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 social mistrust has on the propagation of misinformation on a social network. Agents communicate with each other and observe information sources, changing their opinion with some probability determined by their social trust, which can be low or high. Low social trust agents are less likely to be convinced out of their opinion by their peers and are more likely to observe misinformative information sources. A platform facilitates the creation of a network where users are more likely to connect with agents of the same level of social trust and the same social characteristics. In the case where worldview is relatively important in determining network structure, echo chambers are more pronounced, reducing the probability that agents believe misinformation. At the same time, echo chambers increase polarisation, leading to a trade-off which has implications for the optimal intervention of a platform wishing to reduce misinformation, which we characterise.

Reference Details

2204	Cambridge Working Papers in Economics
2022/02	Janeway Institute Working Paper Series
Published	12 January 2022
Revised	13 April 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†

April 13, 2022

Abstract

We examine the effect social mistrust has on the propagation of misinformation on a social network. Agents communicate with each other and observe information sources, changing their opinion with some probability determined by their social trust, which can be low or high. Low social trust agents are less likely to be convinced out of their opinion by their peers and are more likely to observe misinformative information sources. A platform facilitates the creation of a network where users are more likely to connect with agents of the same level of social trust and the same social characteristics. In the case where worldview is relatively important in determining network structure, echo chambers are more pronounced, reducing the probability that agents believe misinformation. At the same time, echo chambers increase polarisation, leading to a trade-off which has implications

*I would like to thank Matthew Elliott, Alexander Teytelboym, Sanjeev Goyal, Fuhito Kojima, Arjada Barhdi and Akhil Vohra, along with the participants of the Cambridge Economics Micro Theory seminar group for their invaluable contributions to this paper.

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

for the optimal intervention of a platform wishing to reduce misinformation, which we characterise.

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, this thought does not take into account the role of social mistrust in the belief in and sharing of misinformation on online platforms, a role which we examine in detail here.

Social trust and its effect on social media 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 (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, 2021) Furthermore, so-called countermedia information sources, who frequently purvey misinformation, spread a narrative where most people should not be trusted as they are either ignorant or actively nefarious (Rojas, 2010 and Allcott & Gentzkow, 2017), indicating that there is a vicious cycle of low social trust individuals observe content which fosters low social trust, whereby low social trust individuals seek out sources which render them even less trusting of their peers.

Individuals with low social trust, then, have a worldview which has two characteristics relevant to the spread of misinformation on social media. First, they are less likely to be convinced by the opinions of high social trust individuals than the reverse. Second, such users are more likely to believe misinformation in the first place. We construct a model that captures both of these aspects of social mistrust, with the goal of examining how they affect the spread of misinformation.

In the model here, agents interact with each other and information sources on a social network. Users can either be informed or misinformed and they have either low or high social trust. 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 social trust level as them, and bias, in that low social trust users prefer to connect to countermedia sources and high social trust users prefer 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 level of social trust.

We characterise the distribution of opinions of the users as the number of agents tends to infinity. As the social trust of mistrustful agents decreases, the higher the probability that a random agent believes misinformation: misinformation from countermedia news sources is broadcast by users with low social trust to high social trust individuals who are more likely to be convinced out of their belief in the truth than the reverse.

This feature of the model is crucial to our main results. Echo chambers in this context, rather than being a source of the spread of misinformation, protect high social trust users from being convinced by their mistrustful peers. The more agents value homophily with regards to social trust (what we will term "interest-based" networks), the more platform's algorithm amplifies their inherent desire for echo chambers. This reduces the extent to which low social trust agents interact with high social trust agents. The more agents value interacting with agents who are socially similar to them, the less

[1] Of course, there is likely to be a correlation between some social characteristics and social trust. 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.

direct salience their social trust is to network structure, increasing the extent to which low and high social trust agents interact. The probability that a random agent believes misinformation in such friendship networks, then, is greater than the same probability in networks where agents have a direct desire to link to agents who have the same worldview as them.

At the same time, friendship networks tend to be less polarised than interest-based networks, as echo chambers increase the extent to which initial biases remain after communication takes place. As such, we identify a potential trade-off such that reducing the prevalence of echo chambers increases the probability agents are informed but also increases polarisation.

We then turn to the question of interventions in the network to reduce the extent to which misinformation is believed. If the platform intervenes by reducing the probability that users observe one another (a "structural intervention"), then they optimally intervene to reduce the extent to which the most isolated high social trust individuals (i.e. those who are least likely to observe misinformation) interact with the low social trust individuals who are most likely to observe misinformation. Such links are disproportionately costly in terms of the spread of misinformation, as they are most likely to lead to a more convincible agent being converted to believing misinformation.

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 a reduction in social trust increases 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 (2021). Both of these studies provide justification for the asymmetric communication mechanism that plays

a key role in our model.

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 beliefs 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 insulate high social trust users from observing misinformation.

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 (2020) all employ such agents in a naive learning framework, with the latter also considering the limits of a 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 the communication framework in this framework.

Anunrojwong et al (2020) utilise the naive learning framework to examine the case where users observe both a platform and their peers, both of whom share content, influencing beliefs. The focus of their analysis is on the potential for the platform to drive polarisation or consensus on an extreme viewpoint where interactions between agents are symmetric. On the other hand, we examine the case where agents experience asymmetric interactions, which drive polarisation and extremism even without the presence of the platform, though its actions tend to exacerbate the spread of misinformation.

Perhaps the closest of the DeGrootian literature to our paper is Dandekar et al. (2013), which examines the case where agents are more likely to believe evidence which supports their current position, leading to the possibility of polarisation under homophily. By contrast, our analysis focuses on the extent to which homophily drives the direction of beliefs in an environment where asymmetric interactions are independent of belief: low social trust agents are more likely to believe misinformation, but even when they do not, they are still less likely to believe their peers than their more trusting peers are to believe them.

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 in perpetuity 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. In period r, each agent, i, holds an opinion, $v_{ir} \in \{0, 1\}$, where 1 is an informed opinion 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 belief $v_{ir} = 1$ for all r, while if they are countermedia then $v_{ir} = 0$. 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, agents can have two levels of social trust: "low" or "high", such that an agent i has social trust level $\delta_i \in \{\delta_L, \delta_H\} = \Delta$ with $0 < \delta_L < \delta_H \leq 1$. 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

[2]We have 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.

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. Let R_L denote the set of low social trust agents, R_H the set of high social trust agents.

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

Proposition 1. *Suppose G is connected and S is non-empty. Then, the Markov chain $\boldsymbol{v}_t^R$ 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 beliefs.

An example of the communication process with social trust

Suppose that there are three regular agents, A, B and C, with the first two agents having high social trust, the latter having low social trust, and two information sources, one mainstream and one countermedia. We consider two network structures, shown in Figure 1 below.

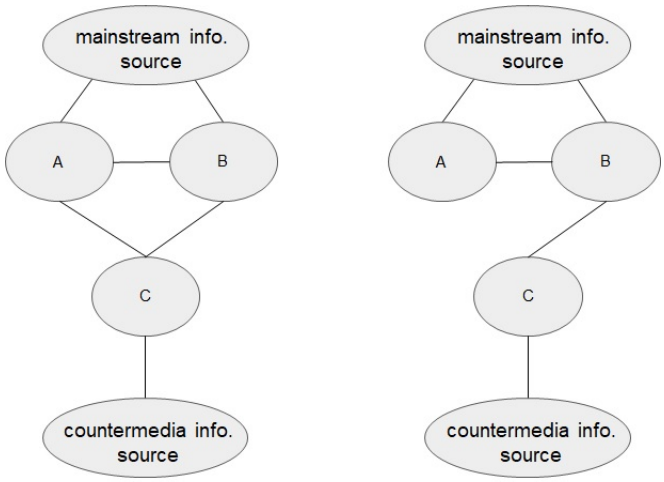

Figure 1: Two realised communication networks, with A, B and C representing regular agents.

Suppose first that $\delta_L = 0.2$ and $\delta_H = 0.8$. In network structure 1, on the right-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 0.52. 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 where $\delta_L = 1$ and $\delta_H = 1$. 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.

3 Network formation and stochastic block models

Having outlined the communication process, we now consider network formation. Agents are made aware of each other and information sources by a platform and choose whether to connect with other agents, forming the network. The network formation process generates a sequence of stochastic block models, which we use to analyse the steady state of the communication process. By modelling network formation explicitly, we are able to consider how network structure is shaped by platform and user preferences, and, in Section 6, how the platform can intervene to affect user beliefs.

Network formation and the platform

Throughout, we will assume that a regular agent, i is associated with both a level of social trust, as discussed above, and a measure which captures social characteristics (e.g. location, schooling, socioeconomic status etc) $\theta_i \in \{\theta_1, ..., \theta_t\} = \Theta$, where $|\theta_i - \theta_j| \in [0, 1]$ measures how socially similar agents are.[3]

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 that they are aware of.

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

First, we consider the connection stage. Regular agents can only connect with other agents who they are aware of, can connect costlessly, and prefer connecting with agents who are similar to them both in terms of social characteristics and social trust (which can be thought of as a proxy for worldview more broadly). We assume also that regular agents have preferences over information sources depending on their level of social trust.

Let $\hat{\theta}_{ij} = -|\theta_j - \theta_i|$ and $\hat{\delta}_{ij} = -1$ if $\delta_i \neq \delta_j$ and 0 otherwise. We assume that conditional of being aware of agent j, an agent i receives the following utility from linking to them:

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

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 and $\alpha \in [0,1]$ measures the relative importance differences in social trust and social characteristics have in determining the utility generated by a link. We assume that ε_{ij}s are i.i.d. Define $\alpha\hat{\delta}_{ij} + (1-\alpha)\hat{\theta}_{ij} = \gamma_{ij} < 0$. Suppose i is aware of j with probability β_{ij}. The total probability that $ij \in R$ are connected is then $w_{ij} = (\beta_{ij} + \beta_{ji})\frac{1+\gamma_{ij}}{2}$ and:

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

Now consider the awareness stage. Agent i 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 induced by the

platform by e.g. suggesting to i that they follow j on a sidebar.[4] Adjusting β_{ij} away from $\hat{\beta}_{ij}$ is costly to the platform - for example, 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 $m_R \times m_R$ matrix whose ijth entry is $\hat{\beta}_{ij}$ and $\chi \in [0, 1]$.

Define $\mathrm{E}[D(\hat{\boldsymbol{\beta}}, \beta)] = \sum_{i \in R} \mathrm{E}[d_i(\hat{\boldsymbol{\beta}}, \beta)] = \sum_j [w_{ij} + w_{ij}^S]$, where d_i is i's degree. 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, $\hat{\boldsymbol{\beta}}$, 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. Once the network is formed, agents then communicate with one another under the process described in Section 2.

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*

[4] 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.

space, $\mathcal{T} = \{\Theta \times \Delta\} = \{T_1, ...T_{2t}\}$*, a* $2t \times 2t$ *matrix of linking probabilities,* $\boldsymbol{W} = \boldsymbol{W}(\alpha)$*, a number of agents,* n*, and a vector* $\boldsymbol{m}(n) = (m_1^R(n), ..., m_{2t}^R(n), m_0^S(n), m_1^S(n))$*, 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 $\delta_i = \delta_j$ 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 social trust, δ_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 $|\hat{\delta}_{ij}| > |\hat{\delta}_{ik}|$ and $\hat{\theta}_{ij} \geq \hat{\theta}_{ik}$ or $\hat{\theta}_{ij} > \hat{\theta}_{ik}$ and $\hat{\delta}_{ij} \geq \hat{\delta}_{ik}$, i.e. there is homophily between groups both in terms of worldview and social characteristics. In terms of information sources, let $\mathcal{T}_L$ and $\mathcal{T}_H$ denote the set of types which contain agents with low and high social trust respectively. Then if $T_i \in \mathcal{T}_H$ and $T_j \in \mathcal{T}_L$ and $k \in S_1$ then $w_{ik}^S > w_{jk}^S$, 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 social trust. How relatively important these measures are for network structure depends on the parameter α. To see this, suppose $i, j \in R_L$ and $k \in R_H$ 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).

Stochastic block models

Throughout, we will consider the expected beliefs 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 beliefs as $n \to \infty$. Doing so allows us to characterise the distribution of beliefs held by agents in steady state, and, given the large number of users of social networks, provide a good approximation of the distribution beliefs that would be held by agents on social block models constructed in the manner described above.

For a fixed n, let $m_i^S(n)$ denote the number of information sources of opinion i, and $m_i^R(n)$ denote the number of regular agents of type T_i. We write:

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

as the limiting fractions of information sources with opinion i and regular agents of type T_i respectively. Throughout, we will maintain the assumption that $q_0^S = q_1^S$, 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.[5]

Let $\tilde{v}_i(n, \alpha)$ be a random variable denoting the beliefs of an agent i and distributed according to the steady state of the model n. Define:

$$z_j(n, \alpha) := \frac{\sum_{i\in T_j} \tilde{v}_i(n, \alpha)}{m_i^R(n)},$$

[5]This assumption simplifies the analysis, but the model could easily incorporate agents who prefer to observe fewer or more information sources than others, as well as correlations between social trust level and these preferences.

as the average opinion of type T_j agents. Let $\mathcal{T}_R$ be the set of all types of regular agents. Public opinion can then be defined as follows:

$$\hat{z}(n,\alpha) = \frac{1}{m_R}[\sum_{T_i \in \mathcal{T}_R} m_i^R(n) z_i(n,\alpha)].$$

We will often be interested in the limit of public opinion, $\lim_{n\to\infty}\hat{z}(n,\alpha) = \bar{z}(\alpha)$. Finally, we write $\bar{G}(\boldsymbol{W}(\alpha)) = \lim_{n\to\infty} G(\boldsymbol{m}(n), \boldsymbol{W}(\alpha))$.

4 Opinion formation and social trust

We state an expression for the limit vector of equilibrium beliefs. We then use that limit vector to examine the effect that our assumption regarding differences in social trust across agent types have on public opinion.

The belief vector

We characterise the limit vector of the beliefs of regular agents. Define $\hat{\boldsymbol{W}}(\alpha)$ as the $2t \times 2t$ trust-adjusted linking probability matrix whose ijth entry is $q_j \delta_i w_{ij}(\alpha)$. Define the normalised degree of a type i regular agent as follows:

$$d_i = \sum_{j\in\mathcal{T}} \delta_i q_j w_{ij}(\alpha) + q_1^S w_{i1}^S(\alpha) + q_0^S w_{i0}^S(\alpha).$$

Let $\boldsymbol{\Lambda}(\alpha)$ denote a diagonal matrix whose ith component is d_i and $M(\alpha) := \boldsymbol{\Lambda}(\alpha) - \hat{\boldsymbol{W}}(\alpha)$. We define a $T \times 1$ column, z^S whose ith entry is $q_1^S w_{i1}^S$ and $\boldsymbol{z}(n,\alpha)$ is a $T \times 1$ column vector whose ith entry is $z_i(n,\alpha)$. The following Theorem holds:

Theorem 1. *The limit vector of regular agent expected opinions, $\boldsymbol{z}(n,\alpha)$, converges almost surely to the expression:*

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

The ith component of the vector $\boldsymbol{z}^{\boldsymbol{S}}$, $q_1^S w_{i1}^S$, measures the direct effect information sources have on the belief probabilities of an agent of type T_i's. 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, T_i, and another, T_j, the larger the effect that the information sources that a given agent of type T_i is connected to have on an agent of type T_j, and vice versa. The result in Theorem 1 is a generalisation of a similar result found in Sadler (2020), which is derived for the case where an interaction agent adopts their partner's view with probability 1.

The role of social trust

We can use the expression in Theorem 1 to understand the effect of a change in the levels of social trust. We do so while holding the worldview parameter fixed, so as not to change the proportion of countermedia sources shown to particular agents. from low social trust individuals.

Proposition 3. *A decrease in δ_L results in an increase in expected public opinion and therefore a decrease in the probability that the average agent believes misinformation; that is, $\frac{d\bar{z}(\alpha)}{d\delta_L} > 0$.*

Proposition 3 highlights the effect that the presence of low social trust agents have on the distribution of equilibrium beliefs. To see why it holds, first consider the case where the network, G, is fixed. As the social trust of these agents decreases, the influence they have on high social trust agents they connect to increases relative to those agents' influence on low social trust agents. The expected proportion of countermedia information sources

observed by low social trust agents is higher than for agents with high social trust. It follows that the probability that misinformation is believed increases in this case, with the opposite being true when δ_L increases.

A change in δ_L also implies a change in optimal network structure in equilibrium: an increase in δ_L increases the proportion of countermedia information sources observed by low social trust agents. This straightforwardly has the effect of increasing the probability that a generic low social trust agent believes misinformation, which in turn increases the probability that high social trust agents believe misinformation as well.

5 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.

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.

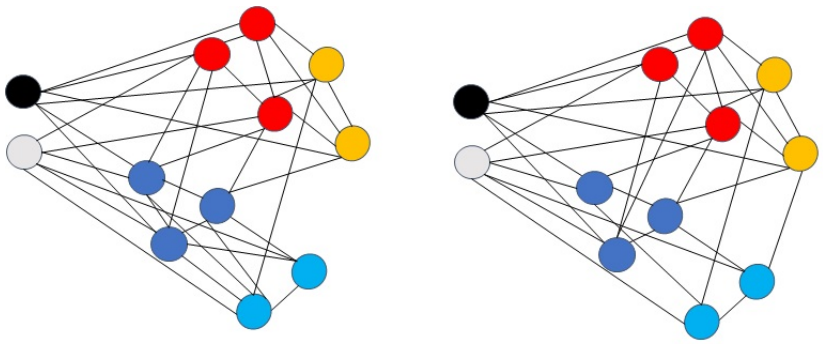

Figure 2: The black node represents a countermedia source, the grey node represents a mainstream media source, and the other colours denote different types of agent. Specifically, the light and dark blue agents are of high social trust, with the other nodes being low social trust, the light blue and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

As highlighted in Section 3, the structure of the network has implications for the expected belief vector, $\boldsymbol{z}(\alpha)$. We can examine how friendship networks differ from interest-based networks by defining two different stochastic block models, $G(\boldsymbol{m}(n), \boldsymbol{W}(\alpha_F))$ and $G(\boldsymbol{m}(n), \boldsymbol{W}(\alpha_I))$, where $\alpha_F < \alpha_I$. The following Theorem holds:

Theorem 2. *Expected public opinion $\bar{z}(\alpha)$ is increasing in α. This implies that the limit probability that a random agent is correctly informed in the friendship stochastic block model is lower than in the interest-based stochastic block model; that is, $\bar{z}(\alpha_F) < \bar{z}(\alpha_I)$.*

The result of Theorem 2 largely relies on the fact that communication between agents with low and high social trust is asymmetric. When low social trust individuals communicate their opinion to high social trust individuals, the interaction is more likely to lead to the latter adopting the former's opinion than the reverse interaction would do. As low social trust agents are also more likely to believe misinformation, it follows that such interactions will increase the average level of belief in misinformation.

In friendship networks, low and high social trust individuals are more likely to be connected to each other, whereas in interest-based networks these agents are more segregated from one another. As communication on the network is asymmetric, such that low social trust types are less convincible than high social trust types, it follows that the expected number of agents who believe misinformation is greater in the friendship network than for the the interest-based network. As such, the interest based network generates an echo chamber more effectively than the friendship network.

Theorem 2 also highlights 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 more salient social characteristics are (i.e. the smaller α is), the more the plat-

form's algorithm encourages users to connect to those who are similar in terms of those characteristics. When $\alpha < \frac{1}{2}$ then, the platform's algorithm acts to reduce the natural echo chamber that exists due to the agents' innate homophily, and in turn increases the probability that agents believe misinformation, with the opposite being true when $\alpha > \frac{1}{2}$.

Echo chambers

A way of conceptualising the result in Theorem 2 is as a statement about the effect of echo chambers on opinions. While the potential role of echo chambers in the polarisation of beliefs and ensuring that a proportion of the population will believe misinformation has been analysed extensively, our results point to a mechanism that is often ignored. As the segregation of different viewpoints becomes more pronounced, the less interaction there is between individuals who believe misinformation and those who do not. If the latter are (on average) more convincible than the former, then echo chambers reduce the spread of misinformation.

To better understand the interaction between echo chambers and social trust, we define $\tilde{\delta} = \delta_H - \delta_L$, and note that $\bar{z}(\alpha, \tilde{\delta})$ is a function of a $\tilde{\delta}$ as a result of the fact that the matrix $\hat{\boldsymbol{W}}(\alpha, \tilde{\delta})$ is a function of $\tilde{\delta}$. The following result holds:

Proposition 4. *Suppose $\alpha_F < \alpha_I$. Then, when $\tilde{\delta} > 0$, $\bar{z}(\alpha_F, \tilde{\delta}) < \bar{z}(\alpha_I, \tilde{\delta})$ and when $\hat{\delta} = 0$, $\bar{z}(\alpha_F, \tilde{\delta}) = \bar{z}(\alpha_I, \tilde{\delta})$.*

If $\tilde{\delta} > 0$, $T_i \in \mathcal{T}_L$ and $T_j \in \mathcal{T}_H$ then $w_{ij}(\alpha_F) > w_{ij}(\alpha_I)$: the friendship model tends to generate graphs that exhibit less homophily with regards to social trust than the interest-based model, which implies the former exhibits less of an "echo chamber" effect. Proposition 4 then states that by protecting high social trust individuals from

misinformation, the interest-based model produces an outcome in which agents are less likely to believe that misinformation.

The two models only become equivalent in the case where agents have the same level of social trust: when this holds, communication between agents more likely to be connected with mainstream information sources and those more likely to observed countermedia sources is as effective at changing opinions as communication in the reverse direction. In the case where there are no differences in social trust then, the fact that agents with mostly mainstream views are more likely to observe misinformed agents in the friendship network than in the interest-based network is canceled out by the fact that misinformed agents are more likely to be convinced out of their position by their mainstream opinion holding peers.

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:

$$P(\alpha) := \lim_{n\to\infty} \sum_i q_i |z_i(n,\alpha) - \bar{z}(n,\alpha)|.$$

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

Proposition 5. *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, Theorem 2 and Proposition 5 imply that, if the platform was incentivised to intervene in the structure of the network, there is a trade-off between polarisation and the probability that misinformation is believed by a random agent. 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.

Note that this trade-off would not exist without differences in social trust. In the case where all agents have the same levels of social trust, network structure affects polarisation, but does not affect the average number of agents who believe misinformation, as highlighted in Proposition 4 and the minimal example in Section 2. The introduction of social trust (or, more broadly, agents who are differentially convincible) results in network structure being a critical element in determining not just who believes what, but the extent to which misinformation is believed. This facet of the model makes the question of network design more pertinent, which we discuss in Section 6.

6 Network interventions

We consider a range of policy interventions which aim to reduce the extent to which misinformation propagates through the network. To do so, we consider a finite network

with large n, such that $\hat{z}(n, \alpha)$ is well approximated by $\bar{z}(\alpha)$, and analye the effect of the platform intervening in the matrix of awareness probabilities. Specifically we consider structural interventions, which involve changing the extent to which regular agents are aware of one another, and interventions which reduce the extent to which agents observe countermedia sources, characterising conditions under which these interventions are particularly effective.

The efficacy of algorithmic interventions

Much of our analysis will focus on the extent to which a change in the platform's algorithm affects public opinion. A natural question before we conduct this analysis would be whether some form of ex-post intervention on the structure of the network would be more effective than intervening on the algorithm prior to the network's realisation. To answer this question, we state the following Theorem:

Theorem 3. *For any* $T \in \mathcal{T}$, $lim_{n\to\infty} max_{i\in T} |E[\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \beta_{ij}}] - E[\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \beta_{kj}}]| = 0$ *for all* $k \in T$.

Theorem 3 states that, as $n \to \infty$, an ex-ante marginal change in the awareness algorithm of a random agent of type k has the same effect on public opinion as the case where the platform knew which β_{ij} would be most effective at changing public opinion at the margin. As such, Theorem 3 validates the effectiveness of interventions in the network generating process, which would tend to require less information than intervention on the realised network.

Structural interventions

First, we compare the relative effectiveness of interventions which directly decrease the probability that one type of regular agent is aware of another type of regular agent,

rather than a source of information. We call these interventions "structural interventions". We focus on structural interv.entions which change the probability that high and low type agents interact with one another: such interventions will necessarily be more effective in reducing the average belief in misinformation than other interventions; a direct consequence of the result in Theorem 2.

We consider a marginal decrease in an individual awareness probability β_{ij}. Both network structure and the number of agents who are of a given type will impact on the overall effect such an intervention will have. To focus on the effects of network structure, we examine a case that we will call type-symmetry, where for all $i \in \mathcal{T}_L$ and $j \in \mathcal{T}_H$, $q_i = q_L$ and $q_j = q_H$. As agents of the same type are identical in expectation, we can draw on the result in Theorem 1 to understand this marginal change in terms of the interaction between types.

Proposition 6. *Suppose $\bar{G}(\boldsymbol{W})$ is type-symmetric. The following two statements hold: (a) if $i, j \in R_L$, $k \in R_H$ and $z_i < z_j$ then $|\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \beta_{ik}}| > |\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \beta_{jk}}|$ and; (b) if $s, t \in R_H$, $v \in R_L$ and $z_s > z_t$, then $|\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \beta_{sv}}| > |\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \beta_{tv}}|$.*

Connections between low social trust agents who are most likely to believe misinformation and their more trusting peers are more influential in spreading misinformation than links between those high social trust agents to other low social trust agents. Agents with the highest belief in misinformation are relatively isolated (both directly and indirectly) from high social trust agents, and therefore truthful information sources. As a result, conditional on being observed such agents are most likely to espouse misinformation which then convinces the high social trust agents they interact with.

Similarly, links between low social trust agents and agents who are most likely to believe the truth are also relatively influential in spreading misinformation compared

with links to other high social trust agents. Again, the type that is least likely to believe misinformation is relatively isolated from low social trust agents, which renders the interactions they do have with such users relatively potent.

As such, the platform can intervene in intra-agent interactions most effectively by lowering the extent to which relatively isolated, high social trust agents come into contact with low social trust individuals who are themselves relatively less likely to come into contact with correctly informed agents and information sources.

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 define the influence of an average type $i \in \mathcal{T}$ agent as $\phi_i := \sum_j \varsigma_{ij}$, where ς_{ij} is the ijth entry in the matrix $K(\alpha)$, a $2t \times 2t$ matrix whose ijth entry is equal to the ijth entry of $M^{-1}(\alpha)$ times $\frac{q_{i\cdot}}{q_j}$. We also define $\tilde{\beta}_i(j,k) := \beta_{ij} - \beta_{ik}$, which allows us to consider a change in the relative proportion of countermedia information sources without changing the amount of information sources i observes overall. We make the following observation:

Proposition 7. *Suppose that $i \in T_i$, $j \in T_j$ and $k_t \in S_t$ with $\phi_i > \phi_j$. Then $|\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \tilde{\beta}_i(k_1,k_0)}| > |\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \tilde{\beta}_j(k_1,k_0)}|$.*

Increasing the proportion of mainstream media sources observed by an agent decreases the probability that every agent believes misinformation. The influence of t, ϕ_t, mea-

sures the effect the information sources observed by an average t 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.

To see a practical implication of Proposition 7, we first note that $\frac{\partial \phi_i}{\partial \delta_L} < 0$ for all $i \in \mathcal{T}_L$: low social trust individuals become relatively less influential as they become more convincible. The intuition for this is that the matrix $K(\alpha)$ (and, equivalently, $M^{-1}(\alpha)$) measures the extent to which type i agents propagate the views of the information sources they observe. When type is are more convincible, they are more likely to adopt the position of their peers, and hence they are less influential overall.

The observations above imply the following result:

Proposition 8. *Suppose $\bar{G}(\boldsymbol{W})$ is type-symmetric such that $q_L = q_H$ and so $q_i = q \; \forall i$. Then for any $j \in \mathcal{T}_H$, $\exists i \in \mathcal{T}_L$ such that $\phi_i > \phi_j$.*

In the case where types only differ by their location in the network, for any high social trust type, there exists some low social trust type who is more influential than them. Specifically, if i and j have the same social characteristics (i.e. $\theta_i = \theta_j$) then if $i \in \mathcal{T}_L$ and $j \in \mathcal{T}_H$, 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.

Optimal interventions

We can use the above analysis to assess when marginal interventions in which the proportion of countermedia information sources agents observe is reduced are more effective than marginal structural interventions. Let $\Delta_S(\alpha) = \max_{i,j \in R}\{|\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \beta_{ij}}|\}$ and $\Delta_I(\alpha) = \max_{i \in R}|\frac{\partial \bar{z}(\alpha;\boldsymbol{\beta}^*)}{\partial \tilde{\beta}_i(k_1,k_0)}|$. We make the following statement:

Theorem 4. *Suppose $\bar{G}(\boldsymbol{W})$ is type-symmetric. Then the most effective structural intervention, $\Delta_S(\alpha)$, is increasing in α. Furthermore, for any α, there exists $\bar{\delta} \in [0,1)$ such that if $\tilde{\delta} < \bar{\delta}$, then $\Delta_I(\alpha,\tilde{\delta}) > \Delta_S(\alpha,\tilde{\delta})$.*

As low and high social trust agents become more connected, the difference in the belief probabilities of the most misinformed and least informed agents becomes smaller. As Proposition 6 implies, this results in the most effective structural intervention being less effective, as agents with opposing opinions are less segregated. Hence, structural interventions are relatively more effective in the interest-based network than in the friendship network.

Now, consider the second statement in Theorem 4. If the difference in social trust levels is sufficiently low, then intervening to reduce the proportion of countermedia sources recommended by the platform's algorithm is more effective than any structural intervention. As $\tilde{\delta}$ decreases, the extent to which agents of different types differ in the probability that they believe misinformation reduces, because, first, the difference in convincibility between low and high social trust agents becomes smaller and secondly the difference in the probability that a low and a high social trust agent observes misinformation also decreases. When $\tilde{\delta} = 0$ both these differences are zero, guaranteeing that intervening to reduce the probability that misinformation is observed in the first place is more effective than a structural intervention.

One further question that we have not yet addressed is the issue of polarisation. It is fairly straightforward to see that structural interventions increase polarisation, while reducing the probability that an agent or agents connect with countermedia sources has the effect of reducing it. Hence, if the goal of the intervener is to both reduce polarisation and the probability that agents believe social mistrust, the latter type of intervention is superior. On the other hand, when α is large, structural interventions are particularly effective, and may be worth the increase in polarisation associated with them in some cases.

7 Concluding remarks

We have analysed an opinion formation model in which some agents have lower social trust than others. Low social trust agents communicate asymmetrically with their high social trust peers, as they are less convincible than them. We consider the case where these mistrustful agents are more likely to propagate the misinformation spread by countermedia sources. The asymmetry in communication results in agents being more likely to believe misinformation in networks in which there is a relatively high level of communication between high and low social trust agents, a result exacerbated by the platform's desire to maximise engagement. However, such networks have less polarisation than networks in which echo chambers are more pronounced, leading to a trade-off between these two features of opinion formation networks.

A key aspect of our analysis is the empirically established link between an agent being mistrustful, believing misinformation and engaging with countermedia sources. One aspect of these links which we have not explored is why low social trust agents are attracted to following countermedia sources. While it is sufficient for our purposes

that such a link exists, one explanation worth highlighting is that these sources present narratives where people with mainstream opinions are either acting in bad faith or have been tricked or hoodwinked into believing those opinions (see Harambam & Aupers, 2014, for example). There is then a feedback loop between viewing countermedia sources and social trust, such that agents who have a slight preference towards countermedia sources become less trusting of mainstream narratives, and as a result seek out countermedia sources.

It should be noted that 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.[6] 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 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.

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 important variables like speed of convergence (e.g. Golub and Jackson, 2010) or polarisation (e.g. Sadler, 2020) but it plays less of a role in determining the average

[6]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.

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 beliefs and status quo bias.

Appendix

Preliminaries

Many of the results in the main text require evaluating various derivatives of the matrix $M^{-1}(\alpha)$. Throughout, we let r_{ij} represent the ijth component of $M^{-1}(\alpha)$. It is useful to define the $2t \times 2t$ matrices $L(\alpha) := -M^{-1}(\alpha)\frac{\partial M(\alpha)}{\partial \alpha}M^{-1}(\alpha)$ and $L_{ik}(\alpha) := -M^{-1}(\alpha)\frac{\partial M}{\partial w_{ik}}M^{-1}(\alpha)$.

It is also worthwhile establishing some facts about the matrices M and M^{-1}. 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 $M(\alpha)$ is of full rank (i.e. has rank of $2t$) and is such that $\sum_j m_{ij} = \bar{q}^S$ for all j. This in turn implies that each row of the matrix $M^{-1}(\alpha)$ sums to $\frac{1}{\bar{q}^S}$. To see this, let v be a $2t \times 1$ vector filled with 1s and hence $M(\alpha)v = v$. Thus, $M^{-1}(\alpha)v = M^{-1}(\alpha)M(\alpha)v = v$. We also let $\frac{1}{1-\bar{q}^S}\boldsymbol{z}^S = \tilde{\boldsymbol{z}}^S$.

We note that we can write $M(\alpha) = AD(\alpha)Q$, where A is a diagonal matrix with ith component δ_i, Q is a diagonal matrix with ith component q_i and $D(\alpha)$ is some symmetric matrix. It follows that $M^{-1} = Q^{-1}D^{-1}A^{-1}$, where $D^{-1}(\alpha)$ is also a symmetric matrix.

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

text:

Lemma 1. *M is an M-matrix and hence M^{-1} is non-negative.*

Proof. By definition, M is a Z-matrix (i.e. a matrix where $m_{ij} \leq 0$ for all $i \neq j$). M 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 is therefore a $M-$matrix. M 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 $\boldsymbol{v}_t^R$ 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:

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

Let T be such that if $i,k \in T$, then $\theta_i = \theta_k$ and $\delta_i = \delta_k$. Then for any and $j \in G$ 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 $\delta_i \in \Delta$, it follows that for each $T_s \in \mathcal{T}$, it follows that if $i,k \in T_s$ 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 $|\Theta| = t$, it follows that $|\Theta \times \Delta| = 2t$, as required.

Proof of Theorem 1

To simplify notation we write $D(\hat{\boldsymbol{\beta}}, \beta) = D$ 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:

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

where $\mu^S_i = \frac{\sum_{j \in S} g_{ij}}{d_i}$, $\mu^R_i = \frac{\sum_{j \in R} g_{ij}}{d_i}$ and d_i is the degree of i in G. 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:

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

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 $R \times R$ matrix $\bar{G}(\alpha, n)$ as a matrix whose ijth entry is written:

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

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

$$h_{ij} = \begin{cases} \frac{\delta_i m_j^R(n) w_{ij}(\alpha)}{\mathrm{E}[D(n,\alpha)]} & \text{if } i \neq j \\ 1 - \frac{\vartheta_i(n,\alpha)}{\mathrm{E}[D(n,\alpha)]} + \frac{\delta_i(m_i^R(n)-1)w_{ii}(\alpha)}{\mathrm{E}[D(n,\alpha)]} & \text{if } i = j \end{cases} .$$

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

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

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

where $\hat{\boldsymbol{z}}^S(n,\alpha)$ is a column vector ith entry is $\frac{\tilde{w}_{ij}(\alpha)m_1^S(n)}{\mathrm{E}[D(n,\alpha)]}$ 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 $i \in T_j$, then the ith entry of $\bar{\boldsymbol{x}}(n,\alpha)$ is equal to the ith entry of $\hat{\boldsymbol{z}}(n,\alpha)$.

We now need to show that $|\boldsymbol{x}(n,\alpha) - \bar{\boldsymbol{x}}(n,\alpha)| \to^{a.s.} 0$ and $\hat{\boldsymbol{z}}(n,\alpha) \to \bar{\boldsymbol{z}}(\alpha)$. 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 $\pi \in [0,1]$. The following statement holds, as shown in the proof of Theorem 1 in Sadler (2020):

Lemma. *(Sadler, 2020) 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 $\boldsymbol{z}(n,\alpha) \to \boldsymbol{z}(\alpha)$, 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.

Proof of Proposition 3

Public opinion, $\bar{z}(\alpha)$, can be written: $M^{-1}(\alpha)\tilde{\boldsymbol{z}}^S\boldsymbol{q}^T = \bar{z}(\alpha)$ where $\boldsymbol{q}^T$ is a $1 \times 2t$ vector with ith entry q_i. The vector $\boldsymbol{q}^T$ is independent of δ_L, but both $M^{-1}(\alpha)$ and $\boldsymbol{z}^S$ are functions of it. We analyse $\frac{\partial M^{-1}(\alpha)}{\partial \delta_L}$ first. We note that $M^{-1}(\alpha)\tilde{\boldsymbol{z}}^S(\alpha)\boldsymbol{q}^T$ can be

rewritten as $QM^{-1}\tilde{\boldsymbol{z}}^S 1^T$. As per the preliminary analysis, $M^{-1}(\alpha) = Q^{-1}D^{-1}(\alpha)A^{-1}$, and so $QM^{-1}\tilde{\boldsymbol{z}}^S\,\mathbf{1^T} = D^{-1}(\alpha)A^{-1}\tilde{\boldsymbol{z}}^S\mathbf{1^T}$.

Note that both A^{-1} and $D^{-1}(\alpha)$ are functions of δ_L. Let $Y(\alpha) = AD$ and so $Y^{-1}(\alpha) = D^{-1}(\alpha)A^{-1}$. Then $\frac{\partial Y^{-1}(\alpha)}{\partial \delta_L} = -Y^{-1}\frac{\partial Y}{\partial \delta_L}Y^{-1}$. Consider $\frac{\partial Y}{\partial \delta_L}$. This matrix has a row of 0s for all rows corresponding to a type, $j \in \mathcal{T}_H$, but $\frac{\partial y_{ii}(\alpha)}{\partial \delta_L} > 0$ and $\frac{\partial y_{ij}(\alpha)}{\partial \delta_L} < 0$ for all j and $i \in \mathcal{T}_L$. Furthermore, $\sum_j y_{ij} = 0$ for all i, which implies that each row sum of $\frac{\partial Y^{-1}(\alpha)}{\partial \delta_L}$ is equal to 0.

As $D(\alpha)$ is symmetric, and $\delta_i = \delta_L$ for all $i \in \mathcal{T}_L$ the above analysis implies that the ith column sum of $\frac{\partial Y^{-1}(\alpha)}{\partial \delta_L}$ is negative when $i \in \mathcal{T}_L$ and positive when $i \in \mathcal{T}_H$. Combined with the fact th.at $\boldsymbol{z}_i^S \leq \boldsymbol{z}_k^S$ for $i \in \mathcal{T}_L$ and all k with the inequality strict when $k \in \mathcal{T}_H$ and the row sum of $\frac{\partial Y^{-1}(\alpha)}{\partial \delta_L}$ is equal to 0, it must be that $\frac{\partial Y^{-1}(\alpha)}{\partial \delta_L}\tilde{\boldsymbol{z}}^S\boldsymbol{q}^T > 0$.

Now consider the $\frac{\partial \boldsymbol{z}^S}{\partial \delta_L}$. By the fact that $\beta_{ij}^* = \frac{1+\delta_i}{4\chi}$ in for$j \in S_1$ and $\boldsymbol{z}_i^S = w_{i1}q_1^S$, $\frac{\partial \boldsymbol{z}_i^S}{\partial \delta_L} > 0$ for $T_i \in \mathcal{T}_H$. Hence, $\frac{\partial \boldsymbol{z}^S}{\partial \delta_L} > 0$ and so $M^{-1}(\alpha)\frac{\partial \tilde{\boldsymbol{z}}^S}{\partial \delta_L}\boldsymbol{q}^T > 0$. The Proposition then immediately follows.

Proof of Theorem 2

To consider the effect of a marginal change in α on $\bar{z}(\alpha)$ we note that $QM^{-1}\tilde{\boldsymbol{z}}^S\mathbf{1^T} = \bar{z}(\alpha)$. We know that $M^{-1}(\alpha) = Q^{-1}D^{-1}(\alpha)A^{-1}$, and so $QM^{-1}\tilde{\boldsymbol{z}}^S 1^T = D^{-1}(\alpha)A^{-1}\tilde{\boldsymbol{z}}^S\mathbf{1^T}$. Again, letting $Y(\alpha) = AD$, we analyse $\frac{\partial Y^{-1}(\alpha)}{\partial \alpha} = -Y^{-1}\frac{\partial Y}{\partial \alpha}Y^{-1}$.

Given the solution to the platform's maximisation problem, the matrix $\frac{\partial D(\alpha)}{\partial \alpha}$ is such that if $i, j \in \mathcal{T}_L$ then $d_{ij} < 0$ and if $k \in \mathcal{T}_H$, $d_{ik} > 0$, with $|d_{ij}| = d_{ik}$ if $\theta_j = \theta_k$. It follows that $d_{ii} = 0$, $\sum_j d_{ij} = 0$ for all i and, as $D(\alpha)$ is symmetric, so too is $\frac{\partial D(\alpha)}{\partial \alpha}$. As A is not a function of α, $V(\alpha) = \frac{\partial Y^{-1}(\alpha)}{\partial \alpha} = -Y^{-1}A\frac{\partial D(\alpha)}{\partial \alpha}Y^{-1}$. Furthermore, each row sum of $\frac{\partial Y^{-1}(\alpha)}{\partial \alpha}$ is equal to 0. Hence:

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

As $\frac{\partial D(\alpha)}{\partial \alpha}$ is symmetric and $\delta_L < \delta_H$ (and noting the negative sign in front of the above expression) implies that $\sum_i v_{ij}(\alpha) < 0$ if $j \in \mathcal{T}_L$ and $\sum_i v_{ik}(\alpha) > 0$ if $k \in \mathcal{T}_H$ and $\sum_{i \in \mathcal{T}_L} |v_{ij}(\alpha)| = \sum_{i \in \mathcal{T}_H} |v_{ij}(\alpha)|$. As $\boldsymbol{z}_i^S \leq \boldsymbol{z}_k^S$ for $i \in \mathcal{T}_L$ and all k with the inequality strict when $k \in \mathcal{T}_H$, it then follows that $\frac{\partial \bar{z}(\alpha)}{\partial \alpha} > 0$.

Proof of Proposition 4

For the first statement, note that the proof of Theorem 2 only relies on $\boldsymbol{z}_i^S < \boldsymbol{z}_j^S$ for all $T_i \in \mathcal{T}_L$ and $T_j \in \mathcal{T}_H$ and $\tilde{\delta} > 0$: the precise values of $\boldsymbol{z}_i^S$ and $\boldsymbol{z}_j^S$ beyond this inequality are not specified. In other words, $\boldsymbol{z}_i^S < \boldsymbol{z}_j^S$ always holds as long as $\tilde{\delta} > 0$. When this inequality holds, that proof then continues to hold: it must be that $\frac{\partial \bar{z}(\alpha, \tilde{\delta})}{\partial \alpha} > 0$ for all values of α, and so $\bar{z}(\alpha_F, \tilde{\delta}) < \bar{z}(\alpha_I, \tilde{\delta})$.

Now, consider the case where $\tilde{\delta} = 0$. The proof of Theorem 2 implies that when $\tilde{\delta} = 0$, $\frac{\partial \bar{z}(\alpha, \tilde{\delta})}{\partial \alpha} = 0$ for all α, and hence, $\bar{z}(\alpha_F, \tilde{\delta}) = \bar{z}(\alpha_I, \tilde{\delta})$ at $\tilde{\delta} = 0$.

Proof of Proposition 5

Note that the result in the proof of Theorem 2 that $L(\alpha) = -Y^{-1} A \frac{\partial D(\alpha)}{\partial \alpha} Y^{-1}$ immediately implies that $\frac{\partial z_j(\alpha)}{\partial \alpha} < 0$ and $\frac{\partial z_k(\alpha)}{\partial \alpha} > 0$ for all $j \in \mathcal{T}_L$ and $k \in \mathcal{T}_H$. Furthermore, equation (1) in the proof of Theorem 2 implies that $\sum_{j \in \mathcal{T}_L} q_j |\frac{\partial z_j(\alpha)}{\partial \alpha}| < \sum_{k \in \mathcal{T}_H} q_k |\frac{\partial z_k(\alpha)}{\partial \alpha}|$.

The derivative $\frac{\partial \bar{z}(\alpha)}{\partial \alpha} = \sum_{i \in \mathcal{T}} q_i \frac{\partial z_i(\alpha)}{\partial \alpha}$. As $\frac{\partial z_j(\alpha)}{\partial \alpha} > 0$ and $\frac{\partial z_k(\alpha)}{\partial \alpha} < 0$ for all $j \in \mathcal{T}_L$ and $k \in \mathcal{T}_H$ and $\sum_{k \in \mathcal{T}_H} \frac{q}{q_k} z_k(\alpha) > \bar{z}(\alpha) > \sum_{j \in \mathcal{T}_L} \frac{q}{q_j} z_j(\alpha)$, it then follows that $\sum_{k \in \mathcal{T}_H} q_k (\frac{\partial z_k(\alpha)}{\partial \alpha} - \frac{\partial \bar{z}(\alpha)}{\partial \alpha}) > 0$ and $\sum_{j \in \mathcal{T}_L} q_j (\frac{\partial \bar{z}(\alpha)}{\partial \alpha} - \frac{\partial z_j(\alpha)}{\partial \alpha}) = \frac{\partial \sum_{j \in \mathcal{T}_L} q_j |z_j - \bar{z}(\alpha)|}{\partial \alpha} > 0$, which implies the result.

Proof of Theorem 3

Note from the proof of Theorem 1, 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 and so $|\boldsymbol{x}(n,\alpha) - \bar{\boldsymbol{x}}(n,\alpha)| \to^{a.s.} 0$. This in turn implies that $|\boldsymbol{z}(n,\alpha) - \bar{\boldsymbol{z}}(n,\alpha)| = 0$. Straightforwardly, for these statements to hold, it must be the case that $\lim_{n\to\infty}\max_{i\in R,T_i}|\mathrm{E}[v_i(n)] - z_{T_i}(n)| = 0$, which then immediately implies that $\lim_{n\to\infty}\mathrm{E}[|v_i(n,\alpha) - v_j(n,\alpha)|] = 0$ for any $i, j \in T$. This holds because if $\lim_{n\to\infty}\max_{i\in R,T_i}|\mathrm{E}[v_i(n,\alpha)] - z_{T_i}(n,\alpha)| = 0$, then both $\lim_{n\to\infty}\max_{i\in R,T_i}(\mathrm{E}[v_i(n,\alpha)] - z_{T_i}(n,\alpha)) = 0$ and $\lim_{n\to\infty}\min_{i\in R,T_i}(\mathrm{E}[v_i(n,\alpha)] - z_{T_i}(n,\alpha)) = 0$.

Noting that $\frac{\partial\bar{\boldsymbol{z}}(n,\alpha)}{\partial\beta_{ij}}$ is continuous for all i, j, it then follows that $\lim_{n\to\infty}\max_{i\in R,T}|\mathrm{E}[\frac{\partial\bar{\boldsymbol{z}}(n,\alpha)}{\partial\beta_{ij}}] - \frac{1}{|T|}\sum_{k\in T}\frac{\partial\bar{\boldsymbol{z}}(n,\alpha)}{\partial\beta_{kj}}]|$ for all $T \in \mathcal{T}$, which implies the result.

Proof of Proposition 6

Letting $l_{uv}(ik)$ represent the uvth component of $L_{ik}(\alpha) = -M^{-1}(\alpha)\frac{\partial M}{\partial w_{ik}}M^{-1}(\alpha)$, then:

$$l_{uv}(ik) = (-q_k\delta_i r_{ui} + q_i\delta_k r_{uk})r_{iv} + (-q_i\delta_k r_{ui} + q_k\delta_i r_{uk})r_{kv}.$$

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 $z_i < z_j$ then $\sum_{u\in\mathcal{T}_L} r_{iu}(\alpha) > \sum_{t\in\mathcal{T}_L} r_{ju}(\alpha)$ and $\sum_{u\in\mathcal{T}_H} r_{iu}(\alpha) < \sum_{u\in\mathcal{T}_H} r_{ju}(\alpha)$. Noting that $\sum_v l_{uv}(ik) = 0\ \forall u$, this then implies that $|\sum_{v\in\mathcal{T}_L} l_{uv}(ik)| > |\sum_{v\in\mathcal{T}_L} l_{uv}(jk)|$ for all u (and so $|\sum_{v\in\mathcal{T}_H} l_{uv}(ik)| > |\sum_{v\in\mathcal{T}_H} l_{uv}(jk)|$ as well).

We know that $M^{-1}(\alpha)\tilde{\boldsymbol{z}}^S\boldsymbol{q}^T = \bar{z}(\alpha)$, and so $\frac{\partial M^{-1}}{\partial w_{uv}}\tilde{\boldsymbol{z}}^S\boldsymbol{q}^T = \frac{\partial\bar{z}(\alpha)}{\partial w_{uv}}$. We have shown that $|\frac{\partial\bar{z}(\alpha)}{\partial w_{ik}}| > |\frac{\partial\bar{z}(\alpha)}{\partial w_{jk}}|$ if $z_i < z_j$ when $z_i, z_j \in \mathcal{T}_L$. To find the the expected effect of a marginal change in the probability that a type u and a type v agent are connected as $n \to \infty$, we just normalise the above expressions by dividing by $\frac{1}{q_u q_v}$. It then follows

that $|\frac{\partial \bar{z}(\alpha)}{\partial \beta_{ik}}| > |\frac{\partial \bar{z}(\alpha)}{\partial \beta_{jk}}|$ if $z_i < z_j$ when $z_i, z_j \in \mathcal{T}_L$ and $q_i = q_j$. This latter condition is guaranteed by type-symmetry.

Now, consider the case where $z_s > z_t$ and $z_s, z_t \in \mathcal{T}_H$. It follows that $\sum_{u \in \mathcal{T}_L} r_{su}(\alpha) < \sum_{u \in \mathcal{T}_L} r_{tu}(\alpha)$ and $\sum_{u \in \mathcal{T}_H} r_{su}(\alpha) > \sum_{u \in \mathcal{T}_H} r_{tu}(\alpha)$. As with the argument above, this implies that $|\sum_{v \in \mathcal{T}_L} l_{uv}(sl)| > |\sum_{v \in \mathcal{T}_L} l_{uv}(tl)|$for all u. It then follows that, when $q_s = q_t$ that $|\frac{\partial \bar{z}(\alpha)}{\partial \beta_{sl}}| > |\frac{\partial \bar{z}(\alpha)}{\partial \beta_{tl}}|$ if $z_s > z_t$.

Proof of Proposition 7

Let $\tilde{w}^0_{s1} = w^S_{s1} - w^S_{s0}$. As $M^{-1}\tilde{\boldsymbol{z}}^S \boldsymbol{q}^T = \bar{z}(\alpha)$ and that 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, $M^{-1}(\alpha)\frac{\partial \boldsymbol{z}^S}{\partial \tilde{w}^0_{s1}}\boldsymbol{q}^T = \frac{\partial \bar{z}(\alpha)}{\partial \tilde{w}^0_{s1}}$. To find the expected effect of a marginal change in the probability that a type i is connected to a single countermedia source, k, we normalise this expression by dividing by $\frac{1}{q_s q^S_0}$. Given the definition of $K(\alpha)$, $\frac{1}{q_s q^S_0}\frac{\partial \bar{z}(\alpha)}{\partial \tilde{w}^0_{s1}} = (K^{-1}(\alpha)\frac{\partial \boldsymbol{z}^S}{\partial \tilde{w}^0_{s1}}\mathbf{1^T})\frac{1}{q^S_0}$, which then immediately implies $\frac{1}{q_s q^S_0}\frac{\partial \bar{z}(\alpha)}{\partial \tilde{w}^0_{s1}\tilde{w}^0_{s1}} > \frac{1}{q_t q^S_0}\frac{\partial \bar{z}(\alpha)}{\partial \tilde{w}^0_{t1}}$. It then follows that $\frac{\partial \bar{z}(\alpha)}{\partial \hat{\beta}_i(k_1,k_0)} > \frac{\partial \bar{z}(\alpha)}{\partial \hat{\beta}_j(k_1,k_0)}$.

Proof of Proposition 8

Note from the preliminaries section that $M(\alpha) = AD(\alpha)Q$. When $q_i = q$, it follows that $K(\alpha) = qAD(\alpha)$ and so $\frac{1}{q}D^{-1}(\alpha)A^{-1}$. Observe that $\delta_L < \delta_H$ and A and therefore A^{-1} are symmetric. Then it must be that $\sum_k \varsigma_{ik} > \sum_k \varsigma_{jk}$, and so the Proposition holds.

Proof of Theorem 4

To see that the first part of the Theorem, note first that, as stated in Theorem 2, $L(\alpha) = -Y^{-1}A\frac{\partial D(\alpha)}{\partial \alpha}Y^{-1}$ and the matrix $\frac{\partial D(\alpha)}{\partial \alpha}$ is such that if $i, j \in \mathcal{T}_L$ then $d_{ij} < 0$

and if $k \in \mathcal{T}_H$, $d_{ik} > 0$, with $|d_{ij}| = d_{ik}$ if $\theta_j = \theta_k$. It follows that for a type $T_i \in \mathcal{T}_L$, $\sum_{u \in \mathcal{T}_L} r_{iu}(\alpha)$ is increasing in α, and $\sum_{u \in \mathcal{T}_H} r_{iu}(\alpha)$ is decreasing, with the opposite being true for any type $T_j \in \mathcal{T}_H$.

As Proposition 6 shows, the above implies that $|\frac{\partial \bar{z}(\alpha)}{\partial w_{ik}}|$ is increasing in α for all $i \in \mathcal{T}_L$ and $k \in \mathcal{T}_H$ (and similarly, $|\frac{\partial \bar{z}(\alpha)}{\partial w_{jt}}|$ where $T_j \in \mathcal{T}_H$ and $T_t \in \mathcal{T}_L$ is increasing in α too). This then implies that $\Delta_S(\alpha)$ is increasing in α when the type-symmetry assumption holds.

For the second statement, we consider the derivative $\frac{\partial M^{-1}(\alpha)}{\partial w_{ik}} = -M^{-1}(\alpha)\frac{\partial M(\alpha)}{\partial w_{ik}}M^{-1}(\alpha)$ where $i \in \mathcal{T}_L$ and $k \in \mathcal{T}_H$. As per the preliminary results, we write $\frac{\partial M(\alpha)}{\partial w_{ik}} = A\frac{\partial D(\alpha)}{\partial w_{ik}}Q$, where $\frac{\partial D(\alpha)}{\partial w_{ik}}$ is such that its iith and jjth component is -1 and its ijth and jith component is 1, with every other entry equal to 0.

We note that $L_{ik}(\alpha) = -D^{-1}(\alpha)A^{-1}\frac{\partial D}{\partial w_{ik}}Q^{-1}D^{-1}(\alpha)A^{-1}$, which then implies that $\sum_{s \in T_H} q_s l_{st}(ij) = -\frac{\delta_H}{\delta_L}[\sum_{s \in T_L} q_s l_{st}(ij)]$. As $M^{-1}(\alpha)\boldsymbol{z}^S\boldsymbol{q}^T = \bar{z}(\alpha)$, this implies that $|\frac{\partial \bar{z}(\alpha,\tilde{\delta})}{\partial w_{ij}}|$ is increasing in $\frac{\delta_H}{\delta_L}$ and therefore decreasing in $\delta_H - \delta_L$, and, furthermore, $|\frac{\partial \bar{z}(\alpha,\tilde{\delta})}{\partial \alpha}| = 0$ when $\tilde{\delta} = 0$.

At the same time, the proof of Proposition 7 shows that $\frac{\partial \bar{z}(\alpha,\tilde{\delta})}{\partial \hat{\beta}_j(k_1,k_0)} > 0$ for $\tilde{\delta} = 0$ as $M^{-1}(\alpha)\frac{\partial \boldsymbol{z}^S}{\partial \bar{w}^0_{s1}}\boldsymbol{q}^T > 0$ for all α. Hence, even if $\Delta_I(\alpha,\tilde{\delta}) < \Delta_S(\alpha,\tilde{\delta})$ for some $\tilde{\delta}$ and α, by the intermediate value theorem, for any α, exists some $\bar{\delta}$ such that if $\tilde{\delta} < \bar{\delta}$ then $\Delta_I(\alpha,\tilde{\delta}) > \Delta_S(\alpha,\tilde{\delta})$.

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] 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.

[24] Sadler, E. Influence campaigns. Tech. rep., Columbia, 2020.

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

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

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

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

[29] 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.

[30] 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.

[31] 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).