation but our finding suggests otherwise.

Our paper provides actionable insights for the platforms and highlights a possibility of democratic content moderation through peer filtering. Platforms can adopt policies to amplify peer-filtering instead of moderating content themselves. There are several advantages of such policies. First, instead of focusing on individual content, these policies work on managing the social network as a whole and thus would be more effective and less expensive. Second, it removes any intent of bias the platforms may introduce because the network itself moderates the content. Third, the control depends upon the aggregate statistics of the misinformation levels in the network that are more accurate that the accuracy of individual content.

1.1 Related Literature

Recent literature has explored strategic models to study the spread of misinformation in networks [Acemoglu et al., 2021, Jackson et al., 2022, Mostagir and Siderius, 2023, Sikder et al., 2020]. The ones closest to our work include [Acemoglu et al., 2021, Jackson et al., 2022]. In particular, [Acemoglu et al., 2021] presents a model of agents homogeneous in prior beliefs and preferences to study the extent of misinformation in networks. We consider heterogeneity in agent preferences. In [Jackson et al., 2022] content is shared but it mutates in each exchange and users observing the history of spread seek to learn about the veracity of the mutated content. In our work, users' prior beliefs are a product of differing spread dynamics for true and false content and posterior beliefs are updated using user's private signals. Thus our model allows us to capture the information aggregation and peer filtering phenomenon absent in earlier work. Other recent work explores how groups learn in networks about the state of the world. [Huang et al., 2024]] studied a setting where users over repeatedly strategically share information with their peers about the state of the world. They studied how the speed of learning is effected by the network structure. [Dasaratha and He, 2019]

studied study a setting where agents express how likely it is that a particular piece of content is true based on a private signal and actions of their neighbors. They showed that there was a confounding effect where one agent can have an outsized impact on another by influencing multiple peers. Like our model these works consider Bayesian agents taking sequential actions in a social network but unlike our model they consider homogeneouos user populations and do not make use of heterogeneous agents with differing types and behaviors. [Mostagir and Siderius, 2022] studied a setting more similar to ours in that it considers heterogeneous agents for whom both the alignment and veracity matter. In their model, there exists both simple and sophisticated typed agents and they examine the contexts in which one type outperforms the other in learning. An important distinction between all of these works and ours is for the above works agents are learning from a local context, the actions of their neighbors, whereas agents in our model use global information about the average veracity of the content. Our paper is also the first to explore the phenomenon of peer filtering and the ability of the network as a whole to filter out more false content than true content thus leading to a higher average veracity of true content in the network as compared to content outside the network.

The outline of the paper is as follows. In Section 2, we present out model of different types of users interacting in a social network and contents varying in their position and veracity. In Section 3, we present the analysis of user belief formations and content sharing decisions. In Section 4 we study the spread dynamics of content given the analysis of user behavior and establish the existence and properties of equilibria under different model parameters. In Section 6 we present our numerical study quantifying the effects of user and content composition, and platform policies on the level of peer filtering, user engagement, and content volume. Finally, in Section 7 we present our conclusions and directions of future work.

2 MODEL

We now introduce a model of users interacting on an online social networking platform. In our model each user may "share" content (articles, videos, posts, etc), which is then included in the media feed of the user's peers by the platform. The model contains description of the content features, the user types, and the network spread dynamics, which we describe in detail next. Subsequent to this, we provide an instance of the model that we will study in this paper where we make specific assumptions on various model.

Content features: We consider a setting where new content arrives periodically to the network through a user chosen uniformly at random (see below for more details). We do not differentiate content based upon its origin. Content could be from news sources, user generate content, or AI generated content from friendly or adversarial sources. Each content's relevant type is characterized by its position on a socially or politically relevant axis, which we refer as its *inclination* I, and by its *veracity* α, namely the degree to which the content is true. For simplicity, we focus on the case where a content's inclination and veracity are independent of each other, with the inclination taking binary values. In particular, each content k's inclination I_k is either *left* or *right* with equal probability. Further, we assume that a content's veracity is drawn independently and identically from a distribution P.

A content's inclination is observable to the users, but its veracity is unobservable. Instead, each user infers the content' veracity from a private signal obtained after the consumption of the content. Specifically, we assume that after consuming a content k, a user i obtains an independent private signal s_{ik} whose distribution depends on the content's veracity, but not on its inclination. In addition to this private signal, the user's belief about a content's veracity is influenced by the aggregate spread dynamics of the content in the network (as described below in Section 4).

User types: The users on the platform derive utility by sharing content with their peers. Similar to the content, the users have an inherent inclination which is equally likely to be either left or right; we say a content is *aligned* with a user if they share the same inclination, and *misaligned* otherwise.

We normalize the users' utility for not sharing a content to zero. The utility users derive from sharing a content depends on its alignment and veracity. More precisely, we assume that all users obtain positive utility for sharing aligned content with high veracity and negative utility for sharing misaligned content with low veracity. On the other hand, as we describe next, users differ in the utility they derive from sharing aligned content with low veracity and misaligned content with high veracity.

Intuitively, we model users as differing along two different axes. First, users differ in their affinity to truth in misaligned content: *impartial* users obtain positive utility from sharing high veracity content even if it is misaligned with their inclination, whereas *partisan* users obtain negative utility from sharing high veracity but misaligned content. Second, users differ in their aversion to misinformation in aligned content: a *fabulist* derives positive utility from sharing aligned content even if it has low veracity, whereas a *truthteller* obtains negative utility from sharing low veracity aligned content. Thus, by combining these two axes, we obtain four types of users: a partisan fabulist (PF), impartial fabulist (IF), a partisan truthteller (PT) and an impartial truthteller (IT). We assume the user population consists of a proportion θ_t with type $t \in \Theta := \{\mathsf{PF}, \mathsf{IF}, \mathsf{PT}, \mathsf{IT}\}$. We let $\theta = (\theta_t)_t$ denote the vector of type proportions.

Network spread model: We consider a large population N of $|N| = n$ users on the platform connected in a sparse social network. We abstract away from the underlying degree distribution, given any admissible degree distribution, and assume that the network is a generated instance from the configuration model [Bender and Canfield, 1978, Chung and Lu, 2002, Chung et al., 2004, Molloy and Reed, 1995, Newman et al., 2001].

In our discrete time model, at each time $t > 0$, one new content from outside the network is arbitrarily introduced to a random user and the user chooses to ignore or share the content based upon her prior belief and the private signal about the veracity of the content. Once a user shares a content, it is received by κ of her neighbors in the network, where κ chosen by the platform is a constant. We assume that the user is not able to differentiate between a novel content and the old content in the network and has the same prior belief for the content's veracity. Therefore, the user's posterior belief after observing her private signal from consuming a content when first encountering it and her subsequent decision about sharing the content are independent of the novelty of the content in the network.

2.1 Binary Veracity Model

Based on the preceding description, we now describe an instance of the model, in particular the assumptions on the content's veracity, the distribution of the private signals, and the users' payoffs. Under the binary veracity model, each content's veracity takes binary values, thus representing the extreme scenario where each content is either true or false. Specifically, each content k's veracity α_k is *true* (T) independently with probability $p > 0$ and *false* (F) otherwise.

Upon consuming the content, a user i obtains an independent signal $s_{ik} \in [0, 1]$, whose distribution depends on whether the content is true or false. Formally, let f_α denote the probability density function of s_{ik} for a content with veracity $\alpha \in \{\mathsf{T}, \mathsf{F}\}$. For simplicity, we consider the following distributional choice:

$$f_{\mathsf{T}}(s) = 2s, \qquad f_{\mathsf{F}}(s) = 2(1-s), \qquad \text{for } s \in [0, 1]. \tag{1}$$

This particular choice simplifies our analysis, and has the property that a user who initially (before consuming the content) believes the content is equally likely to be true or false, after obtaining a private signal $s \in [0, 1]$ after consuming the content, believes that content's veracity is true with probability s.

Finally, the payoff-matrices of the four user types under the binary veracity model are as shown in Figure 1, where we normalize the positive/negative payoffs to ±1 for simplicity.

	T	F
A	1	−1
M	1	−1

(a) impartial truthteller (IT)

	T	F
A	1	−1
M	−1	−1

(b) partisan truthteller (PT)

	T	F
A	1	1
M	1	−1

(c) impartial fabulist (IF)

	T	F
A	1	1
M	−1	−1

(d) partisan fabulist (PF)

Fig. 1. Utility from sharing content for different user types: (a) impartial truthteller, (b) partisan truthteller, (c) impartial fabulist, and (iv) partisan fabulist. Columns T/F in each table stand for true/false content and rows A/M stand for aligned/misaligned content.

3 SINGLE USER BEHAVIOR

Since a content's veracity is unobservable to the users, each user in the network maintains a prior belief regarding the content's veracity. The user then updates this belief after observing the private signal from consuming the content, and arrives at a posterior belief. Based on this posterior belief (and her type), the user then decides whether or not to share the content. To gain insight into this decision, we first seek to understand the setting where the user's prior belief is exogenously fixed; our eventual goal is to apply this understanding to study the multi-user setting where the users' prior beliefs are determined endogenously by the equilibrium dynamics.

Toward that end, let $\delta \in [0, 1]$ denote the prior belief of a user i that any content k she receives is true: $\mathbb{P}(\alpha_k = \mathsf{T}) = \delta$, where $\mathbb{P}$ denotes the belief conditional on receiving a content. After consuming the content, the user receives a private signal s_i according to (1), due to which her belief updates following Bayes' rule, to

$$\mathbb{P}(\alpha_k = \mathsf{T} | s_i = x) = \frac{\delta x}{\delta x + (1-\delta)(1-x)} := \hat{\delta}(x).$$

Using this posterior belief allows us to write the user's expected utility for sharing the content, depending on her type and the content's alignment. For instance, the expected utility of an impartial truthteller (IT) for sharing a content, regardless of its alignment, is given by $\hat{\delta}(x)(1) + (1 - \hat{\delta}(x))(-1) = 2\hat{\delta}(x) - 1$. Figure 2 provides each type's expected utility for sharing the content.

The user shares the content if and only if her expected utility for sharing is non-negative. By noticing that $2\hat{\delta}(x) - 1 > 0$ if and only if $x > 1 - \delta$, we obtain the following characterization of the user behavior in the binary veracity model as a function of her type, her prior belief δ, and her private signal s_i.

	IT	PT	IF	PF
A	$2\hat{\delta}(x) - 1$	$2\hat{\delta}(x) - 1$	1	1
M	$2\hat{\delta}(x) - 1$	-1	$2\hat{\delta}(x) - 1$	-1

Fig. 2. Expected utility of each user type for sharing an aligned (A) or a misaligned (M) content, as a function of her posterior belief $\hat{\delta}(x)$.

Theorem 3.1. *In the binary veracity model, suppose a user's prior belief that the content is true is δ. Then,*

(1) *If the user is an impartial truthteller* (IT)*, then she shares the content if and only if the private signal s_i satisfies $s_i > 1-\delta$, regardless of its alignment. Similarly, if the user is a partisan fabulist* (PF)*, then she shares a content if and only if it is aligned, irrespective of the signal.*
(2) *If the user is an impartial fabulist* (IF)*, then she always shares aligned content, but shares misaligned content if and only if $s_i > 1 - \delta$.*
(3) *If the user is a partisan truthteller* (PT)*, then she never shares misaligned content, but shares aligned content if and only if $s_i > 1 - \delta$.*

A proof may be found in the appendix.

Using the preceding characterization of the users' decisions, along with the signal distribution (1), we can now express the (ex ante) probability $\beta^{\alpha}_{a,t}$ that a user of type $t \in \Theta$ shares a content with alignment $a \in \{\mathsf{A}, \mathsf{M}\}$ and veracity $\alpha \in \{\mathsf{T}, \mathsf{F}\}$. We have

$$\beta^{\alpha}_{a,t} = \begin{cases} 1 & \text{if } a = \mathsf{A} \text{ and } t \in \{\mathsf{IF}, \mathsf{PF}\}; \\ 0 & \text{if } a = \mathsf{M} \text{ and } t \in \{\mathsf{PT}, \mathsf{PF}\}; \\ \overline{\mathrm{F}}_{\alpha}(1-\delta) & \text{otherwise,} \end{cases} \tag{2}$$

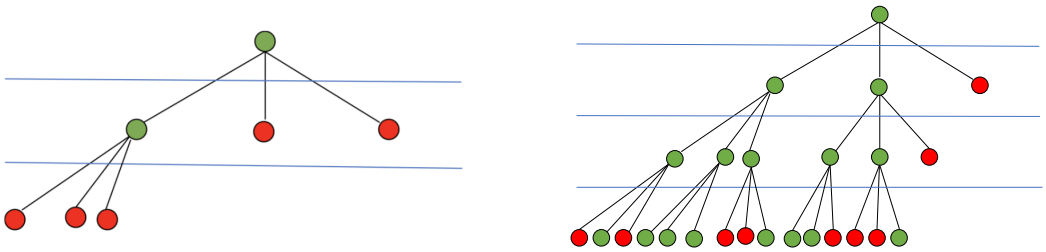
where $\overline{\mathrm{F}}_{\alpha}$ denotes the complimentary cdf of the signal distribution for content with veracity α and hence $\overline{\mathrm{F}}_{\alpha}(1-\delta)$ is the probability that the signal is above $1-\delta$. Furthermore, given the proportions of the types θ in the network, we obtain that a randomly chosen user shares a content with veracity $\alpha \in \{\mathsf{T}, \mathsf{F}\}$ with probability $\beta^{\alpha}(\delta)$ given by

$$\beta^{\alpha}(\delta) \coloneqq \sum_{a \in \{\mathsf{A},\mathsf{M}\}} \frac{1}{2} \cdot \sum_{t \in \Theta} \theta_t \beta^{\alpha}_{a,t} = \frac{1}{2}\left(1 - \theta_{\mathsf{T}} + (\theta_{\mathsf{T}} + \theta_{\mathsf{I}})\overline{\mathrm{F}}_{\alpha}(1-\delta)\right), \tag{3}$$

where $\theta_{\mathsf{T}} = \theta_{\mathsf{IT}} + \theta_{\mathsf{PT}}$ and $\theta_{\mathsf{I}} = \theta_{\mathsf{IT}} + \theta_{\mathsf{IF}}$ denote the marginal proportions of truthtellers and impartial users respectively. Here, we have used the fact that the content is equally likely to be aligned or misaligned with the user. Note that from (2), we have $\beta^{\mathsf{T}}_{a,t} \geq \beta^{\mathsf{F}}_{a,t}$ for all $a \in \{\mathsf{A}, \mathsf{M}\}$ and $t \in \Theta$. Hence, we conclude that $\beta^{\mathsf{T}}(\delta) \geq \beta^{\mathsf{F}}(\delta)$, i.e., the ex ante probability that a random user shares a true content is higher than the probability she shares a false content.

4 SPREAD DYNAMICS AND EQUILIBRIUM

Having described a single user's behavior, we now proceed to investigate the equilibrium in the network with many users. To describe the equilibrium, we first investigate the content spread dynamics in a large network and characterize the proportion of true and false content in the network when all users share content with fixed sharing probabilities. We then impose a consistency requirement in equilibrium, namely that each users' prior belief about the truth of a content they observe in the network must exactly match the proportion of true content in the network. This consistency requirement imposes a restriction on the users' prior belief δ in the preceding section.



(a) (b)

Fig. 3. Spread of content in a large network, green nodes decide to share the content; red nodes decide not to share the content; (a) false content is less likely to be shared and the spread may go bust (b) true content is slightly more likely to be shared and may end up going viral.

4.1 Spread Dynamics

To analyze the spread dynamics in the network, suppose users in the network share a content of veracity α with fixed probability β^α. Given the preceding analysis of a single user's behavior, we assume that $\beta^{\mathsf{T}} \geq \beta^{\mathsf{F}}$. Next, suppose a content with veracity α is introduced at time 0 in the network. The number of users $X^\alpha(\tau)$ that see the content for the first time at time τ is random and depends on how many users choose to share the content at time $\tau - 1$. We have $X^\alpha(0) = 1$, and given the assumptions on the spread dynamics, we have $X^\alpha(1) = \kappa$ if the user chooses to share the content, otherwise $X^\alpha(1) = 0$. In general, we have

$$X^\alpha(\tau + 1) = \sum_{i=1}^{X^\alpha(\tau)} \kappa \cdot Z_i \tag{4}$$

where Z_i is the indicator variable which takes value 1 if the i^{th} user among the $X^\alpha(\tau)$ users exposed to the content at time τ chooses to share the content. The analysis of the single user behavior in the preceding section shows that $Z_i = 1$ with probability β^α for a randomly chosen user. Thus, we conclude that the conditional distribution of $X^\alpha(\tau + 1)$ given $X^\alpha(\tau)$ is κ times a Binomial random variable with parameters $X^\alpha(\tau)$ and β^α. Thus, $X^\alpha(\tau)$ follows a Galton-Watson branching process [Ney and Ney, 2004]. Furthermore, the expected number of users who observe the content for the first time at time τ is given by $\mathbb{E}[X^\alpha(\tau)] = (\kappa\beta^\alpha)^\tau$. As we describe below, the dynamics of this process is captured by $\kappa\beta^\alpha$, called its *branching rate*.

If the branching rate $\kappa\beta^\alpha$ is strictly less than 1, the process eventually equals zero, meaning that the users eventually stop sharing the content. In this case, the total expected number of users who see the content is finite, and is given by

$$\sum_{\tau=0}^{\infty} (\kappa\beta^\alpha)^\tau = \frac{1}{1 - \kappa\beta^\alpha}. \tag{5}$$

On the other hand, if the branching rate is strictly greater than 1, then with positive probability the process never equals zero, implying that the content is shared throughout the network, i.e., the content goes "viral". The probability q^α with which the content goes viral is obtained as the unique positive solution to the following equation:

$$1 - q^\alpha = 1 - \beta^\alpha + \beta^\alpha (1 - q^\alpha)^\kappa. \tag{6}$$

To interpret the equation, the left-hand side denotes the probability that the content does not go viral. This can only happen if (1) either the first user decides not the share the content (which happens with probability $1 - \beta^\alpha$) or (2) the first user shares the content (an event with probability

β^α) but none of the κ subprocesses originating from the κ peers to whom the content is shared go viral (which occurs with probability $(1-q^\alpha)^\kappa$).

Given the preceding analysis of the spread dynamics, we can now write the proportion of true content in the network depending on branching rate of the spread of true vs false content. Recall from the analysis of the single user's behavior that the sharing probability of the true content is larger than that of the false content, i.e., $\beta^{\mathsf{T}} \geq \beta^{\mathsf{F}}$. Thus, we obtain the following cases:

Case 1: $\kappa\beta^{\mathsf{F}} \leq \kappa\beta^{\mathsf{T}} < 1$. Under this condition, neither the true content nor the false content ever go viral, i.e., both types of content reach only a finite number of users. A content with veracity α introduced to a random user in the network reaches $\frac{1}{1-\kappa\beta^\alpha}$ users in total as given by (5). Since the proportion of newly arriving content that is true is given by p, the spread dynamics imply that ratio of the spread volume of true vs false content in the network is given by $\frac{p(1-\kappa\beta^{\mathsf{F}})}{(1-p)(1-\kappa\beta^{\mathsf{T}})}$.

Given the homogeneity assumptions on the network and the spread dynamics, when the volume of content arriving in the network is large, the ratio of the volume of true vs false content seen by a user is same across all users, and equal to the ratio $\frac{p(1-\kappa\beta^{\mathsf{F}})}{(1-p)(1-\kappa\beta^{\mathsf{T}})}$ of true vs false content in the network. This implies that, of all the content seen by a user, the proportion that is true is given by

$$\frac{p\left(\frac{1}{1-\kappa\beta^{\mathsf{T}}}\right)}{p\left(\frac{1}{1-\kappa\beta^{\mathsf{T}}}\right)+(1-p)\left(\frac{1}{1-\kappa\beta^{\mathsf{F}}}\right)}. \tag{7}$$

Case 2: $\kappa\beta^{\mathsf{F}} < 1 < \kappa\beta^{\mathsf{T}}$. Under this condition, the false content never goes viral, but the true content has a positive probability of going viral. In particular, a false content is seen by a finite number of users, whereas a true content has a positive probability of being seen by the entire network. In a large but finite network, it follows that only a negligible fraction of the content seen by a fixed user is false. In the limit as the network size grows, we conclude that the proportion of the content seen by a user that is true approaches 1.

Case 3: $1 < \kappa\beta^{\mathsf{F}} \leq \kappa\beta^{\mathsf{T}}$. Under this condition, both the true and the false content have a positive probability of going viral, and thus seen by the entire network. As above, as the network size grows, most of the content seen by a user has gone viral. Since any new content is true with probability p, and a content with veracity α goes viral with probability q^α, we conclude that of the content seen by a user, the proportion that is true is given by $\frac{pq^{\mathsf{T}}}{pq^{\mathsf{T}}+(1-p)q^{\mathsf{F}}}$.

Putting the three cases together, we obtain that, of all the content seen by a user in a large network, the proportion $\Phi(\beta^{\mathsf{T}},\beta^{\mathsf{F}})$ that is true can be written as

$$\Phi(\beta^{\mathsf{T}},\beta^{\mathsf{F}}) := \begin{cases} \dfrac{\frac{p}{1-\kappa\beta^{\mathsf{T}}}}{\frac{p}{1-\kappa\beta^{\mathsf{T}}}+\frac{1-p}{1-\kappa\beta^{\mathsf{F}}}}, & \text{if } \kappa\beta^{\mathsf{F}} \leq \kappa\beta^{\mathsf{T}} < 1; \\ 1, & \text{if } \kappa\beta^{\mathsf{F}} < 1 \leq \kappa\beta^{\mathsf{T}}; \\ \dfrac{pq^{\mathsf{T}}}{pq^{\mathsf{T}}+(1-p)q^{\mathsf{F}}}, & \text{if } 1 \leq \kappa\beta^{\mathsf{F}} \leq \kappa\beta^{\mathsf{T}}. \end{cases} \tag{8}$$

4.2 Network Equilibrium

Having described the spread dynamics, we are now ready to define our notion of an equilibrium. Specifically, in an equilibrium, we require consistency of the users' prior beliefs with the spread dynamics. In particular, we require that the proportion $\Phi(\beta^{\mathsf{T}},\beta^{\mathsf{F}})$ of true content in the network must be exactly equal to the users' prior belief $\delta = \mathbb{P}(\alpha_k = \mathsf{T})$ that an observed content is true. This yields the following definition of an equilibrium, which is further illustrated in Figure 4.

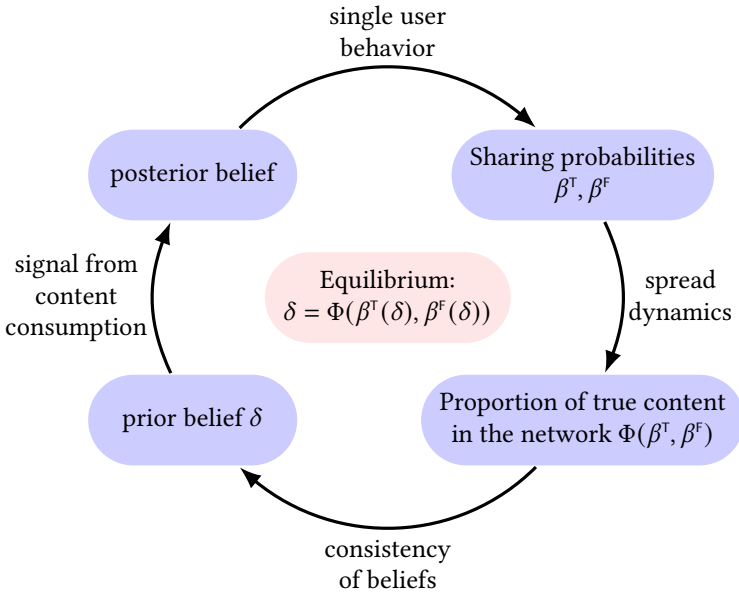


Fig. 4. Equilibrium in the network. Here β^T, β^F are obtained from δ through (3), and the proportion of true content $\Phi(\beta^T, \beta^F)$ is given by (8). An equilibrium corresponds to a value of δ for which the belief is consistent with spread dynamics.

Definition 1. *An equilibrium is given by the sharing probabilities* $(\beta^\alpha : \alpha \in \{T, F\})$ *and the prior belief* $\delta \in (0, 1)$*, such that (i) given the prior belief* δ*, the sharing probabilities* β^α *are obtained from* (2) *and* (3)*; and (ii) given the sharing probabilities* β^α*, the prior belief* δ *satisfies* $\delta = \Phi(\beta^T, \beta^F)$*, where* $\Phi(\beta^T, \beta^F)$ *is given by* (8)*. Taken together, in an equilibrium, the prior belief* δ *satisfies the fixed-point equation* $\delta = \Phi(\beta^T(\delta), \beta^F(\delta))$*.*

We observe that any equilibrium must be one of three types, depending on which of the three cases holds in the definition (8) of the function Φ. In the first type, we have $\kappa\beta^F \leq \kappa\beta^T < 1$ implying that neither content types ever go viral; we call this type a *bust equilibrium*. In the second type, we have $\kappa\beta^F < 1 \leq \kappa\beta^T$, implying that a true content goes viral with positive probability whereas a false content never goes viral; we call this type a *boom-bust* equilibrium. Finally, in the third type, we have $1 \leq \kappa\beta^F \leq \kappa\beta^T$ implying that both content types have a positive probability of going viral; we call this type a *boom* equilibrium. We have obtained the following result that states that there cannot be a boom-bust equilibrium implying that there is no equilibrium in which truth goes viral but no false information does.

Theorem 4.1. *In the binary veracity model, every equilibrium of the network is either a boom equilibrium or a bust equilibrium. In particular, the network does not have a boom-bust equilibrium.*

Proof. Suppose a boom-bust equilibrium exists, and let β^F and β^T denote the sharing probabilities in such an equilibrium. By definition, we have $\kappa\beta^F < 1$ and $\kappa\beta^T > 1$, implying that $\Phi(\beta^F, \beta^T) = 1$ and hence $\delta = \Pi(\delta) = \Phi(\beta^F, \beta^T) = 1$. However, using (3), we then find that $\beta^F = \beta^T = 1$, contradicting the implication that $\kappa\beta^T > 1 > \kappa\beta^F$. □

The preceding theorem states that in any equilibrium, either both true and false content go viral with some positive probability, or neither of them go viral. The main factor driving this result is the

interplay between a user's prior belief about veracity of content in the network, and her sharing decision. In particular, if only true content went viral in the network, then each users, upon coming across a content, believes that it must be true with high probability. Thus, except for really low values of private signals, the user share the content without much consideration for its veracity; this in turn would imply that false content would also go viral with positive probability.

Before showing the existence of equilibrium, we present the following result which establishes the existence of threshold values of κ, which depend solely on the proportions of user types, that categorize the nature of the equilibrium.

Theorem 4.2. *For $\kappa < \frac{2}{1+\theta_{\mathsf{I}}}$, there exists no boom equilibrium. Similarly, for $\kappa > \frac{2}{1-\theta_{\mathsf{T}}}$, there exists no bust equilibrium.*

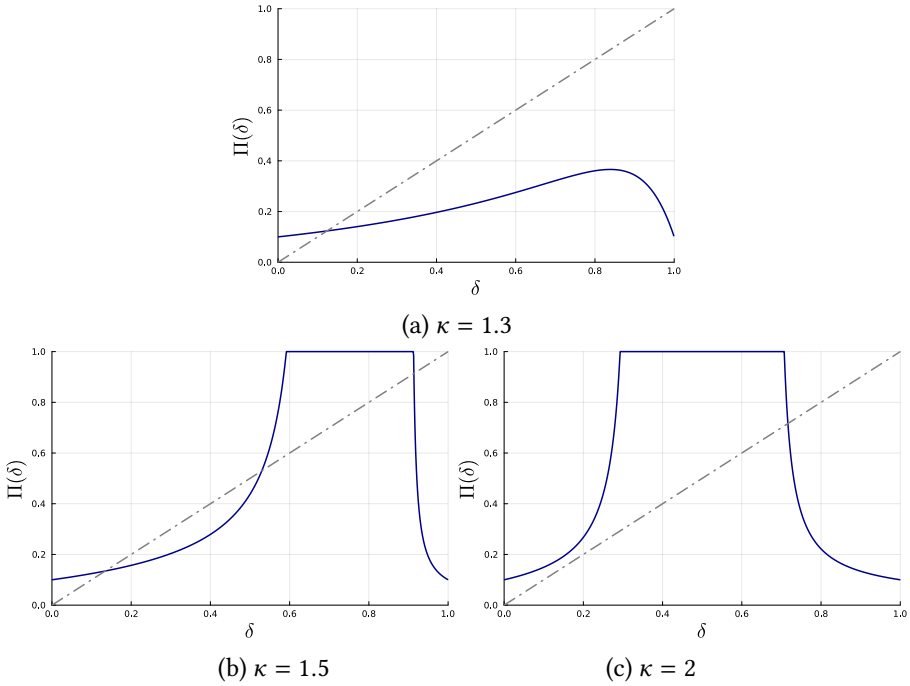
Proof. For a boom equilibrium, it must be the case that $\kappa\beta^{\mathsf{T}} > 1$. From the expression for β^{T}, we obtain that for any $\delta \in [0,1]$, $\beta^{\mathsf{T}}(\delta) \leq \beta^{\mathsf{T}}(1) \leq \frac{1+\theta_{\mathsf{I}}}{2}$. Thus, if $\kappa < \frac{2}{1+\theta_{\mathsf{I}}}$, we obtain $\kappa\beta^{\mathsf{T}} < 1$, implying that there cannot be a boom equilibrium. Similarly, we have $\beta^{\mathsf{F}}(\delta) \geq \beta^{\mathsf{F}}(0) = \frac{1-\theta_{\mathsf{T}}}{2}$. Thus, if $\kappa > \frac{2}{1-\theta_{\mathsf{T}}}$, we have $\kappa\beta^{\mathsf{F}} > 1$, implying that there cannot be a bust equilibrium. □

For example, consider $\theta_t = \frac{1}{4}$ for all for types $t \in \{\mathsf{PF}, \mathsf{IF}, \mathsf{PT}, \mathsf{IT}\}$ of users. For this population, we obatain from Theorem 4.2 that there exists no boom equilibrium if $\kappa < \frac{4}{3}$ and there exists no bust equilibrium if $\kappa > 4$. While Theorem 4.2 states that there exists no boom equilibrium if $\kappa < \frac{2}{1+\theta_{\mathsf{I}}}$ and there exists no bust equilibrium if $\kappa > \frac{2}{1-\theta_{\mathsf{T}}}$, it leaves open the question of equilibrium existence. We answer this question through our next result, which not only establishes the equilibrium existence for all values of κ, but also provides insights into the characteristics of the equilibria in the interval $(\frac{2}{1+\theta_{\mathsf{I}}}, \frac{2}{1-\theta_{\mathsf{T}}})$.

The primary approach for establishing existence of the equilibrium is using fixed-point theorems to show that the equation $\delta = \Phi(\beta^{\mathsf{T}}(\delta), \beta^{\mathsf{F}}(\delta))$ has a solution, where $\beta^{\alpha}(\delta)$ is given by (3) and the function $\Phi(\beta^{\mathsf{T}}, \beta^{\mathsf{F}})$ is defined in (8). To use such an approach, we establish the continuity of the underlying functions; the case-by-case definition of Φ in (8) makes this task technically challenging. Though technical, this step is crucial from an analytical perspective for studying the endogenous impact of model parameters on various network outcomes. We first give an intuition of this analysis through a figure. In Figure 5, we plot the function $\Pi(\delta) := \Phi(\beta^{\mathsf{T}}(\delta), \beta^{\mathsf{F}}(\delta))$ for different values of κ when $p = 0.1$ and $\theta_t = 1/4$ for all $t \in \Theta$. Note that the points where this function intersects the 45° line are the fixed-points of Π, and thus correspond to an equilibrium. We observe that for small κ (Fig. (4a)), the function Π intersects the 45° line once, implying a unique equilibrium. By checking which case holds in (8), we find that this is a bust equilibrium. For moderate values of κ, (Fig. (4b)), the function Π intersects the 45° line three times, implying a multiplicity of equilibria. Here, we find that among the three, the equilibrium with the largest value of δ is a boom equilibrium, whereas the other two are bust equilibria. Upon further increasing the value of κ (Fig. (4c)), we observe that the function Π intersects the 45° line once, implying once again a unique equilibrium. This equilibrium is found to be a boom equilibrium. We now state the main results about the existence of different types of equilibrium under different conditions.

Theorem 4.3. *There exists a boom equilibrium if and only if $\kappa > \frac{2}{1+\theta_{\mathsf{I}}}$. On the other hand, there exists a bust equilibrium if any of the following conditions are met:*

- *$\kappa < \frac{2}{1+\theta_{\mathsf{I}}}$ (a unique bust equilibrium exists);*
- *$\kappa = \frac{2}{1+\theta_{\mathsf{I}}}$ and $\gamma < \frac{1}{2}$ (a unique bust equilibrium exists);*
- *If $\kappa \in \left(\frac{2}{1+\theta_{\mathsf{I}}}, \kappa(\gamma)\right)$ and $\gamma < \frac{1}{2}$ (two bust equilibria exist);*
- *If $\kappa = \kappa(\gamma)$ and $\gamma < \frac{1}{2}$ (a unique bust equilibrium exists),*



(a) $\kappa = 1.3$

(b) $\kappa = 1.5$ (c) $\kappa = 2$

Fig. 5. The function $\Pi(\delta) \coloneqq \Phi(\beta^{\mathsf{T}}(\delta), \beta^{\mathsf{F}}(\delta))$ for $p = 0.1$, $\theta_t = 1/4$ for $t \in \Theta$ and for different values of κ. Here, the function Φ is defined as in (8), and the sharing probabilities $\beta^{\alpha}(\delta)$ for $\alpha \in \{\mathsf{T}, \mathsf{F}\}$ are given by (3).

where $\gamma \coloneqq \frac{p}{1-p}$. *Otherwise, there exists no bust equilibrium. Here,*

$$\kappa(\gamma) \coloneqq \frac{2(\gamma-1)(\gamma+1)^3}{\gamma^4(\theta_{\mathsf{I}}+1) + \gamma^3(-6\theta_{\mathsf{I}} - 8\theta_{\mathsf{T}} + 2) + 12\gamma^2(\theta_{\mathsf{I}} + \theta_{\mathsf{T}}) - 2\gamma(4\theta_{\mathsf{I}} + 3\theta_{\mathsf{T}} + 1) + \theta_{\mathsf{T}} - 1}.$$

Proof Idea. The main difficulty in the proof of the existence of fixed points on $\Pi(\delta)$ comes from the fact that the expression for $\Phi(\beta^{\mathsf{F}}, \beta^{\mathsf{T}})$ in (8) depends on the spread dynamics. In particular, given the case-by-case definition, it is not immediately clear if the function Π is continuous over the interval $[0, 1]$. To overcome this difficulty, we identify a sub-interval (for each value of κ) over which Π is continuous, and on whose endpoints the function $\Pi(\delta) - \delta$ takes positive and negative values, so that an application of intermediate value theorem suffices to show that a fixed point exists. We achieve this by performing a careful perturbative analysis to identify the effect of perturbing δ close to specific values where the definition of the function Φ changes from one case to another. We provide the full details in the Appendix. □

The previous two theorems are summarized in Figure 6.

5 MODEL REDUCTION

In our model, populations is characterized by the proportions of the four types of users in the network. These four variables describing the composition of the population has three degrees of freedom. They don't have four because they are coupled by the constraint they sum to one. It is not the case however that every population most meaningfully differs from all others. Instead, some populations share the same boom and bust equilibria. This characterization is made formal with the following definition of equivalence between two populations.

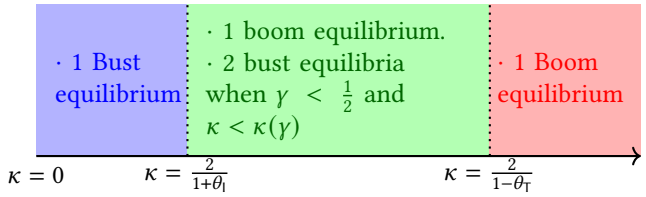


Fig. 6. Visual representation of the characterization of boom and bust equilibria as κ increases. In this diagram, $\gamma := \frac{p}{1-p}$ and $\kappa(\gamma)$ is a rational function of γ, θ_I and θ_T defined in Theorem 4.3.

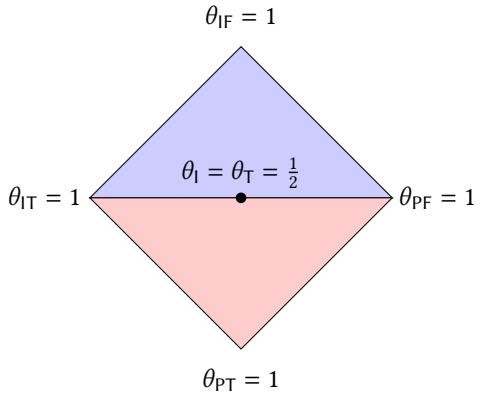


Fig. 7. Parameter space according to Corollary 5.2. The upper blue triangle is the set of populations for which $\theta_{PT} = 0$ and the lower triangle is the set for which $\theta_{IF} = 0$.

Definition 2. *Two user populations with type distributions θ and $\theta' \in \Theta$ respectively are equivalent if for any $p \in (0,1)$ and $\kappa > 0$, the two populations have the same set of equilibria.*

We note that this definition of equivalence does induce equivalence classes on populations. The following result provides sufficient conditions for two populations to be equivalent.

Theorem 5.1. *Let $\theta, \theta' \in \Theta$ be two type distributions satisfying $\theta'_{IF} + \theta'_{IT} = \theta_{IF} + \theta_{IT}$ and $\theta'_{IT} + \theta'_{PT} = \theta_{IT} + \theta_{PT}$. Then two user populations with type distributions θ and θ' respectively are equivalent.*

The preceding result states that from the perspective of the equilibrium analysis, only the *marginal* distribution of the user population's truth-preference and partisanship matters. In particular, two populations with the same marginal distribution of truthtellers and impartial users have the same set of equilibria. This allows us to focus on a subset of user populations where either there are no impartial fabulists or no partisan truthtellers, as highlighted in the following corollary, whose proof is in the Appendix 8.

Corollary 5.2. *Any user population is equivalent to an alternate one for which either $\theta_{IF} = 0$ or $\theta_{PT} = 0$.*

Given this corollary, we can provide a geometric view of the (reduced) parameter space, as shown in Figure 7. Each point in the upper blue triangle represents a user population with type distribution $\theta \in \Theta$ satisfying $\theta_{PT} = 0$, with the three vertices representing extreme user populations with only impartial fabulist, only impartial truthtellers or only partisan fabulists respectively. (For any interior point of the triangle, the type distribution is given by the corresponding barycentric coordinates.) Similarly, each point in the lower red triangle represents a user population with type distribution

$\theta \in \Theta$ satisfying $\theta_{\mathsf{IF}} = 0$, with the lowermost vertex representing a user population consisting of only partisan truthtellers. The points that lie on the horizontal diagonal then represent user populations with both $\theta_{\mathsf{IF}} = \theta_{\mathsf{PT}} = 0$. The implication of Corollary 5.2 is that *any* user population is equivalent to some point in the Figure 7. In particular, the user population where all types are equally distributed is represented by the center of the square.

We leverage this reduction in model parameters for our numerical analysis in the next section.

6 PEER FILTERING

Now that we have developed the techniques to characterize the equilibria, we focus on the outcomes of interest in equilibria and control techniques to help the social network choose the socially desirable equilibrium. We focus in particular on the power of peers in the social network to include their own information and judgement independently and prevent the spread of false information. Peer filtering is a self-correcting phenomenon that moderates the diffusion on misinformation in the social network through an intricate relationship between individual behavior of users and the diffusion dynamics of the social network. We find evidence of this phenomenon in our results below. Two effects determine this phenomenon. First, when the level of misinformation diffusion is high, then users are more skeptical about the veracity of the information in the network and less engaged with sharing information. This reduces the diffusion of all information but more for false information thus reducing the fraction of misinformation in the network that goes viral. Second, when the level of both true and false information is comparable in the social network, then the users weigh their private signal more when deciding about the veracity of information. This leads to more independent decision making by the users. Subsequently the wisdom of the crowds emerges and true content goes more viral than false content.

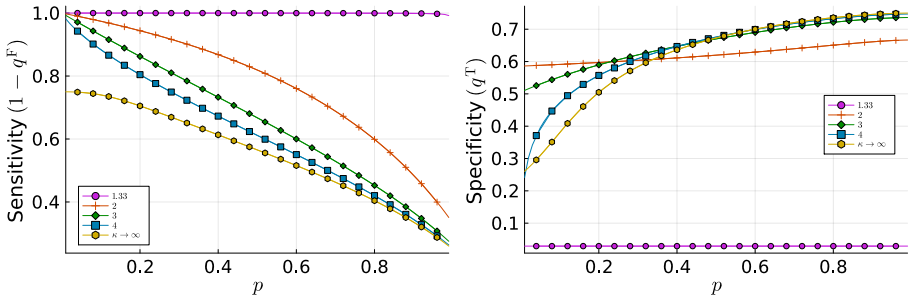
This phenomena could be quite powerful if harnessed appropriately. In a social network where users differ in political alignment and truthfulness and content varying in political position and veracity, we study the *peer filtering effectiveness*, i.e., sensitivity (fraction of false content filtered out) and specificity (fraction of true content not filtered out) and *misinformation level*, i.e., the proportions of true and false content circulating in the network in equilibrium. We examine how the peer-filtering effect depends on the composition of different types of users, informativeness of private signals, average veracity of new content created, and average veracity of the content sources. We also study the non-discriminatory controls that the platform can use to improve the equilibrium peer-filtering effect and mitigate the effects of low veracity of content sources, low signal quality of users, and high fraction of fabulists in the user population. Such platform controls that do not include content moderation, and judgement. We also study the tradeoffs among peer-filtering effectiveness, misinformation level, *engagement level*, i.e., the content sharing frequency of the users, and *volume of content* in the network as a result of the controls used by the platform.

Results: For the binary veracity model, the expressions for the metrics of interest are provided in the following table. We discuss the performance metrics under the boom equilibrium because it is the stable equilibrium for a large range of reasonable values of κ. Since we do observe content going viral, the natural range of κ does admit boom equilibrium. In our numerical analysis of the binary veracity model, we found that in equilibrium peer-filtering has a positive effect and is able to filter out more false news than true news. For the following results, we assume that the population consists of half left and half right leaning users with equal parts each of the four types, i.e., $\theta_{\mathsf{IF}} = \theta_{\mathsf{PF}} = \theta_{\mathsf{IT}} = \theta_{\mathsf{PT}} = 1/4$ and equal parts of left leaning and right leaning content.

Figure 9(a) shows sensitivity and Figure 9(b) shows the specificity as a function of the average new content veracity or the fraction of true content entering the system p for different values of κ. The sensitivity decreases and specificity increases with p because with increasing p users belief about the veracity of the content increases and they are more likely to share the content thus

Quantity	Metric
Peer filtering effectiveness	δ/p
Sensitivity	$1-q^{\mathrm{F}}$
Specificity	q^{T}
Engagement level	$\delta\beta^{\mathrm{T}}+(1-\delta)\beta^{\mathrm{F}}$
Volume	$pq^{\mathrm{T}}+(1-p)q^{\mathrm{F}}$

Fig. 8. Various quantities of interest and the corresponding equilibrium metrics in the binary veracity model.



(a) (b)

Fig. 9. (a) Sensitivity and (b) Specificity in the boom equilibrium as a function of the average veracity p of new content for different values of κ.

increasing the virality of both true and false content. For $\kappa \geq 2$, the sensitivity/specificity is very high/low for small p because almost all content is never shared and is very low/high for high p because almost all content is always shared. $\kappa = 4/3$ is a critical κ given the parameters at which the boom equilibrium emerges. At this value of κ the sensitivity is almost 1 and specificity is 0 because a negligible fraction of the content goes viral irrespective of the average new content veracity. Sensitivity decreases with κ because it increases the virality in general. However, specificity is non-monotonic in κ for low average new content veracity. Increasing κ by a small amount above the critical value of 4/3 gives a big boost to specificity but increasing it further decreases it. This effect is especially important because the sensitivity does not decrease significantly with κ for low values of average new content veracity. Thus by appropriately choosing κ, even for very low values of average new content veracity, peer-filtering effect can be kept very high eliminating a large fraction of false content while allowing a significant fraction of true content to go viral. When the average new content veracity is high then specificity is increasing in κ. For this analysis, we made a conservative assumption that only a quarter of he population is impartial truth teller and only half of the population is truth teller. The critical κ is lower bounded by 1 but would increase with the increasing fraction of partial and impartial truth tellers in the population suggesting that higher values of κ can be sustained with high sensitivity and specificity.

Figure 10 (a) shows the average veracity of the content or the fraction of true content present in the system, δ as a function of the average new content veracity p for different values of κ. This is an indicator of the peer-filtering effectiveness. It shows that δ is at least as big as p suggesting that peer-filtering effectiveness is at least 1. We note that δ is very high even for small values of p for small values of κ. This is because when p is small the users in the social network are more skeptical about the content and only a small fraction of false news goes viral. The difference between δ and p is the highest when the value of p is moderate or the uncertainty about the content veracity is

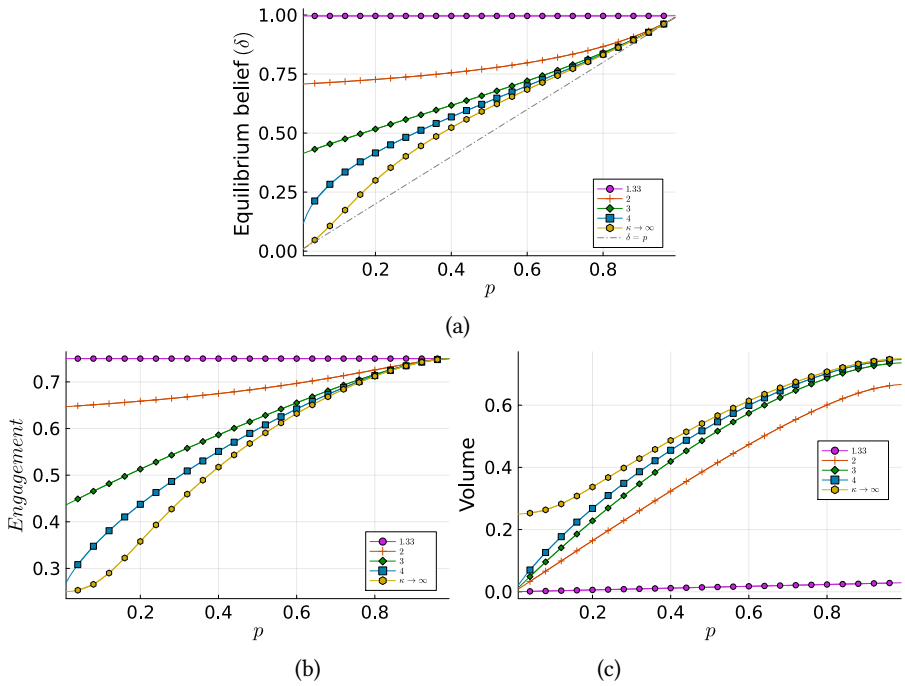


Fig. 10. (a) Average veracity of viral content δ, (b) user engagement level, and (c) volume of viral content in the boom equilibrium as a function of the average veracity p of new content for different values of κ.

high. In this range users weigh their private signal about content veracity more introducing the wisdom of the crowds. This demonstrates the power of peer filtering as a solution for fighting misinformation. For higher value of p δ is naturally high. We also point out that as κ increases the peer filtering effect decreases but does not vanish. Peer filtering effect is still present in the limit as κ approaches infinity. Figures 10 (b) and (c) show the user engagement level and the total content volume in the system respectively as a function of the average new content veracity. We find that the user engagement is consistent with δ suggesting that it may be in platform's interest to keep the misinformation level low on the platform. If κ is the only possible control for the platform then by appropriately choosing κ the platform can both decrease the misinformation level and increase the user engagement without having a significant impact on the volume of content on the platform.

6.1 Peer Filtering under General Type Distributions

Here we study how the type distribution θ affects the key performance metrics. Throughout this section we will fix $\kappa = 3$. This choice is sufficiently large so as to guarantee a boom equilibrium always exists throughout the entire parameter space. This is a result of Theorem 4.3 which says that a boom equilibrium will exist when $\kappa > \frac{2}{1+\theta_{\mathsf{I}}} \geq 2$. We showcase the four cases with $\theta_{\mathsf{IT}} = 1, \theta_{\mathsf{IF}} = 1, \theta_{\mathsf{PT}} = 1, \theta_{\mathsf{PF}} = 1$ respectively. These cases represent the extreme scenarios where the population is homogeneous and consists of entirely one type, and correspond to the four vertices in Figure 7.

Fig. 11(a) shows the sensitivity and Fig. 11(b) shows the specificity as a function of the incoming veracity of content for different settings of θ_{I}. First, we point out that if $\theta_{\mathsf{PF}} = 1$ then no user cares about the truth of the content, only about its alignment, in making the decision whether or not to share. This is why there is no change with respect to changes in p. We see that $\theta_{\mathsf{PT}} = 1$ is very effective at filtering out false content but it comes at the cost that true content is also aggressively

filtered out. On the other side when users are most willing to share it is easier for both true and false content to go viral which is why we see that the $\theta_{\mathsf{IF}} = 1$ curve exhibits low sensitivity and high specificity especially as p increases.

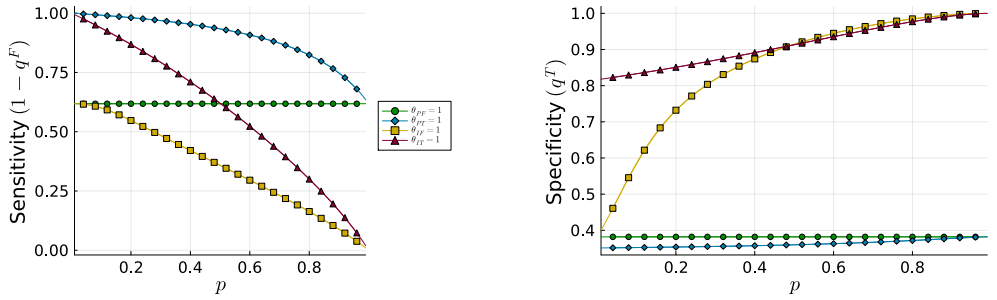


Fig. 11. (a) Sensitivity and (b) Specificity in the boom equilibrium as a function of the average veracity p of new content for different populations. The meaning of each line is as follows: blue diamonds represent $\theta_{PT} = 1$, purple triangles are $\theta_{IT} = 1$, green circles are $\theta_{PF} = 1$, yellow squares are $\theta_{IF} = 1$.

Fig. 12(a) shows the equilibrium δ vs p, Fig. 12(b) shows engagement vs p, and Fig. 12(c) shows volume vs p for various population types. The population with pure impartial fabulists (IF) performs no peer filtering as these users do not base their sharing decisions on the truthfulness of the content. This is the same reason why the volume and the engagement are unaffected by changes in the veracity of content. When all users are truthtellers we see that the equilibrium δ is high which means that of all viral content, the proportion of true content is high. One interesting observation is that the population with purely partisan truthtellers (PT) is more effective in peer filtering than even the population with purely impartial truthtellers (IT). The reason for this is that the partisan aspect makes it less likely that a user chooses to share any particular content but this decreased sharing reduces the probability of false content going viral more than it does true content. The cost of this higher filtering plays out in Figs 12(b) and (c) where purely impartial truthtellers see more volume and more engagement. The population with purely impartial fabulists (IF), who are the most willing to share content, does a poor job of peer filtering, but the resulting equilibrium has high volume and engagement. Curiously, engagement with purely impartial truthtellers is high even for moderate values of p which is due to the high value of δ in equilibrium, leading to high likelihood of the content being shared.

7 CONCLUSIONS

We investigated a democratic method for content moderation that uses the wisdom of crowds and the judgement of users in social networks to filter out false information coming into the network from outside. Through the analysis of a simple yet rich model of users in a social network forming about content veracity and making decisions about sharing the content with their peers, we demonstrated the emergence of the peer filtering phenomena that leads to higher odds of encountering true information in the network than outside the network. We quantified the level of peer filtering using metrics including sensitivity, specificity and effectiveness, i.e.- the odds ratio for true content in the network with true content outside the network. We further studied the impact of user composition, average content veracity, and platform policies on the level of peer filtering as well as user engagement and content volume in the social network. Our results establish the potential of peer filtering to be a natural and democratic solution to the problem of misinformation

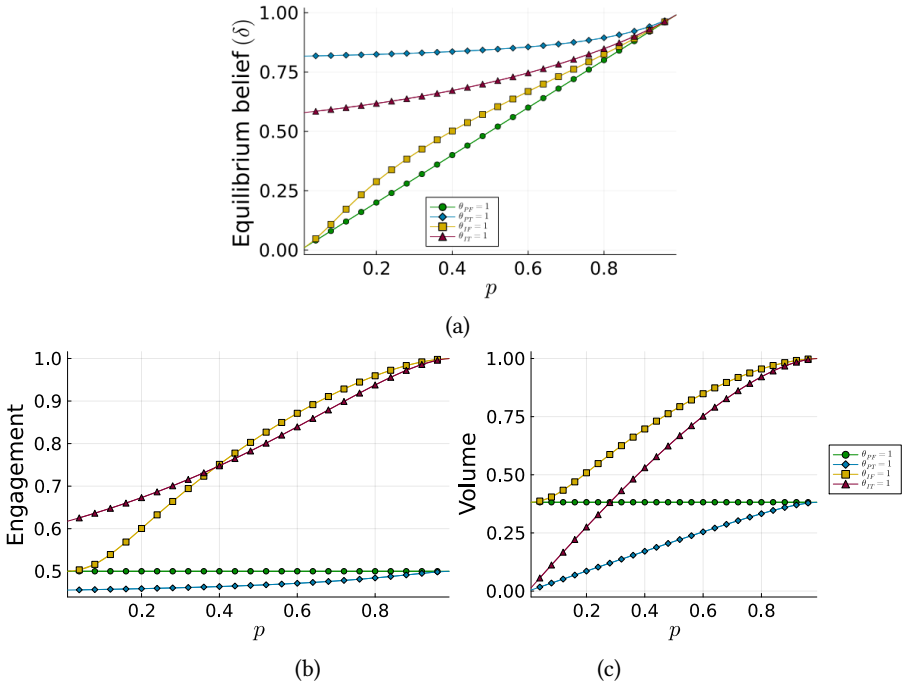


Fig. 12. (a) Average veracity of viral content δ, (b) Engagement level, and (c) volume of viral content as a function p and the population types. The meaning of each line is as follows: blue diamonds represents $\theta_{PT} = 1$, purple triangles are $\theta_{IT} = 1$, green circles are $\theta_{PF} = 1$, yellow squares are $\theta_{IF} = 1$.

spread in online social networks. Our results also highlight that polarization among users does not necessarily increase the level of misinformation. On the contrary, we provided an example in which a more polarized population can have higher peer filtering effect and lower levels of misinformation in the network than a less polarized population.

Our study also hints at possible controls the platforms may develop to mitigate the spread of misinformation. One possible control as our results suggest is the average number of peers who see the content when a user shares it. Online social networking platforms already only display a subset of content shared by the peers in the users news feed. Adjusting the average number of peers who see the news with the changes in the overall level of misinformation in the network can be a powerful non discriminating tool. Further research is needed to identify the right adjustment techniques. Increasing the quality of private signals of the users when exposed to a content can also be a useful tool. This improves information aggregation and user judgements thus increasing the level of peer filtering. Further research is also needed to identify mechanisms for improving signal quality of the users, impact of polarization and homophily in the network. Finally, platform may also publicly display the average level of misinformation to the users. When the misinformation level is high or low, then the users may value the public signal more and take appropriate caution while sharing the information. On the other hand, when the misinformation level is moderate then the users will value their private information more thus proving valuable information aggregation in the network that is important for effective peer filtering.

REFERENCES

Daron Acemoglu, Asuman Ozdaglar, and James Siderius. 2021. *A model of online misinformation.* Technical Report. National Bureau of Economic Research.

John A Banas and Stephen A Rains. 2010. A meta-analysis of research on inoculation theory. *Communication monographs* 77, 3 (2010), 281–311.

Edward A Bender and E Rodney Canfield. 1978. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A* 24, 3 (1978), 296–307.

Puneet Bhargava, Katie MacDonald, Christie Newton, Hause Lin, and Gordon Pennycook. 2023. How effective are TikTok misinformation debunking videos? *Harvard Kennedy School Misinformation Review* (2023).

Nadia M Brashier, Gordon Pennycook, Adam J Berinsky, and David G Rand. 2021. Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences* 118, 5 (2021), e2020043118.

Maria Mercedes Ferreira Caceres, Juan Pablo Sosa, Jannel A Lawrence, Cristina Sestacovschi, Atiyah Tidd-Johnson, Muhammad Haseeb UI Rasool, Vinay Kumar Gadamidi, Saleha Ozair, Krunal Pandav, Claudia Cuevas-Lou, et al. 2022. The impact of misinformation on the COVID-19 pandemic. *AIMS public health* 9, 2 (2022), 262.

John M Carey, Andrew M Guess, Peter J Loewen, Eric Merkley, Brendan Nyhan, Joseph B Phillips, and Jason Reifler. 2022. The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. *Nature Human Behaviour* 6, 2 (2022), 236–243.

Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science* 28, 11 (2017), 1531–1546.

Fan Chung and Linyuan Lu. 2002. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences* 99, 25 (2002), 15879–15882.

Fan Chung, Linyuan Lu, and Van Vu. 2004. The spectra of random graphs with given expected degrees. *Internet Mathematics* 1, 3 (2004), 257–275.

Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political behavior* 42 (2020), 1073–1095.

John Cook, Stephan Lewandowsky, and Ullrich KH Ecker. 2017. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS one* 12, 5 (2017), e0175799.

Krishna Dasaratha and Kevin He. 2019. Aggregative Efficiency of Bayesian Learning in Networks. *arXiv preprint arXiv:1911.10116* (2019).

Ullrich KH Ecker, Stephan Lewandowsky, and Matthew Chadwick. 2020. Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications* 5 (2020), 1–25.

Facebook. 2020. https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/. https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/

Facebook. 2024. https://transparency.fb.com/policies/community-standards/misinformation. https://transparency.fb.com/policies/community-standards/misinformation

Lisa Fazio. 2020. Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review* 1, 2 (2020).

Frances Haugen. 2021. https://facebookpapers.com/. https://facebookpapers.com/

Wanying Huang, Philipp Strack, and Omer Tamuz. 2024. Learning in repeated interactions on networks. *Econometrica* 92, 1 (2024), 1–27.

Matthew O Jackson, Suraj Malladi, and David McAdams. 2022. Learning through the grapevine and the impact of the breadth and depth of social networks. *Proceedings of the National Academy of Sciences* 119, 34 (2022), e2205549119.

Newley Purnell Justin Scheck and Jeff Horwitz. 2021. https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953. https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953

Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Isabella Garcia-Camargo, Emma S Spiro, and Kate Starbird. 2022. Repeat Spreaders and Election Delegitimization: A Comprehensive Dataset of Misinformation Tweets from the 2020 US Election. *Journal of Quantitative Description: Digital Media* 2 (2022).

Stephan Lewandowsky and Sander van der Linden. 2021. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology* 32, 2 (2021), 348–384. https://doi.org/10.1080/10463283.2021.1876983 arXiv:https://doi.org/10.1080/10463283.2021.1876983

Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour* 5, 3 (2021), 337–348.

Ben Lyons, Vittorio Mérola, Jason Reifler, and Florian Stoeckel. 2020. How politics shape views toward fact-checking: Evidence from six European countries. *The International Journal of Press/Politics* 25, 3 (2020), 469–492.

Paul Mena. 2020. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & internet* 12, 2 (2020), 165–183.

Michael Molloy and Bruce Reed. 1995. A critical point for random graphs with a given degree sequence. *Random structures & algorithms* 6, 2-3 (1995), 161–180.

Mohamed Mostagir and James Siderius. 2022. Learning in a post-truth world. *Management Science* 68, 4 (2022), 2860–2868.

Mohamed Mostagir and James Siderius. 2023. Social inequality and the spread of misinformation. *Management Science* 69, 2 (2023), 968–995.

M. E. J. Newman, S. H. Strogatz, and D. J. Watts. 2001. Random graphs with arbitrary degree distributions and their applications. *Pysical Review E* 64 (2001).

Peter E Ney and PE Ney. 2004. *Branching processes*. Courier Corporation.

Jeff Niederdeppe, Kathryn Heley, and Colleen L Barry. 2015. Inoculation and narrative strategies in competitive framing of three health policy issues. *Journal of Communication* 65, 5 (2015), 838–862.

Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

Brendan Nyhan, Jason Reifler, Sean Richey, and Gary L Freed. 2014. Effective messages in vaccine promotion: a randomized trial. *Pediatrics* 133, 4 (2014), e835–e842.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*. Springer, 21–29.

Jessica Paynter, Sarah Luskin-Saxby, Deb Keen, Kathryn Fordyce, Grace Frost, Christine Imms, Scott Miller, David Trembath, Madonna Tucker, and Ullrich Ecker. 2019. Evaluation of a template for countering misinformation—Real-world Autism treatment myth debunking. *PloS one* 14, 1 (2019), e0210746.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.

Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.

Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences* 25, 5 (2021), 388–402.

Michael Pfau and Michael Burgoon. 1988. Inoculation in political campaign communication. *Human Communication Research* 15, 1 (1988), 91–111.

Timothy S Rich, Ian Milden, and Mallory Treece Wagner. 2020. Research note: Does the public support fact-checking social media? It depends who and how you ask. *The Harvard Kennedy School Misinformation Review* 1, 8 (2020).

Adi Robertson. 2021. https://www.theverge.com/2021/9/21/22685863/facebook-safety-security-staff-spending-misinformation-abuse-wall-street-journal-reports. https://www.theverge.com/2021/9/21/22685863/facebook-safety-security-staff-spending-misinformation-abuse-wall-street-journal-reports

Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2021. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *Journal of Public Health* (2021), 1–10.

Salvador Rodriguez. 2020. https://www.cnbc.com/2020/05/28/zuckerberg-facebook-twitter-should-not-fact-check-political-speech.html. https://www.cnbc.com/2020/05/28/zuckerberg-facebook-twitter-should-not-fact-check-political-speech.html

Guy Rosen. 2021. https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/. https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/

Norbert Schwarz, Eryn Newman, and William Leach. 2016. Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy* 2, 1 (2016), 85–95.

Orowa Sikder, Robert E Smith, Pierpaolo Vivo, and Giacomo Livan. 2020. A minimalistic model of bias, polarization and misinformation in social networks. *Scientific reports* 10, 1 (2020), 5493.

Briony Swire-Thompson, Joseph DeGutis, and David Lazer. 2020. Searching for the backfire effect: Measurement and design considerations. *Journal of applied research in memory and cognition* 9, 3 (2020), 286–299.

Sander Van Der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine* 28, 3 (2022), 460–467.

Sander Van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global challenges* 1, 2 (2017), 1600008.

Nathan Walter and Sheila T Murphy. 2018. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication monographs* 85, 3 (2018), 423–441.

Thomas Wood and Ethan Porter. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41 (2019), 135–163.

X.com. 2024. https://help.twitter.com/en/resources/addressing-misleading-info. https://help.twitter.com/en/resources/addressing-misleading-info

Himanshu Zade, Megan Woodruff, Erika Johnson, Mariah Stanley, Zhennan Zhou, Minh Tu Huynh, Alissa Elizabeth Acheson, Gary Hsieh, and Kate Starbird. 2023. Tweet Trajectory and AMPS-based Contextual Cues can Help Users Identify Misinformation. (2023).

8 APPENDIX

8.1 Proof of Theorem 3.1

Proof. We will make frequent use of the observation that $2\hat{\delta}(x) - 1 > 0$ if and only if $x > 1 - \delta$. This follows from simple algebra.

(1) **Case 1: User is an impartial truthteller** Here the expected payoff regardless of alignment is $2\hat{\delta}(x) - 1$ so they will share if and only if $s_i > 1 - \delta$.
(2) **Case 2: User is an partisan truthteller** Here the payoff for sharing misaligned news is -1 so a partisan truthteller will never share misaligned news. But they will share aligned news if and only if $2\hat{\delta}(x) - 1 > 0$ which is true if and only if $s_i > 1 - \delta$.
(3) **Case 3: User is an impartial fabulist** Here the payoff for sharing aligned news is +1 so an impartial fabulist will always share aligned news. But they will share misaligned news if and only if $2\hat{\delta}(x) - 1 > 0$ which is true if and only if $s_i > 1 - \delta$.
(4) **Case 4: User is an partisan fabulist** Here the payoff for sharing aligned news is 1 so a partisan fabulist will always share aligned news. Their payoff for sharing misaligned news is -1 so it won't be ever shared.

□

8.2 Proof of Theorem 4.3

Proof. **Case 1.** $\kappa > \frac{2}{1-\theta_{\mathsf{T}}}$

Using Equation (3), we find that $\beta^{\mathsf{T}}(\delta) \geq \beta^{\mathsf{F}}(\delta) \geq \frac{1-\theta_{\mathsf{T}}}{2}$ for all $\delta \in [0,1]$, and hence, we obtain that $\kappa\beta^{\mathsf{F}}(\delta) > 1$. In turn, this implies that for all $\delta \in [0,1]$, there exists positive solutions $q^{\alpha}(\delta)$ for $\alpha \in \{\mathsf{F},\mathsf{T}\}$ to (6). [ki: can we cite some reference here?] Thus, from (8), we find that $\Pi(\delta) = \Phi(\beta^{\mathsf{F}}(\delta), \beta^{\mathsf{T}}(\delta)) = \frac{pq^{\mathsf{T}}(\delta)}{pq^{\mathsf{T}}(\delta)+(1-p)q^{\mathsf{F}}(\delta)}$ for all $\delta \in [0,1]$. Furthermore, it is straightforward to show that $q^{\alpha}(\delta)$ is continuous in δ for $\alpha \in \{\mathsf{T},\mathsf{F}\}$. Thus, we obtain that $\Pi(\delta)$ is continuous, and furthermore, $\Pi(0) \geq 0$ and $\Pi(1) \leq 1$. Hence, using intermediate value theorem, we conclude that there exists a $\delta^* \in [0,1]$ such that $\Pi(\delta^*) = \delta^*$, and thus corresponds to an equilibrium. That this is a boom equilibrium follows from the fact that $\kappa\beta^{\mathsf{F}}(\delta^*) > 1$.

Case 2. $\kappa \in (\frac{2}{1+\theta_{\mathsf{I}}}, \frac{2}{1-\theta_{\mathsf{T}}})$.

For some fixed small $\epsilon > 0$, let $\bar{\delta} \in (0,1)$ be defined as

$$\bar{\delta} = \sqrt{\frac{2}{\kappa(\theta_{\mathsf{T}} + \theta_{\mathsf{I}})} - \frac{1-\theta_{\mathsf{T}}}{\theta_{\mathsf{T}} + \theta_{\mathsf{I}}} + \epsilon}$$

Then, we have

$$\begin{aligned}
\beta^{\mathsf{F}}(\bar{\delta}) &= \frac{1}{2}\left((1-\theta_{\mathsf{T}}) + (\theta_{\mathsf{T}} + \theta_{\mathsf{I}})\bar{\delta}^2\right) = \frac{1}{\kappa} + \epsilon \\
\beta^{\mathsf{T}}(\bar{\delta}) &= \frac{1}{2}\left((1-\theta_{\mathsf{T}}) + (\theta_{\mathsf{T}} + \theta_{\mathsf{I}})(1-(1-\bar{\delta})^2)\right) \\
&> 1 - \theta_{\mathsf{T}} + (\theta_{\mathsf{T}} + \theta_{\mathsf{I}})\delta - \frac{1}{\kappa} =: h(\kappa)
\end{aligned}$$

Thus, we obtain $\kappa\beta^{\mathsf{F}}(\bar{\delta}) = 1 + \epsilon\kappa$ and $\kappa\beta^{\mathsf{T}}(\bar{\delta}) > \kappa h(\kappa) > 1$. To see this, note that

$$\kappa h(\kappa) = \kappa(1-\theta_{\mathsf{T}}) - 1 + (\theta_{\mathsf{T}} + \theta_{\mathsf{I}})\sqrt{\frac{2\kappa}{\theta_{\mathsf{T}} + \theta_{\mathsf{I}}} - \frac{\kappa^2(1-\theta_{\mathsf{T}})}{\theta_{\mathsf{T}} + \theta_{\mathsf{I}}} + \epsilon}$$

So, $\kappa h(\kappa) > 1$ is equivalent to asking that

$$\begin{aligned}\sqrt{(\theta_{\mathsf{T}}+\theta_{\mathsf{I}})(2-\kappa(1-\theta_{\mathsf{T}}))\kappa} &> 2-\kappa(1-\theta_{\mathsf{T}})\\ \kappa(\theta_{\mathsf{T}}+\theta_{\mathsf{I}})(2-\kappa(1-\theta_{\mathsf{T}})) &> (2-\kappa(1-\theta_{\mathsf{T}}))^2\\ \kappa(\theta_{\mathsf{T}}+\theta_{\mathsf{I}}) &> (2-\kappa(1-\theta_{\mathsf{T}}))\\ \kappa(1+\theta_{\mathsf{I}}) &> 2\\ \kappa &> \frac{1}{1+\theta_{\mathsf{I}}}.\end{aligned}$$

Here, we can divide by $2-\kappa(1-\theta_{\mathsf{T}})$ because it positive under our assumption that $\kappa > \frac{2}{1-\theta_{\mathsf{T}}}$.

Given $\kappa\beta^{\alpha}(\bar{\delta}) > 1$, we conclude that there exists positive solutions $q^{\alpha}(\bar{\delta})$ to (6). Furthermore, we have

$$\begin{aligned}q^{\alpha}(\bar{\delta}) &= \beta^{\alpha}(\bar{\delta})\left(1-(1-q^{\alpha}(\bar{\delta}))^{\kappa}\right)\\ &= \beta^{\alpha}(\bar{\delta})\left(\kappa q^{\alpha}(\bar{\delta}) - \frac{\kappa(\kappa-1)}{2}q^{\alpha}(\bar{\delta})^2 + O(q^{\alpha}(\bar{\delta})^3)\right).\end{aligned}$$

Thus, we get,

$$\frac{1}{\kappa\beta^{\alpha}(\bar{\delta})} = 1 - \frac{\kappa-1}{2}q^{\alpha}(\bar{\delta}) + O\left(q^{\alpha}(\bar{\delta})^2\frac{1}{\kappa\beta^{\alpha}(\bar{\delta})}\right).$$

[ki: Need to make this more formal.] From the fact that $\kappa\beta^{\mathsf{F}}(\bar{\delta}) = 1+\epsilon\kappa$, we obtain that

$$q^{\mathsf{F}}(\bar{\delta}) = \frac{2\kappa}{\kappa-1}\epsilon + O(\epsilon^2).$$

Finally, since $\kappa\beta^{\mathsf{T}}(\bar{\delta}) > \kappa h(\kappa) > 1$, we conclude that $q^{\mathsf{T}}(\bar{\delta}) > \underline{q} > 0$ for all $\epsilon > 0$. Thus, we obtain,

$$\begin{aligned}\Pi(\bar{\delta}) &= \frac{pq^{\mathsf{T}}(\bar{\delta})}{pq^{\mathsf{T}}(\bar{\delta})+(1-p)q^{\mathsf{F}}(\bar{\delta})}\\ &> \frac{p\underline{q}}{p\underline{q}+(1-p)\frac{2\kappa}{\kappa-1}\epsilon+O(\epsilon^2)}\\ &= 1-\frac{2\kappa(1-p)}{p\underline{q}(\kappa-1)}\epsilon+O(\epsilon^2).\end{aligned}$$

Thus, by choosing a small enough $\epsilon > 0$, we obtain that $\Pi(\bar{\delta}) = 1-O(\epsilon) > \bar{\delta}$. Moreover, we obtain $\beta^{\alpha}(\delta) \geq \beta^{\alpha}(\bar{\delta})$ for all $\delta \in [\bar{\delta}, 1]$ and for each $\alpha \in \{\mathsf{T}, \mathsf{F}\}$, implying that $q^{\alpha}(\delta)$ is well-defined, continuous and positive over $\delta \in [\bar{\delta}, 1]$. Thus, we obtain $\Pi(\delta)$ is continuous over $[\bar{\delta}, 1]$. Since $\Pi(1) \leq 1$ and $\Pi(\bar{\delta}) > \bar{\delta}$), we obtain the existence of a $\delta \in [\bar{\delta}, 1]$ with $\Pi(\delta) = \delta$, thereby constituting a boom equilibrium.

Case 3. $\kappa = \frac{2}{1+\theta_{\mathsf{I}}}$.

Since $\beta^{\mathsf{F}}(\delta) = \frac{1}{2}\left((1-\theta_{\mathsf{T}}) + (\theta_{\mathsf{T}}+\theta_{\mathsf{I}})\bar{\delta}^2\right)$ and $\beta^{\mathsf{T}}(\delta) = \frac{1}{2}\left((1-\theta_{\mathsf{T}}) + (\theta_{\mathsf{T}}+\theta_{\mathsf{I}})(1-(1-\bar{\delta})^2)\right)$, we obtain $\kappa\beta^{\alpha}(\delta) > 1$ for all $\delta > 0$ and hence $q^{\alpha}(\delta)$ is well-defined, positive and continuous at all $\delta > 0$. Furthermore, for small enough $\epsilon > 0$, we obtain

$$\begin{aligned}q^{F}(\epsilon) &= \frac{2(\theta_{\mathsf{I}}+\theta_{\mathsf{T}})\epsilon^2}{1+\theta_{\mathsf{T}}} + O(\epsilon^3)\\ q^{T}(\epsilon) &= \frac{4(\theta_{\mathsf{I}}+\theta_{\mathsf{T}})}{\theta_{\mathsf{T}}+1}\epsilon + O(\epsilon^2)\end{aligned}$$

Thus, we obtain

$$\begin{aligned}\Pi(\epsilon) &= \frac{pq^{\mathsf{T}}(\epsilon)}{pq^{\mathsf{T}}(\epsilon) + (1-p)q^{\mathsf{F}}(\epsilon)} \\ &= \frac{2p}{2p + (1-p)\epsilon} + O(\epsilon^2) = 1 - \frac{(1-p)}{2p}\epsilon + O(\epsilon^2).\end{aligned}$$

This implies that for small enough $\epsilon > 0$, we have $\Pi(\epsilon) > \epsilon$. Since $\Pi(\delta) \leq 1$ for all $\delta > \epsilon$, we obtain by continuity that there exists a fixed point of Π in $[\epsilon, 1]$. This fixed point must constitute a boom equilibrium since $\kappa\beta^{\mathsf{F}}(\delta) > 1$ for $\delta \in [\epsilon, 1]$. □

LEMMA 1. *A fixed point δ of Π constitutes a bust equilibrium if and only if*

$$\begin{aligned}f(\delta) = \gamma(\frac{\kappa}{2} - 1 - \frac{\kappa\theta_{\mathsf{T}}}{2}) + \delta\left(1 + \gamma - \frac{\kappa}{2} - \frac{\gamma\kappa}{2} + \frac{\kappa\theta_{\mathsf{T}}}{2} + \frac{\gamma\kappa\theta_{\mathsf{T}}}{2}\right) \\ + \delta^2\left(\frac{\gamma\kappa(\theta_{\mathsf{I}} + \theta_{\mathsf{T}}}{2} - \kappa(\theta_{\mathsf{T}} + \theta_{\mathsf{I}}))\right) + \delta^3\left(\frac{\kappa(\theta_{\mathsf{I}} + \theta_{\mathsf{T}})}{2} - \frac{\gamma\kappa(\theta_{\mathsf{I}} + \theta_{\mathsf{T}})}{2}\right)\end{aligned}$$

and

$$g(\delta) := 4 - \kappa\left(2(1-\theta_{\mathsf{T}}) + 2(\theta_{\mathsf{I}} + \theta_{\mathsf{T}})(1 - (1-\delta)^2)\right) > 0$$

where $\gamma = \frac{p}{1-p} \geq 0$.

PROOF. Suppose that δ constitutes a bust equilibrium. This means that $\kappa\beta^{\mathsf{T}} < 1$ and $\Pi(\delta) = \delta$. Algebra yields that $\kappa\beta^T < 1$ is equivalent to $4 - \kappa\left(2(1-\theta_{\mathsf{T}}) + 2(\theta_{\mathsf{I}} + \theta_{\mathsf{T}})(1-(1-\delta)^2)\right) > 0$.

Taking $\frac{\delta}{1-\delta}$ we get the following derivation

$$\begin{aligned}\frac{\delta}{1-\delta} &= \frac{p}{1-p}\frac{1-\kappa\beta^F}{1-\kappa\beta^T} \\ \delta(1-\kappa\beta^T) &= \gamma(1-\delta)(1-\kappa\beta^F) \\ 0 &= \gamma(1-\delta)(1-\kappa\beta^F) - \delta(1-\kappa\beta^T) \\ 0 &= \gamma(1-\delta)(1-\kappa(\frac{1-\theta_{\mathsf{T}}}{2} + \frac{1}{2}(\theta_{\mathsf{T}} + \theta_{\mathsf{I}})\delta^2) - \delta(1-\kappa(\frac{1-\theta_{\mathsf{T}}}{2} + (\theta_{\mathsf{T}} + \theta_{\mathsf{I}})\delta - \frac{\theta_{\mathsf{I}} + \theta_{\mathsf{T}}}{2}\delta^2))) \\ &= \gamma(\frac{\kappa}{2} - 1 - \frac{\kappa\theta_{\mathsf{T}}}{2}) + \delta\left(1 + \gamma - \frac{\kappa}{2} - \frac{\gamma\kappa}{2} + \frac{\kappa\theta_{\mathsf{T}}}{2} + \frac{\gamma\kappa\theta_{\mathsf{T}}}{2}\right) \\ &\quad + \delta^2(\frac{\gamma\kappa(\theta_{\mathsf{I}} + \theta_{\mathsf{T}}}{2} - \kappa(\theta_{\mathsf{T}} + \theta_{\mathsf{I}}))) + \delta^3(\frac{\kappa(\theta_{\mathsf{I}} + \theta_{\mathsf{T}})}{2} - \frac{\gamma\kappa(\theta_{\mathsf{I}} + \theta_{\mathsf{T}})}{2})\end{aligned}$$

Working the other direction if we assume that $g(\delta) > 0$ and $f(\delta) = 0$ then $g(\delta) > 0$ yields $\kappa\beta^{\mathsf{T}} < 1$. Since $f(\delta) = 0$ we know that $\frac{\delta}{1-\delta} = \frac{p(1+\kappa\beta^F)}{(1-p)(1+\kappa\beta^{\mathsf{T}})}$. Rearrangement yields that $\delta = \frac{\frac{p}{1-\kappa\beta^{\mathsf{T}}}}{\frac{p}{1-\kappa\beta^{\mathsf{T}}} + \frac{1-p}{1-\kappa\beta^F}}$ as desired. □

LEMMA 2. *For $\kappa < \frac{2}{1+\theta_{\mathsf{I}}}$, there exists a bust equilibrium. For $\kappa = \frac{2}{1+\theta_{\mathsf{I}}}$, there exists a (unique) bust equilibrium if and only if $\gamma < \frac{1}{2}$. For $\kappa \in (\frac{2}{1+\theta_{\mathsf{I}}}, \frac{2}{1-\theta_{\mathsf{T}}})$, there exists no bust equilibrium if $\gamma \geq \frac{1}{2}$ or if $\gamma < \frac{1}{2}$ and $\kappa > M(\gamma, \theta_{\mathsf{T}}, \theta_{\mathsf{I}})$; there is a unique bust equilibrium if $\gamma < \frac{1}{2}$ and $\kappa = M(\gamma, \theta_{\mathsf{T}}, \theta_{\mathsf{I}})$; and there are two bust equilibria if $\gamma < \frac{1}{2}$ and $\frac{4}{3} < \kappa < M(\gamma, \theta_{\mathsf{T}}, \theta_{\mathsf{I}})$ where $M(\gamma, \theta_{\mathsf{T}}, \theta_{\mathsf{I}}) := \frac{2(\gamma-1)(\gamma+1)^3}{\gamma^4(\theta_{\mathsf{I}}+1)+\gamma^3(-6\theta_{\mathsf{I}}-8\theta_{\mathsf{T}}+2)+12\gamma^2(\theta_{\mathsf{I}}+\theta_{\mathsf{T}})-2\gamma(4\theta_{\mathsf{I}}+3\theta_{\mathsf{T}}+1)+\theta_{\mathsf{T}}-1}$. Finally, for $\kappa \geq \frac{2}{1-\theta_{\mathsf{T}}}$, there exists no bust equilibrium.*

Proof. **Case 1.** $\kappa < \frac{2}{1+\theta_\mathsf{I}}$. In this case, we have for any $\delta \in [0,1]$,

$$\begin{aligned} g(\delta) &= 4 - \kappa\left(2(1-\theta_\mathsf{T}) + 2(\theta_\mathsf{I}+\theta_\mathsf{T})(1-(1-\delta)^2)\right) \\ &\geq \frac{4}{1+\theta_\mathsf{I}}\left(1+\theta_\mathsf{I} - (\theta_\mathsf{I}+\theta_\mathsf{T})(1-(1-\delta)^2)\right) \\ &= \frac{4(\theta_\mathsf{I}+\theta_\mathsf{T})}{1+\theta_\mathsf{I}}(1-\delta)^2 \geq 0 \end{aligned}$$

Thus, from Lemma 1 it suffices to show the existence of $\delta \in [0,1]$ with $f(\delta) = 0$. To conclude this, observe that $f(0) = -\gamma(1\frac{(1-\theta_\mathsf{T})\kappa}{2}) < 0$ when $\frac{2}{1-\theta_\mathsf{T}} > \kappa$. This is true because for all δ, $\frac{2}{1-\theta_\mathsf{T}} > \frac{2}{1+\theta_\mathsf{I}} > \kappa$. In addition, $f(1) = 1 - \frac{\kappa}{2}(1+\theta_\mathsf{I}) > 0$. We know that that f is continuous over $[0,1]$ therefore there must exist a root to this equation.

We get uniqueness by considering $f'(0)$ and $f'(0)$. $f'(0) = \frac{1}{2}(1+\gamma)(2+\kappa(\theta_\mathsf{T}-1)) > 0$ and $f'(1) = -\frac{1}{2}(1+\gamma)(-2+\kappa+\kappa\theta_\mathsf{I}) < 0$ which follows from the assumption $\kappa > \frac{2}{1+\theta_\mathsf{I}}$. If f were to be a polynomial with more than 1 root in the interval [0,1] and $f(0) < 0, f'(0) > 0, f(1) > 0, f'(1) < 0$ are all true then f has at least three inflection points in the range [0,1]. But this is not possible since f is a cubic. Therefore f has a unique condition.

Case 2. $\kappa = \frac{2}{1+\theta_\mathsf{I}}$. In this case, we can factorize $f(\delta)$ as

$$f(\delta) = -\frac{(\delta-1)^2\left(\gamma+\delta(1-\gamma)\right)(\theta_\mathsf{I}+\theta_\mathsf{T})}{1+\theta_\mathsf{I}}$$

Thus, f has roots at $(\delta = 1)$ and $\delta = \frac{\gamma}{1-\gamma}$.

Note that $g(1) = 0$. This implies that if $\gamma \geq \frac{1}{2}$, there does not exist a bust equilibrium. On the other hand, if $\gamma < \frac{1}{2}$, we obtain $\frac{\gamma}{1-\gamma} \in [0,1]$, and moreover,

$$\begin{aligned} g\left(\frac{\gamma}{1-\gamma}\right) &= \frac{4(1-2\gamma)^2(\theta_\mathsf{I}+\theta_\mathsf{T})}{(\gamma-1)^2(\theta_\mathsf{T}+1)} \\ &= \frac{4(\theta_\mathsf{I}+\theta_\mathsf{T})}{\theta_\mathsf{T}+1}(1-\frac{\gamma}{1-\gamma})^2 \geq 0 \end{aligned}$$

Thus, from Lemma 1, we conclude that if $\gamma < \frac{1}{2}$, there exist a (unique) bust equilibrium at $\delta = \frac{\gamma}{1-\gamma}$.

Case 3. $\kappa \in (\frac{2}{1+\theta_\mathsf{I}}, \frac{2}{1-\theta_\mathsf{T}})$. For $g(\delta) > 0$, we must have

$$\delta < 1 - \sqrt{\frac{\kappa+\kappa\theta_\mathsf{I}-2}{\kappa(\theta_\mathsf{I}+\theta_\mathsf{T})}}.$$

Now, $f(0) = -\gamma(1 - \frac{\kappa(1-\theta_\mathsf{T})}{2} < 0$ and $f(1) = 1 - \frac{\kappa(1+\theta_\mathsf{I})}{2} < 0$. It is straightforward to verify that $f(\delta_0) < 0$.

Moreover, we have

$$f'(\delta) = \frac{(1+\gamma)(2-\kappa+\kappa\theta_\mathsf{T}}{2} + (-2+\gamma)\kappa(\theta_\mathsf{I}+\theta_\mathsf{T})\delta + \frac{3}{2}(1-\gamma)\kappa(\theta_\mathsf{I}+\theta_\mathsf{T}).$$

We have $f'(0) > 0$ and $f'(1) = \frac{1+\gamma}{2}(2-\kappa-\kappa\theta_\mathsf{I}) < 0$. Thus, there exists a (unique) $\delta_+ \in (0,1)$ with $f'(\delta_+) = 0$. We have the following expression for δ_+:

$$\delta_+ = \frac{2-\gamma}{3(1-\gamma)} - \frac{\sqrt{\frac{\gamma^2(-\kappa)+12\gamma^2-8\gamma\kappa+11\kappa-12}{(\gamma-1)^2\kappa}}}{3\sqrt{2}}$$

Thus, we conclude that if $\delta_0 \leq \delta_+$, since $f(0) < 0$ and $f(\delta_0) < 0$, we obtain $f(\delta) < 0$ for all $\delta < \delta_0$, and hence there is no bust equilibrium. Similarly, if $\delta_0 > \delta_+$ and $f(\delta_+) < 0$, then there is

no equilibrium. On the other hand, if $\delta_0 > \delta_+$ and $f(\delta_+) > 0$, then there exists two solutions δ_1 and δ_2 in the interval $[0, \delta_0]$ with $f(\delta_i) = 0$ for $i = 1, 2$. Both these solutions correspond to a bust equilibrium. Finally, if $\delta_0 > \delta_+$ and $f(\delta_+) = 0$, then there exists a unique solution δ_* with $f(\delta_*) = 0$, again corresponding to a bust equilibrium.

Through some algebra, it can be deduced that $\delta_0 \leq \delta_+$ if and only if $\gamma \geq \frac{1}{2}$. For $\gamma < \frac{1}{2}$, we get $f(\delta_+) > 0$ if and only if $\kappa < \frac{2(\gamma-1)(\gamma+1)^3}{\gamma^4(\theta_\mathsf{I}+1)+\gamma^3(-6\theta_\mathsf{I}-8\theta_\mathsf{T}+2)+12\gamma^2(\theta_\mathsf{I}+\theta_\mathsf{T})-2\gamma(4\theta_\mathsf{I}+3\theta_\mathsf{T}+1)+\theta_\mathsf{T}-1} \leq \frac{2}{1-\theta_\mathsf{T}}$. Denote by $M(\gamma, \theta_\mathsf{I}, \theta_\mathsf{T}) := \frac{2(\gamma-1)(\gamma+1)^3}{\gamma^4(\theta_\mathsf{I}+1)+\gamma^3(-6\theta_\mathsf{I}-8\theta_\mathsf{T}+2)+12\gamma^2(\theta_\mathsf{I}+\theta_\mathsf{T})-2\gamma(4\theta_\mathsf{I}+3\theta_\mathsf{T}+1)+\theta_\mathsf{T}-1}$

Summarizing, we obtain that there is no bust equilibrium if $\gamma \geq \frac{1}{2}$ or if $\gamma < \frac{1}{2}$ and $\kappa > M(\gamma, \theta_\mathsf{I}, \theta_\mathsf{T})$; there is a unique bust equilibrium if $\gamma < \frac{1}{2}$ and $\kappa = M(\gamma, \theta_\mathsf{I}, \theta_\mathsf{T})$; and there are two bust equilibria if $\gamma < \frac{1}{2}$ and $M(\gamma, \theta_\mathsf{I}, \theta_\mathsf{T})$.

Case 4. $\kappa \in [\frac{2}{1-\theta_\mathsf{T}}, \infty)$**.** Note that since $\beta^\alpha(\delta) \geq \frac{1-\theta_\mathsf{T}}{2}$ for $\alpha \in \{\mathsf{T}, \mathsf{F}\}$ and $\delta \in [0, 1]$, we have $\kappa\beta^\alpha(\delta) \geq 1$ for $\kappa \geq 4$, and hence a bust equilibrium cannot exist.

□

8.3 Proofs for Section 5

Proof of Lemma 5.1

Proof. It suffices to show that $\beta^\alpha(\delta)$, the overall probability of sharing, under θ and θ' are equivalent as the probability of sharing fully incorporates the type information. Let β^α_θ denote the sharing probability of news with veracity α under a system with type distribution θ.

$$
\begin{aligned}
2\beta^\alpha_\theta &= \sum_{a\in\{A,M\}} \sum_{t\in\theta} \theta_t \beta^\alpha_{a,t} && (9)\\
&= \theta_{IF} + \theta_{PF} + \overline{F}_\alpha(\delta)(\theta_{IT} + \theta_{IF} + \theta_{IF} + \theta_{PT}) && (10)\\
&= \theta_\mathsf{F} + \overline{F}_\alpha(\delta)(\theta_\mathsf{I} + \theta_\mathsf{T}) && (11)\\
&= \theta'_\mathsf{F} + \overline{F}_\alpha(\delta)(\theta'_\mathsf{I} + \theta'_\mathsf{T}) && (12)\\
& && (13)
\end{aligned}
$$

□

Proof of Corollary 8.1

Proof. Suppose we have such θ and θ'. Then

$$
\begin{aligned}
\theta_{IF} + \theta_{IT} &= \theta'_{IT} - t + \theta'_{IT} + t\\
&= \theta'_{IT} + \theta'_{IT}
\end{aligned}
$$

and

$$
\begin{aligned}
\theta_{IF} + \theta_{PF} &= \theta'_{IF} - t + \theta'_{PF} + t\\
&= \theta'_{IF} + \theta'_{PF}
\end{aligned}
$$

□

The following is a useful corollary for writing proofs more compactly.

Corollary 8.1. *Let θ and θ' be two populations. They are equivalent if they can be related as:*

$$\begin{bmatrix}\theta_{IT}\\ \theta_{IF}\\ \theta_{PT}\\ \theta_{PF}\end{bmatrix} = \begin{bmatrix}\theta'_{IT}\\ \theta'_{IF}\\ \theta'_{PT}\\ \theta'_{PF}\end{bmatrix} + t\begin{bmatrix}1\\ -1\\ -1\\ 1\end{bmatrix}$$

For some $t \in \mathbb{R}$.

Proof of Corollary 5.2

Proof. Consider any valid starting θ. It suffices to exhibit a scalar t to fulfill Corollary 8.1. We consider two cases: $\theta_{PT} \geq \theta_{IF}$ and $\theta_{PT} < \theta_{IF}$ to ensure all values remain positive.

Case 1 Take $t = \theta_{IF}$. Then

$$\begin{bmatrix}\theta_{IT}\\ \theta_{IF}\\ \theta_{PT}\\ \theta_{PF}\end{bmatrix} + \theta_{IF}\begin{bmatrix}1\\ -1\\ -1\\ 1\end{bmatrix} = \begin{bmatrix}\theta_{IT}+\theta_{IF}\\ 0\\ \theta_{PT}-\theta_{IF}\\ \theta_{PF}+\theta_{IF}\end{bmatrix}$$

.

We take this right hand side as the constructed θ'.

Case 2 Take $t = \theta_{PT}$. Then

$$\begin{bmatrix}\theta_{IT}\\ \theta_{IF}\\ \theta_{PT}\\ \theta_{PF}\end{bmatrix} + \theta_{PT}\begin{bmatrix}1\\ -1\\ -1\\ 1\end{bmatrix} = \begin{bmatrix}\theta_{IT}+\theta_{PT}\\ \theta_{IF}-\theta_{PT}\\ 0\\ \theta_{PF}+\theta_{PT}\end{bmatrix}$$

.

We take the right hand side as the constructed θ'. □

9 EXTENSIONS

9.1 Non-determinstic spread factor

The first extension we consider is to allow the number of peers content is shared to be be a random variable. In particular, we we will assume that know with probability π_k, a shared content will be shown to k user where $\sum_{k=1}^{\infty} \pi_k = 1$. In this extension, we will consider the effect this has the existence of boom equilbria. The fixed-point equation for the boom case becomes:

$$1 - q^\alpha = 1 - \beta^\alpha + \beta^\alpha \left(\sum_{k=1}^{\infty} \pi_k (1 - q^\alpha)^k \right) \tag{14}$$

We work in cases dependent on $\mathbb{E}[\kappa]$ similar to what was done before in the determinstic case. Specifically, we will consider $\mathbb{E}[\kappa] \geq \frac{2}{1-\theta_T}$, $\mathbb{E}[\kappa] \in (\frac{2}{1+\theta_I}, \frac{2}{1-\theta_T})$, and $\mathbb{E}[\kappa] = \frac{2}{1+\theta_I}$.

Case 1: $\mathbb{E}[\kappa] \geq \frac{2}{1-\theta_T}$.

Using Equation (3), we find that $\beta^{\mathsf{T}}(\delta) \geq \beta^{\mathsf{F}}(\delta) \geq \frac{1-\theta_{\mathsf{T}}}{2}$ for all $\delta \in [0,1]$, and hence, we obtain that $\mathbb{E}[\kappa]\beta^{\mathsf{F}}(\delta) \geq 1$. In turn, this implies that for all $\delta \in [0,1]$, there exists positive solutions $q^\alpha(\delta)$ for $\alpha \in \{\mathsf{F}, \mathsf{T}\}$ to (6). Thus, from (8), we find that $\Pi(\delta) = \Phi(\beta^{\mathsf{F}}(\delta), \beta^{\mathsf{T}}(\delta)) = \frac{pq^{\mathsf{T}}(\delta)}{pq^{\mathsf{T}}(\delta)+(1-p)q^{\mathsf{F}}(\delta)}$ for all $\delta \in [0,1]$. Furthermore, it is straightforward to show that $q^\alpha(\delta)$ is continuous in δ for $\alpha \in \{\mathsf{T}, \mathsf{F}\}$. Thus, we obtain that $\Pi(\delta)$ is continuous, and furthermore, $\Pi(0) \geq 0$ and $\Pi(1) \leq 1$. Hence, using intermediate value theorem, we conclude that there exists a $\delta^* \in [0,1]$ such that $\Pi(\delta^*) = \delta^*$, and thus corresponds to an equilibrium. That this is a boom equilibrium follows from the fact that $\mathbb{E}[\kappa]\beta^{\mathsf{F}}(\delta^*) > 1$.

Case 2: $\mathbb{E}[\kappa] \in (\frac{2}{1+\theta_I}, \frac{2}{1-\theta_T})$

If we let $r^\alpha := 1 - q^\alpha$, then we define $\phi(r) := \sum_{k=0}^{\infty} \pi_k (r^\alpha)^k$ as the probability generating function of r.

From the boom-equilbrium equation written in terms of r we can expand the right-hand side.

$$\begin{aligned}
r^\alpha &= 1 - \beta^\alpha + \beta^\alpha \phi(r) \\
r^\alpha &= \beta^\alpha + \beta^\alpha \left(\phi(1) + \phi'(1)(r^\alpha - 1) + \phi''(1)\frac{(r^\alpha - 1)^2}{2} + O((r^\alpha - 1)^3) \right) \\
&= 1 + \beta^\alpha \left(\phi'(1) + \phi''(1)\frac{r^\alpha - 1}{2} + O((r^\alpha - 1)^2) \right)(r^\alpha - 1) \\
1 &= \beta^\alpha \left(\phi'(1) + \phi''(1)\frac{r^\alpha - 1}{2} + O((r^\alpha - 1)^2) \right) \\
r^\alpha &= \frac{2}{\phi''(1)} \left(\frac{1}{\beta^\alpha} - \theta'(1) - O(\frac{(r^\alpha - 1)^2}{\phi''(1)} \right) + 1
\end{aligned}$$

From here, what we may do is select a small $\epsilon > 0$ and take $\overline{\delta} := \sqrt{\frac{2}{\mathbb{E}[\kappa](\theta_T + \theta_I)} - \frac{1-\theta_T}{\theta_T + \theta_I} + \epsilon}$. What this yields is:

$$\begin{aligned}
\beta^{\mathrm{F}}(\bar{\delta}) &= \frac{1}{2}\left((1-\theta_{\mathrm{T}}) + (\theta_{\mathrm{T}} + \theta_{\mathrm{I}})\bar{\delta}^2\right) = \frac{1}{\mathbb{E}[\kappa]} + \epsilon \\
\beta^{\mathrm{T}}(\bar{\delta}) &= \frac{1}{2}\left((1-\theta_{\mathrm{T}}) + (\theta_{\mathrm{T}} + \theta_{\mathrm{I}})(1 - (1-\bar{\delta})^2)\right) \\
&> 1 - \theta_{\mathrm{T}} + (\theta_{\mathrm{T}} + \theta_{\mathrm{I}})\delta - \frac{1}{\mathbb{E}[\kappa]} =: h(\mathbb{E}[\kappa])
\end{aligned}$$

This implies that $\beta^F(\overline{\delta})\mathbb{E}[\kappa] = 1 + \epsilon\mathbb{E}[\kappa]$. This yields that $q^F = \frac{2\phi'(1)}{\phi''(1)}\epsilon\kappa + O(\epsilon^2)$. Letting $\underline{q}$ be the value of q when $\beta = h(\mathbb{E}[\kappa])$ we have that $q^T > \underline{q} > 0$ for all ϵ. This yields:

$$\begin{aligned}
\pi(\overline{\delta}) &= \frac{pq^T(\overline{\delta}}{pq^T(\overline{\delta} + (1-p)q^F(\overline{\delta}} \\
&> \frac{p\underline{q}}{p\underline{q} + (1-p)\frac{2\phi'(1)}{\phi''(1)} \cdot \epsilon\kappa + O(\epsilon^2)} \\
&= 1 - \frac{2\phi'(1)}{\phi''(1)p\underline{q}}\epsilon\kappa + O(\epsilon^2)
\end{aligned}$$

Case: $\mathbb{E}[\kappa] = \frac{2}{1-\theta_{\mathrm{T}}}$ Since $\beta^{\mathrm{F}}(\delta) = \frac{1}{2}\left((1-\theta_{\mathrm{T}}) + (\theta_{\mathrm{T}} + \theta_{\mathrm{I}})\bar{\delta}^2\right)$ and $\beta^{\mathrm{T}}(\delta) = \frac{1}{2}\left((1-\theta_{\mathrm{T}}) + (\theta_{\mathrm{T}} + \theta_{\mathrm{I}})(1 - (1-\bar{\delta})^2)\right)$, we obtain $\mathbb{E}[\kappa]\beta^\alpha(\delta) > 1$ for all $\delta > 0$ and hence $q^\alpha(\delta)$ is well-defined, positive and continuous at all $\delta > 0$. Furthermore, for small enough $\epsilon > 0$, we obtain

$$\begin{aligned}
q^F(\epsilon) &= \frac{2}{\phi''(1)\phi'(1)}\frac{\theta_{\mathrm{T}} + \theta_{\mathrm{I}}}{1 - \theta_{\mathrm{T}}}\epsilon^2 + O(\epsilon^3) \\
q^T(\epsilon) &= \frac{4(\theta_{\mathrm{I}} + \theta_{\mathrm{T}})}{\theta_{\mathrm{T}} + 1}\epsilon + O(\epsilon^2)
\end{aligned}$$

Thus, we obtain

$$\Pi(\epsilon) = \frac{pq^{\mathsf{T}}(\epsilon)}{pq^{\mathsf{T}}(\epsilon) + (1-p)q^{\mathsf{F}}(\epsilon)} = \frac{2p}{2p + (1-p)\epsilon} + O(\epsilon^2) = 1 - \frac{(1-p)}{2p}\epsilon + O(\epsilon^2).$$

This implies that for small enough $\epsilon > 0$, we have $\Pi(\epsilon) > \epsilon$. Since $\Pi(\delta) \leq 1$ for all $\delta > \epsilon$, we obtain by continuity that there exists a fixed point of Π in $[\epsilon, 1]$. This fixed point must constitute a boom equilibrium since $\kappa\beta^{\mathsf{F}}(\delta) > 1$ for $\delta \in [\epsilon, 1]$.

9.2 General Signal Structures

In this section, we explore how dependent the primary results are to the assumption of a specific "signal structure". By signal structure, what we mean is a set of signal that the users can receive the distributions in which these signals arise for true and false content. Below, is a formal definition.

Definition 3. *A signal structure is a triplet $(\mathcal{S}, \mathbb{P}_{\mathsf{T}}, \mathbb{P}_{\mathsf{F}})$, where $\mathcal{S}$ is a set of signals, and $\mathbb{P}_{\mathsf{T}}$ and $\mathbb{P}_{\mathsf{F}}$ are distributions over $\mathcal{S}$.*

The interpretation of a signal structure $\Sigma = (\mathcal{S}, \mathbb{P}_{\mathsf{T}}, \mathbb{P}_{\mathsf{F}})$ is as follows: the set $\mathcal{S}$ represents the set of values taken by a user's private signal obtained from consuming a content. This signal is assumed to be distributed according to the distribution $\mathbb{P}_{\mathsf{T}}$ if the content is true, and according to the distribution $\mathbb{P}_{\mathsf{F}}$ if the content is false. Specifically, letting $\tilde{s}$ denote the signal received by an user upon consuming a content, we have $\mathbb{P}(\tilde{s} \in B | \alpha = \mathsf{T}) = \mathbb{P}_{\mathsf{T}}(B)$ and $\mathbb{P}(\tilde{s} \in B | \alpha = \mathsf{F}) = \mathbb{P}_{\mathsf{F}}(B)$ for all (measurable) $B \subseteq \mathcal{S}$.

Next, we introduce some notation. Given a signal structure $\Sigma = (\mathcal{S}, \mathbb{P}_{\mathsf{T}}, \mathbb{P}_{\mathsf{F}})$ and $\mu \in [0,1]$, we let $\mathbb{P}_\mu^\Sigma$ denote the measure $\mu(\delta_{\mathsf{T}} \otimes \mathbb{P}_{\mathsf{T}}) + (1-\mu)(\delta_{\mathsf{F}} \otimes \mathbb{P}_{\mathsf{F}})$. When the signal structure Σ is clear from context, we drop the superscript on $\mathbb{P}_\mu^\Sigma$.

The next lemma shows that for any signal exchange there is an equivalent signal structure which has fundamental components used in the section that are needed to justify that true content spreads further than false.

Definition 4. *Given a signal structure $\Sigma = (\mathcal{S}, \mathbb{P}_{\mathsf{T}}, \mathbb{P}_{\mathsf{F}})$ and a mapping $f : \mathcal{S} \to \mathcal{S}'$, we define the pushforward signal structure Σ_f as the signal structure $(\mathcal{S}', \mathbb{P}'_{\mathsf{T}}, \mathbb{P}'_{\mathsf{F}})$, where $\mathbb{P}'_{\mathsf{T}}$ and $\mathbb{P}'_{\mathsf{F}}$ are the pushforwards by f of $\mathbb{P}_{\mathsf{T}}$ and $\mathbb{P}_{\mathsf{F}}$ respectively.*

lemma 3. *For any signal structure $\Sigma = (\mathcal{S}, \mathbb{P}_{\mathsf{T}}, \mathbb{P}_{\mathsf{F}})$ there exists a mapping $f : \mathcal{S} \to [0,1]$ such that the pushforward signal structure $\Sigma_f = ([0,1], \hat{\mathbb{P}}_{\mathsf{T}}, \hat{\mathbb{P}}_{\mathsf{F}})$ satisfies the following properties. (For any $\mu \in [0,1]$, denote $\mathbb{P}_\mu = \mathbb{P}_\mu^\Sigma$. Let $(\tilde{\alpha}, \tilde{s}) \sim \mathbb{P}_\mu$ and $\tilde{u} = f(\tilde{s})$. Note that $(\tilde{\alpha}, \tilde{u}) \sim \mathbb{P}_\mu^{\Sigma_f}$.)*

1. *For all $\mu \in [0,1]$, we have almost surely $\mathbb{P}_\mu(\tilde{\alpha} = \mathsf{T} | \tilde{s} = s) = \mathbb{P}_\mu(\tilde{\alpha} = \mathsf{T} | \tilde{u} = f(s))$*
2. *For almost all $u \in [0,1]$, we have $\mathbb{P}_{1/2}(\tilde{\alpha} = \mathsf{T} | \tilde{u} = u) = u$, and*
3. *$\hat{\mathbb{P}}_{\mathsf{T}}$ first-order-stochastically-dominates $\hat{\mathbb{P}}_{\mathsf{F}}$.*

The preceding result implies that it is without loss of generality to focus on signal structures where the signals take values in $[0,1]$, and for which higher signals are more likely when the content is true than when the content is false (as follows from part (3) of the lemma statement). To see this, part (1) of the statement states that regardless of a user's prior belief about content veracity, her posterior belief upon receiving a signal s under the signal structure Σ is the same as her posterior belief upon receiving the signal $f(s)$ under the pushforward signal structure Σ_f. Thus, one could always focus on Σ_f. This signal structure has the additional property that, if the

content is *a priori* equally likely to be true or false, then the signal represents the posterior belief that the content is true (as stated in part (2)).

Proof. Given a signal structure $\Sigma = (\mathcal{S}, \mathbb{P}_{\mathsf{T}}, \mathbb{P}_{\mathsf{F}})$, let $(\tilde{\alpha}, \tilde{s}) \sim \mathbb{P}^{\Sigma}_{1/2}$, and define f as

$$f(s) \coloneqq \mathbb{P}_{1/2}(\tilde{\alpha} = \mathsf{T} | \tilde{s} = s) \in [0, 1].$$

[ki: the latter expression may not be well-defined for all s.] Let $\Sigma_f = ([0,1], \hat{\mathbb{P}}_{\mathsf{T}}, \hat{\mathbb{P}}_{\mathsf{F}})$ be the corresponding pushforward signal structure. We will next show that Σ_f satisfies the three properties in the lemma statement.

In the following, we consider the coupling $(\tilde{\alpha}, \tilde{s}, \tilde{u})$ with $\tilde{u} = f(\tilde{s})$. Note that if $(\tilde{\alpha}, \tilde{s}) \sim \mathbb{P}^{\Sigma}_{\mu}$ for some $\mu \in [0,1]$, then $(\tilde{\alpha}, \tilde{u}) \sim \mathbb{P}^{\Sigma_f}_{\mu}$.

Part (1) Fix $\mu \in [0,1]$. We have

$$\begin{aligned}\mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{T}, \tilde{s} = s) &= \mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{T}, \tilde{s} = s) \\ &= \mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{T}) \cdot \mathbb{P}_{\mu}(\tilde{s} = s | \tilde{\alpha} = \mathsf{T}) \\ &= \mu \mathbb{P}_{\mathsf{T}}(\tilde{s} = s).\end{aligned}$$

Now,

$$\mathbb{P}_{\mathsf{T}}(\tilde{s} = s) = \mathbb{P}_{1/2}(\tilde{s} = s | \tilde{\alpha} = \mathsf{T}) = \frac{\mathbb{P}_{1/2}(\tilde{\alpha} = \mathsf{T} | \tilde{s} = s)\, \mathbb{P}_{1/2}(\tilde{s} = s)}{\mathbb{P}_{1/2}(\tilde{\alpha} = \mathsf{T})} = 2f(s) \cdot \mathbb{P}_{1/2}(\tilde{s} = s),$$

and hence we have

$$\mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{T}, \tilde{s} = s) = 2\mu f(s) \cdot \mathbb{P}_{1/2}(\tilde{s} = s).$$

Similarly, we have $\mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{F}, \tilde{s} = s) = 2(1-\mu)(1-f(s)) \cdot \mathbb{P}_{1/2}(\tilde{s} = s)$, and hence

$$\mathbb{P}_{\mu}(\tilde{s} = s) = 2\left(\mu f(s) + (1-\mu)(1-f(s))\right) \cdot \mathbb{P}_{1/2}(\tilde{s} = s).$$

Thus, we obtain,

$$\mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{T} | \tilde{s} = s) = \frac{\mu f(s)}{\mu f(s) + (1-\mu)(1-f(s))}. \tag{15}$$

Next, using similar arguments, we have

$$\begin{aligned}\mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{T}, \tilde{u} = f(s)) &= \mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{T}, f(\tilde{s}) = f(s)) \\ &= \mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{T}) \cdot \mathbb{P}_{\mu}(f(\tilde{s}) = f(s) | \tilde{\alpha} = \mathsf{T}) \\ &= \mu \mathbb{P}_{\mathsf{T}}(f(\tilde{s}) = f(s)).\end{aligned}$$

Now, for any $s' \in \mathcal{S}$ with $f(s') = f(s)$, we have

$$\begin{aligned}\mathbb{P}_{\mathsf{T}}(\tilde{s} = s') &= 2f(s') \cdot \mathbb{P}_{1/2}(\tilde{s} = s') \\ &= 2f(s) \cdot \mathbb{P}_{1/2}(\tilde{s} = s').\end{aligned}$$

Thus, upon integrating over all such s', we obtain

$$\mathbb{P}_{\mathsf{T}}(f(\tilde{s}) = f(s)) = 2f(s) \cdot \mathbb{P}_{1/2}(f(\tilde{s}) = f(s)),$$

and hence, we have

$$\mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{T}, \tilde{u} = f(s)) = 2\mu f(s) \cdot \mathbb{P}_{1/2}(f(\tilde{s}) = f(s)).$$

Similarly, we have $\mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{F}, f(\tilde{s}) = f(s)) = 2(1-\mu)(1-f(s)) \cdot \mathbb{P}_{1/2}(f(\tilde{s}) = f(s))$, and thus we get

$$\mathbb{P}_{\mu}(\tilde{\alpha} = \mathsf{T} | \tilde{u} = f(s)) = \frac{\mu f(s)}{\mu f(s) + (1-\mu)(1-f(s))}.$$

Comparing with (15), we conclude that, almost surely,

$$\mathbb{P}_\mu\left(\tilde{\alpha} = \mathsf{T} | \tilde{s} = s\right) = \mathbb{P}_\mu\left(\tilde{\alpha} = \mathsf{T} | \tilde{u} = f(s)\right).$$

Part (2) From Part (1), we have for almost all s,

$$\mathbb{P}_{1/2}\left(\tilde{\alpha} = \mathsf{T} | \tilde{u} = f(s)\right) = \mathbb{P}_{1/2}\left(\tilde{\alpha} = \mathsf{T} | \tilde{s} = s\right) = f(s).$$

Since this holds for almost all s, we obtain that $\mathbb{P}_{1/2}\left(\tilde{\alpha} = \mathsf{T} | \tilde{u} = u\right) = u$ for almost all $u \in [0,1]$.
Part (3) We have

$$\begin{aligned}
\mathbb{P}_{1/2}\left(\tilde{\alpha} = \mathsf{T}, \tilde{u} \geq u\right) &= \mathbb{E}_{1/2}\left[\mathbf{I}\{\tilde{\alpha} = \mathsf{T}, \tilde{u} \geq u\}\right] \\
&= \mathbb{E}_{1/2}\left[\mathbb{P}_{1/2}\left(\tilde{\alpha} = \mathsf{T} | \tilde{u}\right) \mathbf{I}\{\tilde{u} \geq u\}\right] \\
&= \mathbb{E}_{1/2}\left[\tilde{u}\mathbf{I}\{\tilde{u} \geq u\}\right] \\
&\geq \mathbb{E}_{1/2}\left[\tilde{u}\right] \mathbb{E}_{1/2}\left[\mathbf{I}\{\tilde{u} \geq u\}\right] \\
&= \frac{1}{2}\left(\mathbb{P}_{1/2}\left(\tilde{\alpha} = \mathsf{T}, \tilde{u} \geq u\right) + \mathbb{P}_{1/2}\left(\tilde{\alpha} = \mathsf{F}, \tilde{u} \geq u\right)\right)
\end{aligned}$$

Here, the third equality follows from Part (2), the inequality follows from the monotonicity considerations, and the final equality follows from the fact that $\mathbb{E}_{\frac{1}{2}}[\tilde{u}] = \mathbb{P}_{1/2}(\tilde{\alpha} = \mathsf{T}) = \frac{1}{2}$.

Thus, we obtain

$$\mathbb{P}_{1/2}\left(\tilde{\alpha} = \mathsf{T}, \tilde{u} \geq u\right) \geq \mathbb{P}_{1/2}\left(\tilde{\alpha} = \mathsf{F}, \tilde{u} \geq u\right).$$

Upon dividing by $\mathbb{P}_{1/2}(\tilde{\alpha} = \mathsf{T}) = \mathbb{P}_{1/2}(\tilde{\alpha} = \mathsf{F}) = \frac{1}{2}$, we obtain for almost all u,

$$\hat{\mathbb{P}}_{\mathsf{T}}\left(\tilde{u} \geq u\right) \geq \hat{\mathbb{P}}_{\mathsf{F}}\left(\tilde{u} \geq u\right).$$

Thus, we conclude that $\hat{\mathbb{P}}_{\mathsf{T}}$ first-order-stochastically-dominates $\hat{\mathbb{P}}_{\mathsf{F}}$. □

The preceding lemma is sufficient to establish many of the structural results on an equilibrium. Specifically, part (2) of the lemma statement implies that the single user behavior, as described in Theorem 3.1 continue to hold. Similarly, the first-order-stochastic-dominance result implies that $\beta_{\mathsf{T}}(\delta) \geq \beta_{\mathsf{F}}(\delta)$, and hence $q^{\mathsf{T}} \geq q^{\mathsf{F}}$. This in turn implies that in any equilibrium, we observe peer filtering, i.e., $\delta \geq p$. [ki: What else can we add here?]