UNIVERSITY OF CAMBRIDGE
Faculty of Economics

JANEWAY INSTITUTE CAMBRIDGE

CAMBRIDGE WORKING PAPERS IN ECONOMICS
JANEWAY INSTITUTE WORKING PAPERS

In platforms we trust:
misinformation on social networks in the presence of social mistrust

George Charlson
University of Cambridge

Abstract

We examine the effect of social trust on a network in which agents communicate with each other and information sources, changing their opinion with some probability. Agents whose peers are more likely to spread misinformation are consequently less trusting than agents whose neighbours are more informed, and therefore change their views with less probability. When echo chambers are strong, weakening them results in there being more interaction between high and low social trust agents, increasing the spread of misinformation. When echo chambers are weak, however, weakening them further reduces the differences in social trust, decreasing the asymmetries in communication and hence the probability agents are misinformed. As a result of the non-linear relationship between the strength of echo chambers and the spread of misinformation, optimal interventions in network structure depend on why agents form links in the first place.

Reference Details

2204 Cambridge Working Papers in Economics
2202 Janeway Institute Working Paper Series

Published 12 January 2022
Revised 23 August 2022

Key Words communication, networks, network design, misinformation, platforms
JEL Codes D82, D83, D85

Websites www.econ.cam.ac.uk/cwpe
www.janeway.econ.cam.ac.uk/working-papers

In platforms we trust: misinformation on social networks in the presence of social mistrust*

George Charlson†

August 23, 2022

Abstract

We examine the effect of social trust on a network in which agents communicate with each other and information sources, changing their opinion with some probability. Agents whose peers are more likely to spread misinformation are consequently less trusting than agents whose neighbours are more informed, and therefore change their views with less probability. When echo chambers are strong, weakening them results in there being more interaction between high and low social trust agents, increasing the spread of misinformation. When echo chambers are weak, however, weakening them further reduces the differences in social trust, decreasing the asymmetries in communication and hence the probability agents are misinformed. As a result of the non-linear relationship between the strength of

*I would like to thank Matthew Elliott, Alexander Teytelboym, Sanjeev Goyal, Fuhito Kojima, Arjada Barhdi, Ben Golub, Bryony Reich, Alireza Tahbaz-Salehi, Andrea Prat and Akhil Vohra, along with the participants of the Cambridge Economics Micro Theory seminar group and the Seventh Annual Conference on Network Science and Economics for their invaluable contributions to this paper.

†Cambridge Janeway Institute, Austin Robinson Building, Sidgwick Ave, Cambridge CB3 9DD, gc556@cam.ac.uk

echo chambers and the spread of misinformation, optimal interventions in network structure depend on why agents form links in the first place.

KEYWORDS: communication, networks, network design, misinformation, platforms.
JEL classification: D82, D83, D85.

1 Introduction

It is well-documented that social media platforms, like Facebook, Reddit and Twitter, are hotbeds of misinformation on matters ranging from politicians (Allcott and Gentzkow, 2017), scientific discoveries (Naeem et al, 2020) and celebrity news (Arnold et al, 2019). One aspect of this multi-faceted problem that has been well studied is the role of echo chambers in the propagation of misinformation (see, for example, Acemoglu, Ozdaglar and Siderius, 2021). Agents who believe misinformation are more likely to be connected to others who also believe it, and so misinformation is able to propagate, at least amongst a subset of the population. Reducing the prevalence of echo chambers is therefore often seen as a key component of the battle against misinformation.

Here, I consider the other side of breaking echo chambers: people who are correctly informed are exposed to falsehoods when doing so. On its face, this might not be a concern - communication on social networks is commonly bidirectional, and, hence, at the very least, there is less polarisation when such communication occurs. However, when some agents are less trusting than their peers, bidirectional communication is not symmetric. I examine the effect this asymmetry has on the effectiveness of reducing the prevalence of echo chambers on social networks.

Social trust and its effect on communication has become of increasing interest to social scientists (Jennings and Stroud, 2021; Ognyanova, 2021; Hopp et al, 2020 and

Kwon and Barone, 2020). Here, we define social trust as the extent to which a person believes that the speech or actions of others are true or motivated by good intentions (Gambetta, 1988). Individuals with low social trust are thus less likely to be convinced by the opinions of others than those with high levels of social trust.

Recent research shows that there is a link between social trust and misinformation - specifically, followers and sharers of misinformative sources and content are more likely to have low levels of trust in both other citizens and the mainstream media (see Zimmerman and Kohring, 2020 and Hopp et al, 2020). Experimental evidence suggests that people who believe misinformation are less likely to be convinced out of their opinion even after being shown the truth (Rhodes, 2022) Furthermore, so-called countermedia information sources, who frequently purvey misinformation, foster and a support negative worldview in which most people should not be trusted as they are either ignorant or actively nefarious (Rojas, 2010 and Allcott & Gentzkow, 2017).

There is empirical evidence, then, that those who believe misinformation also less likely to trust their peers, and therefore they are less likely to be convinced out of their opinions. We construct a model in which those who have misinformed peers are, as a consequence, less likely to trust their peers. We then examine the effect that this endogenous process of social formation affects the spread of misinformation on a network.

In the model here, agents interact with each other and information sources on a social network. Users can either be informed or misinformed. Users are connected both with each other and information sources; one type ("mainstream" information sources) which espouses the informed opinion, the other, "countermedia" information sources, espouses the misinformed opinion. Users exhibit both homophily, in the sense that they prefer to connect with users of the with the same social characteristics and

worldview, and bias, in that users with a conspiratorial worldview prefer to connect to countermedia sources and high social trust users prefer to connect to mainstream sources.[1]

The agents communicate on a network that is shaped by a platform's algorithm, which suggests which users an agent should follow, taking into account users' preference for homophily and biases. The platform wishes to maximise the degree of the network, and hence chooses an algorithm that reinforces these preferences. At the optimum, this algorithm generates a stochastic block model, with types distinguished by both social characteristics and their worldview.

Agents communicate on two separate issues. On the first issue, they communicate, eventually finding out the truth. This allows them to observe the probability that their neighbours spread misinformation, which fixes their level of social trust: the more misinformative their peers, the less trustful they are. Agents whose neighbours are more likely to be misinformed thus are less likely to update their views after an interaction with their peers.

Agents then communicate about the second issue. We characterise the distribution of opinions on the issue as the number of agents tends to infinity. Communication between agents in this model is therefore endogenously asymmetric: low social trust agents, who are more likely to observe and therefore believe misinformation, are less convincible than their peers.

This feature of the model is crucial to the main results. When echo chambers are strong in this context, they protect high social trust users from being convinced by their misinformed and mistrustful peers, and hence strengthening them further decreases the

[1] Of course, there is likely to be a correlation between some social characteristics and worldview. Our analysis is agnostic as to the extent and direction of this link, as none of the results depend on any particular relationship between these two variables.

amount of misinformation on the platform. However, when echo chambers are weak, weakening them further reduces the difference in social trust levels, which also serves to decrease misinformation.

Hence, echo chambers have a non-linear effect, with stronger echo chambers being most effective in combatting misinformation in contexts where worldview is the dominant determinant of connection patterns (on networks such as Twitter), while weaker echo chambers are preferred when agents most value interacting with those who are socially similar to them (on friendship networks, like Facebook). Strong echo chambers also increase polarisation, leading to their being a trade-off for a social planner wishing to locally reduce polarisation and misinformation propagation on platforms with high levels of ideological segregation.

We then turn to the question of interventions in the network to reduce the prevalence of misinformation. If a social planner incentivises a platform to reduce misinformation by intervening in the structure of the network (a "structural intervention"), then when echo chambers are strong they optimally intervene to reduce the extent to which the most isolated individuals who have a mainstream worldview (i.e. those who are least likely to observe misinformation) interact with the conspiratorial individuals who are most likely to observe misinformation. When echo chambers are weak however, the platform prefers to increase even further the interaction between that the mainstream and conspiratorial groups who communicate the most across the ideological divide.

Reducing the extent to which users observe countermedia sources is also a way of reducing misinformation propagation. We characterise an influence measure that captures the optimal users to target with such an intervention, finding that having low social trust results in relatively large amounts of influence, as does being well-connected in the network.

Literature review

Social trust is a well-contested term within the sociological literature (see Verducci and Schröer, 2010 for an overview), but broadly can be thought of as being the belief that other citizens (as opposed to political, social or media elites) will, for one reason or another, act in a way that is, at best, to our benefit, and at worse not to our detriment (see Gambetta, 1988 and Warren, 1999 as examples). Of particular interest from our perspective is Gambetta's (1988) observation that trust "is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action": the socially mistrustful agents in our model are less likely to believe their peers than the socially trustful ones.

Our analysis fits into a growing literature on the role of social trust as a driver of polarisation and misinformation on social networks and in public life more generally. People who are socially mistrustful are more likely to vote for a populist political candidate (Hooghe and Dassonneville, 2018), spread countermedia content (Hopp et al, 2020), were less likely to socially distance during the Covid-19 pandemic (Woelfert and Kunst, 2020) and are more likely to believe conspiracy theories in general (Pierre, 2020).

Relatedly, a number of empirical papers have highlighted how exposure to opposing viewpoints may have differential effects on different users. For example, Bail et al (2018) find that exposure to a Twitter bot with an opposing viewpoint actually increased political polarisation amongst ideologically extreme right-wing subjects. These subjects are also less likely to reduce their belief in false stories that support their political position than their left-wing counterparts according to Rhodes (2022).

A number of economic theory papers have tackled the question of fake news, which

can be broadly put into two categories: Bayesian agent approaches, in which fully rational agents choose whether to share a piece of content, often with the input of a benevolent (Candogan and Drakopoulos, 2020 and Papanastasiou, 2020) or manipulative (Chen and Papanastasiou, 2021 and Keppo et al, 2019) platform, and bounded rationality or naive learning approaches (Nguyen et al., 2012, Toernberg, 2018 and Mostagir, Ozdaglar, and Siderius 2020) in which agents update their opinions heuristically on the basis of the opinion's of their neighbours.

Within the former strand of the literature, Acemoglu, Ozdagalar and Siderius (2021) is the closest to this paper. There, echo chambers generate an incentive to share misinformation, as it is less likely to be identified as such, with the platform exacerbating this issue by selectively displaying misinformation to create a filter bubble, which contrasts with our finding that echo chambers have the potential to insulate high social trust users from observing misinformation in some cases.

Our approach fits more closely with those employing boundedly rational agents and the naive learning on networks literature more broadly, which largely employs a DeGroot-based social learning approach (see Golub and Jackson, 2010). Specifically, we examine the case where there are agents who are naive learners who are influenced by users who do not update their opinion, namely information sources. Yildiz et al (2013), Vohra (2021) and Sadler (2022) all employ such agents in a naive learning framework, with the latter also considering the limits of the distribution of opinions on a stochastic block model. We examine the effect the interaction between heterogeneous levels of social trust and network structure has on communication in this framework.

Both Dandekar et al. (2013) and Anunrojwong et al (2020) utilise a naive learning framework in the context of misinformation, with the former examining the case where agents are more likely to believe evidence which supports their current position, leading

to the possibility of polarisation under homophily, while the latter analyses the effect of users who are more likely to believe evidence which supports their current position, leading to the possibility of polarisation under homophily. In both cases, communication is symmetric.

2 Communication

The model is in two parts: a network formation stage, in which users choose who they are connected with, which generates a network G, and a communicate stage, in which agents communicate on G. We consider the latter process first, before examining the network formation process in Section 3.

Communication and social trust

We consider agents interacting on a social network. Agents take two forms: "information sources" and "regular agents". Agents are linked by a graph G. Let S denote the set of information sources and R denote the set of regular agents, with $|S| = m_S$ and $|R| = m_R$. If $i, j \in R$ then if there exists an edge $ij \in G$, it is undirected, while if $i \in R$ and $j \in S$, ij is directed (there are assumed to be no links between information sources).

Suppose that there are n agents (i.e. both regular agents and information sources) and time is discrete. We consider communication between these agents which relates to an issue on which no agent knows the truth, but every agent has an opinion on some issue, I. Specifically, in period r, each agent, i, holds an opinion on issue I, $v_{ir}^I \in \{0, 1\}$, where 1 is an informed opinion (i.e. it aligns with the truth) and 0 is a misinformed opinion. In each period, a single regular agent is chosen uniformly at random. The

regular agent, i, observes a single agent, j, chosen uniformly at random from their neighbourhood (i.e. any agent j where $ij \in G$).

Information sources are either "mainstream" or "countermedia". If i is a mainstream information source, they have opinion $v_{ir}^I = 1$ for all r, while if they are countermedia then $v_{ir}^I = 0$. Let S_0 and S_1 be the set of countermedia and mainstream sources respectively. We assume that S_0, $S_1 \neq \emptyset$.

If i observes $j \in S$ (i.e. an information source) in period r then i adopts j's opinion with probability 1 in $r+1$.[2]

Meanwhile, if $j \in R$ then the probability that i adopts j's opinion depends on i's social trust. Specifically, let δ_i denote i's social trust level, which will ultimately endogenously determined in equilibrium, and $\boldsymbol{\delta}$ be the $n \times 1$ vector of social trusts. If the link ij is realised in period r and $j \in R$, then an agent i adopts j's opinion in $r+1$ with probability δ_i: that is, agents with low social trust are less likely to adopt the opinion of an agent they observe with less probability than a high social trust agent is.

The opinion forming process then forms a Markov chain. Define v^S as the vector of opinions of information sources, and $\boldsymbol{v}_t^I$ the vector of opinions of regular agents at time t on issue I. The following statement holds:

Proposition 1. *Suppose G is connected and S is non-empty. Then, the Markov chain $\boldsymbol{v}_t^I(G)$ has a unique steady-state distribution.*

This result is similar to the one found in Yildiz et al (2013), but holds for the more general case where agents are not convinced by the agent they observe with probability 1. We exploit the result in Proposition 1 throughout to examine how network structure affects the steady-state distribution of opinions.

[2]We adopt this assumption to focus our attention on interactions between regular agents. Agents may well differ in the extent to which they are convinced by information sources (and indeed, this may also depend on the type of information source), and this could be incorporated into the model easily.

An example of the communication process with exogenous social trust

To fix ideas about our analysis, we consider a stylised example in which social trust is exogenous. Suppose that there are three regular agents, A, B and C, with the first two agents having a social trust level of 0.8, the latter having a social trust level of 0.2, and two information sources, one mainstream and one countermedia. We consider two network structures, shown in Figure 1 below.

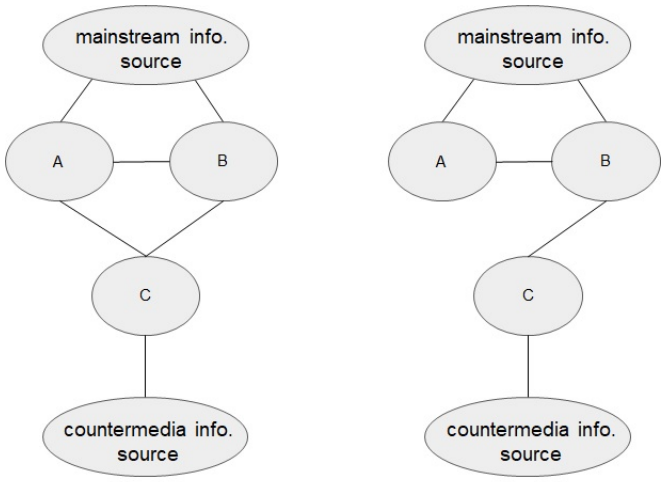

Figure 1: Two realised communication networks, with A, B and C representing regular agents.

In network structure 1, on the left-hand side of Figure 1, where A, B and C are connected, the unique steady-state distribution is such that agents A, B and C believe misinformation with probabilities 0.36, 0.36 and 0.81 respectively, and hence the probability that a random agent believes misinformation is 0.51. Compare this result to network structure 2. In that case, A, B and C believe misinformation with probabilities 0.14, 0.32 and 0.87 respectively, and so the probability a random agent believes misinformation is now 0.45. Reducing the extent to which high social trust types are

connected to low social trust types reduces the propagation of misinformation, the result of the fact that communication between agents A and C (who are connected in network structure 1 but not 2) is asymmetric, such that A is convinced by C more often than the reverse.

Compare this result to the case every agent has a social trust level of 1. Under network structure 1 and 2, agents A, B and C believe misinformation with probabilities $\frac{1}{4}$ and $\frac{1}{8}$, $\frac{1}{4}$ and $\frac{1}{4}$, and $\frac{1}{2}$ and $\frac{5}{8}$ respectively. In both cases, however, a random agent believes misinformation with probability $\frac{1}{3}$. Hence, in the case where there is no social mistrust, while the probability that different agents believe misinformation is affected by network structure (and therefore polarisation is too), average public opinion is not. These observations will be formalised by our model, along with a notion of endogenous social trust.

Social trust formation

As mentioned above, we will be considering the case where social trust levels are endogenously determined by the structure of the network. We assume that while agents are persistently biased towards the information sources they observe (believing them with probability 1), they are able to observe the likelihood that their neighbours, on average, spread misinformation.

There will be two stages of communication in the model. First, agents communicate about an issue I_1, each having some common social trust level, which we normalise such that $\delta_i = 1$ for all i. After a many-period communication process, agents will learn the truth about I_1, allowing them to change their belief about the extent to which their neighbours, on average, can be trusted, in a manner we describe below.[3] Agents then

[3]We are, then, implicitly assuming that agents are not sophisticated enough to observe which of

communicate on a new issue, I_2 in perpetuity, without ever learning the truth. We will ultimately be interested in the distribution of beliefs regarding issue I_2.[4]

Specifically, we will assume an agent i's social trust level while communicating about I_2 on a graph G will be $\delta_i = \sum_j \frac{g_{ij}}{(\sum_j g_{ij})} \mathrm{E}[\tilde{v}_j | G, \boldsymbol{\delta} = 1] \; \forall i$. That is, the probability that i changes their opinion after communicating with a regular agent, j, about issue I_2 depends on the probability that all of i's regular agent neighbours had an opinion reflecting the truth during communication about I_1.

3 Network formation

Having outlined the communication process, we now consider network formation. The network formation process takes place in two stages: an awareness stage, in which a platform partially determines the extent to which agents are aware of each other; and a connection stage, in which agents choose to connect with agents of whom they are aware. We consider these two stages in reverse order.

The connection stage

Throughout, we will assume that a regular agent, i is associated with both a **worldview** and **social characteristics**. Suppose that $\varrho_i \in \{\varrho_c, \varrho_m\}$ (conspiratorial and mainstream) denotes the two possible worldviews. Let $\theta_i \in \{\theta_1, ..., \theta_y\} = \Theta$ be a measure which captures social characteristics (e.g. location, schooling, socioeconomic status

their friends are more reliable, forming instead a general impression of the veracity of the information passed onto them by their regular agent connections.

[4] Of course, real world agents communicate on many different issues, and will likely update their social trust accordingly. The qualitative results of our analysis would not fundamentally change if there were many rounds of communication on different issues, as opposed to just two, though the analysis would become less tractable.

etc), where $|\theta_i - \theta_j| \in [0,1]$ measures how socially similar agents are.[5]

Agents prefer to connect to sources of information that cohere with their worldview and agents who hold the same worldview as them and/or share their social characteristics. Define $\hat{\theta}_{ij} = -|\theta_j - \theta_i|$ and $\hat{\varrho}_{ij} = -1$ if $\varrho_i \neq \varrho_j$ and 0 otherwise. We assume that conditional of being aware of agent j, an agent $i \in R$ receives the following utility from linking to them:

$$u_i(\hat{\theta}_{ij}, \hat{\delta}_{ij}) = \begin{cases} \alpha\hat{\varrho}_{ij} + (1-\alpha)\hat{\theta}_{ij} + \varepsilon_{ij} & \text{if } j \in R \\ (1 - f(\varrho_i)) + \varepsilon_{ij} & j \in S_0 \\ f(\varrho_i) + \varepsilon_{ij} & j \in S_1 \end{cases}$$

where $\varepsilon_{ij} \sim U[-1,1]$ is an idiosyncratic shock which captures other benefits (for example, financial) i receives from being connected with j, $\alpha \in [0,1]$ measures the relative importance differences in social trust and social characteristics have in determining the utility generated by a link and $f(\varrho_c) < f(\varrho_m) = 1 - f(\varrho_c)$. The final statement guarantees that those with a conspiratorial worldview receive greater utility, in expectation, from being connected to a countermedia source than those with a mainstream worldview. We assume that ε_{ij}s are i.i.d.

The platform and the awareness stage

Now, consider the awareness stage. Agent $i \in R$ is aware of agent j with probability $\beta_{ij} = \beta + \hat{\beta}_{ij}$ where $\beta \in [0,1)$ is the probability that i is aware of j without platform intervention and $\hat{\beta}_{ij} \in [0, 1-\beta]$ represents an increase in the awareness probability

[5]Our analysis does not preclude there being a correlation between worldview and social characteristics; we merely allow for the possibility that there are agents of both worldviews within each demographic group.

induced by the platform by e.g. suggesting to i that they follow j on a recommendation list.[6] We assume that $\beta_{ij} \perp \beta_{\iota k}$ for all i, j, k and ι.

Adjusting β_{ij} away from $\hat{\beta}_{ij}$ is costly to the platform: for example, because, making agents more aware of each other decreases the prominence of advertisements. Specifically, we assume that the platform's cost function is $C(\hat{\boldsymbol{\beta}}) = \chi \sum_i \sum_j \hat{\beta}_{ij}^2 = \chi \sum_i \sum_j (\beta_{ij} - \beta)^2$, where $\hat{\boldsymbol{\beta}}$ is a $m_R \times n$ matrix whose ijth entry is $\hat{\beta}_{ij}$ for $i \in R$ and a cost parameter $\chi \in [0, 1]$.

Define $\alpha\hat{\varrho}_{ij} + (1-\alpha)\hat{\theta}_{ij} = \gamma_{ij} < 0$. Suppose i is aware of j with probability β_{ij}. The total probability that $ij \in R$ are connected is then $w_{ij}(\hat{\boldsymbol{\beta}}, \beta) = (\beta_{ij} + \beta_{ji})\frac{1+\gamma_{ij}}{2}$ (as the realisations of β_{ij} and β_{ji} are independent) and:[7]

$$w_{ij}^S(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \begin{cases} (\beta_{ij})[1 - \frac{f(\varrho_i)}{2}] & i \in R \text{and } j \in S_0 \\ (\beta_{ij})[\frac{1}{2} + \frac{f(\varrho_i)}{2}] & i \in R \text{and } j \in S_1 \end{cases}.$$

Now we can state the platform's optimisation problem. The platform's payoff is determined by the following function:

$$\mathrm{E}[D(G)|\hat{\boldsymbol{\beta}}, \beta)] = \sum_{i \in R} \mathrm{E}[\varphi_i(G)|\hat{\boldsymbol{\beta}}, \beta] = \sum_j [w_{ij}(\hat{\boldsymbol{\beta}}, \beta) + w_{ij}^S(\hat{\boldsymbol{\beta}}, \beta)],$$

where $\varphi_i(G) = \sum_j^n g_{ij}(G)$ is i's degree in G. The platform's payoff is increasing in the expected number of edges in the network, as this is a proxy for the amount of time users spend on the platform, which in turn determines platform revenues. The platform then solves the maximisation problem: $\max_{\hat{\boldsymbol{\beta}}}[D(\hat{\boldsymbol{\beta}}, \beta) - C(\hat{\boldsymbol{\beta}})]$.

Network formation then takes place as follows. The platform chooses the matrix, $\hat{\boldsymbol{\beta}}$,

[6] Of course on real-world platforms, the innate probability that i is aware of j would itself be correlated with i and j's characteristics, as well as the number of users on the platform. This could easily be incorporated into the model, but would not materially affect the conclusions, so we maintain this assumption for simplicity.

[7] For clarity, we will use the notation w_{is}^S to refer to the probability that an agent i is connected with an information source of opinion s throughout.

determining the matrix of awareness probabilities $\boldsymbol{\beta}$. The pattern of awareness and the idiosyncratic shocks are then realised and each agent simultaneously chooses whether to connect to each of the agents they are aware of in the linkage phase of the game.

4 Optimal networks and equilibria

Solving the platform's problem

We state the following result regarding the solution to the platform's optimisation problem described above:

Proposition 2. *Holding n fixed, the unique solution to the platform's optimisation problem, $\hat{\boldsymbol{\beta}}$, generates a stochastic block model, $G(\boldsymbol{m}(n), \boldsymbol{W}(\alpha))$, with discrete type space, $\mathcal{T} = \{\Theta \times \Delta, \} = \{T_1, ...T_{2y}, S_0, S_1\}$, a $2(y+1)\times 2(y+1)$ matrix of linking probabilities, $\boldsymbol{W} = \boldsymbol{W}(\alpha)$, a number of agents, n, and a vector $\boldsymbol{m}(n) = (m_1^R(n), ..., m_{2y}^R(n), m_0^S(n), m_1^S(n))$, where $|T_i| = m_i^R(n)$ and $|S_i| = m_i^S(n)$.*

The unique solution to the platform's problem is such that if $\varrho_i = \varrho_j$ and $\theta_i = \theta_j$, then $w_{ik} = w_{jk}$ for all k. Hence, the solution to the platform's problem generates a single stochastic block model, with types determined by both an agent's worldview, ϱ_i, and the social characteristics measure, θ_i. The ijth component of the linking probability matrix, $\boldsymbol{W}(\alpha)$, is then the probability that a type i agent will observe a type j agent. Upon the realisation of the idiosyncratic shock terms and the pattern of awareness, the agents' linkage choice determine the realised network of this stochastic block model.

At this optimum, if $i, j, k \in R$ then $w_{ij} > w_{ik}$ if $|\hat{\varrho}_{ij}| > |\hat{\varrho}_{ik}|$ and $\hat{\theta}_{ij} \geq \hat{\theta}_{ik}$ or $\hat{\theta}_{ij} > \hat{\theta}_{ik}$ and $\hat{\varrho}_{ij} \geq \hat{\varrho}_{ik}$, i.e. there is homophily between groups both in terms of worldview and social characteristics. Let R_c and R_m denote the set of agents with conspiratorial and

mainstream worldviews respectively. Then if $i \in R_m$ and $j \in R_c$ and $k \in S_1$ then $w^S_{ik} = w^S_{i1} > w^S_{j1} = w^S_{jk}$ with the reverse being true when $k \in S_0$.

Homophily between groups takes two forms here: one relating to the social characteristics measure θ_i and the other relating to worldview. How relatively important these measures are for network structure depends on the parameter α. To see this, suppose $i, j \in R_c$ and $k \in R_m$ with $\hat{\theta}_{ij} > \hat{\theta}_{ik}$. Then $w_{ij}(\alpha)$ is increasing in α and $w_{ik}(\alpha)$ is decreasing in α. As α increases, the relative salience of similarities in social trust increases and the importance of social similarities decrease. The optimal network structure from the platform's point of view reflects this, and hence low social trust individuals become more (less) connected in expectation as α increases (decreases).

Public opinion in the limit

Throughout, we will consider the expected opinions of agents on a stochastic block model generated by the platform's choice of the awareness matrix, $\hat{\boldsymbol{\beta}}$, prior to the realisation of both the pattern of awareness and the idiosyncratic shock terms.

Formally, we take a sequence of stochastic block models $\{G(\boldsymbol{m}(n), \boldsymbol{W}(\alpha))\}_{n\in\mathbb{N}}$ in order to analyse the distribution of opinions as $n \to \infty$. Doing so allows us to characterise the distribution of opinions held by agents in steady state, and, given the large number of users of social networks, provide a good approximation of the distribution opinions that would be held by agents on social block models constructed in the manner described above.

For a fixed n, recall that $m^S_i(n)$ denotes the number of information sources of opinion i, and $m^R_s(n)$ denotes the number of regular agents of type s. We write:

$$\lim_{n\to\infty}\frac{m_i^S(n)}{n} = q_i^S > 0, \;\; \lim_{n\to\infty}\frac{m_s^R(n)}{n} = q_s^R > 0,$$

as the limiting fractions of information sources with opinion i and regular agents of type s respectively. Throughout, we will maintai n the assumption that $q_0^S = q_1^S = q$, which, given the optimal expressions for w_{i0}^S and w_{i1}^S, implies that each type observes the same proportion of information sources, differing only in the relative amount of misinformative sources they observe.[8] We let $\bar{q}^S = w_{i0}^S q + w_{i1}^S q$.

Abusing notation, let $\tilde{v}_i(\boldsymbol{m}(n), \alpha, \boldsymbol{\delta})$ be a random variable denoting the opinions of an agent i and distributed according to the steady state of the a stochastic block model $G(\boldsymbol{m}(n), \boldsymbol{W}(\alpha))$. Define:

$$z_s(n, \alpha, \boldsymbol{\delta}) := \frac{\sum_{i\in s}\tilde{v}_i(\boldsymbol{m}(n), \alpha, \boldsymbol{\delta})}{m_s^R(n)},$$

as the average opinion of type s agents for the model $G(\boldsymbol{m}(n), \boldsymbol{W}(\alpha))$ regarding issue I_2. Let $\mathcal{T}_R$ be the set of all types of regular agents, with $\mathcal{T}_m$ and $\mathcal{T}_c$ being the set of types of regular agents with mainstream and conspiratorial worldviews respectively. Public opinion regarding I_2 can then be defined as follows:

$$\bar{z}(n, \alpha, \boldsymbol{\delta}) = \frac{1}{m_R}[\sum_{s\in\mathcal{T}_R} m_s^R(n) z_s(n, \alpha, \boldsymbol{\delta})].$$

We be considering the limit distributions of both public opinion, $\lim_{n\to\infty}\bar{z}(n, \alpha, \boldsymbol{\delta}) = \bar{z}(\alpha, \boldsymbol{\delta})$, and the steady state opinions of type js, $\lim_{n\to\infty} z_j(n, \alpha, \boldsymbol{\delta}) = z_j(\alpha, \boldsymbol{\delta})$.

We will also be interested in the average social trust level for type s agents, which we denote, somewhat abusing notation, as follows:

$$\delta_s := \frac{\sum_{i\in s}\delta_i(n, \alpha)}{m_s^R(n)},$$

[8] This assumption simplifies the analysis, but the model could easily incorporate agents who prefer to observe fewer or more information sources than others.

Our results going forward will be stated for the $2y \times 1$ vector of average social trust levels when social trust levels are consistent, which we denote $\boldsymbol{\delta}^*(n, \alpha)$.

We note that $\boldsymbol{\delta}^*(n, \alpha)$ is not random with respect to the communication process (as it considers the expectation of this process) but is random prior to the realisation of the graph, G, whereas the opinion vector, $z_s(n, \alpha, \boldsymbol{\delta})$ is random both because the graph generating process is stochastic and agents' opinions change due to communication.

Timing of the model, a summary

We summarise the timing of the model as follows:

1. The platform chooses the awareness matrix, $\boldsymbol{\beta}$;
2. Agents make their connection choices, generating a graph G;
3. Communication takes place on the graph G about issue I_1, which yields a social trust vector, $\boldsymbol{\delta}(n, \alpha)$;
4. Communication takes place on the graph G about issue I_2,with agents communicating with social trusts $\boldsymbol{\delta}(n, \alpha)$.

5 Opinion formation and social trust

We state an expression for the limit vector of equilibrium opinions for a fixed set of social trust levels. We then state a result concerning the existence and uniqueness of the social consistent equilibrium of the model.

The opinion vector

Define $\hat{\boldsymbol{W}}(\alpha)$ as the $2y \times 2y$ trust-adjusted linking probability matrix whose jkth entry is $q_k \delta_j w_{jk}(\alpha)$, where $w_{jk}(\alpha)$ is the probability that a type j regular agent is connected with a type k regular agent. Define the normalised expected degree of a type j whose agents whose average social trust level is δ_j as follows:

$$d_j = \sum_{k \in \mathcal{T}_R} \delta_j q_j w_{jk}(\alpha) + q_1^S w_{j1}^S(\alpha) + q_0^S w_{j0}^S(\alpha).$$

Let $\boldsymbol{\Lambda}(\alpha)$ denote a diagonal matrix whose jth component is d_j and $M(\alpha, \delta) := \boldsymbol{\Lambda}(\alpha, \boldsymbol{\delta}) - \hat{\boldsymbol{W}}(\alpha, \boldsymbol{\delta})$. We define a $2y \times 1$ column, z^S whose jth entry is $q_1^S w_{j1}^S$. The following Theorem holds:

Theorem 1. *Suppose that for all $\forall s \in \mathcal{T}$, $\delta_i = \delta_j$ if $i, j \in s$. Then, for any $2y \times 1$ social trust vector, $\boldsymbol{\delta}$, the limit vector of the expected opinions of regular agents converges almost surely to the expression:*

$$\boldsymbol{z}(\alpha, \boldsymbol{\delta}) = \boldsymbol{M}^{-1}(\alpha, \boldsymbol{\delta}) \boldsymbol{z^S}.$$

The jth component of the vector $\boldsymbol{z^S}$, $q_1^S w_{j1}^S$, measures the direct effect information sources have on the belief probabilities of an agent of type j. The matrix $\boldsymbol{M}^{-1}(\alpha)$ then measures the amplification of information sources by regular agents on social media: the higher the expected number of links between one agent type, j, and another, k, the larger the effect that the information sources that a given agent of type j is connected to have on an agent of type k, and vice versa.

The result in Theorem 1 is consistent with that found in Stadler (2022), with the key difference being that the matrix $\boldsymbol{M}^{-1}(\alpha, \boldsymbol{\delta})$ is dependent on the levels of social

trust of the different members of population. As we will see, this feature of the opinion vector $\boldsymbol{z}(\alpha, \boldsymbol{\delta})$ results in the peer-to-peer interaction structure having an important role in determining public opinion and polarisation.

We have stated Theorem 1 assuming that $\forall s \in \mathcal{T}$, $\delta_i = \delta_j$ if $i, j \in s$. Let $\tilde{\boldsymbol{W}}(\alpha)$ be a $2y \times 2y$ matrix whose ith entry is $\frac{w_{ij}q_j}{\sum_{j\in\mathcal{T}_R} w_{ij}}$ for $i, j \in \mathcal{T}_R$. Having stated the vector of the expected opinions of regular agents, we validate this assumption as follows:

Proposition 3. *The $2y \times 1$ social trust vector, $\boldsymbol{\delta}^*(n, \alpha) \rightarrow_{a.s.} \tilde{\boldsymbol{W}}(\alpha)\boldsymbol{z}(\alpha, \mathbf{1})$. Furthermore, for all $s \in \mathcal{T}$, the following result holds:*

$$lim_{n\to\infty} max_{i\in R} |E[v_i(n)|G(n)] - z_s(\alpha, \mathbf{1})| = 0 \text{ for } i \in s.$$

As communication when $\boldsymbol{\delta} = \mathbf{1}$ is a special case of the situation analysed in Theorem 1, it follows that the consistent value of δ_i for agent i of type s converges almost surely to an average value which can be expressed $\delta_s^* = \sum_{t\in\mathcal{T}_R}(\frac{w_{st}q_t}{\sum_{t\in\mathcal{T}_R} w_{st}} z_t(\alpha, \mathbf{1}))$. As $n \rightarrow \infty$, then, the maximum deviation from this expression for the vector $\boldsymbol{\delta}^*(n, \alpha)$ tends to zero. From now on, then, we will analyse the limit of the communication process letting $\boldsymbol{\delta}^*(\alpha) = \tilde{\boldsymbol{W}}(\alpha)\boldsymbol{z}(\alpha, \mathbf{1})$.

The effect of social trust on opinions

To fix ideas about the effect of social trust, we consider the effect of a hypothetical exogenous change in social trust:

Proposition 4. *Suppose $i \in \mathcal{T}_c$ and $j \in \mathcal{T}_m$. Then, $\frac{d\bar{z}(\alpha)}{d\delta_i} > 0$ and $\frac{d\bar{z}(\alpha)}{d\delta_j} < 0$.*

Proposition 4 highlights the effect social trust has on the belief vector. If the social trust of the agents with a conspiratorial worldview increases, then the countermedia

sources they tend to follow become relatively less influential over their opinions, and they listen to their peers more. This leads to an increase in the probability that they believe the truth, which also increases that probability for other types of agent as well.

The opposite holds for agents with a mainstream worldview: in this case, increasing δ_j results in agents who are more likely to have a mainstream view than average being more receptive to the opinions of their peers. This ultimately results in agents of that type putting more weight on the opinions of those who share misinformation more, which leads to a reduction in the probability that the average agent is informed.

Endogenous social trust

The equilibrium social trust vector, $\boldsymbol{\delta}^*(\alpha)$, is tied down by the nature of the communication process about issue I_1. We first state a result linking social trust in equilibrium and social characteristics:

Proposition 5. *Suppose $i \in T_s \in \mathcal{T}_c$ and $j \in T_k \in \mathcal{T}_m$ be such that $\theta_i = \theta_j$. Then if $\alpha > 0$, $\delta_i^*(\alpha) > \delta_j^*(\alpha)$.*

Groups with the same social characteristics can be ordered in terms of their social trust levels according to their worldview, with those of a conspiratorial worldview having lower social trust. This is the result of the fact that an agent i with the conspiratorial worldview are more likely to be connected with others with that worldview, who spread misinformation during the first communication phase with higher probability. As a result, i is less trusting in the second. This provides an account as to why those who are more likely to believe misinformation in the first place are also more likely to have lower social trust (see, e.g. Hopp, 2020).

We will be interested in the effect of the ideological preference parameter has on

public opinion. It is therefore worth stating the following result:

Proposition 6. *Suppose* $i \in T_s \in \mathcal{T}_c$, $j \in T_k \in \mathcal{T}_m$ *and* $\theta_i = \theta_j$. *Then* $\delta_i^*(\alpha) - \delta_j^*(\alpha)$ *is increasing in* α.

An increase in α increases the extent to which agents interact with those who share their worldview. As a result, agents with a conspiratorial worldview are less trusting in equilibrium, because they are more likely to encounter misinformation than when echo chambers are less pronounced.

6 Network structure and echo chambers

We examine the effect that different network types have on misinformation. Specifically, we compare a network where there is more homophily with regards to social trust with a network where social similarity is more important in determining who connects with whom. In doing so, we also analyse the effect of echo chambers on the spread of misinformation across the network.

The effect of echo chambers

Our analysis in Section 5 alludes to a potential trade-off in the effect of echo chambers on public opinion. On the one hand, echo chambers reduce the extent to which relatively low social trust and misinformation believing agents communicate with their relatively high social trust peers, but they also reinforce the differences in social trust between the agents with the two different worldviews, with misinformed agents becoming more distrustful as echo chambers become stronger. These two channels through which network structure affects public opinion, which we term the "asymmetric com-

munication" and "change in social trust" channels respectively, affect average public opinion in opposite directions.

We resolve this trade-off in the following way:

Theorem 2. *There exists a $\bar{\alpha}$ such that, if $\alpha < \bar{\alpha}$ then public opinion, $\bar{z}(\alpha, \delta^*)$, is decreasing in α and if $\alpha > \bar{\alpha}$ then $\bar{z}(\alpha, \delta^*)$ is increasing in α.*

Theorem 2 shows that echo chambers have a non-linear effect in our model. Weak echo chambers result in the social trust parameters of agents with different worldviews being relatively similar. In which case, communication between the two groups is relatively symmetric, and so the asymmetric communication channel is weak. Meanwhile, when echo chambers are weak, an increase in α results in a relatively large decrease (increase) in the social trust levels of conspiratorial (mainstream) agents. Hence, increasing α has the net effect of increasing the probability misinformation is believed.

In the case when echo chambers are strong, the reverse is true. In this case, the social trust levels of conspiratorial agents are significantly lower than those of their mainstream peers, and hence the asymmetric communication channel is strong, dominating the change in social trust channel. As a result, strengthening already strong echo chambers results in a decrease in misinformation.

Network types and misinformation

We observe that many large-scale online social networks can be categorised into two broad classes. The first type, which we call friendship networks, are such that users tend to be connected with others they have met, to at least some extent, offline, and are thus associated with them by friendship, work or education. Examples of friendship networks include Facebook and Snapchat. Interest-based networks, on the other hand,

involve agents interacting with people with similar worldviews or interests to them. The most prominent example of an interest-based network is Twitter, but forum networks and Reddit work in a similar way.

In terms of our model, interest-based networks would be generated by a relatively high value of α, the relative weight differences in social trust have on the probability that an agent connects with another conditional on being aware of them. The platform's algorithm would then be more likely to show agents with low social trust other agents with this worldview. On the other hand, friendship networks would be generated by a relatively low value of α, and as a result i's social demographic measure, θ_i, has more of an impact on the strength of i's linking probabilities than in the interest-based network. An example of these two networks is displayed below.

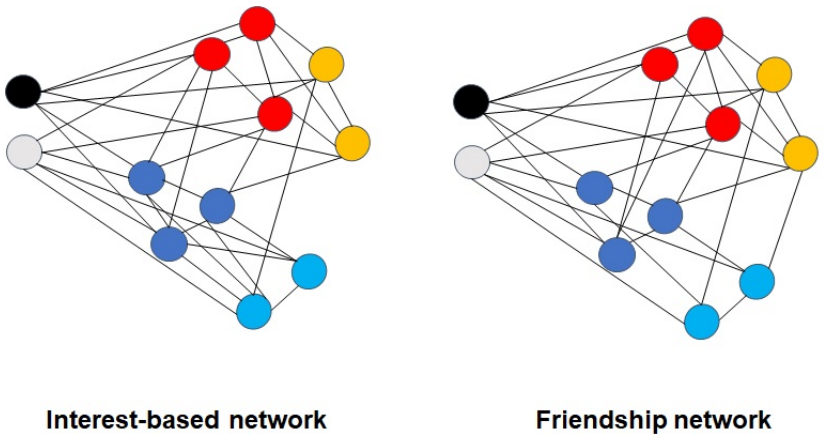

Figure 2: The black node represents a countermedia source, the grey node represents a mainstream media source, and the other colours denote different types of agent. Specifically, the light and dark blue agents are of the mainstream worldview, with the other nodes being conspiratorial, the light blue and orange types are maximally socially similar, as are the dark blue and red types.

To provide an interpretation of the result in Theorem 2, we define a friendship stochastic block model as one in which $\alpha < \bar{\alpha}$, while if $\alpha \geq \bar{\alpha}$ then the model is a interest-based

stochastic block model. Then, the following result holds:

Corollary 1. *If the network is generated by a interest-based stochastic block model then* $\bar{z}(\alpha)$ *is increasing in* α*, while if it is generated by a friendship stochastic block model then* $\bar{z}(\alpha)$ *is decreasing in* α*.*

Theorem 2 and Corollary 1 highlight the role of the platform's algorithm in the spread of misinformation. By recommending countermedia sources to low social trust types, the platform's incentive to maximise engagement straightforwardly increases the probability that misinformation is believed. The role of platforms' algorithms in propagating sources that are misinformative, which this mechanism within the model captures, is well established.

However, the role of the platform goes beyond recommending countermedia sources. The platform's incentive to maximise engagement means that, if they are operating in an environment in which agents have a strong desire for homophily with regards to worldview, then the optimal algorithm strengthens and reinforces this desire, reducing the amount of misinformation spread on the platform. If, on the other hand, agents have a stronger desire for homophily with regards to social characteristics, then the platform's algorithm weakens echo chambers further, which also serves to reduce the amount of misinformation on average. Hence, the platform's incentives to increase engagement with regards to peer-to-peer connections may somewhat counteract the negative effect information sources suggestions have on misinformation.

Polarisation

While our main focus here is on the average belief in misinformation, it is also worth commenting on polarisation. We define polarisation as follows:

$$P(\alpha) := \lim_{n\to\infty} \sum_{j\in\mathcal{T}_R} q_j |z_j(n,\alpha) - \hat{z}(n,\alpha)|.$$

Polarisation then measures the expected deviation of the expected belief of a type j agent from the average belief of a generic agent as $n \to \infty$.

Theorem 3. *Polarisation $P(\alpha)$ is increasing in α. The level of polarisation in the friendship network is then less than in the interest-based network, $P(\alpha_F) < P(\alpha_I)$.*

Echo chambers in this setting increase the probability that a random agent is informed in steady state, but they increase polarisation. This follows simply from the fact that the stronger the echo chamber is in equilibrium, the less interaction there is between agents who observe countermedia sources with different probabilities. If agents who are likely to have different opinions to one another do not interact as much, low social trust agents are more likely to believe misinformation, and high social trust agents are more likely to be informed.

Taken together, Theorems 2 and 3 imply that, if the platform was incentivised to intervene in the structure of the network, there can be a trade-off between polarisation and the probability that misinformation is believed by a random agent when echo chambers are relatively pronounced. Often polarisation is discussed as being a fundamental part of the spread of misinformation. Here, as more pronounced echo chambers protect high social trust individuals from being as exposed to misinformation, the problem of polarisation and the spread of misinformation are two different issues, and solutions to combat them may be contradictory when echo chambers are strong to begin with. We consider interventions in the structure of the network below.

7 Policy implications

We consider policy interventions in which a social planner is able to intervene in the platform's awareness algorithm to reduce the extent which misinformation propagates through the network.

To keep the analysis at the level of the group, rather than assessing interventions which affect individual agents, we will assume that any intervention requires the platform to maintain the stochastic block structure which is, at any rate, optimal (as per Proposition 2). That is, we let $\bar{\beta}_{st}$ be the equilibrium awareness level of all agents i of type s and j of type t, and will examine the derivative $\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \bar{\beta}_{st}}$.

Benchmark: unconstrained optimal interventions

We first examine a benchmark case where a social planner has full control of the platform's algorithm and wishes to reduce the spread of misinformation on the network. We also assume $\beta = 0$, which implies that the planner has full control over the awareness matrix $\boldsymbol{\beta}$. That is, the planner solves the following maximisation problem: $\max_{\hat{\boldsymbol{\beta}}}[\bar{z}(\alpha, \boldsymbol{\beta}; \hat{\boldsymbol{\beta}})]$, the solution to which we characterise as follows:

Proposition 7. *Suppose $\alpha < 1$. Any $\boldsymbol{\beta}^*$ which solves the misinformation minimisation problem is such that $\beta^*_{ij} = 0$ if $i \in \mathcal{T}_c$ and $j \in \mathcal{T}_m$.*

The probability that a random agent believes misinformation when echo chambers are maximally strong (e.g. those generated by $\alpha = 1$) is equal to that probability when $\alpha = 0$, i.e. where there is no homophily in ideology. However, the planner is not able to control α, and hence, they cannot compel agents to connect across the ideological divide as frequently as they do within their own worldview. Of these two cases, the

planner can only implement maximally strong echo chambers, reducing the probability that conspiratorial and mainstream agents interact to zero.

Structural interventions

Clearly, the above benchmark is unlikely to be implementable. We now consider targeted algorithmic interventions which are the most effective in reducing misinformation at the margin. We call these interventions "structural".

We consider a marginal change in an individual awareness probability β_{ij}. Network structure, the number of agents who are of a given type and the existing strength of echo chambers will impact on the overall effect such an intervention will have. We will consider two types of marginal structural intervention: interventions in the peer-to-peer network and in information source recommendations.

Interventions in the peer-to-peer network

There are two types of peer-to-peer marginal interventions to examine. One involves marginally decreasing some β_{st} with the aim of reducing the extent to which asymmetric communication occurs, whereas the other involves marginally increasing some β_{ij} in order to reduce differences in levels of social trust between the two types. To consider these two cases formally, we define $B = -\boldsymbol{\beta}$, so that we can more easily consider marginal decreases in $\boldsymbol{\beta}$. The following statement holds:

Theorem 4. *Suppose $i, j \in \mathcal{T}_m$ and $s, t \in \mathcal{T}_c$, $z_i > z_j$ and $z_s < z_t$, with $\delta_i > \delta_j$ and $\delta_s < \delta_t$. The following two statements hold:*

(1) $\frac{\partial \bar{z}(\alpha;\boldsymbol{B}^*)}{\partial \bar{B}_{sk}} > \frac{\partial \bar{z}(\alpha;\boldsymbol{B}^*)}{\partial \bar{B}_{tk}}$ *for* $k \in \mathcal{T}_m$ *and* $\frac{\partial \bar{z}(\alpha;\boldsymbol{B}^*)}{\partial \bar{B}_{ik}} > \frac{\partial \bar{z}(\alpha;\boldsymbol{B}^*)}{\partial \bar{B}_{jk}}$ *for* $k \in \mathcal{T}_c$*;*

(2) $\frac{\partial \bar{z}(\alpha;\boldsymbol{B}^*)}{\partial \bar{\beta}_{sk}} < \frac{\partial \bar{z}(\alpha;\boldsymbol{B}^*)}{\partial \bar{\beta}_{tk}}$ *for* $k \in \mathcal{T}_m$ *and* $\frac{\partial \bar{z}(\alpha;\boldsymbol{B}^*)}{\partial \bar{\beta}_{ik}} > \frac{\partial \bar{z}(\alpha;\boldsymbol{B}^*)}{\partial \bar{\beta}_{jk}}$ *for* $k \in \mathcal{T}_c$.

The first claim in Theorem 4 states that if the platform were to reduce the probability of a connection between a group of conspiratorial agents and a group of mainstream agents, then choosing a conspiratorial group with low social trust and a high probability of belief in misinformation is more effective at reducing misinformation than a group with higher social trust and a lower probability of believing misinformation. Such isolated agents spread misinformation more effectively than those agents who are more likely to observe mainstream agents.

Similarly, reducing the extent to which an isolated group of mainstream agents - that is, a group with relatively high levels of social trust and low probability of believing misinformation - is connected with conspiratorial agents also disproportionately increases the probability a random agent believes misinformation.

The same applies in reverse if the platform were to increase the likelihood of cross-ideological linkages. In this case, a social planner wishing to reduce misinformation would prefer to further strengthen the links between less isolated mainstream and conspiratorial agents, as doing so maximises the extent to which social trust levels of the two groups become more similar, and minimises the costs of asymmetric communication.

Theorem 4, then, holds because of the same forces as those that explain the result in Theorem 2: the effects of echo chambers in the model are non-linear, such that strengthening relatively strong links and weakening relatively weak ones across the ideological divide both have the effect of reducing the spread misinformation because the former results in a weakening of the asymmetry in social trust levels while the latter reduces the extent to which asymmetric communication occurs in the first place.

Furthermore, Theorem 2 directly implies that there is some $\exists\hat{\alpha}$ such that if $\alpha < (\geq)\hat{\alpha}$ then the most effective intervention which reduces the probability of links between a

group of conspiratorial and a group of mainstream agents is less (more) effective than the most effective intervention which increases that probability.

Intervening in information source recommendations

We now consider a marginal change in the mix of information sources users observe. Clearly, to reduce misinformation it would be necessary to reduce the proportion of countermedia sources observed by a given agent. However, which user(s) to target with such an intervention is not trivial and will depend on network structure, as we show below.

We note that we can write $M(\alpha, \boldsymbol{\delta^*}) = A(\boldsymbol{\delta^*})V(\alpha)Q = Y(\alpha, \boldsymbol{\delta^*})Q$ where $A(\boldsymbol{\delta^*})$ is a diagonal matrix with sth component δ_s^*, Q is a diagonal matrix with s component q_s and $V(\alpha)$ is a symmetric matrix whose ijth entry is equal to $\frac{1}{q_j \delta_i} m_{ij}$. Define the influence of a type $s \in \mathcal{T}_R$ agent as $\phi_j := \sum_j y_{ij}(\alpha)$, where $y_{ij}(\alpha)$ is the ijth entry of $Y^{-1}(\alpha)$. We also define $\tilde{\beta}_i(S_1, S_0) := \bar{\beta}_{iS_1} - \bar{\beta}_{iS_0}$, where $\bar{\beta}_{iS_j}$ is the probability that a type i is aware of an information source of type S_j. We make the following observation:

Proposition 8. *Suppose that $k_t \in S_t$ with $\phi_i > \phi_j$. Then $|\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta^*})}{\partial \tilde{\beta}_i(S_1,S_0)}| > |\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta^*})}{\partial \tilde{\beta}_i(S_1,S_0)_:}|$. Furthermore, for any $j \in \mathcal{T}_m$, $\exists i \in \mathcal{T}_c$ such that $\phi_i > \phi_j$.*

Increasing the proportion of mainstream media sources observed by an agent decreases the probability that every agent believes misinformation. The influence of i, ϕ_i, measures the effect the information sources observed by an average i type agent have on the opinions of other regular type agents, weighted by the total proportion of agents who are of that type. Hence, if $\phi_i > \phi_j$, then the marginal effect of increasing the relative probability type is observe mainstream sources on public opinion is greater than marginally increasing the same probability for type js.

The second result in Proposition 8 indicates that for any mainstream type, there exists conspiratorial type which is more influential than them. Specifically, if i and j have the same social characteristics (i.e. $\theta_i = \theta_j$) then if $i \in \mathcal{T}_c$ and $j \in \mathcal{T}_m$, then i is more influential than j. Both i and j occupy an equivalent position in the network when $q_i = q \; \forall i$. As a result, what determines the relative influence between the two types is the extent to which they change their opinions when confronted with agents with a differing viewpoint, and, as low social trust types are less convincible than high social trust types, it follows that i is more influential than j.

8 Concluding remarks

We have analysed an opinion formation model in which some agents have lower social trust than others, with an agent's social trust being determined by the extent to which their neighbours are to spread misinformation. The prescence of social trust results in asymmetric communication between agents. Echo chambers have the effect of increasing differences in social trust levels but decrease the extent to which misinformed and informed agents interact. Hence, weakening echo chambers can increase or decrease the spread of misinformation depending on the characteristics of the network to start with.

The model set-up leans heavily towards the current discourse in Western countries like the United States and the UK where mainstream sources are relatively trustworthy and countermedia sources are often misinformative.[9] Mainstream sources may not necessarily be trustworthy in other countries, where mainstream media sources may echo government propaganda. For example, there is evidence that Facebook was used to spread of pro-government and anti-Muslim misinformation during the 2017 Myanmar

[9]This, of course, does not hold all the time even in Western countries, where, for example, mainstream sources can be, for example, captured by corporate interests.

genocide (see Whitten-Woodring et al. 2020). In this case, countermedia sources would counteract rather than propagate misinformation. We have not actively explored this possibility here, but note that our model provides a general framework to analyse such questions.

One aspect of the effect of trust on misinformation propagation which is ignored here is trust levels in information sources. Misinformation reduces trust in the media in general, and there is evidence (see, e.g. Hopp et al) that those who spread misinformation are less trusting of mainstream media sources. We do not model the extent to which trust levels in media affect misinformation spread, but it could incorporate these factors, which would serve to strengthen the mechanisms observed here.

We have focused largely on the implications of social trust on the spread of misinformation. However, the model here provides general insights as to how differences in social trust interact with network structure in determining opinion formation in network models. In models in which agents communicate symmetrically, network structure shapes variables like speed of convergence (e.g. Golub and Jackson, 2010) or polarisation (e.g. Sadler, 2022) but it plays less of a role in determining the average belief of agents on the network.

Here, network structure, and specifically links between high and low social trust agents have a crucial part to play in determining the extent to which misinformation is believed. This opens up questions regarding the effect network structure has on opinion formation when agents are susceptible to social biases, such as confirmation bias, stubborn opinions and status quo bias.

Appendix

Preliminaries

Many of the results in the main text require evaluating various derivatives of the matrix $M^{-1}(\alpha, \boldsymbol{\delta}^*)$. Throughout, we let r_{ij} represent the ijth component of $M^{-1}(\alpha, \boldsymbol{\delta}^*)$.

It is also worthwhile establishing some facts about the matrices $M(\alpha, \boldsymbol{\delta}^*)$ and $M^{-1}(\alpha, \boldsymbol{\delta}^*)$. By assumption, $q_i^S = q$ for $i = 0, 1$ and $w_{i0}^S q + w_{i1}^S q = \bar{q}^S$ for all i. As $\delta_i \neq \delta_j$ and each type by definition have different values for θ_k, then, the matrix $M(\alpha, \boldsymbol{\delta}^*)$ is of full rank (i.e. has rank of $2y$) and is such that $\sum_j m_{ij} = \bar{q}^S$ for all j. This in turn implies that each row of the matrix $M^{-1}(\alpha, \boldsymbol{\delta}^*)$ sums to $\frac{1}{\bar{q}^S}$. To see this, let v be a $2y \times 1$ vector filled with 1s and hence $M(\alpha, \boldsymbol{\delta}^*)v = v$. Thus, $M^{-1}(\alpha, \boldsymbol{\delta}^*)v = M^{-1}(\alpha, \boldsymbol{\delta}^*)M(\alpha, \boldsymbol{\delta}^*)v = v$. We also let $\frac{1}{1-\bar{q}^S}\boldsymbol{z}^S = \tilde{\boldsymbol{z}}^S$ and $\mathcal{T}_c$ and $\mathcal{T}_m$ denote the set of types which contain agents with conspiratorial and mainstream worldviews respectively.

We note that we can write $M(\alpha, \boldsymbol{\delta}^*) = A(\boldsymbol{\delta}^*)V(\alpha)Q$, where $A(\boldsymbol{\delta}^*)$ is a diagonal matrix with ith component δ_i and $V(\alpha)$ is some symmetric matrix. It follows that $M^{-1}(\alpha, \boldsymbol{\delta}^*) = Q^{-1}V^{-1}(\alpha)A^{-1}(\boldsymbol{\delta}^*)$, where $V^{-1}(\alpha)$ is also a symmetric matrix. We also let $Y(\alpha, \boldsymbol{\delta}^*) = AV(\alpha, \boldsymbol{\delta}^*)$.

We establish a further result which will be useful for proving the statements in the text:

Lemma 1. *$M(\alpha, \boldsymbol{\delta}^*)$ is an M-matrix and hence $M^{-1}(\alpha, \boldsymbol{\delta}^*)$ is non-negative.*

Proof. By definition, $M(\alpha, \boldsymbol{\delta}^*)$ is a Z-matrix (i.e. a matrix where $m_{ij} \leq 0$ for all $i \neq j$). $M(\alpha, \boldsymbol{\delta}^*)$ is also strictly diagonally dominant by construction. It follows from the Gershgorin circle theorem that the real parts of M's eigenvalues are positive. $M(\alpha, \boldsymbol{\delta}^*)$

is therefore a $M-$matrix. $M(\alpha, \boldsymbol{\delta}^*)$ is also non-singular by construction. The inverse of a non-singular M-matrix is non-negative. □

Proof of Proposition 1

Suppose first that there is a connected subgraph of regular agents, that is, a connected graph such that for all $i \in R$, $\exists j \in R$ such that $ij \in G$. As the graph is connected and $S_0, S_1 \neq \emptyset$, then it follows that there exists at least one link $ij \in G$, where $i \in R$ and $j \in S_j$ for $j = 0, 1$. The fact that the subgraph of regular agents is connected and each regular agent i changes their opinion with some positive probability if the link $ik \in G$ is realised in the communication game, it follows that $\boldsymbol{v}_t^I$ is irreducible, and thus has a unique steady state distribution. If the subgraph of regular agents is not connected, then the same argument holds for each component of the regular agent subgraph.

Proof of Proposition 2

As per the expressions for w_{ij} and w_{ij}^S in the text, and noting that $|\gamma_{ij}| < 1$, $D(\hat{\boldsymbol{\beta}}, \beta)$ is a (weakly) increasing and linear function in $\hat{\beta}_{ij}$ for all i, j. The cost function is $C(\hat{\boldsymbol{\beta}})$ is convex and separable in each $\hat{\beta}_{ij}$. It follows that either $\hat{\beta}_{ij}^* = 1 - \beta$ (i.e. the solution is not interior) or the solution to the first-order condition $\frac{\partial D(\hat{\boldsymbol{\beta}}, \beta)}{\partial \beta_{ij}} - \frac{\partial C(\hat{\boldsymbol{\beta}})}{\partial \beta_{ij}} = 0$. If this first-order condition is satisfied, then:

$$\hat{\beta}_{ij}^* = \begin{cases} \frac{1+\gamma_{ij}}{4\chi} \text{ if } & i, j \in R \\ \frac{2-f(\varrho_i)}{4\chi} & i, \in R \text{and } j \in S_0 \\ \frac{1+f(\varrho_i)}{4\chi} & i, \in R \text{and } j \in S_1 \end{cases}.$$

Let T be such that if $i, k \in T$, then $\theta_i = \theta_k$ and $\varrho_i = \varrho_k$. Then for any $j \in G$ and

$i, k \in T$, whether the solution is interior or not, $\beta^*_{ij} = \beta^*_{kj}$ and $\beta^*_{ji} = \beta^*_{jk}$ (and, in fact, it is simple to see that $\beta^*_{ij} = \beta^*_{ji}$). As this applies to any $\theta_i \in \Theta$ and $\varrho_i \in \Delta$, it follows that for each $T \in \mathcal{T}$, it follows that if $i, k \in T$ then, for the solution to the platform's problem, $w_{ij} = w_{kj}$ for all $j \in G$, including when $j \in S_0$ or S_1. As $|\Delta| = 2$ and $|\Theta| = y$, it follows that $|\Theta \times \Delta| + 2 = 2(y + 1)$, as required.

Proof of Theorem 1 and Proposition 3

To simplify notation we write $D = \sum_i \varphi_i(G)$ as the total degree of G. In the communication game, an edge is realised with probability $\frac{1}{D}$. Let $\tilde{G}(n)$ denote a Markov matrix whose ijth entry can be written:

$$\tilde{g}_{ij} = \begin{cases} \frac{\delta_i}{D} & \text{if } i, j \in R \text{ and } j \in G_i \\ 1 - \frac{\varphi_i(G)(\mu_i^S + \delta_i \mu_i^R)}{D} & \text{if } i, j, \in R \text{and } i = j \\ \frac{1}{D} & i, \in R \text{and } j \in S \end{cases} .$$

where $\mu_i^S = \frac{\sum_{j \in S} g_{ij}}{d_i}$, $\mu_i^R = \frac{\sum_{j \in R} g_{ij}}{d_i}$. Let $\tilde{G}^R(\alpha, n)$ denote the submatrix of interactions between regular agents and $\tilde{G}^S(n)$ denote the submatrix of interactions between stubborn agents and regular agents. Let $\boldsymbol{x}^R = \mathrm{E}[v_i | G(n, \boldsymbol{W}(\alpha))]$. At steady state, it must be that:

$$\boldsymbol{x}^R = \tilde{G}^S(n)\boldsymbol{v}^S + \tilde{G}^R(n, \alpha)\boldsymbol{v}^R$$

where $\boldsymbol{v}^S$ is the vector of information source opinions and $\boldsymbol{v}^R$ is the vector of regular agent opinions. It follows that:

$$\boldsymbol{x}^R = (I - \tilde{G}^R(n, \alpha)^{-1}\tilde{G}^S(n)\boldsymbol{v}^S.$$

We define the $m_R \times m_R$ matrix $\bar{G}(\alpha, n, \boldsymbol{\delta})$ as a matrix whose ijth entry is written:

$$\bar{g}_{ij} = \begin{cases} \frac{\delta_i w_{ij}(\alpha)}{\mathrm{E}[D(n,\alpha)]} & \text{if } i,j \in R \text{ and } j \in G_i \\ 1 - (\frac{\sum_{j \in R} \delta_i w_{ij}(\alpha) + \sum_{k \in S} w_{ij}^S(\alpha)}{\mathrm{E}[D(n,\alpha)]}) & \text{if } i,j, \in R \text{and } i = j \\ \frac{w_{ij}(\alpha)}{\mathrm{E}[D(n,\alpha)]} & i, \in R \text{and } j \in S \end{cases} .$$

Let $\vartheta_i(n,\alpha) = \sum_{j \in R} \delta_i w_{ij}(\alpha) m_j^R(n) + \sum_{j \in S} w_{ij}^S(\alpha) m_j^S(n)$. Define $H(n, \alpha, \boldsymbol{\delta})$ as a $2t \times 2t$ matrix with entry jk :

$$h_{jk} = \begin{cases} \frac{\delta_j m_k^R(n) w_{jk}(\alpha)}{\mathrm{E}[D(n,\alpha)]} & \text{if } j \neq k \\ 1 - \frac{\vartheta_j(n,\alpha)}{\mathrm{E}[D(n,\alpha)]} + \frac{\delta_j (m_j^R(n) - 1) w_{jj}(\alpha)}{\mathrm{E}[D(n,\alpha)]} & \text{if } j = k \end{cases} .$$

$H(n,\alpha)$ is then a representative type matrix, with its jkth entry representing the expected interaction between an agent of type j and a random type k agent. Define:

$$\bar{\boldsymbol{x}}^R(\sigma) = (I - \bar{G}^R(n,\alpha))^{-1} \bar{G}^S(n) \boldsymbol{v}^S; \text{ and}$$

$$\boldsymbol{z}(n, \alpha, \boldsymbol{\delta}) = (I - H(n, \alpha, \boldsymbol{\delta}))^{-1} \hat{\boldsymbol{z}}^S(n, \alpha)$$

where $\boldsymbol{z}^S(n, \alpha, \boldsymbol{\delta})$ is a column vector ith entry is $\frac{w_{ij}(\alpha) m_1^S(n)}{\mathrm{E}[D(n,\alpha)]}$ and $\bar{G}^R(n,\alpha)$ and $\bar{G}^S(n)$ denote the submatrices of interactions between regular agents and other regular and information source respectively corresponding to the stochastic matrix $\tilde{G}(n,\alpha)$. It is clear that if i is of type k, then the kth entry of $\bar{\boldsymbol{x}}(n,\alpha)$ is equal to the jth entry of $\boldsymbol{z}(n, \alpha, \boldsymbol{\delta})$.

We now need to show that $|\boldsymbol{x}(n,\alpha) - \bar{\boldsymbol{x}}(n,\alpha)| \to^{a.s.} 0$ and $\boldsymbol{z}(n, \alpha, \boldsymbol{\delta}) \to \boldsymbol{z}(\alpha, \boldsymbol{\delta})$. For the first statement, we let A_n be a random square matrix where $a_{ii} = 0$ and $a_{ij} = D(n,\alpha)\tilde{g}_{ij}$. A_n is then the sum of an upper and a lower triangular matrix, both

of which have independent entries for all $\delta_i, \delta_j \in [0,1]$. The following statement holds, as shown in the proof of Theorem 1 in Sadler (2022):

Lemma. *(Sadler, 2022) There exist constants $c, C > 0$ such that: $Pr(|\boldsymbol{x}(n,\alpha) - \bar{\boldsymbol{x}}(n,\alpha)| > \frac{k|\boldsymbol{v}^S|}{n^{3/2}}) \leq Ce^{-ck^2}$ for all sufficiently large k. It follows from the Borel-Cantelli lemma that $|\boldsymbol{x}^R(n,\alpha) - \bar{\boldsymbol{x}}^R(n,\alpha)| \to^{a.s.} 0$ for all sufficiently large k.*

To see that $\boldsymbol{z}(n,\alpha,\boldsymbol{\delta}) \to \boldsymbol{z}(\alpha,\boldsymbol{\delta})$, note the following:

$$(I - H(n,\alpha))^{-1}\hat{\boldsymbol{z}}^S(n,\alpha,) = (\frac{\mathrm{E}[D(n,\alpha)]}{n}I - \frac{\mathrm{E}[D(n,\alpha)]}{n}H(n))^{-1}\frac{\mathrm{E}[D(n,\alpha)]}{n}\hat{\boldsymbol{z}}^S(n,\alpha),$$

which in turn equals $(\tilde{\Lambda}(n,\alpha) - \hat{\boldsymbol{W}}(n))^{-1}\boldsymbol{z}^S(n)$,where $\tilde{\Lambda}(n,\alpha)$ is a diagonal matrix with ith component $\frac{\mathrm{E}[\vartheta_i(n,\alpha)] - w_{ii}}{n}$. The limit of this expression as $n \to \infty$ is then the statement in Theorem 1.

To prove Proposition 3 we note that the case where $\delta_i = 1 \ \forall i \in \mathcal{T}$ is a special case of the above analysis. We note that the above analysis implies that $\delta_i(n,\alpha) \to_{a.s.} \sum_t \frac{w_{st}}{(\sum_t w_{st})} z_t = \delta_s^*$. Almost sure convergence automatically implies the second result in Proposition 3.

Proof of Proposition 4

Public opinion, $\bar{z}(\alpha,\boldsymbol{\delta^*})$, can be written: $M^{-1}(\alpha,\boldsymbol{\delta^*})\tilde{\boldsymbol{z}}^S\boldsymbol{q}^T = \bar{z}(\alpha,\boldsymbol{\delta})$ where $\boldsymbol{q}^T$ is a $1 \times 2t$ vector with ith entry q_i. The vectors $\boldsymbol{q}^T$ and $\boldsymbol{z}^S$ are independent of δ_i, but $M^{-1}(\alpha,\boldsymbol{\delta^*})$ is a function of it. We analyse $\frac{\partial M^{-1}(\alpha,\boldsymbol{\delta^*})}{\partial \delta_i} = -M^{-1}\frac{\partial M(\alpha,\boldsymbol{\delta^*})}{\partial \delta_i}M^{-1}$.

Note that each row of M^{-1} sums to $\frac{1}{\bar{q}^S}$. Hence, each row sum of $\frac{\partial M^{-1}(\alpha,\boldsymbol{\delta^*})}{\partial \delta_i}$ equals 0. Consider $\frac{\partial M(\alpha,\boldsymbol{\delta^*})}{\partial \delta_i}$. This matrix has a row of 0s for all rows corresponding to a type, $k \neq i$, but $-\frac{\partial m_{ii}(\alpha,\boldsymbol{\delta^*})}{\partial \delta_i} > 0$ and $-\frac{\partial m_{ij}(\alpha,\boldsymbol{\delta^*})}{\partial \delta_i} < 0$ for all j. Note that $\boldsymbol{z}_i^S \leq \boldsymbol{z}_k^S$

for i (as $i \in \mathcal{T}_c$) and all k with the inequality strict when $k \in \mathcal{T}_m$. Hence, if $i \in \mathcal{T}_c$, $M^{-1}(\alpha, \boldsymbol{\delta^*})\tilde{\boldsymbol{z}}^S \boldsymbol{q}^T > 0$

Precisely the same argument in reverse applies to $\frac{\partial M^{-1}(\alpha, \boldsymbol{\delta^*})}{\partial \delta_j}$, for $j \in \mathcal{T}_m$, and so $\frac{\partial M^{-1}(\alpha, \delta^*)}{\partial \delta_j} \boldsymbol{z}^S \boldsymbol{q}^T < 0$.

Proof of Proposition 5 and 6

We first prove the following Lemma:

Lemma 2. *Suppose $i \in \mathcal{T}_c$ and $j \in \mathcal{T}_m$ and $\theta_i = \theta_j$. Then $z_i(\alpha,\mathbf{1}) > z_i(\alpha, \mathbf{1})$ when $\alpha > 0$.*

To see this, we note first that that each row of the matrix $M^{-1}(\alpha)$ sums to $\frac{1}{\bar{q}^S}$. Given the equilibrium stochastic block model, it must be the case that when $\theta_i = \theta_j$, $w_{ik} > w_{jk}$ if $k \in \mathcal{T}_c$. This implies that $r_{ik}(\alpha, \mathbf{1}) > r_{ij}(\alpha, \mathbf{1})$ (the ikth and ijth entries in the matrix $M^{-1}(\alpha)$ respectively) when $\alpha > 0$. As $z_i^S = z_k^S > z_j^S$, the result holds. Again, $\theta_i = \theta_j$, $w_{ik} > w_{jk}$ if $k \in \mathcal{T}_c$. As $\boldsymbol{\delta}^*(\alpha) = \tilde{\boldsymbol{W}}(\alpha)\boldsymbol{Q}\boldsymbol{z}(\alpha, \mathbf{1})$, the result in Proposition 5 follows immediately. Furthermore, the fact that $\sum_j^{2y} r_{ij} = 0$ for all i implies that $r_{ik}(\alpha, \mathbf{1}) - r_{ij}(\alpha, \mathbf{1})$ is increasing in α, which implies Proposition 6.

Proof of Theorem 2 and Proposition 7

To assess the function $\bar{z}(\alpha, \boldsymbol{\delta^*})$ and its derivative, we write $\boldsymbol{\delta^*}(\alpha)$ where it is useful to do. We will show that $\bar{z}(0, \boldsymbol{\delta^*}(0)) = \bar{z}(1, \boldsymbol{\delta^*}(1))$. To see this, first note that $z_i^S = z_j^S$ for $i, j \in \mathcal{T}_k$ $k = c, m$: when $\alpha = 1$, $z_i(1, \boldsymbol{\delta^*}(1)) = \tilde{z}_i^S$ for any arbitrary $\boldsymbol{\delta^*}$ and so $\bar{z}(1, \boldsymbol{\delta^*}(1)) = \sum_{i=1}^{2y} q_i \tilde{z}_i^S$.

Recall that $QM^{-1}\tilde{\boldsymbol{z}}^S \mathbf{1}^{\boldsymbol{T}} = \bar{z}(\alpha, \boldsymbol{\delta}^*)$. We know that $M^{-1}(\alpha, \boldsymbol{\delta}^*) = Q^{-1}V^{-1}(\alpha)A^{-1}(\boldsymbol{\delta}^*(\alpha))$,

and so:

$$QM^{-1}(\alpha, \boldsymbol{\delta}^*)\tilde{\boldsymbol{z}}^S 1^T = V^{-1}(\alpha)A^{-1}(\boldsymbol{\delta}^*)\tilde{\boldsymbol{z}}^S \mathbf{1}^{\boldsymbol{T}}.$$

When $\alpha = 0$, if $\theta_i = \theta_j$ then $\delta_i = \delta_j$ and $w_{ik} = w_{jk}\ \forall k$. It follows that we can define a new type space, $\hat{\mathcal{T}} = \{\hat{T}_1, ..., \hat{T}_t\}$ where $\hat{T}_i = \{T_j \cup T_k\}$ for some $j, k \in \mathcal{T}$ with $\theta_j = \theta_j$. As the product $V^{-1}(\alpha)A^{-1}(\boldsymbol{\delta}^*)$ is independent of $\boldsymbol{Q}$, we define $\hat{V}^{-1}(\alpha, \boldsymbol{\delta}^*)$ and $\hat{A}^{-1}(\alpha, \boldsymbol{\delta}^*)$ as $y \times y$ matrices whose ijth entry are equal to the ijth entry of $V^{-1}(\alpha, \boldsymbol{\delta}^*)$ and $A^{-1}(\alpha, \boldsymbol{\delta}^*)$ respectively and $\hat{\boldsymbol{z}}^S$ as a $y \times 1$ vector whose ith entry is equal to $\tilde{z}_j^S + \tilde{z}_k^S$ for $j \in \mathcal{T}_c$ and $k \in \mathcal{T}_m$. Hence, $\bar{z}(0, \boldsymbol{\delta}^*(0)) = \bar{z}(1, \boldsymbol{\delta^*}(1))$.

To prove Theorem 2, we consider the effect of a marginal change in α on $\bar{z}(\alpha)$, recall that $Y(\alpha) = A(\boldsymbol{\delta}^*)V(\alpha, \boldsymbol{\delta}^*)$. We will analyse $\frac{\partial Y^{-1}(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} = -Y^{-1}(\alpha, \boldsymbol{\delta}^*)\frac{\partial Y(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha}Y^{-1}(\alpha, \boldsymbol{\delta}^*)$.

We note that:

$$\frac{\partial \bar{z}(\alpha, \boldsymbol{\delta^*})}{\partial \alpha} = -Y^{-1}(A(\boldsymbol{\delta}^*)\frac{\partial V(\alpha)}{\partial \alpha} + \frac{\partial A(\boldsymbol{\delta}^*)}{\partial \alpha}V(\alpha))Y^{-1}\tilde{\boldsymbol{z}}^S \mathbf{1}^{\boldsymbol{T}}. \tag{1}$$

We analyse the derivatives $\frac{\partial V(\alpha)}{\partial \alpha}$ and $\frac{\partial A(\boldsymbol{\delta}^*)}{\partial \alpha}$ in turn.

Let v_{ij} denote the ijth entry of the matrix $\frac{\partial V(\alpha)}{\partial \alpha}$. Given the solution to the platform's maximisation problem, the matrix $\frac{\partial V(\alpha)}{\partial \alpha}$ is such that if $i, j \in \mathcal{T}_c$ then $v_{ij} < 0$ and if $k \in \mathcal{T}_m$, $v_{ik} > 0$, with $|v_{ij}| = v_{ik}$ if $\theta_j = \theta_k$. It follows that $v_{ii} = 0$, $\sum_j v_{ij} = 0$ for all i and, as $V(\alpha)$ is symmetric, so too is $\frac{\partial V(\alpha)}{\partial \alpha}$.

Consider a type pair, i, j where $i \in \mathcal{T}_c$ and $j \in \mathcal{T}_m$. By Proposition 4, $\delta_i^*(\alpha) \leq \delta_j^*(\alpha)$ with the equality strict if $\alpha > 0$. As $z_i^S \leq z_j^S$ for $i \in \mathcal{T}_c$ and all k with the inequality strict when $j \in \mathcal{T}_m$, the symmetry of $\frac{\partial V(\alpha)}{\partial \alpha}$ implies that ith and jth entry of the vector $-Y^{-1}A(\boldsymbol{\delta}^*)\frac{\partial V(\alpha)}{\partial \alpha}Y^{-1}\boldsymbol{z}^S \mathbf{1}^T$ is weakly greater than 0, increasing in $\delta_j^* - \delta_i^*$ and equal to zero when $\delta_j^* - \delta_i^* = 0$. As these observations hold for all i, j pairs, $-Y^{-1}A(\boldsymbol{\delta}^*)\frac{\partial V(\alpha)}{\partial \alpha}Y^{-1}\boldsymbol{z}^S \geq 0$.

We note that $\boldsymbol{\delta}^*(\alpha) = \tilde{\boldsymbol{W}}(\alpha)\boldsymbol{Q}\boldsymbol{z}(\alpha, \mathbf{1})$, and $v_{ij}(\alpha) = -w_{ij}(\alpha)$ for $i \neq j$, with $v_{ii}(\alpha)$ also linearly decreasing in w_{ij}. Again, $i \in \mathcal{T}_c$ and all k with the inequality strict when $j \in \mathcal{T}_m$: it follows that $\frac{\partial z_i(\alpha, \mathbf{1})}{\partial \alpha} < 0$ if $i \in \mathcal{T}_c$ and increasing otherwise. Hence $\frac{\partial \delta_i}{\partial \alpha} < 0$ if $i \in \mathcal{T}_c$ and increasing otherwise. By Proposition 4, this implies that $-Y^{-1}\frac{\partial A(\boldsymbol{\delta}^*(\alpha))}{\partial \alpha}V(\alpha)Y^{-1}\boldsymbol{z}^S\mathbf{1}^{\boldsymbol{T}} < 0$.

We let $G(\alpha) = -Y^{-1}A(\boldsymbol{\delta}^*(\alpha))\frac{\partial V(\alpha)}{\partial \alpha}Y^{-1}\boldsymbol{z}^S\mathbf{1}^{\boldsymbol{T}}$ and $F(\alpha) = -Y^{-1}\frac{\partial A(\boldsymbol{\delta}^*(\alpha))}{\partial \alpha}V(\alpha)Y^{-1}\boldsymbol{z}^S\mathbf{1}^{\boldsymbol{T}}$, and so $\frac{\partial \bar{z}(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} = F(\alpha) + G(\alpha)$. Let $X(\alpha, \boldsymbol{\delta}^*) = A(\boldsymbol{\delta}^*)\frac{\partial V(\alpha)}{\partial \alpha}$ and $X'(\alpha, \boldsymbol{\delta}^*) = V(\alpha)\frac{\partial A(\boldsymbol{\delta}^*)}{\partial \alpha}$. Then:

$$\frac{\partial G(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} = [2Y^{-1}X(\boldsymbol{\delta}^*, \alpha)Y^{-1}X(\boldsymbol{\delta}^*, \alpha)Y^{-1} - Y^{-1}(\frac{\partial A}{\partial \alpha}\frac{\partial V}{\partial \alpha})Y^{-1}]\boldsymbol{z}^S\mathbf{1}^T.$$

The first term is weakly positive as per the above analysis. The ith entry of $\frac{\partial A}{\partial \alpha}$ is negative if $i \in \mathcal{T}_c$ and positive otherwise. The ijth entry of $\frac{\partial V}{\partial \alpha}$ is negative if $i, j \in \mathcal{T}_k$ for $k = c, m$ and $j \neq i$ and positive otherwise. Hence, $Y^{-1}(\frac{\partial A}{\partial \alpha}\frac{\partial V}{\partial \alpha})Y^{-1}\boldsymbol{z}^S\mathbf{1}^T < 0$, implying the second term (noting the negative sign) is (weakly) positive. Hence, $\frac{\partial G(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} \geq 0$ for all $\boldsymbol{\delta}^*$. Now consider:

$$\frac{\partial F(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} = [2Y^{-1}X'(\boldsymbol{\delta}^*, \alpha)Y^{-1}X'(\boldsymbol{\delta}^*, \alpha)Y^{-1} - Y^{-1}(\frac{\partial A}{\partial \alpha}\frac{\partial V}{\partial \alpha} + V(\alpha)\frac{\partial^2 A}{\partial \alpha^2})Y^{-1}]\boldsymbol{z}^S\mathbf{1}^T.$$

The first term is always strictly positive, with the second also being positive, as per the above. To analyse $\frac{\partial^2 A}{\partial \alpha^2}$, we examine the vector $-M^{-1}(\alpha, \mathbf{1})\frac{\partial M(\alpha, \mathbf{1})}{\partial \alpha}M^{-1}(\alpha, \mathbf{1})\boldsymbol{z}^S$ and its derivative:

$$2[M^{-1}(\alpha, \mathbf{1})\frac{\partial M(\alpha, \mathbf{1})}{\partial \alpha}M^{-1}(\alpha, \mathbf{1})\frac{\partial M(\alpha, \mathbf{1})}{\partial \alpha}M^{-1}(\alpha, \mathbf{1})]\boldsymbol{z}^S \tag{2}$$

The matrix $\frac{\partial M(\alpha, \mathbf{1})}{\partial \alpha}$ is a zero row-sum matrix whose ijth entry is positive if $i, j \in \mathcal{T}_k$ for

$k = c, m$ and negative otherwise. Hence the term in the square brackets of (2) is such that its ijth entry is positive if $i, j \in \mathcal{T}_k$ and negative otherwise. This then implies that $\frac{\partial^2 \delta_i(\alpha)}{\partial \alpha^2} < 0$ for $i \in \mathcal{T}_c$ and $\frac{\partial^2 \delta_j(\alpha)}{\partial \alpha^2} > 0$. It then follows that $-Y^{-1}(V(\alpha)\frac{\partial^2 A}{\partial \alpha^2})Y^{-1}\boldsymbol{z}^S\mathbf{1}^T$ is positive, which implies that $\frac{\partial F(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} > 0$

Thus, $\bar{z}(\alpha, \boldsymbol{\delta}^*)$ is convex in $\alpha \in [0, 1]$. As per the initial analysis that $\bar{z}(0, \boldsymbol{\delta}^*(0)) = \bar{z}(1, \boldsymbol{\delta}^*(1))$,we know $\exists \bar{\alpha}$ such that $G(\bar{\alpha}) = -F(\bar{\alpha})$, and the two functions do not across again for $\alpha \in [0, 1]$, completing the proof of Theorem 2.

For proof of Proposition 7, notice that even when $\boldsymbol{\beta} = 1$, if $\alpha < 1$, then the equilibrium stochastic block model is such that $w_{ij}(\alpha) > w_{ik}(\alpha)$ if $\theta_j = \theta_k$ and $i, j \in \mathcal{T}_s$ for $s = c, m$ and $k \in \mathcal{T}_l$ for $l \neq s$: the statement in Lemma 2 holds for all $\boldsymbol{\beta}$. Hence, the highest possible level of public opinion, $\bar{z}(1, \boldsymbol{\delta}^*(1))$, is only achieved when the matrix $\boldsymbol{\beta}$ satisfies the condition in stated in the Proposition.

Proof of Theorem 3

Recall that $\frac{\partial z(\alpha)}{\partial \alpha} = -M^{-1}(\alpha, \boldsymbol{\delta}^*)\frac{\partial M(\alpha)}{\partial \alpha}M^{-1}(\alpha, \boldsymbol{\delta}^*)z^S$, and, as $A(\boldsymbol{\delta}^*(\alpha))V(\alpha)Q$, the analysis in the proof of Theorem 2 implies that: (1) $\frac{\partial z_j(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} > 0$ and $\frac{\partial z_k(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} < 0$ for all $j \in \mathcal{T}_c$ and $k \in \mathcal{T}_m$; and (2) $\sum_{k \in \mathcal{T}_m} q_k \frac{\partial z_k(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} > \frac{\partial \bar{z}(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} > \sum_{j \in \mathcal{T}_c} q_j \frac{\partial z_j(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha}$. It then follows that $\sum_{k \in \mathcal{T}_m} q_k (\frac{\partial z_k(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha} - \frac{\partial \bar{z}(\alpha, \boldsymbol{\delta}^*)}{\partial \alpha}) > 0$ and $\sum_{j \in \mathcal{T}_N} q_j (\frac{\partial \bar{z}(\alpha)}{\partial \alpha} - \frac{\partial z_j(\alpha)}{\partial \alpha}) = \frac{\partial \sum_{j \in \mathcal{T}_N} q_j |z_j - \bar{z}(\alpha)|}{\partial \alpha} > 0$, which implies the result.

Proof of Theorem 4

Modifying equation (1), we write:

$$\frac{\partial \bar{z}(\alpha, \boldsymbol{\delta}^*)}{\partial \bar{B}_{ij}} = -Y^{-1}(A(\boldsymbol{\delta}^*)\frac{\partial V(\alpha)}{\partial \bar{B}_{ij}} + \frac{\partial A(\boldsymbol{\delta}^*)}{\partial \bar{B}_{ij}}V(\alpha))Y^{-1}\tilde{\boldsymbol{z}}^S\mathbf{1}^{\boldsymbol{T}}. \tag{3}$$

To analyse the first term of (3), we can rewrite the first term in this expression in terms of w_{ij} and $M_{\boldsymbol{\delta}^*}(\alpha)$ - a matrix whose ijth entry is equal to the ijth entry of $M(\alpha, \boldsymbol{\delta}^*)$ for all $\boldsymbol{\delta}$ - as follows:

$$-Y^{-1}A(\boldsymbol{\delta}^*)\frac{\partial V(\alpha)}{\partial \bar{B}_{ij}}Y^{-1}\tilde{\boldsymbol{z}}^S\mathbf{1}^{\boldsymbol{T}} = -M^{-1}_{\boldsymbol{\delta}^*}(\alpha)\frac{\partial M_{\boldsymbol{\delta}^*}(\alpha)}{\partial w_{ik}}M^{-1}_{\boldsymbol{\delta}^*}(\alpha)\tilde{\boldsymbol{z}}^S\mathbf{1}^{\boldsymbol{T}}.$$

Let $l_{uv}(sk)$ represents the uvth component of $L_{sk}(\alpha) = -M^{-1}_{\boldsymbol{\delta}^*}(\alpha)\frac{\partial M_{\boldsymbol{\delta}^*}(\alpha)}{\partial w_{sk}}M^{-1}_{\boldsymbol{\delta}^*}(\alpha)$, then:

$$l_{uv}(sk) = (-q_k\delta_s r_{us} + q_s\delta_k r_{uk})r_{sv} - (q_s\delta_k r_{us} - q_k\delta_s r_{uk})r_{kv}$$

where r_{ij} is the ijth entry of $M^{-1}_{\boldsymbol{\delta}^*}(\alpha)$. Note that, as stated in the preliminary section above, $\sum_v r_{uv} = \frac{1}{\bar{q}^S}$ for all $u \in \mathcal{T}$. Therefore, if $z_s < z_t$ and $\delta_s < \delta_t$ then $\sum_{u\in\mathcal{T}_c} r_{su} > \sum_{u\in\mathcal{T}_c} r_{tu}$ and $\sum_{u\in\mathcal{T}_m} r_{su} < \sum_{u\in\mathcal{T}_m} r_{su}$. Noting that $\sum_v l_{uv}(sk) = 0\ \forall u$, this then implies that $|\sum_{v\in\mathcal{T}_c} l_{uv}(sk)| > |\sum_{v\in\mathcal{T}_c} l_{uv}(tk)|$ for all u (and so $|\sum_{v\in\mathcal{T}_m} l_{uv}(sk)| > |\sum_{v\in\mathcal{T}_m} l_{uv}(tk)|$ as well).

Now, consider the case where $z_i > z_j$, $\delta_i > \delta_j$ and $i, j \in \mathcal{T}_m$. Then, $\sum_{k\in\mathcal{T}_c} r_{ik}(\alpha) < \sum_{k\in\mathcal{T}_c} r_{jk}(\alpha)$ and $\sum_{k\in\mathcal{T}_m} r_{ik}(\alpha) > \sum_{k\in\mathcal{T}_m} r_{jk}(\alpha)$. As per the argument above, this implies that $|\sum_{k\in\mathcal{T}_c} l_{tk}(is)| > |\sum_{k\in\mathcal{T}_c} l_{tk}(js)|$ for all k.

We now need to consider the second term in (3). As per the proof of Theorem 2 above, this term is negative, while its derivative is positive. If $\delta_s < \delta_t$ this implies that:

$$-Y^{-1}\frac{\partial A(\boldsymbol{\delta}^*)}{\partial \bar{B}_{sk}}V(\alpha)Y^{-1}\tilde{\boldsymbol{z}}^S\mathbf{1}^{\boldsymbol{T}} > -Y^{-1}\frac{\partial A(\boldsymbol{\delta}^*)}{\partial \bar{B}_{tk}}V(\alpha)Y^{-1}\tilde{\boldsymbol{z}}^S\mathbf{1}^{\boldsymbol{T}}\ \forall k.$$

Hence, the first statement in the Theorem is proven. Note also that $\frac{\partial A(\boldsymbol{\delta}^*)}{\partial \bar{B}_{ij}} = -\frac{\partial A(\boldsymbol{\delta}^*)}{\partial \bar{\beta}_{ij}}$ and $\frac{\partial V(\boldsymbol{\delta}^*)}{\partial \bar{B}_{ij}} = -\frac{\partial V(\boldsymbol{\delta}^*)}{\partial \bar{\beta}_{ij}}$, which, given the analysis above, immediately implies the second statement.

Proof of Proposition 8

Let $\tilde{w}^0_{s1} = w^S_{s1} - w^S_{s0}$. As $M^{-1}\tilde{\boldsymbol{z}}^S\boldsymbol{q}^T = \bar{z}(\alpha)$, then M^{-1} is independent of $\tilde{w}^0_{s1}$ (though it is not independent of w^S_{s1} or w^S_{s0}, $\frac{\partial M^{-1}(\alpha)}{\partial w^S_{s1}} = \frac{\partial M^{-1}(\alpha)}{\partial w^S_{s1}}$, and so $\frac{\partial M^{-1}(\alpha)}{\partial \tilde{w}^0_{s1}} = 0$). Then, $M^{-1}(\alpha)\frac{\partial \boldsymbol{z}^S}{\partial \tilde{w}^0_{s1}}\boldsymbol{q}^T = \frac{\partial \bar{z}(\alpha)}{\partial \tilde{w}^0_{s1}}$. It then follows that if $\phi_i > \phi_j$, then $\frac{\partial \bar{z}(\alpha)}{\partial \hat{\beta}_i(S_1,S_0)} > \frac{\partial \bar{z}(\alpha)}{\partial \hat{\beta}_j(S_1,S_0)}$.

For the second statement, $Y(\alpha, \boldsymbol{\delta}^*) = A(\boldsymbol{\delta}^*)V(\alpha)$. $V(\alpha)$ and thus $V^{-1}(\alpha)$ are symmetric and by Proposition 4 for $\theta_i = \theta_j$ and $i \in \mathcal{T}_c$ and $j \in \mathcal{T}_m$ $\delta_i < \delta_j$. Then it must be that $\sum_k y_{ik}(\alpha) > \sum_k y_{jk}(\alpha)$, completing the proof.

References

[1] Acemoglu, D., Ozdaglar, A., and Siderius, J. Misinformation: Strategic sharing, homophily, and endogenous echo chambers. Working Paper 28884, National Bureau of Economic Research, June 2021.

[2] Allcott, H., and Gentzkow, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*, 2 (May 2017), 211–36.

[3] Anunrojwong, J., Iyer, K., and Manshadi, V. Information design for congested social services: Optimal need-based persuasion, 2020.

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

[5] Bin Naeem, S., Bhatti, R., and Khan, A. An exploration of how fake news is taking over social media and putting public health at risk. *Health Information Libraries Journal 38* (07 2020).

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

[7] Chen, L., and Papanastasiou, Y. Seeding the herd: Pricing and welfare effects of social learning manipulation. *Management Science 67*, 11 (2021), 6734–6750.

[8] Dandekar, P., Goel, A., and Lee, D. T. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences 110*, 15 (2013), 5791–5796.

[9] Gambetta, D. Can we trust trust? In *Trust: Making and Breaking Cooperative Relations*. Blackwell, 1988.

[10] Golub, B., and Jackson, M. O. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics 2*, 1 (February 2010), 112–49.

[11] Harambam, J., and Aupers, S. Contesting epistemic authority: Conspiracy theories on the boundaries of science. *Public understanding of science (Bristol, England) 24*, 4 (May 2015), 466â480.

[12] Hooghe, M., and Dassonneville, R. A spiral of distrust: A panel study on the relation between political distrust and protest voting in belgium. *Government and Opposition 53*, 1 (2018), 104â130.

[13] Hopp, T., Ferrucci, P., and Vargo, C. J. Why Do People Share Ideologically Extreme, False, and Misleading Content on Social Media? A Self-Report and Trace DataâBased Analysis of Countermedia Content Dissemination on Facebook and Twitter. *Human Communication Research 46*, 4 (05 2020), 357–384.

[14] Jennings, J., and Stroud, N. J. Asymmetric adjustment: Partisanship and correcting misinformation on facebook. *New Media Society* (2021).

[15] Keppo, J., Kim, M. J., and Zhang, X. Learning manipulation through information dissemination. *Operations Research* (2021).

[16] Kwon, M., and Barone, M. J. A World of Mistrust: Fake News, Mistrust Mind-Sets, and Product Evaluations. *Journal of the Association for Consumer Research 5*, 2 (2020), 206–219.

[17] Mostagir, M., Ozdaglar, A. E., and Siderius, J. When is society susceptible to manipulation? Tech. rep., SSRN Scholarly Paper ID 3474643, 2020.

[18] Nansen, B., O'Donnell, D., Arnold, M., Kohn, T., and Gibbs, M. Death by twitter: Understanding false death announcements on social media and the performance of platform cultural capital. *First Monday 24*, 12 (Dec. 2019).

[19] Nguyen, N. P., Yan, G., Thai, M. T., and Eidenbenz, S. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference* (New York, NY, USA, 2012), WebSci '12, Association for Computing Machinery, pp. 213–222.

[20] Ognyanova, K. Network approaches to misinformation, evaluation and correction. In *Networks, Knowledge Brokers, and the Public Policy-making Process*. Palgrave Macmillan, 2021.

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

[22] Pierre, J. M. Mistrust and misinformation: A two-component, socio-epistemic model of belief in conspiracy theories. *Journal of Social and Political Psychology 8*, 2 (Oct. 2020), 617–641.

[23] Rhodes, S. C. Filter bubbles, echo chambers, and fake news: How social media conditions individuals to be less critical of political misinformation. *Political Communication 39*, 1 (2022), 1–22.

[24] Rojas, H. Corrective actions in the public sphere: How perceptions of media and media effects shape political behaviors. *International Journal of Public Opinion Research 22*, 3 (08 2010), 343–363.

[25] Sadler, E. Influence campaigns. *American Economic Journal: Microeconomics (Forthcoming)* (2022).

[26] Toernberg, P. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLOS ONE 13*, 9 (09 2018), 1–21.

[27] Verducci, S., and Schröer, A. *Social Trust*. Springer US, New York, NY, 2010, pp. 1453–1458.

[28] Vohra, A. Strategic influencers and the shaping of belief. Tech. rep., University of Cambridge, 2021.

[29] Warren, M. *Democracy and Trust*. Cambridge University Press, 1999.

[30] Whitten-Woodring, J., Kleinberg, M. S., Thawnghmung, A., and Thitsar, M. T. Poison if you donât know how to use it: Facebook, democracy, and human rights in myanmar. *The International Journal of Press/Politics 25*, 3 (2020), 407–425.

[31] Woelfert, F. S., and Kunst, J. R. How political and social trust can impact social distancing practices during covid-19 in unexpected ways. *Frontiers in Psychology 11* (2020), 3552.

[32] Yildiz, E., Ozdaglar, A., Acemoglu, D., Saberi, A., and Scaglione, A. Binary opinion dynamics with stubborn agents. *ACM Trans. Econ. Comput. 1*, 4 (dec 2013).

[33] Zimmermann, F., and Kohring, M. Mistrust, disinforming news, and vote choice: A panel survey on the origins and consequences of believing disinformation in the 2017 german parliamentary election. *Political Communication 37*, 2 (2020), 215–237.