Peer Filtering

Anonymous Author(s)

ABSTRACT

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 (fraction of types of users), average content veracity and platform control (number of peers who observe the content) on the power of peer filtering effect.

ACM Reference Format:
Anonymous Author(s). 2023. Peer Filtering. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 [1–3].

Social networking platforms spend a lot of effort on curtailing misinformation and have clear policies about removing harmful misinformation [4, 5]. Facebook spent over $13 billion between 2016 and 2021 on "safety and security", employs 40,000 employees to reduce misinformation [6, 7] and its employees and contractors spent 3.2 million hours alone on searching, labelling and removing misinformation in 2020 [8]. 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 [9–13], (ii) *Nudging* users to verify content before sharing [14–17], (iii) *Debunking* content after it has been spreading in the network [18–20], and (iv) *Prebunking* or *inoculating* users before misinformation gets introduced in the network [18, 21–24]. While some have shown small or moderate effectiveness [25–29], in general they have been inadequate in combating ideological reasoning [30–36], only have short term benefits [30, 37–40], and require readily available reliable information [14–17].

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 [41]. 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[42] 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. Recent literature has

explored strategic models to study the spread of misinformation in networks [43–46]. The ones closest to our work include [44, 45]. In particular, [44] 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 [45] 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.

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. 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.

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.

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 3, we present the analysis of user belief formations and content sharing decisions. In 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 5 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 6 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). 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, we assume that the network is a generated instance from the configuration model [47–51].

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.

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

	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)

Figure 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.

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.

	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

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

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.

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* $s_i > 1 - \delta$.

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

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

where $\overline{\mathsf{F}}_{\alpha}$ denotes the complimentary cdf of the signal distribution for content with veracity α and hence $\overline{\mathsf{F}}_{\alpha}(1-\delta)$ 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

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

where $\theta_{\mathsf{F}} = \theta_{\mathsf{IF}} + \theta_{\mathsf{PF}}$, $\theta_{\mathsf{T}} = \theta_{\mathsf{IT}} + \theta_{\mathsf{PT}}$ and $\theta_{\mathsf{I}} = \theta_{\mathsf{IT}} + \theta_{\mathsf{IF}}$. 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.

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

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

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 [52]. 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*.

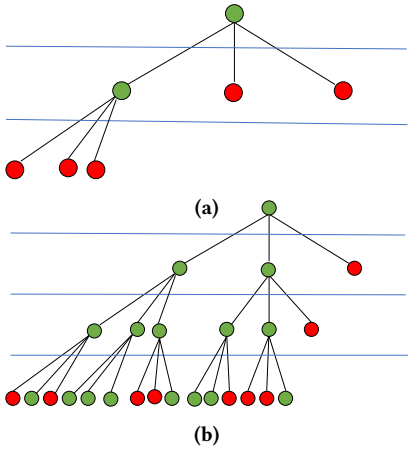


Figure 3: Spread of content in a large network, green nodes decide to share the content; red nodes decide not to share the content; (a) false content is less likely to be shared and the spread may go bust (b) true content is slightly more likely to be shared and may end up going viral.

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

$$1 - q^{\alpha} = 1 - \beta^{\alpha} + \beta^{\alpha}(1 - q^{\alpha})^{\kappa}. \tag{4}$$

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. Fix a user i and consider a content with veracity α that is first seen by a peer who is ℓ hops away in the network. For this content to reach the user i, each of the ℓ users on the connecting path should choose to share the content, an event with probability $(\beta^{\alpha})^{\ell}$. Since there are κ^{ℓ} such peers who are ℓ hops away from i and since new content arrives uniformly at random over the network, the total probability that a content with veracity α that arrives ℓ hops away and reaches user i is proportional to $(\kappa\beta^{\alpha})^{\ell}$. Summing over all ℓ, we obtain that the total probability that a content with veracity α is seen by user i is proportional to $\frac{1}{1-\kappa\beta^{\alpha}}$. Since any new content is true with probability p and false with probability $1-p$, we obtain via Bayes' rule that of the content seen by a user, the proportion that is true is given by

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

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

$$\Phi(\beta^{\mathsf{T}},\beta^{\mathsf{F}}) := \begin{cases} \dfrac{\frac{p}{1-\kappa\beta^{\mathsf{T}}}}{\frac{p}{1-\kappa\beta^{\mathsf{T}}}+\frac{1-p}{1-\kappa\beta^{\mathsf{F}}}}, & \text{if } \kappa\beta^{\mathsf{F}} \leq \kappa\beta^{\mathsf{T}} < 1; \\ 1, & \text{if } \kappa\beta^{\mathsf{F}} < 1 \leq \kappa\beta^{\mathsf{T}}; \\ \dfrac{pq^{\mathsf{T}}}{pq^{\mathsf{T}}+(1-p)q^{\mathsf{F}}}, & \text{if } 1 \leq \kappa\beta^{\mathsf{F}} \leq \kappa\beta^{\mathsf{T}}. \end{cases} \tag{5}$$

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 Fig. 4.

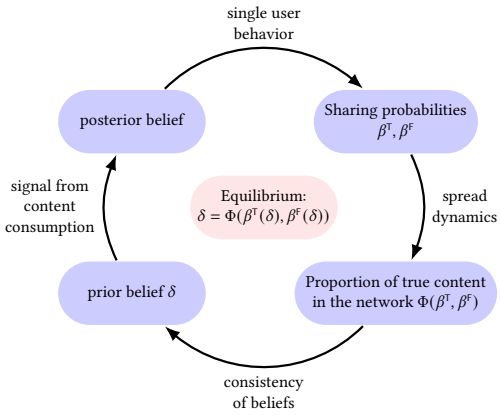


Figure 4: Equilibrium in the network. Here $\beta^{\mathsf{T}}, \beta^{\mathsf{F}}$ are obtained from δ through (3), and the proportion of true content $\Phi(\beta^{\mathsf{T}}, \beta^{\mathsf{F}})$ is given by (5). An equilibrium corresponds to a value of δ for which the belief is consistent with spread dynamics.

Definition 1. *An equilibrium is given by the sharing probabilities $(\beta^{\alpha} : \alpha \in \{\mathsf{T},\mathsf{F}\})$ 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 $\delta = \Phi(\beta^{\mathsf{T}},\beta^{\mathsf{F}})$, where $\Phi(\beta^{\mathsf{T}},\beta^{\mathsf{F}})$ is given by (5). Taken together, in an equilibrium, the prior belief δ satisfies the fixed-point equation $\delta = \Phi(\beta^{\mathsf{T}}(\delta),\beta^{\mathsf{F}}(\delta))$.*

We observe that any equilibrium must be one of three types, depending on which of the three cases holds in the definition (5) of the function Φ. In the first type, we have $\kappa\beta^{\mathsf{F}} \leq \kappa\beta^{\mathsf{T}} < 1$ implying that neither content types ever go viral; we call this type a *bust equilibrium*. In the second type, we have $\kappa\beta^{\mathsf{F}} < 1 \leq \kappa\beta^{\mathsf{T}}$, 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 $1 \leq \kappa\beta^{\mathsf{F}} \leq \kappa\beta^{\mathsf{T}}$ 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 $\kappa\beta^{\mathsf{F}} < 1$ and $\kappa\beta^{\mathsf{T}} > 1$, implying that $\Phi(\beta^{\mathsf{F}},\beta^{\mathsf{T}}) = 1$ and hence $\delta = \Pi(\delta) = \Phi(\beta^{\mathsf{F}},\beta^{\mathsf{T}}) = 1$. However, using (3), we then find that $\beta^{\mathsf{F}} = \beta^{\mathsf{T}} = 1$, contradicting the implication that $\kappa\beta^{\mathsf{T}} > 1 > \kappa\beta^{\mathsf{F}}$. □

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.

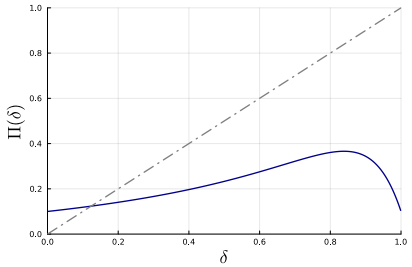
Before showing the existence of equilibrium, we present the following result which establishes the existence of threshold values of κ, which depend solely on the proportions of user types, that categorize the nature of the equilibrium.

Theorem 4.2. *For $\kappa < \frac{2}{1+\theta_I}$, there exists no boom equilibrium. Similarly, for $\kappa > \frac{2}{\theta_F}$, there exists no bust equilibrium.*

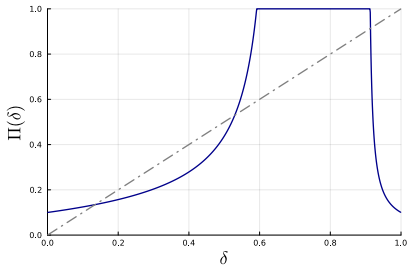
Proof. For a boom equilibrium, it must be the case that $\kappa\beta^T > 1$. From the expression for β^T, we obtain that for any $\delta \in [0,1]$, $\beta^T(\delta) \leq \beta^T(1) \leq \frac{1+\theta_I}{2}$. Thus, if $\kappa < \frac{2}{1+\theta_I}$, we obtain $\kappa\beta^T < 1$, implying that there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

A full characterization of the equilibria for general θ is complex. For simplicity of presentation in the rest of the paper we consider $\theta_t = \frac{1}{4}$ for all for types $t \in \{PF, IF, PT, IT\}$ of users. The methodology used and the qualitative results are the same for other values of θ. With this choice of θ_t, we obtain from Theorem 4.2 that there exists no boom equilibrium if $\kappa < \frac{2}{1+\theta_I} = \frac{4}{3}$ and there exists no bust equilibrium if $\kappa > \frac{2}{\theta_F} = 4$. Our results establish the existence of equilibrium for all values of κ, and also provide insight into the character of the equilibria for the intermediate range $\kappa \in (\frac{4}{3}, 4)$.

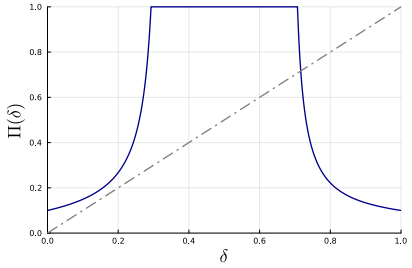
The primary approach for establishing existence of the equilibrium is using fixed-point theorems to show that the equation $\delta = \Phi(\beta^T(\delta), \beta^F(\delta))$ has a solution, where $\beta^\alpha(\delta)$ is given by (3) and the function $\Phi(\beta^T, \beta^F)$ is defined in (5). To use such an approach, we establish the continuity of the underlying functions; the case-by-case definition of Φ in (5) 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 $\Pi(\delta) := \Phi(\beta^T(\delta), \beta^F(\delta))$ 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 (5), 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.



(a) $\kappa = 1.3$



(b) $\kappa = 1.5$



(c) $\kappa = 2$

Figure 5: The function $\Pi(\delta) := \Phi(\beta^T(\delta), \beta^F(\delta))$ 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 (5), and the sharing probabilities $\beta^\alpha(\delta)$ for $\alpha \in \{T, F\}$ are given by (3).

Theorem 4.3. *There exists a boom equilibrium if and only if $\kappa > \frac{4}{3}$. On the other hand, there exists a bust equilibrium if any of the following conditions are met:*

- *$\kappa < \frac{4}{3}$ (a unique bust equilibrium exists);*
- *$\kappa = \frac{4}{3}$ and $\gamma < \frac{1}{2}$ (a unique bust equilibrium exists);*
- *If $\kappa \in \left(\frac{4}{3}, \kappa(\gamma)\right)$ and $\gamma < \frac{1}{2}$ (two bust equilibria exist);*
- *If $\kappa = \kappa(\gamma)$ and $\gamma < \frac{1}{2}$ (a unique bust equilibrium exists).*

Otherwise, there exists no bust equilibrium. (Here, $\kappa(\gamma) := (4 + 8\gamma - 8\gamma^3 - 4\gamma^4)/(1 + 18\gamma - 24\gamma^2 + 10\gamma^3 - 3\gamma^4) \in (\frac{4}{3}, 4]$ for $\gamma \in [0, \frac{1}{2})$.)

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 $\Phi(\beta^F, \beta^T)$ in (5) depends on the spread dynamics. In particular, given the case-by-case definition, it is not immediately clear if the

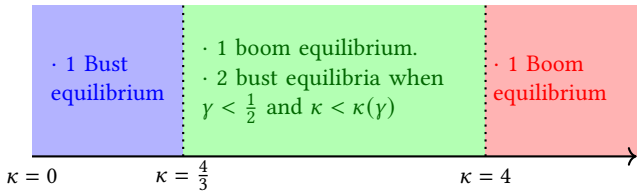


Figure 6: Visual representation of the characterization of boom and bust equilibria when $\theta_t = \frac{1}{4}$.

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 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.

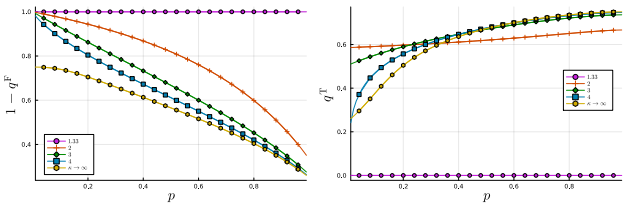
This phenomena could be quite powerful if harnessed appropriately. In a social network where users differ in political alignment and truthfulness and content varying in political position and veracity, we study the *peer filtering effectiveness*, i.e., sensitivity (fraction of false content filtered 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

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

Figure 7: Various quantities of interest and the corresponding equilibrium metrics in the binary veracity model.

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., $\theta_{\mathsf{lF}} = \theta_{\mathsf{rF}} = \theta_{\mathsf{lT}} = \theta_{\mathsf{rT}} = 1/4$ and equal parts of left leaning and right leaning content.



(a) (b)

Figure 8: (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.

Fig. 8(a) shows sensitivity and Figure 8(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 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.

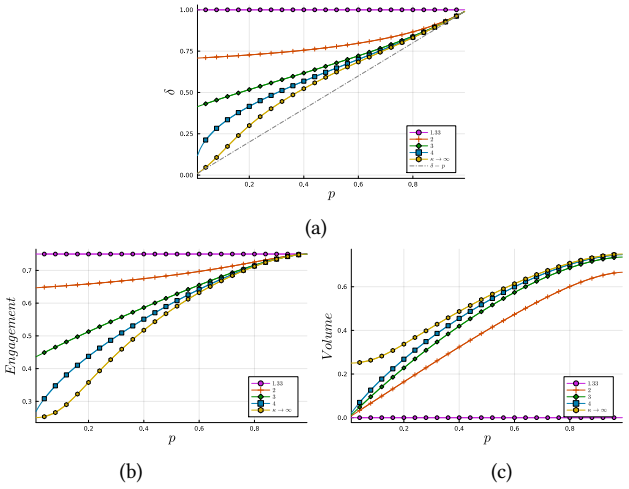


Figure 9: (a) Average veracity of viral content δ, (b) user engagement level, and (c) volume of viral content in the boom equilibrium as a function of the average veracity p of new content for different values of κ.

Fig. 9 (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 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 9 (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 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 spread in online social networks.

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. 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. Further research is also needed to characterize optimal signaling mechanisms for the platform.

REFERENCES

[1] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature human behaviour*, vol. 5, no. 3, pp. 337–348, 2021.

[2] M. M. F. Caceres, J. P. Sosa, J. A. Lawrence, C. Sestacovschi, A. Tidd-Johnson, M. H. U. Rasool, V. K. Gadamidi, S. Ozair, K. Pandav, C. Cuevas-Lou, *et al.*, "The impact of misinformation on the covid-19 pandemic," *AIMS public health*, vol. 9, no. 2, p. 262, 2022.

[3] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, and L. D. de Figueiredo Nicolete, "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*, pp. 1–10, 2021.

[4] "https://transparency.fb.com/policies/community-standards/misinformation."

[5] "https://help.twitter.com/en/resources/addressing-misleading-info."

[6] "https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/."

[7] "https://www.theverge.com/2021/9/21/22685863/facebook-safety-security-staff-spending-misinformation-abuse-wall-street-journal-reports."

[8] "https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953."

[9] "https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/."

[10] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "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*, pp. 21–29, Springer, 2021.

[11] H. Zade, M. Woodruff, E. Johnson, M. Stanley, Z. Zhou, M. T. Huynh, A. E. Acheson, G. Hsieh, and K. Starbird, "Tweet trajectory and amps-based contextual cues can help users identify misinformation," 2023.

[12] I. Kennedy, M. Wack, A. Beers, J. S. Schafer, I. Garcia-Camargo, E. S. Spiro, and K. Starbird, "Repeat spreaders and election delegitimization: A comprehensive dataset of misinformation tweets from the 2020 us election," *Journal of Quantitative Description: Digital Media*, vol. 2, 2022.

[13] G. Pennycook and D. G. Rand, "Fighting misinformation on social media using crowdsourced judgments of news source quality," *Proceedings of the National Academy of Sciences*, vol. 116, no. 7, pp. 2521–2526, 2019.

[14] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, "Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention," *Psychological science*, vol. 31, no. 7, pp. 770–780, 2020.

[15] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand, "Shifting attention to accuracy can reduce misinformation online," *Nature*, vol. 592, no. 7855, pp. 590–595, 2021.

[16] G. Pennycook and D. G. Rand, "The psychology of fake news," *Trends in cognitive sciences*, vol. 25, no. 5, pp. 388–402, 2021.

[17] L. Fazio, "Pausing to consider why a headline is true or false can help reduce the sharing of false news," *Harvard Kennedy School Misinformation Review*, vol. 1, no. 2, 2020.

[18] S. Van Der Linden, "Misinformation: susceptibility, spread, and interventions to immunize the public," *Nature Medicine*, vol. 28, no. 3, pp. 460–467, 2022.

[19] M.-p. S. Chan, C. R. Jones, K. Hall Jamieson, and D. Albarracín, "Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation," *Psychological science*, vol. 28, no. 11, pp. 1531–1546, 2017.

[20] B. Nyhan, J. Reifler, S. Richey, and G. L. Freed, "Effective messages in vaccine promotion: a randomized trial," *Pediatrics*, vol. 133, no. 4, pp. e835–e842, 2014.

[21] S. Van der Linden, A. Leiserowitz, S. Rosenthal, and E. Maibach, "Inoculating the public against misinformation about climate change," *Global challenges*, vol. 1, no. 2, p. 1600008, 2017.

[22] M. Pfau and M. Burgoon, "Inoculation in political campaign communication," *Human Communication Research*, vol. 15, no. 1, pp. 91–111, 1988.

[23] J. Niederdeppe, K. Heley, and C. L. Barry, "Inoculation and narrative strategies in competitive framing of three health policy issues," *Journal of Communication*, vol. 65, no. 5, pp. 838–862, 2015.

[24] J. Cook, S. Lewandowsky, and U. K. Ecker, "Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence," *PloS one*, vol. 12, no. 5, p. e0175799, 2017.

[25] K. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan, *et al.*, "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*, vol. 42, pp. 1073–1095, 2020.

[26] N. Walter and S. T. Murphy, "How to unring the bell: A meta-analytic approach to correction of misinformation," *Communication monographs*, vol. 85, no. 3, pp. 423–441, 2018.

[27] N. M. Brashier, G. Pennycook, A. J. Berinsky, and D. G. Rand, "Timing matters when correcting fake news," *Proceedings of the National Academy of Sciences*, vol. 118, no. 5, p. e2020043118, 2021.

[28] P. Mena, "Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook," *Policy & internet*, vol. 12, no. 2, pp. 165–183, 2020.

[29] T. Wood and E. Porter, "The elusive backfire effect: Mass attitudes' steadfast factual adherence," *Political Behavior*, vol. 41, pp. 135–163, 2019.

[30] B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behavior*, vol. 32, no. 2, pp. 303–330, 2010.

[31] B. Lyons, V. Mérola, J. Reifler, and F. Stoeckel, "How politics shape views toward fact-checking: Evidence from six european countries," *The International Journal of Press/Politics*, vol. 25, no. 3, pp. 469–492, 2020.

[32] T. S. Rich, I. Milden, and M. T. Wagner, "Research note: Does the public support fact-checking social media? it depends who and how you ask," *The Harvard Kennedy School Misinformation Review*, vol. 1, no. 8, 2020.

[33] N. Schwarz, E. Newman, and W. Leach, "Making the truth stick & the myths fade: Lessons from cognitive psychology," *Behavioral Science & Policy*, vol. 2, no. 1, pp. 85–95, 2016.

[34] P. Bhargava, K. MacDonald, C. Newton, H. Lin, and G. Pennycook, "How effective are tiktok misinformation debunking videos?," *Harvard Kennedy School Misinformation Review*, 2023.

[35] U. K. Ecker, S. Lewandowsky, and M. Chadwick, "Can corrections spread misinformation to new audiences? testing for the elusive familiarity backfire effect," *Cognitive Research: Principles and Implications*, vol. 5, pp. 1–25, 2020.

[36] J. Paynter, S. Luskin-Saxby, D. Keen, K. Fordyce, G. Frost, C. Imms, S. Miller, D. Trembath, M. Tucker, and U. Ecker, "Evaluation of a template for countering misinformation—real-world autism treatment myth debunking," *PloS one*, vol. 14, no. 1, p. e0210746, 2019.

[37] J. M. Carey, A. M. Guess, P. J. Loewen, E. Merkley, B. Nyhan, J. B. Phillips, and J. Reifler, "The ephemeral effects of fact-checks on covid-19 misperceptions in the united states, great britain and canada," *Nature Human Behaviour*, vol. 6, no. 2, pp. 236–243, 2022.

[38] J. A. Banas and S. A. Rains, "A meta-analysis of research on inoculation theory," *Communication monographs*, vol. 77, no. 3, pp. 281–311, 2010.

[39] S. Lewandowsky and S. van der Linden, "Countering misinformation and fake news through inoculation and prebunking," *European Review of Social Psychology*, vol. 32, no. 2, pp. 348–384, 2021.

[40] B. Swire-Thompson, J. DeGutis, and D. Lazer, "Searching for the backfire effect: Measurement and design considerations," *Journal of applied research in memory and cognition*, vol. 9, no. 3, pp. 286–299, 2020.

[41] "https://facebookpapers.com/."

[42] "https://www.cnbc.com/2020/05/28/zuckerberg-facebook-twitter-should-not-fact-check-political-speech.html."

[43] O. Sikder, R. E. Smith, P. Vivo, and G. Livan, "A minimalistic model of bias, polarization and misinformation in social networks," *Scientific reports*, vol. 10, no. 1, p. 5493, 2020.

[44] D. Acemoglu, A. Ozdaglar, and J. Siderius, "A model of online misinformation," tech. rep., National Bureau of Economic Research, 2021.

[45] M. O. Jackson, S. Malladi, and D. McAdams, "Learning through the grapevine and the impact of the breadth and depth of social networks," *Proceedings of the National Academy of Sciences*, vol. 119, no. 34, p. e2205549119, 2022.

[46] M. Mostagir and J. Siderius, "Social inequality and the spread of misinformation," *Management Science*, vol. 69, no. 2, pp. 968–995, 2023.

[47] F. Chung, L. Lu, and V. Vu, "The spectra of random graphs with given expected degrees," *Internet Mathematics*, vol. 1, no. 3, pp. 257–275, 2004.

[48] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Pysical Review E*, vol. 64, 2001.

[49] F. Chung and L. Lu, "The average distances in random graphs with given expected degrees," *Proceedings of the National Academy of Sciences*, vol. 99, no. 25, pp. 15879–15882, 2002.

[50] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random structures & algorithms*, vol. 6, no. 2-3, pp. 161–180, 1995.

[51] E. A. Bender and E. R. Canfield, "The asymptotic number of labeled graphs with given degree sequences," *Journal of Combinatorial Theory, Series A*, vol. 24, no. 3, pp. 296–307, 1978.

[52] P. E. Ney and P. Ney, *Branching processes*. Courier Corporation, 2004.

7 APPENDIX

LEMMA 1. *For $\kappa > \frac{4}{3}$, there exists a boom equilibrium.*

PROOF. **Case 1.** $\kappa > 4$ Using (3), we find that $\beta^{\mathsf{T}}(\delta) \geq \beta^{\mathsf{F}}(\delta) \geq \frac{1}{4}$ 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 (4). Thus, from (5), 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. $\kappa \in (\frac{4}{3}, 4)$. In this case, we have $2\left(\frac{1}{\kappa} - \frac{1}{4}\right) \in (0,1)$. For some fixed small $\epsilon > 0$, let $\bar{\delta} \in (0,1)$ be defined as

$$\bar{\delta} = \sqrt{2\left(\frac{1}{\kappa} - \frac{1}{4}\right) + \epsilon}$$

Then, we have

$$\begin{aligned}
\beta^{\mathsf{F}}(\bar{\delta}) &= \frac{1}{4} + \frac{1}{2}\bar{\delta}^2 = \frac{1}{\kappa} + \epsilon \\
\beta^{\mathsf{T}}(\bar{\delta}) &= \frac{1}{4} + \frac{1}{2}\left(1 - (1-\bar{\delta})^2\right) \\
&> \frac{1}{2} - \frac{1}{\kappa} + \sqrt{2\left(\frac{1}{\kappa} - \frac{1}{4}\right)} := h(\kappa).
\end{aligned}$$

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

$$\kappa h(\kappa) = \frac{\kappa}{2} - 1 + \sqrt{2\kappa - \frac{\kappa^2}{2}}.$$

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

$$\begin{aligned}
\sqrt{2\kappa - \frac{\kappa^2}{2}} &> 2 - \frac{\kappa}{2} \\
2\kappa - \frac{\kappa^2}{2} &> 4 - 2\kappa + \frac{\kappa^2}{4} \\
0 &> 4 - 4\kappa + \frac{3}{4}\kappa^2.
\end{aligned}$$

The last inequality follows from the fact that since $\kappa > \frac{4}{3}$, we have $4 - 4\kappa + \frac{3}{4}\kappa^2 > 4 - 3\kappa > 0$.

Given $\kappa\beta^{\alpha}(\bar{\delta}) > 1$, we conclude that there exists positive solutions $q^{\alpha}(\bar{\delta})$ to (4). 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,

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

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. $\kappa = 4$. Since $\beta^{\mathsf{F}}(\delta) = \frac{1}{4} + \frac{1}{2}\delta^2$ and $\beta^{\mathsf{T}}(\delta) = \frac{1}{4} + \delta - \frac{1}{2}\delta^2$, 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

$$\begin{aligned}
q^{\mathsf{F}}(\epsilon) &= \frac{2}{3}\left(1 - \frac{1}{1+2\epsilon^2}\right) + O(\epsilon^3) = \frac{4}{3}\epsilon^2 + O(\epsilon^3) \\
q^{\mathsf{T}}(\epsilon) &= \frac{2}{3}\left(1 - \frac{1}{1+4\epsilon-2\epsilon^2}\right) + O(\epsilon^2) = \frac{8}{3}\epsilon + O(\epsilon^2).
\end{aligned}$$

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

$$\begin{aligned}
f(\delta) &:= \delta^3\frac{\kappa(1-\gamma)}{2} + \delta^2\frac{\kappa(\gamma-2)}{2} \\
&\quad + \delta\frac{(1+\gamma)(4-\kappa)}{4} + \frac{\gamma(\kappa-4)}{4} = 0 \\
g(\delta) &:= 4 - \kappa(1 + 2(1-(1-\delta)^2)) > 0
\end{aligned}$$

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 $\kappa\beta^{\mathsf{T}} < 1$ is equivalent to $4 - \kappa(1 + 2(1-(1-\delta)^2)) > 0$. Taking $\frac{\delta}{1-\delta}$ we get the

following derivation

$$\frac{\delta}{1-\delta} = \frac{p}{1-p}\frac{1-\kappa(\frac{1}{4}+\frac{\delta^2}{2})}{1-\kappa(\frac{1}{4}+\frac{1}{2}(2-\delta)\delta)}$$

$$\delta(1-\kappa(\frac{1}{4}+\frac{1}{2}(2-\delta)\delta)) = \gamma(1-\delta)(1-\kappa(\frac{1}{4}+\frac{\delta^2}{2}))$$

$$\delta - \frac{\kappa\delta}{4} - \kappa\delta^2 + k\delta^3 = \gamma(1-\frac{\kappa}{4}-\frac{\kappa\delta^2}{2}-\delta+\frac{\kappa\delta}{4}+\frac{\kappa\delta^3}{2})$$

$$0 = \frac{\delta^3\kappa}{2}(1-\gamma) + \kappa\delta^2(\frac{\gamma}{2}-1) + \delta(1+\gamma)(1-\frac{\kappa}{4}) + \frac{\kappa\gamma}{4} - \gamma$$

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 3. *For $\kappa < \frac{4}{3}$, there exists a bust equilibrium. For $\kappa = \frac{4}{3}$, there exists a (unique) bust equilibrium if and only if $\gamma < \frac{1}{2}$. For $\kappa \in (\frac{4}{3}, 4)$, there exists no bust equilibrium if $\gamma \geq \frac{1}{2}$ or if $\gamma < \frac{1}{2}$ and $\kappa > \frac{4+8\gamma-8\gamma^3-4\gamma^4}{1+18\gamma-24\gamma^2+10\gamma^3-3\gamma^4}$; there is a unique bust equilibrium if $\gamma < \frac{1}{2}$ and $\kappa = \frac{4+8\gamma-8\gamma^3-4\gamma^4}{1+18\gamma-24\gamma^2+10\gamma^3-3\gamma^4}$; and there are two bust equilibria if $\gamma < \frac{1}{2}$ and $\frac{4}{3} < \kappa < \frac{4+8\gamma-8\gamma^3-4\gamma^4}{1+18\gamma-24\gamma^2+10\gamma^3-3\gamma^4}$. Finally, for $\kappa \geq 4$, there exists no bust equilibrium.*

PROOF. **Case 1.** $\kappa < \frac{4}{3}$**.** In this case, we have for any $\delta \in [0,1]$,

$$\begin{aligned} g(\delta) &= 4 - \kappa(1 + 2(1-(1-\delta)^2)) \\ &> 4 - \frac{4}{3}(1 + 2(1-(1-\delta)^2)) \\ &= \frac{8}{3}(1-\delta)^2) \geq 0. \end{aligned}$$

Thus, from Lemma 2 it suffices to show the existence of $\delta \in [0,1]$ with $f(\delta) = 0$. To conclude this, observe that $f(0) = \frac{\gamma(\kappa-4)}{4} < 0$ and $f(1) = 1 - \frac{3\kappa}{4} > 0$ and that f is continuous over $[0,1]$.

Case 2. $\kappa = \frac{4}{3}$**.** In this case, we can factorize $f(\delta)$ as

$$f(\delta) = \frac{2}{3}(1-\delta)^2\left(\delta(1-\gamma)-\gamma\right).$$

Thus, f has two roots ($\delta = 1$) and $\delta = \frac{\gamma}{1-\gamma}$. Note that $g(1) = 4-3\kappa = 0$. This implies that if $\gamma \geq \frac{1}{2}$, there exists bust equilibrium. On the other hand, if $\gamma < \frac{1}{2}$, we obtain $\frac{\gamma}{1-\gamma} \in [0,1]$, and moreover,

$$g\left(\frac{\gamma}{1-\gamma}\right) = \frac{4}{3}\left(1-\frac{\gamma}{1-\gamma}\right)^2 \geq 0$$

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

Case 3. $\kappa \in (\frac{4}{3}, 4)$**.** For $g(\delta) > 0$, we must have

$$\delta < 1 - \sqrt{1 - 2\left(\frac{1}{\kappa} - \frac{1}{4}\right)} := \delta_0.$$

Now, $f(0) = \frac{\gamma(\kappa-4)}{4} < 0$ and $f(1) = 1 - \frac{3\kappa}{4} < 0$. It is straightforward to verify that $f(\delta_0) < 0$.

Moreover, we have

$$f'(\delta) = \delta^2\frac{3\kappa(1-\gamma)}{2} + \delta\kappa(\gamma-2) + \frac{(1+\gamma)(4-\kappa)}{4}.$$

We have $f'(0) > 0$ and $f'(1) = (1+\gamma)\left(1-\frac{3\kappa}{4}\right) < 0$. Thus, there exists a (unique) $\delta_+ \in (0,1)$ with $f'(\delta_+) = 0$. We have the following expression for δ_+:

$$\delta_+ = \frac{-\kappa(\gamma-2) - \sqrt{\kappa^2(\gamma-2)^2 - (3/2)\kappa(4-\kappa)(1-\gamma^2)}}{3\kappa(1-\gamma)}$$

Now, we have

$$\begin{aligned} f(\delta_+) =& \frac{72 - 144\gamma + 144\gamma^2 - 72\gamma^3 - 50\kappa + 84\gamma\kappa - 60\kappa\gamma^2 + 22\kappa\gamma^3}{108(\gamma-1)^2} \\ &+ \frac{(\frac{-((12-12\gamma^2-11\kappa+8\gamma\kappa+\gamma^2\kappa)^3}{((-1+\gamma)^4\kappa))})(\sqrt{2}-2\sqrt{2}\gamma+\sqrt{2}\gamma^2)}{108(\gamma-1)^2} \end{aligned}$$

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 $\kappa < \frac{4+8\gamma-8\gamma^3-4\gamma^4}{1+18\gamma-24\gamma^2+10\gamma^3-3\gamma^4} \leq 4$.

Summarizing, we obtain that there is no bust equilibrium if $\gamma \geq \frac{1}{2}$ or if $\gamma < \frac{1}{2}$ and $\kappa > \frac{4+8\gamma-8\gamma^3-4\gamma^4}{1+18\gamma-24\gamma^2+10\gamma^3-3\gamma^4}$; there is a unique bust equilibrium if $\gamma < \frac{1}{2}$ and $\kappa = \frac{4+8\gamma-8\gamma^3-4\gamma^4}{1+18\gamma-24\gamma^2+10\gamma^3-3\gamma^4}$; and there are two bust equilibria if $\gamma < \frac{1}{2}$ and $\frac{4}{3} < \kappa < \frac{4+8\gamma-8\gamma^3-4\gamma^4}{1+18\gamma-24\gamma^2+10\gamma^3-3\gamma^4}$.

Case 4. $\kappa \in [4, \infty)$**.** Note that since $\beta^\alpha(\delta) \geq \frac{1}{4}$ 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.

□

LEMMA 4. *For $\kappa < \frac{4}{3}$ there is a unique bust equilibrium.*

PROOF. To prove this claim we as a temporary measure express γ in terms of κ and δ. Algebra yields that $\gamma = \frac{-4\delta+\delta\kappa+4\kappa\delta^2-2\kappa\delta^3}{-4+4\delta+\kappa-\kappa\delta+2\kappa\delta^2-2\kappa^3k}$ is the only solution if (κ, δ, γ) together represent a bust equilibrium. To show that this implies there is a unique δ we need that $\frac{-4\delta+\delta\kappa+4\kappa\delta^2-2\kappa\delta^3}{-4+4\delta+\kappa-\kappa\delta+2\kappa\delta^2-2\kappa^3k}$ is invertible for $\delta \in (0,1)$. This is done by computing the derivative with respect to δ which is

$$\frac{-8\left(4\delta^3-6\delta^2+4\delta+1\right)\kappa + \left(4\delta^4+8\delta^3-12\delta^2+8\delta+1\right)\kappa^2 + 16}{(\delta-1)^2\left(2\delta^2\kappa+\kappa-4\right)^2}.$$

It can be shown that this expression is never zero $\delta \in (0,1)$ and $\kappa \in (0, \frac{4}{3})$ and is monotonically increasing. Therefore an inverse exists from which we conclude that there is a unique root. □