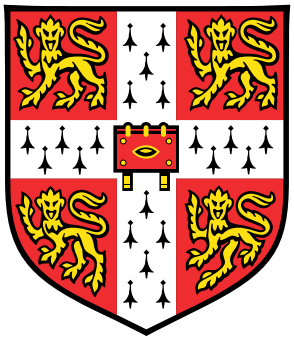
Essays on Social Networks



Yu Yang Tony To

Faculty of Economics

University of Cambridge

This thesis is submitted for the degree of

Doctor of Philosophy

King's College

May 2022

Declaration

This thesis is submitted to the University of Economics in accordance with the requirements of the degree of Doctor of Philosophy. I hereby declare that the thesis is my own and original work. Chapter 2 is joint work with Syngjoo Choi, Seoul National University, Sanjeev Goyal, University of Cambridge, and Frederic Moisan, Emlyon Business School. Chapter 3 is joint work with Sanjeev Goyal, University of Cambridge. All of them can attest to my significant contribution to these projects in terms of original idea, modelling, model-solving, experimental design, statistical analysis, and writing. I also declare that this thesis has not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation contains fewer than 60,000 words including tables, footnotes, bibliography and appendices.

Yu Yang Tony To

May 2022

Acknowledgements

I am indebted to Sanjeev Goyal, my supervisor, and to Matthew Elliott, my advisor, for their continued support and guidance. I have learnt a huge amount from the countless conversations. It has been a privilege and rewarding experience to work with them. Additionally, I would like to thank my co-authors, Syngjoo Choi and Frederic Moisan. They have been a constant source of inspiration, motivation, and knowledge. I am also grateful to Mikhail Safronov, Hamid Sabourian, George Charlson, Jörg Kalbfuss, Fulin Gao, Yi Wei, and other colleagues for their time, useful advice, and valuable input.

I gratefully acknowledge the administrative support of the Faculty of Economics and King's College, and the Cambridge International & King's College Scholarship for funding my doctoral studies. I would also like to thank my teachers and professors who helped build a scientific curiosity in me. They are the unsung heroes that guided my journey to this day. Their passion and dedication pushed me to pursue my interests and dreams.

I have been supported by a large number of friends and colleagues in Cambridge during my time here, and I am grateful to them all. A special word of gratitude goes to my long-time friends: Clemence, Cynthia, Nicolas, Luke, Michael, Keith, Micha, Praniya, Preeti, Angellina, and Jen. Your friendship and companionship have supported me throughout this challenging journey. Thank you for always being there and for being interested in my work.

Finally, I am grateful to my parents, without whom none of this would have been possible. Thank you, Mom and Dad, for giving me strength and always believing in me. Your trust and love have made me who I am today. I dedicate my doctoral thesis to my grandpas and grandmas.

Abstract

This thesis consists of three essays on the economics of social networks. It broadly deals with understanding the value of social connections on favour exchange and information exchange. Social networks facilitate trust, learning, and communication, all crucial in the modern online environment. Examining the effects of network structure provides new tools and insights on decision-making and behaviour.

Chapter 1 develops a model of repeated favour exchange on social networks where individuals choose between allocating the opportunity to the expert (market action) or a friend (favouritism action). Assuming favouring a friend reduces one's payoff, favouritism cannot be sustained in a stage game. However, by introducing a grim-trigger strategy where a selective group of individuals favour each other, favouritism can be sustained in an infinitely repeated game. In particular, the maximum clique of the network defines favouritism behaviour that is coalition-proof where no group of individuals have incentives to deviate collectively. While aggregate surplus increases with network connectivity, it decreases with the number of favouritism-practising agents. Additionally, favouritism exacerbates payoff inequality that arises from degree inequality: Favouritism players cooperate to extract a large portion of the aggregate surplus at the expense of market players, creating a negative externality on the economy.

Chapter 2 conducts an experiment to study the impact of network structure on opinion formation. At the start, subjects observe a private signal and then make a guess. In subsequent periods, subjects observe their neighbours' guesses before guessing again. Inspired by empirical research, we consider three canonical networks: Erdös-Rényi, Stochastic Block and Royal Family. We find that a society with 'influencers' is more likely to arrive at an

incorrect consensus and that one with ‘network homophily’ is more likely to persist with diverse beliefs. These aggregate patterns are consistent with individuals following a DeGroot updating rule.

Chapter 3 studies incentives for verifying information in social networks. Agents derive value from sharing correct information and suffer a reputational loss from sharing false information. So agents can undertake costly verification prior to sharing information. We show incentives for verification are increasing in degree. This implies that information quality is increasing in average degree and is higher in more egalitarian networks. We then introduce an external agent whose goal is to maximise views through a choice of news source accuracy. We find that denser networks lead to higher accuracy when information accuracy is either expensive or cheap, and sparse networks lead to more accurate information otherwise.

JEL Codes: C73, D63, D83, D85

Table of contents

Chapter 1

Repeated Favouritism on Networks

1.1 Introduction

Favouritism refers to the practice of giving unfair preferential treatment to a favoured person or group. This idea encompasses concepts like nepotism (favouritism based on kinship) and cronyism (favouritism from positions of authority). While cronyism is commonly a feature endemic to developing countries, this view is challenged by recent financial and political scandals based on collusion networks. Examples of favouritism in this setting include referring incompetent friends to job vacancies or hiring poorly-qualified family members in managerial roles.[1] These behaviours distort the matching process and are only beneficial to a few selected individuals. The loss in social welfare and inequality generated by favouritism motivates this research.

Across different cultures and economies, social connections play a key role in facilitating favouritism. Examples include "guanxi" in Chinese society, "blat" in Russian culture, and "old boy's network" among the British elite.[2] As the law prohibits favouritism, these arrangements are enforced without explicit contracts or regulations. One form of informal

[1] Barr and Oduro (2002), Bandiera et al. (2009) find referrals that favour friends and relatives distort the recruitment process and are a source of inefficiency and inequity.

[2] McDonald (2011) shows that social capital flows through gendered and racialised networks, creating labour market inequalities. Ledeneva and Ledeneva (1998) discuss how blat favours are normally provided to "svoim", one of us. Karhunen et al. (2018) address corruption as a negative reciprocal practice in social networks such as guanxi and blat/svyzai.

enforcement is through quid pro quo where favours are reciprocated. This paper aims to examine the effect of network connections in enforcing favour exchange.

I propose a model of favouritism on a social network. Economic opportunities arrive over time to a random individual, the *principal*, who must realise this opportunity with another player. The match quality of each opportunity differs among players. One individual, the *expert*, yields the most productive outcome while all others are non-experts. If the principal has a neighbouring expert, she always matches with the expert. If the principal does not have a neighbouring expert, she must decide whether to practise *market behaviour*, matching with the non-neighbouring expert, or practise *favouritism*, matching with an inefficient neighbour. I define favouritism as allocating resources towards friends, in particular, diverting resources away from their efficient uses.

In the stage game, principals will not practise favouritism because they earn lower payoffs than matching with the expert. In the repeated game, favouritism can be sustained based on the expectations of neighbours returning favours. I propose a grim trigger strategy with two types of players: Market players who always practise market behaviour; Favouritism players who only provide favours among other favouritism-practising neighbours and revert to market behaviour if any favouritism players deviated in the past. The network has three effects on the incentives of a favouritism-practising agent: First, a higher degree (and more favouritism neighbours) reduces the inefficiency from one sustaining favouritism. Second, having poorly connected (favouritism-practising) neighbours reduce the number of competitors for the same favour. Third, having fewer non-neighbouring favouritism players reduce the number of opportunities redirected away from oneself.

All players practising market behaviour is always a subgame perfect equilibrium. However, in a network with multiple equilibria, this *Pure market* strategy profile is dominated: highly connected players can communicate with their highly connected neighbours to collectively deviate to favouritism, leaving the poorly-connected players to practise market behaviour. We find that the aggregate surplus increases in the total number of links but decreases in the proportion of favouritism players. Moreover, favouritism players cooperate

to extract a large portion of aggregate surplus at the expense of market players. As a result, favouritism amplifies income inequality due to heterogeneous degree endowment.

Related Literature. There is a vast economic literature studying the impact of social network on favour exchange (Möbius, 2001, Hauser et al., 2008), risk-sharing (Bramoullé and Kranton, 2007, Bloch et al., 2008, Ambrus et al., 2014) and trust (Karlan et al., 2009). Network measures such as clustering, closure, and support have emerged to discuss the role of the network in fostering cooperation. The existing literature studies coordination on the social optimal action where sustaining favour exchange increases aggregate surplus. This paper, on the other hand, aims to address the negative consequences of cooperation. Favouritism redirects opportunities from their efficient destination and lowers the aggregate social surplus.[3] My paper contributes to this strand of literature by discussing how favour exchange that reduces aggregate surplus is facilitated by network degree inequality.

The grim trigger strategy in sustaining cooperation has been widely studied.[4] A recent paper by Jackson et al. (2012) studies a game of favour exchange where only connected players are asked to provide a favour. They propose a grim trigger strategy where the connection is deleted if the request is rejected. Players are allowed to renegotiate after the punishment phase to reform links. They show that a *minimal-connected cliques* network is the *renegotiation-proof equilibrium* robust against *social contagion* — where one defect causes link deletions to propagate through the network. The favouritism game in this paper differs by studying the strategic choice of whom to match with. The principal's decision is between favouring inefficient neighbours or matching with the expert for the efficient output. I propose a grim trigger strategy where the community punishes the defectors through market reversion.[5] I find that the favouritism sustaining subgroup is characterised by the opposing

[3]The welfare implications of favour exchange among social connections are mixed in the existing empirical literature. For example, Brogaard et al. (2014) suggest that economics journals editors use their connections to identify higher-quality papers; Zinovyeva and Bagues (2015) find professors in Spain who were connected to their promotion jury publish less after promotion; Bramoullé and Huremović (2017) find both (distortionary) reciprocal favour exchanges and information efficiencies at work.

[4]For a detailed literature review in repeated games on networks, see Nava (2016).

[5]An alternate grim trigger punishment is briefly discussed in our paper called ostracism. In the literature of network-based cooperation in a repeated game, ostracism is defined as targeted link deletions against an individual (Haag and Lagunoff, 2006, Lippert and Spagnolo, 2011, Ali and Miller, 2016).

tension between wanting more (favouritism-practising) neighbours but less (favouritism-practising) neighbours of neighbours. This yields novel findings on the role of *maximum cliques* — the largest complete subgraphs — in sustaining favouritism.

The closest paper to my model is Bramoullé and Goyal (2016), which studies favouritism as an exchange of favours between two distinct groups. Players practising favouritism are willing to bear the cost of lower productivity (from hiring less competent group members instead of hiring productive outsiders) because they fear the threat of "losing out on favours" from their group in the future. My research extends upon their paper by studying favouritism in a network setting. In my model, an individual's direct neighbours constitute as her potential favouritism group and individuals are heterogeneous in degree. This generalises Bramoullé and Goyal (2016)'s model beyond groups of homogeneous players. Their model provides insights on the effect of group sizes on favouritism, whereas this paper examines the effect of players' degrees. In equilibrium, high degree centrality and low degree inequality emerge as the main preconditions for favouritism.

The main finding of the cooperative core network can be seen in the networks literature.[6] Gagnon and Goyal (2017) show that when the network action and market action are strategic substitutes, players within the q-core of the network adopt the network (cooperative) action while others adopt the market action. In my model, the cooperative core structure emerges from the inequality in the underlying network where (i) the highly-connected individuals coordinate to form the favouritism-sustaining community against the outsiders, and (ii) the (poorly connected) outer community cannot credibly punish the community for deviating. This is in line with the result from Kets et al. (2011) where they demonstrate how inequality is directly influenced by the inability of the poor to form viable coalitions.[7]

The paper proceeds as follows. Section 1.2 introduces the stage game and the infinitely repeated game of favouritism on a network. Section 1.3 explores the relationship between

[6]Haag and Lagunoff (2006) find that when the discount factors are known to the planner, the optimal network design for a repeated Prisoner's Dilemma game is a cooperative "core" and an uncooperative "fringe".

[7]Bernheim et al. (1987) introduced Coalition-Proof Nash equilibrium as a refinement of the Nash set. Kahn and Mookherjee (1992) extend the definition using stable sets to characterise equilibria in infinite games. Ambrus et al. (2014) then apply the technique to a network game of risk-sharing. They show that the outer community can punish the coalition by ostracizing the coalition. The maximum punishment the outer community can inflict equals the link capacity of the perimeter of the coalition.

the network structure and the incentives for practising favouritism. Section 1.4 introduces coalition deviations as a method of equilibrium selection. Section 1.5 examines the aggregate surplus and inequality implications of favouritism. Lastly, Section 1.6 discusses potential research directions, limitations and concludes.

1.2 Model

A finite set of n players, $N = \{1, ..., n\}$, are connected in a social network described by an undirected graph $g \in G$, with links $g_{ij} \in \{0, 1\}$. $g_{ij} = 1$ indicates that player i and j are linked under the network g and $g_{ij} = 0$ otherwise. The neighbourhood of player i denotes the set of players that are connected to i, $N_i(g) = \{j | g_{ij} = 1\}$. $g_{ii} = 0$ by convention. The degree of player i is the number of neighbours she has, denoted by $d_i(g) = |N_i(g)|$. Non-neighbours denote the set of players that individual i is not connected to, $N_i^c(g) = \{j \neq i | g_{ij} = 0\}$. On the network g, I study an infinitely repeated game in which players play the stage game described below.

1.2.1 Stage game

At the start of the game, an opportunity arises. Nature randomly selects a *principal* $m \in N$ to receive this opportunity. Nature then selects an *expert* e from the remaining players $N \backslash \{m\}$. The model assumes that each player has an equal and independent chance of being selected. The probability of any pair of principal and expert being selected equals $p = \frac{1}{n} \frac{1}{n-1}$.

In order to realise this opportunity, principal m must choose a *respondent* to offer the opportunity to, $a_m \in N \backslash \{m\}$. The principal incurs a search cost $c \geq 0$ to offer the opportunity to non-neighbours. The cost is otherwise waived when offering to neighbours.[8] The respondent then decides whether to accept or reject the offer, $r_{a_m} \in \{1, 0\}$. If the respondent rejects, $r_{a_m} = 0$, the opportunity is lost resulting in no output and zero payoffs. If

[8]The search cost is modelled from the employers' perspective. Firms incur a search cost or monitoring cost to determine the worker's productivity unless the referee is connected to the employee. Existing literature has discussed the role of social contacts in minimizing search costs for jobs (Calvó-Armengol, 2004, Mortensen and Vishwanath, 1994, Galeotti and Merlino, 2014).

the respondent accepts, the principal and respondent are *matched*. If the principal matches with the expert, they produce the efficient production output, normalised to 1. If the principal matches with a non-expert, output equals $L \leq 1$ which reflects the importance of match quality. Under a general rule of surplus division, the share of output given to an expert equals $\alpha > 0$, and for a non-expert, $\beta > 0$. Thus, the payoff of a principal m equals:

$$u^m = \begin{cases} 1-\alpha & \text{if matching with the neighbouring expert} \\ 1-\alpha-c & \text{if matching with the non-neighbouring expert} \\ L-\beta & \text{if matching with a neighbouring non-expert} \\ L-\beta-c & \text{if matching with a non-neighbouring non-expert} \end{cases} \tag{1.1}$$

Suppose the principal earns less matching with the expert, $1-\alpha \leq L-\beta$, then she has no incentive to offer the opportunity to a neighboring expert, let alone searching for a non-neighboring expert. So her dominant strategy is to match with a neighboring non-expert (who then always accepts the offer). This equilibrium is not of interest as "favouritism" is the efficient strategy and it is always sustained in equilibrium. Instead, for the rest of the paper, assume (i) the expert earns more than the non-experts, $\alpha > \beta$, and (ii) the principal earns more when matched with an expert, $1-\alpha > L-\beta$.

Assumption 1.1. *The output from an efficient match is 1, and otherwise* $L \leq 1$.
The wage to the expert is higher than non-experts, $\alpha > \beta$.
The payoff of the principal is higher when matches with the expert, $1-\alpha > L-\beta$.

Proposition 1.1. *Suppose Assumption 1.1 holds. In the stage game:*

(i) If $c < (1-\alpha)-(L-\beta)$*, the unique subgame perfect equilibrium is* $\{a_m = e, r_{a_m} = 1\}$.

(ii) If $c \geq (1-\alpha)-(L-\beta)$*, the unique subgame perfect equilibrium is:*

$$a_m = \begin{cases} e & \text{if } e \in N_m(g) \\ j \in N_m & \text{otherwise} \end{cases}, \quad r_{a_m} = 1.$$

Proof. In a stage game, only the principal m and the respondent a_m have an action to take. The respondent earns 0 if she rejects the offer so she always accepts, $r_{a_m} = 1$. Given that the respondent always accepts, we look at the principal's incentives: if the principal has a neighbouring expert, it is optimal to match with him; if the principal has no neighbouring expert, searching for the expert would earn $1-\alpha-c$ whereas matching with a neighbour would earn $L-\beta$. If searching for the expert is cheaper, then the principal always matches with the expert (case (i)). Otherwise, she only matches with her neighbours, only prioritizing the expert when he is a neighbour (case (ii)). □

I call "favouring a friend" matching with a neighbouring non-expert instead of searching for the (non-neighbouring) expert. Favouring a friend results in an aggregate social surplus of L, while searching for and matching with the expert results in a surplus of $1-c$. If $L \geq 1-c$, then favouring a friend is social surplus maximising, and favouritism can be sustained in a stage game ($L \geq 1-c$ implies $c \geq (1-\alpha)-(L-\beta)$). Instead, this paper is interested in the opposite case where favouritism reduces social surplus, $L < 1-c$. The rest of the paper assumes $c < (1-\alpha)-(L-\beta)$ and proposes a strategy where favouritism can be sustained through reciprocity in a repeated game despite it being aggregate surplus reducing.

Assumption 1.2. *Search cost is sufficiently small,* $c < (1-\alpha)-(L-\beta)$.

1.2.2 Repeated game

This section studies the infinitely repeated game in which players play the stage game described above in discrete periods indexed by $t = 1, 2, \ldots$. Players discount future payoffs at a common discount factor $\delta \in (0,1)$. In each period t, players seek to maximise the discounted sum of expected payoffs,

$$u_{i,t} + E\left[\sum_{s=1}^{\infty} \delta^s u_{i,t+s}\right]$$

where $u_{i,t}$ is the payoff received by player i and time t, conditional on the strategy profile of all players.

Let us first define some notations and terminologies for the repeated game. In any period $t \geq 1$, nature selects a principal $m_t \in N$ and an expert $e_t \in N \backslash \{m_t\}$ uniformly at random. The principal then offers the opportunity to a respondent, $a_{m_t} \in N \backslash \{m_t\}$, who accepts or rejects the offer, $r_{a_{m_t}} \in \{1,0\}$. Payoffs are realised before the next period arrives and the process repeats. Define $p_t = \{m_t, e_t, a_{m_t}, r_{a_{m_t}}\}$. At time t, the history of the game consists of nature's choice of principal and expert, and the decisions of principals and respondents in all prior periods. Define history at time t as $h_t = \{p_1, p_2, ..., p_{t-1}\}$. Let H_t be the set of possible histories at time t. The set of all possible histories is $H \equiv \bigcup_{t=0}^{\infty} H_t$. The strategy of a principal at time t is $s_{m_t} : H_t \times G \to N \backslash \{m_t\}$. The strategy of a respondent chosen by m_t is $s_{a_{m_t}} : H_t \times G \to \{1,0\}$. All other players have no choice of action at time t.

I propose a strategy in the repeated game called *market behaviour* (M): the principal m_t offers the opportunity to the expert for all histories, $s_{m_t}(.) = e_t$; the respondent a_{m_t} accepts the opportunity for all histories, $s_{a_{m_t}}(.) = 1$. In the stage game, since $c < (1-\alpha) - (L-\beta)$, the principal matching with the expert is a subgame perfect equilibrium. Therefore, in the infinitely repeated game, all players practising market behaviour is a subgame perfect equilibrium. I call this equilibrium the *Pure market equilibrium*.

Proposition 1.2. *Suppose Assumption 1.1 and 1.2 hold. For all networks g, all players practising market behaviour is a subgame perfect equilibrium of the repeated game.*

I propose another strategy in the repeated game called *favouritism* (F): the respondent a_{m_t} accepts the opportunity for all histories, $s_{a_{m_t}}(.) = 1$; when there is a neighbouring expert, the principal m_t offers the opportunity to the expert for all histories; but when there is no neighbouring expert, she offers the opportunity to one of her neighbours for all histories. Formally,

$$s_{m_t}(.) = \begin{cases} e_t & \text{if } e_t \in N_{m_t}(g) \\ j \in N_{m_t} & \text{otherwise} \end{cases}$$

When the principal has no neighbouring expert, deviating from favouritism to market behaviour increases her current payoff from $L-\beta$ to $1-\alpha-c$. Hence, all players practising favouritism strategy is *not* a subgame perfect equilibrium.

So instead, I propose a grim trigger strategy profile denoted by s^*. The players are partitioned into two groups, S_M and S_F. For all histories, the respondent accepts the offer, and the principal in S_M practises market behaviour, as defined previously. The principal in S_F practises a grim trigger favouritism strategy: In period $t = 1$, the principal in S_F practise favouritism — offering the opportunity to the expert when she has a neighbouring expert, and otherwise, randomly offering the opportunity to a neighbour in S_F. Formally,

$$s_{m_t}(.) = \begin{cases} e_t & \text{if } e_t \in N_{m_t}(g) \\ j \in N_{m_t} \cap S_F & \text{otherwise, with probability } \frac{1}{|N_{m_t} \cap S_F|}. \end{cases}$$

In all subsequent periods, if all principals in S_F have practised favouritism in all prior periods, the principal in S_F will continue to practise favouritism. If any principal in S_F has ever deviated to practising market behaviour in the history, she will practise market behaviour for the rest of the game. Any other deviation will not trigger the punishment phase.[9]

Intuitively, if no player has deviated so far, principals in S_F will practise favouritism with others within S_F, principals in S_M will practise market behaviour, and respondents will accept the offers. If a principal in S_F ever deviates from favouritism, i.e., not matching with a favouritism-practising neighbour when she has no neighbouring expert, all principals will revert to market behaviour.

There are multiple sophisticated strategies in selecting a neighbour to favour. It is reasonable for players to provide favours only to those who will return these favours. Additionally, favours are often exchanged indirectly among a collective rather than exclusively with one individual. So, I restrict the attention to the symmetric strategy where favouritism-practising principals randomly favour neighbours who will return their favours.[10] The punishment of reversion to market behaviour is motivated by the fact that if one refuses to coordinate and offers no favours, this could collapse the social norm of exchanging favours. The society

[9] Note that s^* encompasses multiple strategy profiles depending on the partition S_F, S_M. For example, $S_F = \varnothing$ is the Pure market equilibrium.

[10] An alternative grim trigger strategy can be constructed using *pairwise favouritism* which redefines favouritism as only favouring a single neighbour. In Appendix A.1, I argue that it is easier to sustain random favouritism (in the baseline model) than pairwise favouritism. Individuals also earn more by practising random favouritism.

then reverts to the socially optimal, strategically dominant action, namely the Pure market equilibrium.[11]

Theorem 1.1. *Suppose Assumption 1.1 and 1.2 hold. For all networks g, under the strategy profile s^*, there exists a non-empty set of players practising favouritism in a subgame perfect equilibrium, $S_F \neq \varnothing$, if and only if all players in S_F satisfy the following condition:*

$$-x+\frac{\delta}{1-\delta}p\left[(n-1-d_i)(-x)-|N_i^c\cap S_F|\alpha+\beta\sum_{j\in N_i\cap S_F}\frac{n-1-d_j}{|N_j\cap S_F|}\right]\geq 0 \tag{1.2}$$

where $x=(1-\alpha-c)-(L-\beta)$.

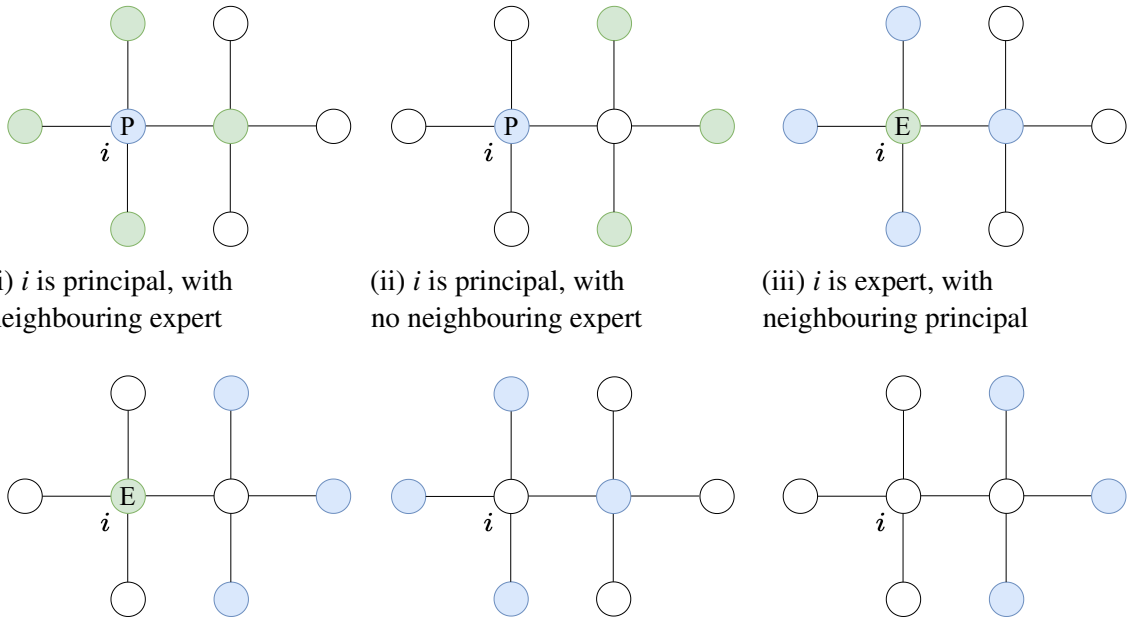
Condition (1.2) is satisfied when a principal in S_F has no profitable one-shot deviation from the strategy profile s^*. If all principals in S_F satisfy this condition, the strategy profile s^* is a subgame perfect equilibrium where favouritism is sustained. To prove this, I first calculate the probabilities of principal-expert allocation, then evaluate the expected payoffs between practising favouritism and market behaviour, and lastly show that there is no profitable one-shot deviation for any principal or respondent.

Proof. In each period, from the perspective of player i, there are six mutually exclusive and collectively exhaustive cases of principal and player allocations:

(i) Player i is the principal, with a neighbouring expert
(ii) Player i is the principal, with no neighbouring expert
(iii) Player i is the expert, with a neighbouring principal j
(iv) Player i is the expert, with a non-neighbouring principal j
(v) Player i is not the principal nor the expert, but has a neighbouring principal j
(vi) Player i is not the principal nor the expert, and has a non-neighbouring principal j

To illustrate, consider an 8-player network with individual i of degree 4 (Figure 1.1). In case (i), i is the principal (highlighted in blue). There are four possible locations of the expert

[11] One alternative punishment is *ostracism* where the deviator is removed from the favouritism group and never receives favours again. Because the rest of the favouritism group continues to sustain favouritism in the ostracism punishment phase, opportunities are also redirected away from the deviator when she is the expert. Thus, it is easier to sustain favouritism under the ostracism punishment than the market-reversion punishment. However, ostracism punishment reduces the size of S_F, which means the condition (1.2) to sustain favouritism may no longer hold. This could lead to a cascading collapse of the favouritism-practising groups.



(i) i is principal, with neighbouring expert

(ii) i is principal, with no neighbouring expert

(iii) i is expert, with neighbouring principal

(iv) i is expert, with non-neighbouring principal

(v) i is non-expert, with neighbouring principal

(vi) i is non-expert, with non-neighbouring principal

Fig. 1.1 Networks illustrating principal and expert locations from player i's perspective

Table 1.1 Principal-expert allocation, player i's payoffs, and probability of receiving payoff

	Cases	$a_i = M$	$a_i = F$	$a_j = M$	$a_j = F$	Probability
i)	i is principal, with neighbouring expert	$1-\alpha$	$1-\alpha$	—	—	pd_i
ii)	i is principal, with no neighbouring expert	$1-\alpha-c$	$L-\beta$	—	—	$p(n-1-d_i)$
iii)	i is expert, with neighbouring principal j	—	—	α	α	pd_i
iv)	i is expert, with non-neighbouring principal j	—	—	α	—	$p\lvert N_i^c \cap S_M\rvert$
v)	i is not expert, with neighbouring principal j	—	—	—	β	$p \sum_{j \in N_i \cap S_F} \frac{n-1-d_j}{\lvert N_j \cap S_F\rvert}$
vi)	i is not expert, with non-neighbouring principal j	—	—	—	—	—

that would give i a neighbouring expert (highlighted in green). So the probability of case (i) occurring is $\frac{1}{8} \cdot \frac{4}{7}$. We next calculate the probabilities of each case for a general network.

Probabilities. The probability of player i being selected as the principal equals $\frac{1}{n}$ and the probability of her having a neighbouring expert equals $\frac{d_i}{n-1}$. Therefore, case (i) occurs with probability pd_i where $p = \frac{1}{n}\frac{1}{n-1}$. Similarly, case (iii) occurs with probability pd_i. In both cases, regardless of whether the principal is practising favouritism or market behaviour, she always offers the opportunity to the neighbouring expert. Player i receives $1-\alpha$ as the principal in case (i) and received α as the expert in case (iii).

In case (ii), principal i has no neighbouring expert with probability $p(n-1-d_i)$, where $(n-1-d_i)$ is the number of non-neighbouring players that are potentially the expert. She earns $1-\alpha-c$ if she practises market behaviour and earns $L-\beta$ if she practises favouritism.

In case (iv), player i is the expert with no neighbouring principal. If the non-neighbouring principal practises favouritism, he will favour his friends over the expert and player i will not receive any payoffs. Only when the non-neighbouring principal practises market behaviour will expert i receive payoff α. This occurs with probability $p|N_i^c \cap S_M|$, where $|N_i^c \cap S_M|$ is the total number of non-neighbours of i that is practising market behaviour.

In case (v), player i as the non-expert only receives payoffs when she has a neighbouring principal j who has no neighbouring expert, practises favouritism, and matches with i. If j has a neighbouring expert or if he practises market behaviour, he will always match with the expert. The principal j has no neighbouring expert with probability $\frac{n-1-d_j}{n-1}$. If he practises favouritism by selecting a respondent randomly among his favouritism neighbours $N_j \cap S_F$, he matches with i with probability $\frac{1}{|N_j \cap S_F|}$. Overall, i receives favours from j and earns payoff β with probability $\frac{1}{n}\sum_{j \in N_i \cap S_F} \frac{n-1-d_j}{n-1}\frac{1}{|N_j \cap S_F|} = p\sum_{j \in N_i \cap S_F} \frac{n-1-d_j}{|N_j \cap S_F|}$.

Lastly, in case (vi), the non-expert i will earn no payoff as there is no neighbouring principal. Table 1.1 summarises the probabilities of each case and the associated payoffs received by player i conditional on the principal's action. Note that if instead the respondent rejects, both the principal and the respondent earn no payoffs.

Expected payoffs. We now compute the expected payoff of players conditional on their actions. Suppose a subset of players S_F practises favouritism and S_M practises market

behaviour. The stage game expected payoff of a player $i \in S_F$ equals:

$$p\left[d_i(1-\alpha)+(n-1-d_i)(L-\beta)+d_i\alpha+|N_i^c\cap S_M|\alpha+\beta\sum_{j\in N_i\cap S_F}\frac{n-1-d_j}{|N_j\cap S_F|}\right]. \quad (1.3)$$

Whereas the expected payoff of a player in $i \in S_M$ equals:

$$p\Big[d_i(1-\alpha)+(n-1-d_i)(1-\alpha-c)+d_i\alpha+|N_i^c\cap S_M|\alpha\Big]. \quad (1.4)$$

Incentives to deviate. Next, I prove that no player has a profitable one-shot deviation from the strategy profile s^* when condition (1.2) holds. First, if the respondent deviates and rejects the offer, her payoff in the current period lowers from α/β to zero. Second, if the principal (in S_F or S_M) with a neighbouring expert deviates to match with a non-expert, her immediate payoff lowers from $1-\alpha$ to $L-\beta$. Third, the market principal (in S_M) with no neighbouring expert earns $1-\alpha-c$ in the current period (from matching with the non-neighbouring expert). If she deviates to matching with a neighbouring non-expert (or with a non-neighbouring non-expert), she will earn lower immediate payoffs $L-\beta$ (or $L-\beta-c$). Since all these deviations do not trigger the punishment phase, the expected future payoffs are unchanged. Therefore, for all these players, there is no profitable one-shot deviation.

At last, we look at the incentives of a favouritism principal (in S_F) when she has no neighbouring expert. If she practises favouritism, she earns the unproductive payoff, $L-\beta$, but expects her favouritism-practising neighbours, $N_i\cap S_F$, to return favours in the future. Fixing the strategy profiles of others s^*_{-i}, her expected payoff at period t of not deviating equals:

$$L-\beta+\sum_{t'=1}^{\infty}\delta^{t'}\underbrace{p\left[d_i+(n-1-d_i)(L-\beta)+|N_i^c\cap S_M|\alpha+\beta\sum_{j\in N_i\cap S_F}\frac{n-1-d_j}{|N_j\cap S_F|}\right]}_{\text{Expression (1.3) when } S_F \text{ is not an empty set}} \quad (1.5)$$

Instead, if she deviates and practises market behaviour by matching with a non-neighbouring expert, she earns $1-\alpha-c$ in the current period but forgoes all future favours as all players

revert to market behaviour. Her expected payoff at period t for deviating to M equals:

$$1-\alpha-c+\sum_{t'=1}^{\infty}\delta^{t'}\underbrace{p\Big[d_i+(n-1-d_i)(1-\alpha-c)+|N_i^c|\alpha\Big]}_{\text{Expression (1.4) when } S_F \text{ is an empty set,}} \tag{1.6}$$

This principal has no incentive to deviate from favouritism to market behaviour if and only if the expected payoff of continuing to practise F is greater or equal to the expected payoff of deviating, i.e., eq. (1.5) is weakly greater than eq. (1.6). Any other one-shot deviation is not profitable because it reduces payoffs in the current period and triggers the punishment phase, thus eliminating all future favours.

The punishment of reverting to the Pure market equilibrium is credible because it is a subgame perfect equilibrium. By the one-shot deviation principle, the strategy profile s^* is a subgame perfect equilibrium as long as all players in S_F satisfy inequality (1.2). □

1.3 Networks and Favouritism incentives

In this section, we explore how the network affects the incentives for practising favouritism. I denote $x \equiv (1-\alpha-c)-(L-\beta)$ as the difference in current payoffs between practising favouritism and market behaviour for a principal with no neighbouring expert. The subgame perfect equilibrium (SPE) condition (1.2) can be rewritten as:

$$-\underbrace{x(1+\frac{\delta}{1-\delta}p(n-1-d_i))}_{\text{current and future inefficiency losses}}-\underbrace{\frac{\delta}{1-\delta}p|N_i^c\cap S_F|\alpha}_{\text{future wage lost}}+\underbrace{\frac{\delta}{1-\delta}p\beta\sum_{j\in N_i\cap S_F}\frac{n-1-d_j}{|N_j\cap S_F|}}_{\text{future favours gained from favouritism}}\geq 0 \tag{1.2'}$$

The first term is the efficiency loss from practising favouritism when one has no neighbouring expert. It includes the loss in the current period and losses in all future periods when principal i has no neighbouring expert (with probability $p(n-1-d_i)$). A higher degree reduces this term in magnitude because each additional connection reduces the likelihood of one having no neighbouring expert.

The second term is the wages lost while player i is the expert due to non-neighbours favouring their friends. The favouritism players outside of i's neighbourhood will redirect opportunities towards themselves and away from i, which lowers i's incentive to sustain favouritism. This highlights an important intuition of practising favouritism: opportunities only flow into the local favouritism group but not outwards. If there are many non-neighbours practising favouritism, i would rather deviate and revert to Pure market equilibrium.

The third term is the expected payoff of future favours gained from practising favouritism. Having more favouritism-practising neighbours increases one's likelihood to receive a favour. However, neighbour j having a higher degree reduces the chance of him favouring a friend because it increases the chance of him having a neighbouring expert. Additionally, since favours are randomly distributed among favouritism-practising neighbours, neighbour j having more favouritism-practising friends $|N_j \cap S_F|$ would increase the competition for his favour and dilute the chance of j matching with i.

Proposition 1.3. *Suppose Assumption 1.1 and 1.2 hold. For all networks g, $S_F = N$ is not a subgame perfect equilibrium of the repeated game.*

All players practising favouritism cannot be sustained in equilibrium. The formal proof is left in Appendix A.3. The intuition is that the least connected player earns the fewest favours from her neighbours due to high competition. Additionally, her poor connectivity means that most opportunities are redirected towards other favouritism players, away from her. Therefore she has incentives to deviate, trigger the punishment phase, and revert the network to practising market behaviour instead.

Overall, a player's incentive to sustain favouritism increases in her degree and the number of neighbours practising favouritism. However, the incentive decreases in the number of non-neighbours practising favouritism, the degree of her favouritism-practising neighbour, and the number of favouritism friends said neighbour has. Note that the incentive of having more friends (practising favouritism) but fewer friends of friends (practising favouritism) is symmetric across all players practising favouritism. The tension created by these opposing incentives means that linked players with highly unequal degrees cannot practise favouritism together. As a result, the favouritism-sustaining group S_F is often a *regular* subgraph — a

graph induced by the subset of nodes where all have constant degree — or a *clique* — all nodes in the subgroup are connected.

1.3.1 Examples of equilibria

I illustrate the relationship between network structure and favouritism incentives with the help of some examples. The parameters are as follows: $n = 10, L = 0.8, \alpha = 0.5, \beta = 0.4, c = 0.01, \delta = 0.9$. Let us look at core-periphery networks and regular networks.

Core-periphery network. Consider a 10-player core-periphery network where players 1 to 5 form a completely connected core and the rest are the periphery players, each linked with one distinct core player. There are only two equilibria under strategy profile s^*: the Pure market equilibrium and the equilibrium where $S_F = \{1,2,3,4,5\}$. They are represented in Figure 1.2 with $S_F = \{1,2,3,4,5\}$ highlighted in blue.

Let us compare the payoffs of players between the two equilibria. In the Pure market equilibrium, the ex-ante one-period payoff of player i equals:

$$p\Big[(n-1)-(n-1-d_i)c\Big] = p\Big[(n-1)(1-c)+d_ic\Big]. \tag{1.7}$$

The core players earn $8.96p$ while periphery players earn $8.92p$, where $p = \frac{1}{n}\frac{1}{n-1}$. Poorly-connected players earn lower payoffs because they have a lower chance of having a neighbouring expert, hence they incur the search cost more often. In contrast, in the favouritism-sustaining equilibrium, core F-players earn $10.2p$ while periphery M-players earn $6.92p$. This is because the favouritism group S_F redirects opportunities towards themselves which would have otherwise gone to an outsider expert in the Pure market equilibrium. Note that the aggregate payoff over all players equals $89.4p$ in the Pure market equilibrium but only $85.6p$ in the favouritism-sustaining equilibrium. The favouritism group extracts a large proportion of the reduced aggregate surplus at the expense of the outsiders. I further explore this inequality in Section 1.5.

Regular network. Now suppose there are two 10-player regular networks of degree 4. One is a regular ring lattice — a graph where vertices are connected to four neighbours, two on each

side. The other network comprises two disjointed 5-player complete subgraphs — a graph where every pair of vertices is connected by an edge. The Pure market equilibrium is the unique equilibrium of the regular ring lattice network. But for the 5-complete network, apart from the Pure market equilibrium, there are two more equilibria, namely $S_F = \{1,2,3,4,5\}$ and $S_F = \{6,7,8,9,10\}$, i.e., one of the 5-complete subgraphs (Figure 1.4). In both networks, all players earn $8.95p$ in the Pure market equilibrium. In contrast, all players in S_F earns $10.5p$ while others in S_M earns $6.45p$.

Despite the two networks having the same degree and degrees of neighbours, the ring lattice network cannot sustain favouritism in equilibrium while the 5-complete network can. The intuition is that we cannot form a sufficiently dense subgraph with low degree inequality from the ring lattice network. As shown in Proposition 1.3, the entire network cannot practise favouritism together in equilibrium. So either the favouritism group is too small for it to be profitable to sustain, or the favouritism group has highly unequal degrees which cannot practise favouritism together. On the other hand, in the 5-complete network, the group $\{1,2,3,4,5\}$ forms a clique with high connectivity and low degree inequality. Players within the clique have high incentives to sustain favouritism in equilibrium.

1.4 Coalition-proof equilibrium

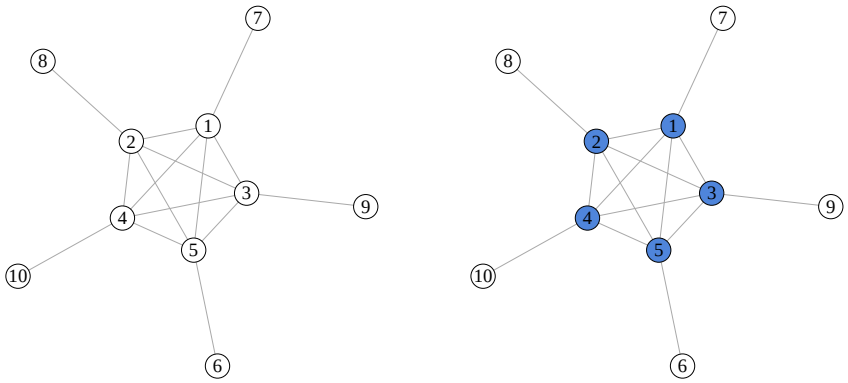
Previously, the equilibrium under s^* only considers an individual's incentive to deviate. Some equilibria may not be stable when players can collectively renegotiate their strategy. Hence, I propose a refinement of the subgame perfect equilibrium such that no group of players has incentives to deviate collectively.

Definition 1.1. *A coalition is a non-empty connected set of players. A strategy profile s' dominates strategy profile s, $s' \succ s$, if and only if there exists a coalition $C \subseteq N$ such that:*

(i) $s_j = s'_j$ *for all* $j \in N \backslash C$,

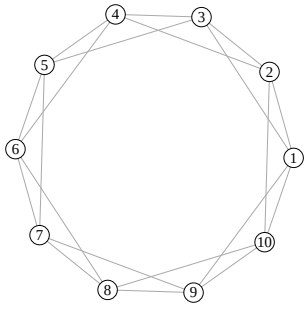
(ii) $E[u_i(s')] \geq E[u_i(s)]$ *for all* $i \in C$, *with strict equality for some* $i \in C$.

Note that an individual also constitutes a coalition. Hence, all non-equilibrium strategies are dominated: there exists at least one player who has a profitable (coalition) deviation.



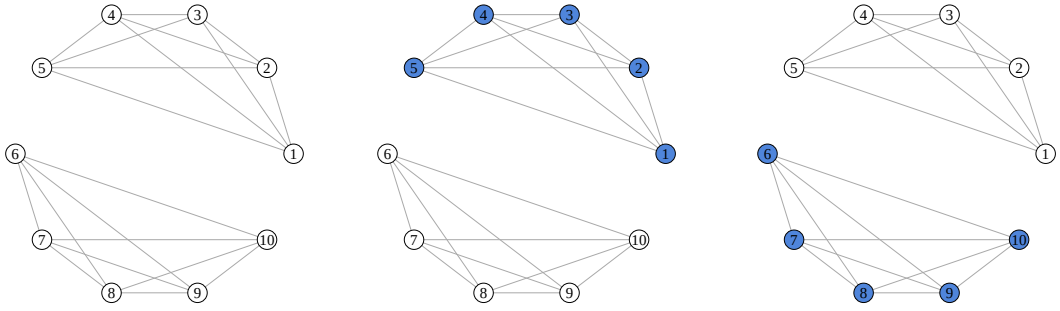
(a) Pure Market equilibrium (b) $S_F = \{1,2,3,4,5\}$

Fig. 1.2 Equilibria under s^* on a core-periphery network.



(a) Pure Market equilibrium

Fig. 1.3 Equilibria under s^* on a regular ring lattice network of degree 4.



(a) Pure Market equilibrium (b) $S_F = \{1,2,3,4,5\}$ (c) $S_F = \{6,7,8,9,10\}$

Fig. 1.4 Equilibria under s^* on a 5-complete network.

Since an equilibrium strategy can be dominated by a non-equilibrium strategy (which is also dominated), let us focus on the set of equilibrium strategy profiles that are undominated, namely *Coalition-proof*.

Proposition 1.4. *Suppose Assumption 1.1 and 1.2 hold. Under strategy profile s^* and network g, consider any two distinct equilibria, S and S', where $S'_F \neq \varnothing$. If $S_F \subsetneq S'_F$, then $S' \succ S$.*

Suppose there are two equilibria S and S' where the favouritism group S_F is a strict subset of S'_F. Even though all players have no individual incentive to deviate from equilibrium S, there exists a coalition $(S'_F \backslash S_F)$ who can earn high expected payoffs by collectively switching to equilibrium S' where they practise favouritism. Hence, S is dominated by S'. The formal proof is left in Appendix A.3.

Corollary 1.1. *Suppose there exists at least two equilibria under strategy profile s^* and network g, the Pure market equilibria are always dominated. Thus, the Pure market equilibrium is coalition-proof if and only it is the unique equilibrium.*

Proof. Recall that the Pure market equilibrium always exists for all networks. If there exists at least two equilibria S^0 and S', $S^0 = \varnothing \subseteq S'_F$, the Pure market equilibria is always dominated. This implies that the Pure market equilibrium is coalition-proof if and only it is the unique equilibrium. □

Corollary 1.2. *Suppose there is a set of equilibria $\mathbb{S}$ under strategy profile s^* and network g. An equilibrium S is coalition-proof if and only if for all $S' \in \mathbb{S}$, $S_F \nsubseteq S'_F$.*

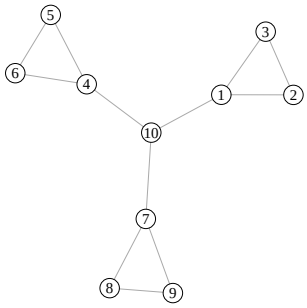
The formal proof is left in Appendix A.3. Instead, I provide a sketch of the proof. Suppose equilibrium favouritism group S_F is a proper subset of another equilibrium favouritism group S'_F, then S is dominated (Proposition 1.4). Suppose for all $S' \in \mathbb{S}$, S_F is not a proper subset of S'_F. Assume there exists an equilibrium S' which dominates S, in other words, there exists a profitable coalition deviation from S to S'. If this coalition comprises only market players under equilibrium S, then the new favouritism group S'_F is a proper superset of S_F, thus reaching a contradiction. If this coalition comprises at least one favouritism player

under equilibrium S, the favouritism player has no incentive to deviate individually or collectively to market behaviour because her earns higher expected payoff under favouritism. Therefore, equilibrium S is coalition-proof if and only if S_F is not a proper subset of any other equilibrium favouritism group.

To illustrate the result, consider a connected-triad network and suppose $n = 10, L = 0.8, \alpha = 0.5, \beta = 0.4, c = 0.01, \delta = 0.95$. There are three types of equilibria: Pure market equilibrium S^0 (Figure 1.5), single-F-group equilibria (Figure 1.6), and multiple-F-groups equilibria (Figure 1.7). Let us focus on the equilibria where $S_F^0 = \varnothing$, $S_F' = \{7,8,9\}$ and $S_F'' = \{4,5,6,7,8,9\}$. Note that each favouritism group is a proper subset of the next — $S_F^0 \subsetneq S_F' \subsetneq S_F''$. Suppose equilibrium S^0 is reached, players 4 to 9 will collectively deviate from S^0 to S''. Suppose equilibrium S' is reached, players 4 to 6 will collectively deviate from S' to S''. Both S^0 and S' are dominated and S'' is coalition-proof. Now suppose players are less patient, $\delta = 0.9$. S'' is no longer an equilibrium while S' becomes the undominated equilibrium. It is harder to sustain favouritism in S'' because the competing favouritism group $\{7,8,9\}$ redirects opportunities towards themselves. The larger this competing favouritism group is, the fewer opportunities one would receive as the expert.

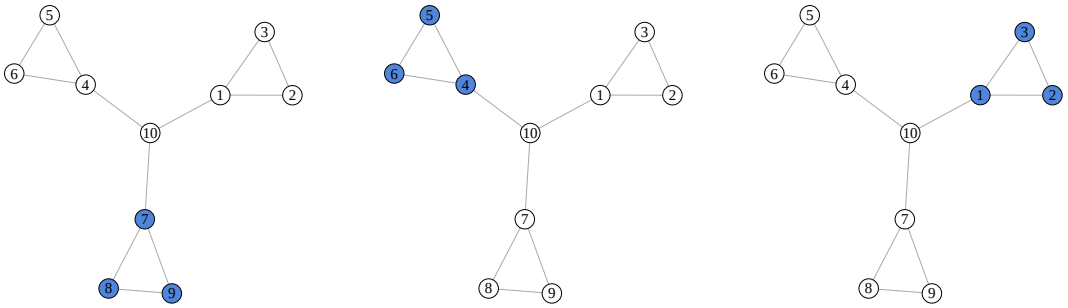
Maximum clique. An observation is that the coalition-proof equilibrium is related to the *maximum clique* of a graph, $K(g)$ — the clique in a graph with the most vertices. The coalition-proof equilibria of the networks illustrated before all correspond to their respective maximum cliques (Figure 1.8). Intuitively, players have equal and high degrees in a maximum clique. Additionally, all maximum cliques are *maximal* — cannot be extended by including one more adjacent vertex. Maximum cliques cannot be a subset of another clique. Therefore, the maximum clique is the coalition-proof equilibrium.

When players are sufficiently patient, the coalition-proof equilibrium is a union of the maximum cliques in a network (see the connected-triad network example in Figure 1.7). As discussed before, the presence of non-neighbouring favouritism players makes it harder to sustain favouritism. If the competing favouritism group is sufficiently small (or if players are sufficiently patient), both favouritism cliques can co-exist in the coalition-proof equilibrium.



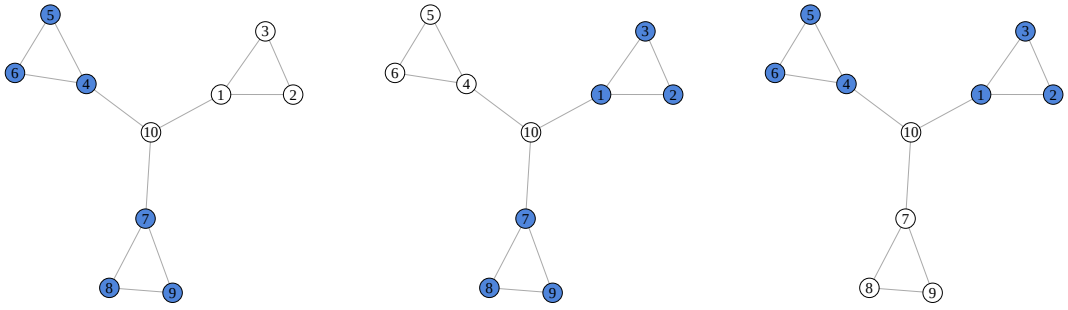
(a) $S_F^0 = \varnothing$

Fig. 1.5 Unqiue Pure market equilibrium under s^* when $\delta = 0.8$



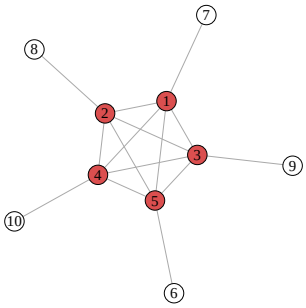
(a) $S'_F = \{7,8,9\}$ (b) $S'_F = \{4,5,6\}$ (c) $S'_F = \{1,2,3\}$

Fig. 1.6 Equilibria additional to Figure 1.5 under s^* when $\delta = 0.9$

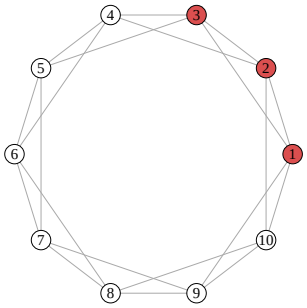


(a) $S''_F = \{4,5,6,7,8,9\}$ (b) $S''_F = \{1,2,3,7,8,9\}$ (c) $S''_F = \{1,2,3,4,5,6\}$

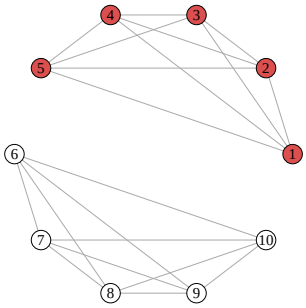
Fig. 1.7 Equilibria additional to Figures 1.5 and 1.6 under s^* when $\delta = 0.95$



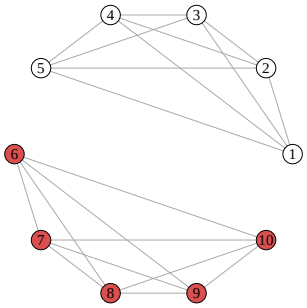
(a) Core-periphery network
$K = \{1,2,3,4,5\}$



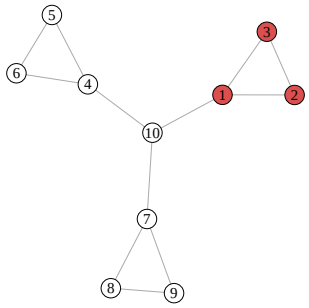
(b) Regular Ring lattice
$K = \{1,2,3\}$



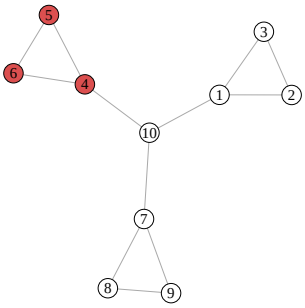
(c) Regular 5-complete
$K = \{1,2,3,4,5\}$



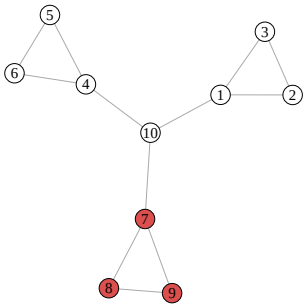
(d) Regular 5-complete
$K = \{6,7,8,9,10\}$



(e) Connected-triad
$K = \{1,2,3\}$



(f) Connected-triad
$K = \{4,5,6\}$



(g) Connected-triad
$K = \{7,8,9\}$

Fig. 1.8 Maximum cliques on networks

1.5 Aggregate surplus and Inequality

Definition 1.2. *Aggregate surplus is defined as the expected net output in one period:*

$$p\sum_{i=1}^{n} d_i + \Big(1 - p\sum_{i=1}^{n} d_i\Big)\Big(\frac{|S_F|}{n}L + (1 - \frac{|S_F|}{n})(1-c)\Big) \tag{1.8}$$

The total number of possible links in a network of size n is $\frac{n(n-1)}{2}$. The number of connected pairs equals to the total number of links in the network, $\frac{1}{2}\sum_{i=1}^{n} d_i$. With probability $\frac{2}{n(n-1)}\frac{1}{2}\sum_{i=1}^{n} d_i = p\sum_{i=1}^{n} d_i$, the pair of principal and expert are neighbours. Both market and favouritism principals would produce efficient output 1 and incur no search cost. With probability $1 - p\sum_{i=1}^{n} d_i$, the principal and expert are not neighbours. With probability $\frac{|S_F|}{n}$, the principal practises favouritism and produces inefficient output L. With probability $1 - \frac{|S_F|}{n}$, the principal practises market behaviour, incurs the search cost, and produces net output $1-c$.

Proposition 1.5. *For all networks g, the aggregate surplus is strictly increasing in the total number of links, and strictly decreasing in the number of players practising favouritism.*

Proof. Since $1 > \frac{|S_F|}{n}L + (1 - \frac{|S_F|}{n})(1-c)$, a higher network connectivity means a higher chance of principal and expert being connected. This waivers the search cost when practising market behaviour and reduces the inefficiency cost when practising favouritism. Since $c < (1-\alpha) - (L-\beta)$ and $\alpha > \beta$, favouritism reduces aggregate surplus, $1 - c > L$. Consequently, minimising the number of players in S_F increases aggregate surplus. □

Proposition 1.6. *Suppose Assumption 1.1 and 1.2 hold. Under strategy profile s^*, if there exists an equilibrium where $S_F \neq \varnothing$, then players in S_F earn a higher expected payoff in equilibrium S than in Pure market equilibrium; the opposite is true for players in S_M.*

Proof. When favouritism is sustained in equilibrium, favouritism players earn more than they would in the Pure market equilibrium (by condition (1.2)). The opposite is true for market players: they receive fewer opportunities when they are the expert than in the Pure market equilibrium because of others practising favouritism ($p|N_i^c \cap S_M| < p|N_i^c|$). □

In the Pure market equilibrium, poorly-connected players earn lower expected payoffs due to the search costs they are more likely to incur. The heterogeneity in degrees induces this payoff inequality. In contrast, when favouritism is sustained in equilibrium, favouritism players cooperate to extract a large portion of the reduced aggregate surplus. As a result, favouritism exacerbates the payoff inequality.

Social Planner. Consider a social planner who can design the network to maximise aggregate surplus under the coalition-proof equilibrium. In a complete network, all principals and experts are neighbours which waivers the search cost for the expert. Hence, the expected output equals 1. If links are free, the social planner will construct the complete network. However, social connections require maintenance and the linking cost is not zero. What network would maximise aggregate surplus given costly links?

Recall that if there exist at least two equilibria under network g, the coalition-proof equilibrium always sustains favouritism. Since favouritism reduces aggregate surplus ($L < 1-c$ by Assumption 1.2), the social planner would construct the network g such that the Pure market equilibrium is the unique equilibrium. The aggregate surplus of Pure market equilibrium with $\sum_{i=1}^{n} d_i = D$ connections equals $1-(1-pD)c$. An example of such a network would be the ring-lattice regular network. However, this socially optimal network is drastically different from the equilibrium hub-spoke network with endogenous linking among heterogeneous agents (see Appendix A.2).

1.6 Conclusion

The paper proposes a model of informal favour exchange where cooperation lowers aggregate surplus. Individuals can favour their neighbours or locate the most efficient expert at a search cost. Favouritism cannot be sustained in a stage game because it is inefficient but can be sustained in a repeated game through reciprocity. The opposing tension between favouritism-practising agents in wanting high degrees and low neighbours' degrees means that the cliques of a network are likely to sustain favouritism. In particular, the maximum clique of the network collectively prefers to sustain favouritism even though all players practising market

behaviour is the socially optimal equilibrium. I argue that favouritism is a mechanism for individuals to extort surplus from the society towards their favouritism subgroup.

The model yields novel insights on the relationship between favouritism behaviour and a network feature — the maximum clique. An interesting empirical research question is whether highly connected individuals are more likely to practise favouritism. Anecdotally, favour exchange networks like the British "old boy's network" are only among the elite. These individuals are endowed with connections and practise favouritism with other well-endowed individuals. But since explicit nepotism is illegal, it is difficult to obtain empirical data on favour exchange patterns within these communities. One suggestion is to analyse job referrals within the labour market where a manager refers someone from her social network to a job opening.

I conclude with a few remarks on some limitations of the model. First, individuals in my model perform favour exchange on a predetermined social network. While some social connections like family ties are stable, real-life social networks can evolve endogenously and change exogenously. Second, the model suppresses other forms of heterogeneity (outside of degree) to isolate the effects of the network. These heterogeneities can amplify or diminish the network effects on favouritism incentives. Combining endogenous network formation with player heterogeneity could offer insights into the relationship between network homophily and favouritism. I argue that the equilibrium network is of the hub-spoke structure under endogenous linking which promotes favouritism (Appendix A.2). Finally, the model assumes payoffs are exogenous for principals and experts. The basic assumptions on output and wages are insufficient in modelling complex negotiation within the favouritism game. Degree heterogeneity can be factored into the bargaining power of a principal as she has more options on whom to favour.

Chapter 2

Learning in Canonical Networks

2.1 Introduction

> In these democratic days, any investigation into the trustworthiness and peculiarities of popular judgements is of interest. *Galton (1907), pages 450-451.*

More than a hundred years after Galton's discovery of the "wisdom of crowds" (Galton, 1907), as democratic politics became more common across the world, our collective opinions and beliefs matter for an ever-widening range of subjects. Pioneering work on the role of social networks was carried out by sociologists in the mid-twentieth century (Lazarsfeld and Merton, 1954, Katz and Lazarsfeld, 1966, Coleman et al., 1966). More recently, with the growing usage of social media, there has been renewed interest in the role of social networks in shaping opinion formation and behaviour. Existing studies have highlighted two features of real-world social networks: (i) deep inequalities in the number of connections where the average is small but the variance is very large, and (ii) network homophily — tendency of people with similar traits to form links with each other (Barabási and Albert, 1999, Newman, 2010, McPherson et al., 2001, Currarini et al., 2009). The theory of social learning shows that these network features have powerful effects on opinions and behaviour (Bala and Goyal (1998), Bala and Goyal (2001), DeMarzo et al. (2003), Mossel et al. (2014) and Golub and Jackson (2010)); for a survey of this research see Golub and Sadler (2016) and Goyal (forthcoming). This paper aims to experimentally test these theoretical predictions in large

canonical networks, i.e., networks that are rich and complex and that reflect inequality and homophily.

We consider a model taken from Gale and Kariv (2003) in which individuals receive noisy signals about the true state of the world and make a guess repeatedly over time. We consider a binary state setting with a binary guess where the optimal guess is to match the true state. Individuals also observe the guesses of their neighbours, which in principle allows information to flow across paths of the social network. We examine how the network shapes the long-run process of information dissemination.

We study learning in three networks: Erdös-Rényi (a baseline for connections among homogeneous individuals), Stochastic Block (reflecting network homophily) and Royal Family network (that accommodates 'influential individuals' along with local interactions). Figure 2.1 presents these three networks and Figures 2.2a and 2.2b present the learning dynamics under DeGroot updating (DeGroot, 1974): at any period t, an individual guesses the state that corresponds to the majority guess in her neighbourhood in the previous period $t-1$. We are led to three hypotheses: (i) individual behaviour converges; (ii) the presence of network homophily leads to the persistence of diverse opinions/guesses; (iii) the presence of influential individuals gives rise to incorrect consensus and sub-optimal behaviour. In real life, people are diverse in preferences, capacities for information processing, and decision-making rules. It is therefore unclear if these theoretical predictions will obtain in practice.

We conduct a laboratory experiment to test these predictions.[1] Our experiments yield three findings. First, learning occurs in all the networks so rapidly that most of the consensus level achieved happens early. Second, breakdown of consensus and persistence of diverse opinions is more likely in the Stochastic Block network as compared to the other two networks. Third, incorrect consensus is much more likely in the Royal Family network as

[1]With observational data, it would be difficult to test these theoretical predictions about network effects because of identification issues. One reason is that network structures are often endogenous and a second reason is that network structures are rarely fully observable in real life; this creates the possibility that there is a gap between what players observe in a network and what a researcher observes. Thus it would be difficult to attribute the change in behaviours to a learning process in a network. Given these concerns with observational data, we resort to controlled laboratory experiments with large-scale networks.

compared to the other two networks. Finally, we show that the vast majority of individual guesses are consistent with DeGroot updating rule.

Related Literature. There is a large body of experimental research on opinion formation and behaviour. Early contributions include Choi et al. (2005), Mobius et al. (2015), Kearns et al. (2012). For a survey of the experimental research in economics see Choi et al. (2016), Breza (2016). Our paper is closest to two recent papers by Grimm and Mengel (2020) and Chandrasekhar et al. (2020) who use a model of binary states and repeated guessing. Their experiments use stylized small networks to disentangle the updating rules of subjects. They find that subjects' behaviour is close to that predicted by DeGroot updating.

The empirical literature on networks has highlighted the complex and rich structures and brought out the salience of network homophily and connection inequality. It is unclear what rules of behaviour individuals will follow when confronted by such complex environments. To address this concern, we propose an experiment with large networks that can accommodate key features of empirical networks. This leads us to study three canonical networks: Erdös-Rényi representing a baseline of decentralized contacts (Newman (2018)), Stochastic Block network representing network homophily (see McPherson et al. (2001), Newman (2018)) and Royal Family network capturing highly influential nodes together with local influence (Acemoglu et al. (2011), Bala and Goyal (1998), Mossel et al. (2014)). Our contribution is therefore twofold: one, we propose a new experimental design with canonical networks and two, we show that the learning patterns of our subjects are consistent with predictions of a model where agents follow the DeGroot updating rule.

Our paper is also related to Becker et al. (2017) and Becker et al. (2019) and the ongoing work of Agranov et al. (2020). Specifically, Agranov et al. (2020) consider a star and a core-periphery network, while Becker et al. (2017) study a hub-spoke network. Our paper differs from these papers in the canonical networks we study: these networks are complex and they accommodate salient features of empirical networks like inequality and homophily. To the best of our knowledge, our paper offers the first experimental evidence supporting strong network effects in such a setting and on the consistency of decision making by subjects with DeGroot updating rule.

2.2 Theory and Hypotheses

We use a model with two states, two signals, and two guesses that is taken from Gale and Kariv (2003). There is a set of individuals $N = \{1, .., n\}$, with $n \geq 2$. There are two possible states of the world, $\omega \in \{0, 1\}$, which individuals believe to be equally likely a priori.

Time is discrete and proceeds as $t = 0, 1, 2....$ In period 0, individuals observe a noisy but informative signal on the true state: individual i receives a binary signal $s_i \in \{0, 1\}$. The probability of receiving the correct signal corresponding to the true state is $p \in (1/2, 1]$. From period $t \geq 1$, an individual chooses a binary guess $a_{i,t} \in \{0, 1\}$. Guessing the true state correctly yields a payoff of 1, and guessing incorrectly yields 0. Thus upon receiving a signal of $s_i = 1$, the expected payoff of an individual guessing $a_{i,t} = 1$ is p and the payoff from guessing $a_{i,t} = 0$ is $1 - p$. Individuals follow their signal in period 1 (note that this guess is also optimal for a myopic individual who seeks to maximise one period payoff).

Individuals are located in an information network, g. We allow for both directed and undirected networks. A link $g_{ij} \in \{0, 1\}$ reflects information access. If $g_{ij} = 1$ then individual i observes the guesses of individual j. $g_{ii} = 0$ by convention. The neighbours of individual i are given by $N_i(g) = \{j | g_{ij} = 1\}$. We will suppose that an individual i gets to observe the guesses of everyone in her neighbourhood. In particular, at time t, individual i observes the guesses of her neighbours from period 1 until period $t - 1$. These observations on neighbours' guesses and the signal in period 0 are inputs into individual i's belief at time t about the likelihood of state $\omega = 1$, denoted as $\mu_{i,t}$.

In principle, in period 2, an individual can infer a signal from the first period guess of a neighbour; moreover, in subsequent periods, she can also potentially make inferences on the signals of the neighbours of neighbours, and so forth. These inferences are challenging even in simple situations, but in complex networks, they appear to be even less plausible. With these concerns in mind, building on the literature on majority dynamics (Benjamini et al., 2016) and DeGroot updating (DeGroot, 1974), we propose the following simple rule of thumb for individuals: In period $t = 1$, individual i makes a guess that mimics her signal s_i; in subsequent periods $t \geq 2$, she guesses $a_{i,t}$ that corresponds to the majority guess in her neighbourhood in the previous period (which includes her last period guess $a_{i,t-1}$). To

facilitate learning, let us suppose that individuals randomize (with equal probability) between the two states in case of no majority (Grimm and Mengel, 2020). To summarize, an individual i updates her guess $a_{i,t}$ at time t in the following way:

$$a_{i,t} = \begin{cases} 1 & \text{if } \mu_{i,t} > \frac{1}{2}, \\ 0 & \text{if } \mu_{i,t} < \frac{1}{2}, \\ \{0,1\} & \text{if } \mu_{i,t} = \frac{1}{2} \end{cases} \tag{2.1}$$

$$\text{where } \mu_{i,t} = \frac{1}{|N_i(g)|+1}\{\sum_{j=1}^{n} a_{j,t-1} \cdot g_{ij} + a_{i,t-1}\}.$$

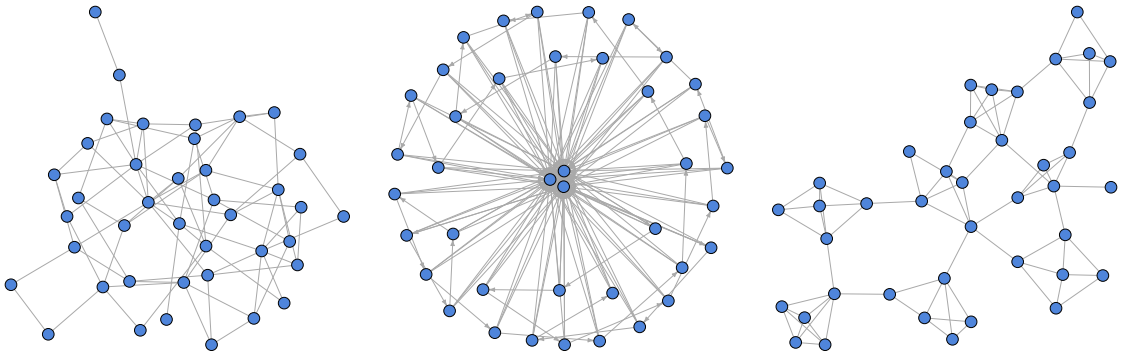
We shall refer to this rule as *DeGroot updating* in the rest of the paper.

We study the learning dynamics and long-run outcomes in three archetypal networks: i) the Erdös-Rényi network; ii) the Stochastic Block network (that reflects network homophily); and iii) the Royal Family network (that represents networks with highly influential individuals and local interaction). Figure 2.1 presents these networks; we selected these networks as they are representative in their respective classes and have distinct theoretical predictions (see Section B.1.1 for elaborations on the network generation process).

To formulate our hypotheses, we ran simulations of DeGroot learning rule on 1000 sets of signals for each network. The signals are drawn i.i.d. for 40 players with signal quality $p = 0.7$. Players then update their beliefs and guesses under the DeGroot updating rule. We organize the simulation results by defining a variable c_t:

$$c_t = \begin{cases} (n_t - n_0)/(n - n_0) & \text{if } n_t \geq n_0, \\ (n_t - n_0)/n_0 & \text{if } n_t < n_0, \end{cases} \tag{2.2}$$

where n_0 denotes the number of correct signals received at time 0 and n_t denotes the number of correct guesses made at time t. To account for variations in n_0 (as signals are randomly selected with quality $p = 0.7$), c_t measures the extent to which the average guess at time t move toward correct consensus ($n_t \geq n_0$) or towards incorrect consensus ($n_t < n_0$) relative to



(a) Erdös-Rényi network (b) Royal Family network (c) Stochastic Block network

Fig. 2.1 Canonical networks

the initial assignment of signals. Note that the potential amount of learning towards incorrect consensus is much larger than correct consensus. So the extent of learning is normalized by the maximum margin of learning towards correct consensus ($n - n_0$) or towards incorrect consensus ($n_0 - 0$). Together, c_t ranges between -1 (incorrect consensus) and 1 (correct consensus) with $c_t = 0$ representing no learning.

Figure 2.2a shows that learning occurs rapidly and the consensus is achieved within the first few periods of the game. This is also reflected in the frequency of switching behaviour: Figure 2.2b shows that roughly 25% of the individuals switch their guesses in period 2 after observing the guesses of their neighbours. This frequency falls to less than 5% by period 4 and becomes negligible eventually.

We next note that the network has powerful effects on consensus levels. The Royal Family network achieves complete consensus ($c_t = 1$ or -1) by period 4 in almost all simulation runs. By contrast, the Stochastic Block network attains only 60% of potential learning by period 4 and then remains at that level afterwards. Learning in the Erdös-Rényi network continues for longer: the network attains 87% of potential learning by period 7. To separate learning towards correct from learning toward incorrect consensus, Figure 2.2c presents the distribution of c_t averaged across periods 7-12. In the Erdös-Rényi network, correct consensus obtains in 61% of the cases. In the Royal Family network, consensus obtains in all cases: 79% on correct consensus and 21% on incorrect consensus. In the Stochastic Block

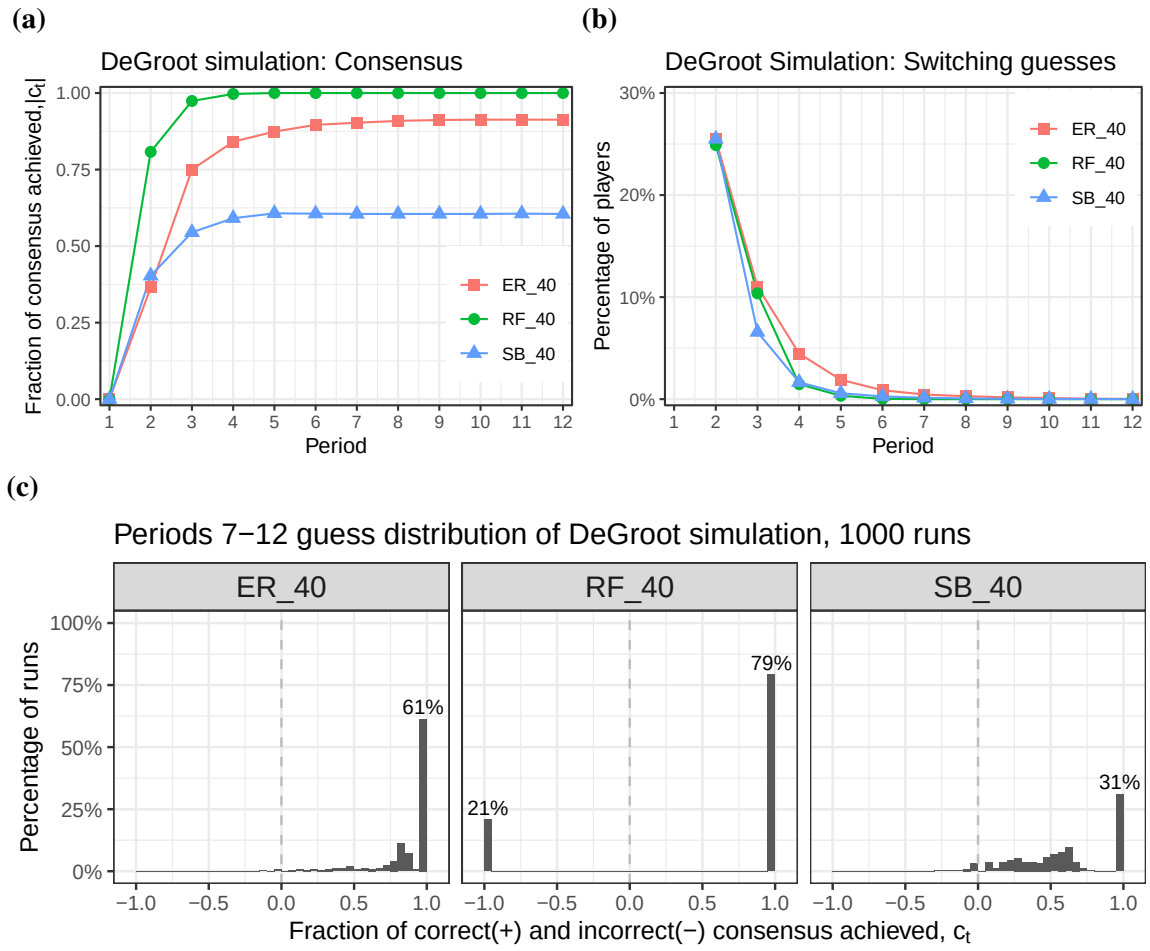


Fig. 2.2 DeGroot simulations. (a) In period 1, $|c_t|$ equals 0 because all individuals guess their signal. By period 4, RF (green) achieves $|c_t| = 100\%$ in almost all cases. SB (blue) attains $|c_t| = 60\%$. ER (red) attains $|c_t| = 87\%$ by period 7. (b) After period 4, less than 5% of individuals switch their guesses from the previous period. By period 7, this frequency is negligible. (c) In periods 7-12, ER network reaches correct consensus in 61% of cases, RF in 79% of cases, and SB in 31% of cases. Almost all remaining cases yield breakdown of consensus in ER and SB (39% and 69%, respectively) or incorrect consensus in RF (21%) (n=1000 per network).

model, correct consensus obtains only in 31% of the cases. We obtain similar predictions if we consider variations of the DeGroot updating rule (see Section B.1.3).

We use these theoretical results to formulate three hypotheses:

H.1 Individual guesses converge to a limit guess in all networks.

H.2 The breakdown of consensus is more likely in the Stochastic Block network as compared to the Erdös-Rényi and Royal Family network.

H.3 Incorrect consensus is more likely in the Royal Family network as compared to the Erdös-Rényi and Stochastic Block network.

Let us provide some intuition underlying these hypotheses. The Stochastic Block network is comprised of smaller communities that have a greater density of ties within and fewer ties across them. Since a community is smaller in size than the whole network and has access to fewer signals, it is less likely to reach the correct consensus independently. To illustrate this, consider a scenario where the entire network guesses 1 except for a community that guesses 0. Suppose there is only one link between an individual X (in the community) and the rest of the network, let us say that this link is with individual Y (outside the community). Since X observes herself and other members of her community, she observes a majority guess of 0, while Y observes a majority guess of 1. Under the DeGroot updating rule, X's community therefore agrees upon an incorrect consensus and cannot learn about the external majority (Chandrasekhar et al., 2020). This insulation of communities is more likely in the Stochastic Block than the Erdös-Rényi network because of higher network homophily.

We next discuss why the rate of convergence is higher and why incorrect consensus is so common in the Royal Family network. Observe that, in this network, the 3 members of the 'royal family' (i) constitute a clique among themselves with only one source of information from the outside world, (ii) are observed by everyone in the network, and (iii) constitute a majority in the neighbourhood of everyone. The first property means that the 'royal family' converge to the same guess by period 2. The second and third properties taken together with the DeGroot updating rule imply that everyone outside the 'royal family' imitates the guesses of the 'Royal Family' clique thereby leading to a quick convergence. However, if the

majority of the ‘Royal Family’ happen to get incorrect signals then the consensus will be on the wrong guess.

2.3 Experimental Design

We recruited 480 participants from the Laboratory for Research in Experimental and Behavioral Economics (LINEEX) at the University of Valencia to take part in a learning game. Subjects were randomized to one of three experimental conditions, each associated with a distinct network structure: Erdös-Rényi, Stochastic Block, and Royal Family network. We ran a total of 12 sessions, 4 sessions for each experimental condition. Each session consisted of a group of 40 subjects on a social network who played 6 rounds of the learning game. No subject participated in more than one session.

In each round of the game, subjects were randomly assigned a position in a social network. Subjects’ positions were reshuffled from one round to the next to reduce potential repeated game effects during the experiment (subjects could not keep track of a participant’s position across rounds). Subjects in the same session saw the network structure along with different IDs associated with different nodes. Because subjects in the network conditions were not statistically independent, all analyses of collective estimates in the network conditions were conducted at the round level such that each network provided 24 observations. Moreover, because each session completed multiple rounds of the learning game within an experimental trial, we cluster our main analysis at the session level (see Fréchette (2012) for the discussion on dealing with session effects in the laboratory).

Subjects were informed about a bag containing 10 balls. They were told that the bag contains either 7 Red and 3 Green balls (we will refer to this as the RED bag) or 7 Green and 3 Red balls (the GREEN bag). Each of these two combinations is a priori equally likely. At the start of a round, each subject drew a ball from the bag and saw its colour. There was a 70% chance of getting the ‘correctly’ coloured ball (representing the signal) corresponding to the colour of the bag (representing the true state).

For 12 periods, subjects were asked to guess whether the bag was RED or GREEN. At period $t = 1$, subjects' guess was based on their prior and the colour of the ball initially drawn by them. From period $t \geq 2$ until $t = 12$, subjects also observed guesses of neighbours in previous periods from which they could update their beliefs and revise their guesses. At the end of the round, one period (from 1 to 12) was picked at random to determine actual payoffs in the round: subjects earned 3 euros if their guess matched the colour of the bag (GREEN or RED), and 0 euro otherwise. Total earnings for a subject corresponded to the sum of earnings in each round and a 5 euro show-up fee.

The experiment lasted approximately 1.5 hrs. The average payment per subject was 19.3 euros (including the 5 euro show-up fee). The details of the experimental procedures, including sample instructions, are presented in Section B.4.

2.4 Findings

We start with a presentation of the learning dynamics. We then compare the level of correct and incorrect consensus and the breakdown of consensus achieved by each network. Lastly, we study whether subjects' behaviour matches various updating rules.

Dynamics of Learning

We begin by discussing the dynamics of learning and the stability of long-run behaviour. Figures 2.3a and 2.3b summarize the data. In line with the DeGroot simulation, most of the learning occurs in the early phase of the dynamics: More than three-quarters of the final consensus achieved by period 12 is attained by period 4. In particular, the Royal Family and Stochastic Block networks have more rapid learning than the Erdös-Rényi network. The rapid convergence is also supported by evidence on switching frequency: 20% of subjects switched their guess in period 2 after observing the first-period guess of their neighbours; this switching frequency falls to 10% towards the end of the experiment in period 12. In addition, there are large learning effects across rounds: as a result, the switching probability falls significantly across rounds — only 5% of subjects switched their guess in the last

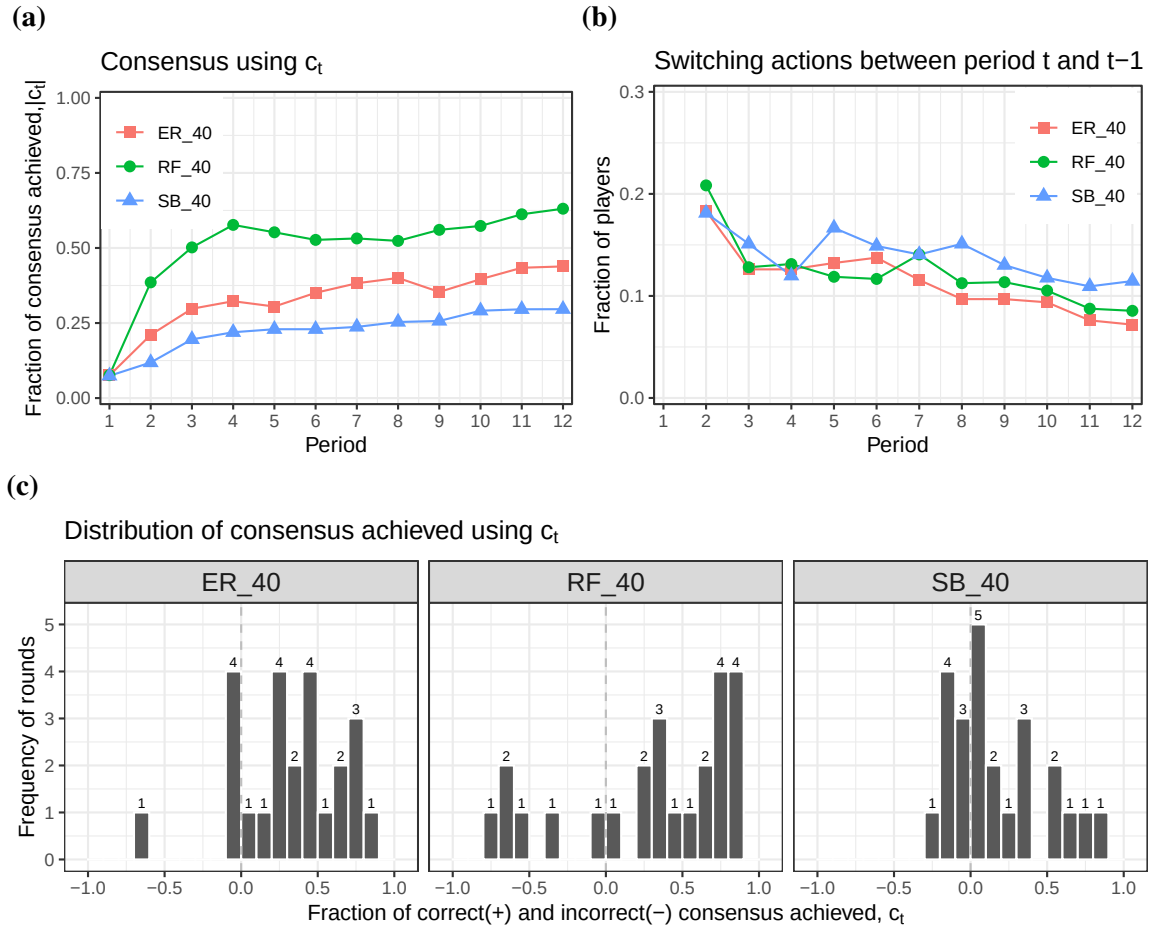


Fig. 2.3 Learning and consensus building. (a) For ER, RF and SB, by period 4, the average $|c_t|$ equals 35%, 58% and 22% respectively. By period 12, ER, RF, and SB, average $|c_t|$ equals 44%, 63%, 30%, respectively. (b) Roughly 20% of subjects switch their guesses in period 2; switching reduces to 10% by period 12. (c) Distribution of c_t is almost uniform between 0 and 1 for ER, bimodal around 1 and -0.7 for RF, and modal around 0 for SB. (n=72: 24 per network).

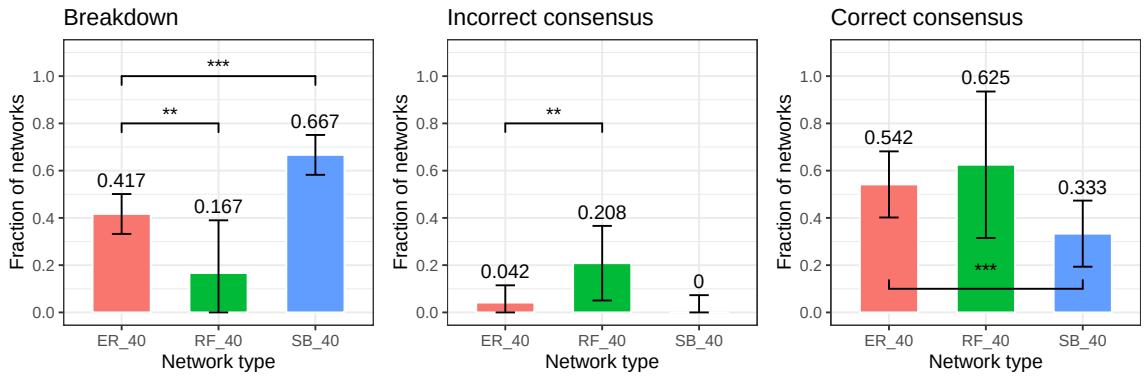


Fig. 2.4 Network effects on consensus. Error bars display standard 95% confidence interval around the mean. Compared to ER: (i) Breakdown of consensus is 25 pp (percentage points) more likely under SB (n=48, 95% CI [0.17,0.33], p-value<0.01), and 25 pp less likely under RF (n=48, 95% CI [-0.47,-0.03], p-value<0.05); (ii) Incorrect consensus is 17 pp more likely under RF (n=48, 95% CI [0.01,0.32], p-value<0.05), and 4 pp less likely under SB (n=48, 95% CI [-0.11,0.03]); (iii) Correct consensus is 21 pp less likely under SB (n=48, 95% CI [-0.35,-0.07], p-value<0.01), and 8 pp more likely under RF (n=48, 95% CI [-0.23,0.39]).

three rounds (Appendix Figure B.5). This evidence supports our first hypothesis: *individual guesses converge in all networks.*

Turning to consensus, we note that the level of consensus attained in the experiment is lower than the theoretical prediction (we examine these factors more closely in the Updating Rule section below and in the Appendix). However, the ranking of consensus dynamics across networks is consistent with the DeGroot simulation: the Royal Family network achieves the highest level of consensus from period 2 onward; the Stochastic Block network attains consistently the lowest level of consensus; the Erdös-Rényi network attains level of consensus in between the other two networks.

Consensus Outcomes

We examine the character of long-run outcomes through the measurement of c_t for each network averaged over the last 6 periods, i.e. between periods 7-12, averaged across all rounds and the 4 sessions (similar patterns are obtained if we consider fewer periods or rounds, see Appendix Figure B.7). In line with the DeGroot simulation reported in Figure 2.2c, Figure 2.3c shows that the distribution in the Royal Family network is bi-modal near $c_t =$

1 and $c_t = -1$, with a higher likelihood on $c_t = 1$ representing correct consensus. The Stochastic Block network has a mode around $c_t = 0$, indicating a greater likelihood of no learning and hence the persistence of diverse opinions. The Erdös-Rényi network leads to a fairly uniform spread of c_t between 0 and 1.

To make a statistical evaluation of the effects of networks on consensus, we proceed as follows: for each round, we average c_t across the last 6 periods. Thus for each network, there are a total of 24 data points (4 sessions with 6 rounds each). Then we categorize each round by whether the averaged c_t is above k (indicating the round achieving correct consensus), below $-k$ (incorrect consensus), or between k and $-k$ (breakdown of consensus). For concreteness, we choose k to be 0.3, so correct consensus is defined as the round achieving more than 30% of the maximum possible learning. Our main findings are robust to different widths k and an alternative, continuous, definition of consensus (Section B.2.2).

In Figure 2.4, we report the proportion of rounds that achieve correct or incorrect consensus or exhibit a breakdown of consensus for each network (and the corresponding 95% confidence interval). The estimates are derived from the following regression model: for group g in round r,

$$y_{g,r}^{correct} = \beta_0 + \mathbf{1}_g^{RF}\beta_1 + \mathbf{1}_g^{SB}\beta_2 + \varepsilon_{g,r}$$

where $\mathbf{1}_g^{RF}$ is an indicator function of whether the group g is playing on the Royal Family network. $y_{g,r}^{correct}$ is an indicator function of whether the round r achieved correct consensus: $\frac{1}{6}\sum_{t=6}^{12} c_{g,r,t} > k$. To account for session effects, we cluster the analysis at the session level (see Fréchette (2012) for the discussion on dealing with session-effects in the laboratory). β_0 can be interpreted as the proportion of Erdös-Rényi networks that reaches correct consensus, whereas β_1 (β_2) can be interpreted as the difference in proportion of networks that reaches correct consensus between Royal Family and Erdös-Rényi network (Stochastic Block and Erdös-Rényi network). Regression results are presented in the Appendix (Table B.7).

First, we find that breakdown of consensus is more likely in the Stochastic Block network than the Erdös-Rényi network (n=48, p-value<0.01), whereas it is less likely in the Royal Family network (n=48, p-value<0.05). Out of 24 rounds and 3 networks (72 data points in total), 22 arrive at breakdown of consensus: 14 in Stochastic Block, 6 in Erdös-Rényi, and 2 in

Royal Family network. Recall that there are 8 communities (consisting of 5 individuals each) in the Stochastic Block network. In period 12, 52% of the communities obtain consensus in the Stochastic Block network. This suggests that it is the disagreement across communities that is an important source of the breakdown in consensus in the Stochastic Block network. This is illustrated in 1 round of the Stochastic Block network where more than 7 communities reach complete consensus (5 out of 5 subjects agree) and yet there is breakdown of consensus in the society as a whole. These observations support our second hypothesis: *network homophily leads to breakdown of consensus sustains diverse opinions in a network.*

Second, we find that incorrect consensus is more likely in the Royal Family network than in the Erdös-Rényi network and Stochastic Block network (n=48, p-value<0.05). To appreciate the impact of the 'royal family', note that when 70% of the network receives the correct signal, incorrect consensus is defined as more than half the network guesses incorrectly. Thus the Royal Family network achieves incorrect consensus in 5 rounds (out of 24) as compared to 1 round in Erdös-Rényi and 0 round Stochastic Block network. This supports our third hypothesis: *the presence of highly influential individuals reflected in the Royal Family network, raises the likelihood of incorrect consensus*.

Lastly, we note that correct consensus is less likely in the Stochastic Block network than in Erdös-Rényi (n=48, p-value<0.01) and Royal Family network. Our estimation results are robust to alternative model specifications such as the logit model (Appendix Table B.9).

Updating Rule

The environment faced by individuals is complex, so individuals may use different and possibly time-varying updating rules. In this section, we examine how closely individual behaviour matches DeGroot updating.

At every period $t \geq 1$, DeGroot learning predictions are made based on guesses in period $t-1$. We define a binary variable for 'matching DeGroot prediction': it equals 1 when the subject i's guess in period t coincides with the DeGroot prediction, and 0 otherwise. In the case of DeGroot predicting indifference, the variable equals 1 regardless.

Figure 2.5a presents the percentage of individual guesses that were consistent with DeGroot predictions (in orange colour). We see that, on average, 88% of guesses match with the DeGroot rule (n=11,520 per network). This is higher than the baseline of how well guessing the signal matches with DeGroot predictions: simulations show that only 75% guesses of pseudo subjects (if guessing only their signal) match with DeGroot (see Appendix B.2.3 for detailed comparison). Figure 2.5a also presents the fraction of guesses that were contrary to DeGroot prediction but were in line with the signal (in purple colour). In the case when subjects' guesses do not match with DeGroot, about 70% guesses follow their signals. Taken together, DeGroot and persisting with own signal explain more than 95% of the variation in guesses.

By analyzing the fraction of agents that fail to (correctly) guess their signal in period 1, we estimate that about 10% of guesses are made randomly. Indeed, across the networks, the level of consensus achieved in the experiment is comparable to the consensus attained if subjects follow DeGroot updating rule with a 0.1 probability of trembles (see Appendix Figure B.10a). This shows that small deviations from DeGroot updating rule at the individual level can have a significant impact on the level of consensus reached. It also suggests that subjects are unlikely to be using other updating rules that are more sophisticated than DeGroot.

Figure 2.5b presents the time series of the fraction of guesses that matched the prediction of DeGroot updating across rounds. The increase in the match with DeGroot prediction suggests that there is learning across rounds. In particular, as subjects play more rounds, they are more likely to guess their signal in period 1, and they are less likely to persist with their signal in later periods (see Appendix Table B.12).

We next turn to heterogeneity in updating rules *across subjects*. The percentage of guesses matching the DeGroot prediction at the subject level is presented in Appendix Figure B.11. We see that a substantial fraction of subjects in each network follows DeGroot rule. For instance, 80% of subjects in the Erdös-Rényi network match with DeGroot predictions at least 80% of the time; these fractions are 72% in the Royal Family network and 76% in the Stochastic Block network, respectively. This is again compared to the baseline of how well guessing signal matches with DeGroot predictions: simulations show that only 44%

guesses of pseudo subjects (if guessing only signal) in the Erdös-Rényi network match with DeGroot predictions at least 80% of the time (37% in the Royal Family network, and 41% in the Stochastic Block network).

Testing how data matches with other learning rules is generally difficult in large networks. Here we briefly comment on Bayesian learning (for a discussion of variants of DeGroot and other updating rules see Chandrasekhar et al. (2020) and Grimm and Mengel (2020)). As Bayesian rules cannot be computed for large networks, following Chandrasekhar et al. (2020), we consider the role of *information dominant* players. We shall say that player X is an information dominant leader of player Y if X observes Y and all neighbours of Y. A Bayesian player X should ignore guesses of Y (after period 1) while Y should imitate X in all periods. In our experiment, when DeGroot prediction conflicts with the information leader's guess, only around 10% of subjects follow Bayesian prediction (ER:10%, RF:4%, SB:14%), while the rest follow DeGroot prediction. Similarly, when the DeGroot prediction contradicts the signal received, less than 30% of subjects follow their signal (ER:25%, RF:29%, SB:29%), while the rest follow DeGroot. Regression estimates are presented in the Appendix Tables B.11 and B.12. To sum up, the vast majority of guesses are consistent with the predictions of the DeGroot updating rule.

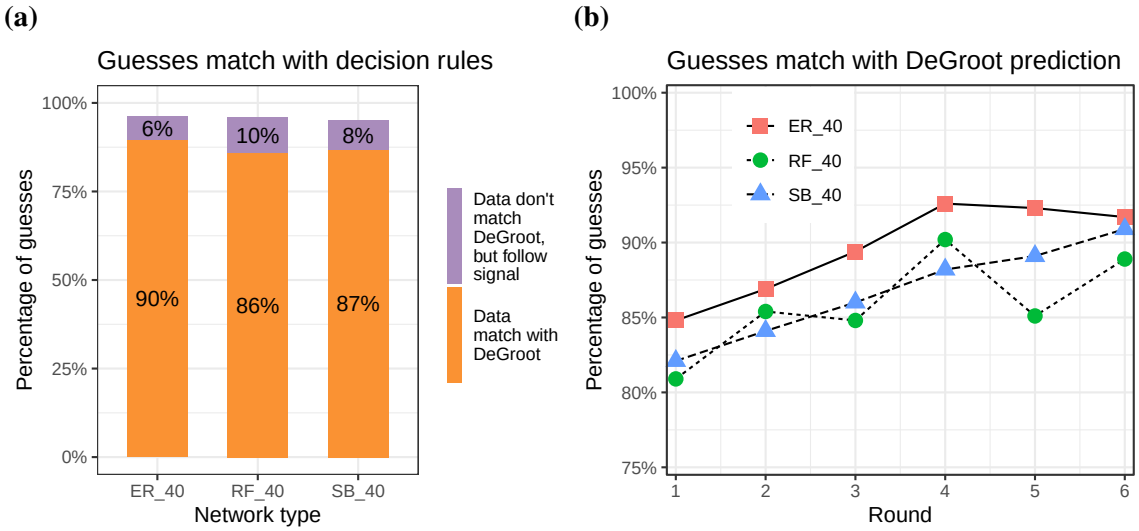


Fig. 2.5 Comparing actual guesses with DeGroot prediction. (a) 88% of guesses match with DeGroot prediction and 6∼10% match with signal. Together they explain 95% of variation in guesses. (b) 80∼85% of guesses match with DeGroot prediction in round 1; this increases to 88∼92% by round 6 (n=34,560: 11,520 per network).

Chapter 3

Information Verification and Sharing in Networks

3.1 Introduction

Social connections have always been essential in information dissemination. Recently, the spread of misinformation on social networks has raised serious societal concerns. In particular, more and more people are receiving information and sharing content on social media sites. In 2016, 14% of Americans said they use social media as their primary sources of news (Allcott and Gentzkow, 2017) with over 70% of Americans getting at least some of their news from social media (Levy, 2021). Social media sites such as Facebook, Twitter and YouTube have provided hotbeds for the spread of fabricated information, commonly referred to as "fake news".

Verification of content is central to preventing misinformation in traditional news media. However, with consumers shifting towards social media for news and information (Shearer and Gottfried, 2017), centralized fact-checking (third-party identification of inaccuracies before or after content dissemination) faces challenges in scalability due to the growing volume of online contents posted every day. Furthermore, the lack of trust in centralized fact-checking diminishes its benefits. According to a Pew Research Center survey, 70% of Republicans and 48% of Americans believe that fact-checkers are biased (Walker and

Gottfried, 2019). All these factors leave verification of information in the hands of the consumers. In this paper, we study incentives of individuals to verify and share information in social networks. In particular, we explore how the network affects verification incentives.

There are two types of agents: a *seed* who receives news directly from the source, and *non-seeds* who receive news indirectly from another agents. When an agent receives news, she first decides whether to verify it (at a cost) and then whether to share the news. A piece of news has some exogenous probability of being true.[1] Agents derive sharing benefits from sharing truthful news but incur an exogenous reputational loss for sharing news that is false.[2] Given that verification reveals the veracity of news perfectly, agents who verify will only share true news and not share false news. This restricts an individual's action set to (i) sharing without verification, (ii) not sharing without verification, and (iii) verifying and only sharing true news.

Individuals are situated in a social network which determines who they can share the news with. An individual's payoff is proportional to her *degree* — the number of agents she is connected to. Since an individual's degree is private information and the identity of her sender is unknown, sophisticated agents must form beliefs on other agents' strategies. Our solution concept is Bayesian Nash equilibrium. We are interested in two aggregate outcomes: the *quality of indirect news* — the percentage of news circulating round the network that is true, and the *spread* of news — the average probability of an agent receiving the news.

We show that there exists a unique equilibrium. If the information accuracy is poor (compared to the reputational damage from sharing false news), non-seeds always share news without verification because seeds will not share unverified news. If the information accuracy is good, both seeds and non-seeds will share unverified news. Overall, verification likelihood is increasing in degree. Our game features strategic substitutes: when other agents verify more, indirect news is more likely to be true, which lowers incentives to personally verify the news. We then examine the comparative static with respect to the key variables. First, as

[1] We abstract from the presences of preferential bias or political leaning to focus on how agents' and their neighbours' connectedness impact their verification and sharing behaviour. See Acemoglu et al. (2021) for a model that examines bias in this setting.

[2] This type of reputational cost has been documented in Altay et al. (2020), who find that Facebook users who had passed on misinformation in the past experienced bad reputation in their future sharing activity.

reputational loss increases, both seeds and non-seeds verify more because sharing unverified false news incurs a larger loss. This improves the quality of indirect news. However, later on, the improvement in indirect news quality lowers the need for non-seeds to personally verify the news so non-seeds verify less. Second, as the information accuracy improves, seeds verify less when they share unverified news. Initially, the quality of indirect news to deteriorate so non-seeds verify more. But later on, the quality of indirect news improves with the information accuracy so non-seeds verify less. Consider finally the effects of the network. The probability of seed verification is invariant to changes in the conditional degree distribution. However, under a denser network or a more egalitarian network (assuming a concave cumulative verification cost function), there are more seeds with higher probability of verification. Thus, the quality of indirect news is higher under a denser network which reduces all non-seeds' verification.

We then extend the model to endogenize the information accuracy served to a network. We consider the game with an external agent (a social media platform) whose goal is to maximise the number of views by engaging in costly investment in information accuracy. We show that the spread — the probability of news reaching an agent — is increasing in the information accuracy. When information accuracy is not expensive, the platform would invest in a minimum level of accuracy such that the agents verify with certainty. But if the accuracy required is expensive, it would switch to not investing. We find that a denser network requires a lower level of information accuracy to sustain consumer verification which can be afforded at higher investment costs. Therefore, denser networks lead to higher accuracy when information accuracy is expensive (or cheap).

Related Literature. Our paper is a contribution to the study of information generation and diffusion in social networks. This has been the subject of a great deal of recent research, see e.g., Charlson (2022), Mostagir et al. (2022), Candogan and Drakopoulos (2020), Chen and Papanastasiou (2021), Keppo et al. (2022), Törnberg (2018), Nguyen et al. (2012), Acemoglu et al. (2010). We spell out the new aspects of our paper by relating it to three recent papers, Kranton and McAdams (2022), Acemoglu et al. (2021), Papanastasiou (2020).

Kranton and McAdams (2022) study the incentives of media producers to invest in content quality. They present a model where consumers care about the quality of news (which the media producer invests in at a cost), and share news amongst their neighbours, but cannot verify the news. Their main finding is that, in highly-connected networks, all news will be widely viewed regardless of its quality, so a producer has no incentive to invest in information accuracy. As pointed out in our introductory discussion above, due to scalability constraints on external providers, verification by consumers is essential in large-scale networks. Our model introduces verification by network agents and studies the relationship between the incentives of the platform to improve the information accuracy and the incentive of network agents to verify and share news. A key result is that the platform either invests in at least a minimal level of information accuracy to sustain some network verification or does not invest at all. Similar to the finding in Kranton and McAdams (2022), platform invests less in denser (and more egalitarian) networks under intermediate investment costs. However, contrary to their finding, under high investment costs, we find that the platform invests more in denser networks and does not invest at all in sparser networks.

Papanastasiou (2020) studies a model where agents sequentially decides on verification and sharing based on heterogeneous ideological beliefs. The paper finds that the posterior belief in the news quality is improving with time (how far the content has travelled) which leads to a sharing cascade — agents after a certain period share content without verifying. The sharing process is therefore prone to the proliferation of fabricated content. Then the platform wishes to reduce the spread of misinformation by choosing when to inspect the content. Their results compare whether platform inspection is more effective than self-policing, and when is external inspection optimal.

Acemoglu et al. (2021) builds upon Papanastasiou (2020) by incorporating strategic complementarity in the verification behaviour. They also introduce a profit-driven media platform. The platform can select which article to introduce to the network, and design the sharing patterns among its users. They show that the platform will propagate extreme articles amongst the most extremist users and incorporate homophily in the sharing algorithm. This creates endogenous ideological echo chambers.

Our contribution to the literature is in two ways. First, we study the network effect on verification and sharing incentives. In Papanastasiou (2020), all agents can only share to one neighbour downstream, and in Acemoglu et al. (2021), a fixed number of neighbours (in the viral phase of an article's lifetime). Instead, agents in our model are placed on a social network which determines the neighbours they can share with. We also introduce a connection between an agent's network degree and payoff. By explicitly modelling the sharing network using a degree distribution, we find that the quality of indirect news is higher for denser and more egalitarian networks. This has negative implications on the quality of news circulating on social media sites (such as Twitter) which often exhibit strong connection inequalities.[3]

Second, we study the incentives of a views-maximising platform investing in information accuracy. We show that the spread of news is increasing in information accuracy (quality of news source). However, the proportion of news in circulation that is true may decrease in information accuracy. Investments in information accuracy can crowd out verification by consumers and lead to poorer quality of indirect news. To the best of our knowledge, our model is the first to study the network effect on platform investment. One novel finding is: when information accuracy is expensive, the platform only has an incentive to invest in information accuracy for denser (or more egalitarian) networks and not invest otherwise.

There is a growing literature on information design by platforms. One strand of literature on news markets studies how the revenue-generating process of media producers could bias content. Gentzkow and Shapiro (2006) find that news producers who benefit from having a reputation for accuracy slant their news towards consumers' initial beliefs. Besley and Prat (2006) and Gentzkow et al. (2006) find that producers who earn revenue from advertising reduce bias; In contrast, Ellman and Germano (2009) shows that newspapers bias their news towards their advertisers. Our paper does not consider political slants. Instead, we study the incentives of a platform investing in information accuracy and how the network affects the incentives.

[3]Empirical work shows that a large majority of individuals get most of their information from a very small subset of the group, viz., the influencers. Information networks tend to exhibit the law of the few (Galeotti and Goyal, 2010).

The rest of the paper is organized as follows. Section 3.2 introduces our model of verification and sharing on a network. We characterize the equilibrium and highlight its key features. Section 3.3 provides the comparative static results with respect to the key variables and the network. Section 3.4 endogenizes the information accuracy by introducing a views-maximising platform. Section 3.5 concludes the findings and discusses potential for future research.

3.2 Model

We consider a set of individuals $N = \{1, 2, ..., n\}$, $n \geq 2$, located in a sharing network g where $g_{ij} \in \{0, 1\}$. A link $g_{ij} = 1$ indicates that individual j receives news shared by i, $g_{ij} = 0$ otherwise. $g_{ii} = 0$ by convention. The neighbours/followers of individual i are given by $N_i(g) = \{j : g_{ij} = 1\}$, and her degree is denoted by $d_i(g) = |N_i(g)|$. The degree distribution $f(d)$ represents the fraction of agents with degree d.[4] We denote the corresponding cumulative degree distribution by F.

Time is discrete: at the start, $t = 0$, a single agent is chosen (uniformly at random) as the *seed* and receives a piece of news. This piece of news is either *true* or *false*. Individuals have a common prior that news is true with probability μ, where $0 < \mu < 1$. Once individual i receives a piece of news, she first decides on whether or not to verify the news to determine its validity. Verification perfectly reveals to the agent whether the news is false. The cost of verification is c_i; these costs are identically and independently drawn from a continuous probability distribution $h(c)$ with the corresponding cumulative distribution function $H(c)$ which has full support $[\underline{c}, \overline{c}]$. We assume that $h(c)$ is atomless, continuously differentiable, and positive for all $c \in [\underline{c}, \overline{c}]$.

After the seed decides whether or not to verify, she then decides whether or not to transmit the news depending on the costs and benefits of sharing news. Note that an individual gets only one period to verify and share the news, right after the receiving it. If the receiving

[4]For algebraic proofs, we assume that the degree distribution can be approximated by a continuous probability density function which has full support $\mathbb{R}^+$ and is continuously differentiable. This approximation is reasonable when the population size n is large. For example, the degree distribution of an undirected Erdös-Rényi graph of size n with linking probability p is approximated by the normal distribution $N(np, np)$.

agent decides not to share the news, the sharing of the news downstream is discontinued indefinitely. Whereas if she shares, her neighbours receive the news in the next period and repeat the decision process of verifying and sharing news. The game ends either when everyone received the news or when no one is sharing news anymore.

In period t, if the news shared by agent i is true, she earns a sharing benefit of 1 for each neighbour; If the news shared is false, she earns 0.[5] After every agent who freshly received the news has decided whether to verify and whether to share, the news is examined externally with probability $\phi > 0$ which reveals the underlying veracity of news.[6] If the news is revealed to be false, then those who shared the news in the immediate previous period are punished as they suffer a reputational loss of $r \geq 0$ for each neighbour.[7] In short, conditional on the news being false, an agent sharing false news is expected to suffer a loss of $R = \phi r$ for each neighbour.

Model Assumptions. Following Papanastasiou (2020) and Acemoglu et al. (2021), we assume that the sharing network g is a tree network — there exists only one unique path from one node to another. The seed is then the root of the network, with information flowing through the branches to the non-seeds. A denser network implies nodes having more branches on average. Given that a tree network is acyclic, individuals would not receive the same information twice or from multiple sources. This implies that having more neighbours will not increase the chance of receiving (unverified) news. Thus, the probability of news being true conditional on receiving it, denoted by z, is independent of own-degree.

To make the problem tractable, we also assume *degree independence* as defined in Galeotti et al. (2010). Let $\tilde{f}_i(d_j) = f(d_j|g_{ij} = 1)$ denote the conditional degree density for neighbour j of agent i, that is the degree density for agent j given that i and j are connected. In particular, degree independence in our game implies that, non-seeds have identical beliefs about the degree of their sender, independent of their own-degrees. This

[5] An alternative payoff structure is that individuals receive sharing benefits of 1 for each neighbour, independent of the veracity of news. We show that the two models yield the same equilibrium strategies subject to redefining the reputational loss parameter R (see Appendix C.1).

[6] In Appendix C.3, we allow the examination probability to be determined endogenously by verification downstream.

[7] This type of reputational cost has been documented empirically in Altay et al. (2020) where Facebook users who passed on misinformation in the past experienced bad reputation in their future shares.

implies that $f(d_j|g_{ij}=1) = f(d_j|d_j \geq 1)$. By assuming the network has no agent with less than one connection, $f(d_j|d_j \geq 1) = f(d_j)$. As a result, $\tilde{f}_i(d_j) = f(d_j)$. Given the strategies of senders with different degrees, a non-seed can form expectations based on the degree distribution f.

Before we examine the payoffs of the strategies, we discuss the assumptions regarding beliefs. We call news received from the source *direct* news and news received from other agents *indirect* news. Following Galeotti et al. (2010), we assume that agents' degrees are private information. Thus, agents know their own degrees and whether or not they are the seed (whether they received direct news). However, agents have no knowledge on their neighbours' degree, whether or not their sender is the seed, and whether or not the news they received has been verified. Consequently, non-seeds must form beliefs on their neighbours' degrees, their strategies, and the likelihood of indirect news being true, z.[8] Our solution concept is Bayesian Nash equilibrium (BNE).

Strategies. Upon receiving the news, each agent first decides whether to verify and then decides whether to share the news. Using backward induction, we start with the sharing incentives. Sharing true news earns a positive payoff, whereas sharing false news only incurs losses. Given that inspection reveals the veracity of news perfectly, the dominant sharing strategy conditional on verifying news is to share true news and not share false news. In contrast, if the agent chooses to not verify, her sharing strategy is based on her degree d, her verification cost c, and her belief in the veracity of the news z. Therefore, for a seed with degree d and cost c, the action set is restricted to $a^{seed}_{d,c} \in \{\mathscr{S}, \mathscr{K}, \mathscr{V}\}$, where

- $\mathscr{S}$ represents *sharing without verification*,
- $\mathscr{K}$ represents *killing (not sharing) the news without verification*, and
- $\mathscr{V}$ represents *verifying and only sharing true news*.

Similarly, for a non-seed with degree d and cost c, the action set is $a^{non-seed}_{d,c} \in \{\mathscr{S}, \mathscr{K}, \mathscr{V}\}$.

[8] Another variable of interest is the quality of news in period t, denoted as y_t. Note that y_t is increasing with time t. At the end of the game, false news would have been verified and no longer shared, so y_t by definition equals 1. We instead focus on z because it represents the fraction of true news in circulation at any given time.

The payoffs for the seed and non-seed are as follows:

$$U_{d,c}^{seed} = \begin{cases} \mu d - (1-\mu)Rd & \text{if } a_{d,c}^{seed} = \mathscr{S} \\ 0 & \text{if } a_{d,c}^{seed} = \mathscr{K} \\ \mu d - c & \text{if } a_{d,c}^{seed} = \mathscr{V} \end{cases} \tag{3.1}$$

$$U_{d,c}^{non-seed} = \begin{cases} zd - (1-z)Rd & \text{if } a_{d,c}^{non-seed} = \mathscr{S} \\ 0 & \text{if } a_{d,c}^{non-seed} = \mathscr{K} \\ zd - c & \text{if } a_{d,c}^{non-seed} = \mathscr{V} \end{cases} \tag{3.2}$$

A seed believes the news to be true with probability μ. A non-seed, conditional on receiving news, must form beliefs about the likelihood of indirect news being true, denoted by z. Therefore, when sharing unverified news, the ex-ante payoffs for seeds and non-seeds with degree d are $\mu d - (1-\mu)Rd$ and $zd - (1-z)Rd$. In contrast, the payoff from killing (i.e., action $a_{d,c} = \mathscr{K}$) is normalized at 0 for both seeds and non-seeds. As a result, a seed prefers sharing unverified news than killing it if and only if $\mu \geq \frac{R}{R+1}$. Likewise, a non-seed prefers sharing unverified news than killing the story if and only if $z \geq \frac{R}{R+1}$. Therefore, the preferences over sharing or killing without verification are independent of the network.[9]

Note that for a non-seed to receive indirect news, the news either has been verified to be true ($z = 1$) or has yet to be verified ($z = \mu$). Conditional on receiving news, the probability that indirect news is true must be weakly higher than the quality of unverified news, $z \geq \mu$. It follows that if a seed shares unverified news so will a non-seed. This yields us the following observation: When $\mu \geq \frac{R}{R+1}$, both the seeds and non-seeds will share unverified news over killing the story.

First, consider when $\mu < \frac{R}{R+1}$: the seeds will kill unverified news so the decision is between verifying and killing the news. Not verifying (and hence not sharing) earns 0, whereas verifying (and only sharing true news) earns ex-ante payoff $\mu d - c$. Therefore, a

[9] In Appendix C.2, we suppose the total reputational loss from sharing false news is invariant to degree. Then in equilibrium, high degree agents share unverified news rather than killing, and vice versa. Given that the results under the baseline model are rich, we leave this extension as a discussion.

seed with degree d and cost c will verify if and only if $c \leq \mu d$, and she will only share news that she has verified to be true. Given that only true news is shared by the seed, the non-seed correctly anticipates that news she received must be true, $z = 1$. Therefore, it is optimal for a non-seed with degree d to never verify, and share all news received to earn payoff d.

Second, consider when $\mu \geq \frac{R}{R+1}$: everyone will share unverified news so the decision is between verifying and sharing without verification. Suppose an agent with degree d does not verify, she earns ex-ante payoffs $\mu d - (1-\mu)Rd$ as a seed and $zd - (1-z)Rd$ as a non-seed. Now suppose the agent verifies, she earns ex-ante payoffs $\mu d - c$ as the seed and $zd - c$ as a non-seed. Hence, a seed with degree d and cost c will verify if and only if $c \leq (1-\mu)Rd$, whereas a non-seed will verify if and only if $c \leq (1-z)Rd$. Note that the verification strategy for agent i is a cutoff strategy. Because agent i's cost is drawn random from the cost distribution H, this induces a probability distribution over her actions. Dropping the "c notation", we define the verification strategy for a seed with degree d as $p_d : \mathbb{R}^+ \to [0,1]$ and for a non-seed $q_d : \mathbb{R}^+ \to [0,1]$. p_d can be interpreted as the ex-ante probability of seed verifying given degree d, and q_d is the probability of non-seed verifying given degree d.

In summary, when $\mu < \frac{R}{R+1}$, a seed with degree d will verify with probability $p_d = H(\mu d)$, will only share true news after verification, and will kill unverified news; a non-seed will never verify and always shares unverified news. When $\mu \geq \frac{R}{R+1}$, the equilibrium is then characterized as follows: all seeds and non-seeds will only share true news after verification, and will share unverified news; for all degree d, the verification probabilities of a seed and a non-seed with degree d are $p_d = H((1-\mu)Rd)$ and $q_d = H((1-z)Rd)$ respectively, where

$$1 - z = \frac{(1-\mu)(1-\sum_k f(k)p_k)}{(1-\mu)(1-\sum_k f(k)p_k) + \mu(1+(n-2)\sum_k f(k)q_k)}. \tag{3.3}$$

This last expression of z is obtained as follows: Recall z to be the probability belief that the indirect news is true conditional on receiving it. Let ω be the state of receiving indirect

news, and let $v \in \{T, F\}$ be the veracity of news. Using the Bayes rule,

$$z = Pr(v = T|\omega) = \frac{Pr(\omega|v = T) \times Pr(v = T)}{Pr(\omega|v = T) \times Pr(v = T) + Pr(\omega|v = F) \times Pr(v = F)}. \tag{3.4}$$

As previously stated, when $\mu \geq \frac{R}{R+1}$, both verified and unverified true news are always shared. So conditional on the news being true, individuals will receive indirect news with certainty, $Pr(\omega|v = T) = 1$. Moreover, $Pr(v = T) = \mu$. The expression of z can therefore be written as

$$z = \frac{\mu}{\mu + (1 - \mu)Pr(\omega|v = F)}. \tag{3.5}$$

To find the expression of $Pr(\omega|v = F)$, we first solve for $Pr(\omega_k|v = F)$ — the probability of receiving news from an individual of degree k conditional on the news being false. There are two ways to receive false news: either the sender is the seed who did not verify, or she is a non-seed who received unverified false news and did not verify. Recall that verified false news is not shared. Conditional on the news being false, an individual only receives news from others when the news is unverified — the probability of a non-seed receiving unverified false news equals $Pr(\omega|v = F)$. Therefore,

$$Pr(\omega_k|v = F) = \frac{1}{n-1}(1 - p_k) + \frac{n-2}{n-1}Pr(\omega|v = F)(1 - q_k). \tag{3.6}$$

Since neighbours' degrees are unknown, in expectation, the probability of receiving news given that the news is false equals

$$\begin{aligned} Pr(\omega|v = F) &= \sum_k f(k)Pr(\omega_k|v = F) \\ &= \frac{1}{n-1}\sum_k f(k)(1 - p_k) + \frac{n-2}{n-1}Pr(\omega|v = F)\sum_k f(k)(1 - q_k). \end{aligned} \tag{3.7}$$

Solving for

$$Pr(\omega|v = F) = \frac{\sum_k f(k)(1 - p_k)}{(n-1) - (n-2)\sum_k f(k)(1 - q_k)} \tag{3.8}$$

and substituting it into the definition of z gives us eq. (3.3).

3.2.1 Equilibrium

Our solution concept is Bayesian Nash equilibrium (BNE). The sharing strategies conditional on verification decision described above are dominant strategies; the verification strategies described above are optimal given beliefs on the veracity of indirect news. Beliefs are consistent with strategies: In equilibrium, the probability z is both the likelihood of indirect news being true (under strategies $(\mathbf{p},\mathbf{q})$ where $\mathbf{p}=(p_d)_{d\in\mathbb{N}}$ and $\mathbf{q}=(q_d)_{d\in\mathbb{N}}$) and the ex-ante belief of non-seeds that the news they received is true.

Proposition 3.1. *There exists a unique equilibrium:*

(i) Suppose $\mu < \frac{R}{R+1}$. For all degree d:

- *Seeds verify ($\mathscr{V}$) with probability $p_d^* = H(\mu d)$ and kill ($\mathscr{K}$) otherwise;*
- *Non-seeds never verify and always share ($\mathscr{S}$);*
- *$z^* = 1$.*

(ii) Suppose $\mu \geq \frac{R}{R+1}$. For all degree d:

- *Seeds verify ($\mathscr{V}$) with probability $p_d^* = H((1-\mu)Rd)$ and share ($\mathscr{S}$) otherwise;*
- *Non-seeds verify ($\mathscr{V}$) with probability $q_d^* = H((1-z^*)Rd)$ and share ($\mathscr{S}$) otherwise, where*

$$z^* = 1 - \frac{(1-\mu)(1-\sum_k f(k)p_k^*)}{(1-\mu)(1-\sum_k f(k)p_k^*) + \mu(1+(n-2)\sum_k f(k)q_k^*)}. \tag{3.9}$$

We name the case where seeds kill unverified news (when $\mu < \frac{R}{R+1}$) as the *Killing Equilibrium*, and the case where all agents share unverified news (when $\mu \geq \frac{R}{R+1}$) as the *Sharing Equilibrium*.

Proof. In order to prove that there exists an equilibrium, we show that there exists a fixed point for the verification strategies $(\mathbf{p},\mathbf{q})$ such that the beliefs are consistent with the strategies. To do so, we check the conditions for Brouwer's fixed-point theorem of the verification strategies. p_d and q_d both lie in $[0,1]$ so $\mathbf{p}$ and $\mathbf{q}$ both lie in the set $[0,1]^n$. Fixing degree d, p_d is a trivial function of the parameters, meanwhile, q_d is a function of $(\mathbf{p},\mathbf{q})$. Thus, the vector of best responses $(\mathbf{p}^*,\mathbf{q}^*)$ is a mapping from the compact and convex set $[0,1]^n \times [0,1]^n$ to itself. This mapping is continuous because $h(c)$ is atomless with full support on $[\underline{c},\overline{c}]$ and the

ex-ante payoffs are continuous on $[0,1]$ in p_d, q_d for all d. Existence follows from Brouwer's fixed-point theorem.

The uniqueness of equilibrium when $\mu < \frac{R}{R+1}$ is straightforward: all agents have a unique best response in verification and sharing strategy. Whereas in the case when $\mu \geq \frac{R}{R+1}$, we prove that there exists a unique fixed-point by contradiction. Suppose there exists two distinct equilibria $(\mathbf{p}, \mathbf{q})$ and $(\mathbf{p}', \mathbf{q}')$. Fixing all parameters and distributions, $\mathbf{p} = \mathbf{p}'$ because $p_d = p'_d$ for all d. Consequently, for the two equilibria to be distinct, there must exists a degree d such that $q_d \neq q'_d$. First, suppose $\sum_k f(k) q_k = \sum_k f(k) q'_k$. Then under eq. (3.9), $z = z'$. By the expression $q_d = H((1-z)\mu d)$, it implies that $q_d = q'_d$ for all d, thus reaching a contradiction that the two equilibrium are distinct. Second, suppose $\sum_k f(k) q_k > \sum_k f(k) q'_k$. Then under eq. (3.9), $z > z'$ which implies that $q_d < q'_d$ for all d. Thus $\sum_k f(k) q_k < \sum_k f(k) q'_k$, reaching a contradiction. Therefore, there exists a unique equilibrium. □

Since direct news has never been verified while indirect news could have been, indirect news is more likely to be true than direct news ($z^* \geq \mu$). Quality of indirect news improves the further it travels. Hence, indirect news requires less verification than direct news, i.e., $p^*_d \geq q^*_d$ for all d. Note that $p^*_d = 0$ implies $q^*_d = 0$ — if an individual never verifies as the seed, she would never verify as a non-seed as well.

A seed's strategy only depends on her own degree, the reputational loss, and the information accuracy μ. It is independent of the strategy of other agents. When $\mu < \frac{R}{R+1}$, a non-seed's strategy is also independent of others. However, when $\mu \geq \frac{R}{R+1}$, a non-seed has no information about the identity of the sender and must form expectations of others' strategies based on the network. These beliefs on the verification probabilities are reflected in the probability of receiving true news indirectly z^*. If seeds or non-seeds verify more, the quality of indirect news improves, and the non-seed can verify less. Therefore, the verification of other agents is a strategic substitute for a non-seed's verification.[10]

Fixing the network, it is straightforward to see that verification probability is increasing in degree. In the Killing Equilibrium, only verified true news are shared. As d increases,

[10] Appendix C.3 incorporates strategic complementarity of the verification behaviour by allowing non-seeds to punish seeds. We find similar results as the baseline model.

the potential benefits of revealing and sharing true news is increasing for the seed, so verifying news becomes more attractive and p_d^* increases. Non-seeds never verify because they only receive true news, so q_d^* is constant at 0. In the Sharing Equilibrium, all agents share unverified news. As d increases, the ex-ante punishments for sharing unverified news ($(1-\mu)Rd$ and $(1-z^*)Rd$) increase, so not verifying becomes more costly. Overall, the verification probability p_d^* and q_d^* are both increasing in degree d.

Remark 3.1. *Suppose $\mu \geq \frac{R}{R+1}$. For all agents with degree d:*
If $(1-\mu)Rd_{max} < \underline{c}$, then $p_d^* = 0$, $q_d^* = 0$ and $z^* = \mu$ (Unverified Sharing Equilibrium).
If $(1-\mu)Rd_{min} \geq \bar{c}$, then $p_d^* = 1$, $q_d^* = 0$ and $z^* = 1$ (Verified Sharing Equilibrium).

We identify that there are two extreme scenarios where non-seeds will not verify when $\mu \geq \frac{R}{R+1}$. One, if the punishment for the most connected seed (with degree d_{max}) is not enough to incentivize verification (i.e., $(1-\mu)Rd_{max} < \underline{c}$), then no seed will verify and hence non-seeds will also not verify. Since the news is unverified, non-seeds believe the quality of indirect news to be their prior μ. We denote this as the *Unverified Sharing Equilibrium*. Two, if the punishment for the least connected seed is large enough to incentivize verification (when $(1-\mu)Rd_{min} \geq \bar{c}$), then all seeds will verify. Non-seeds will only receive verified true news, so they will never verify. We denote this as the *Verified Sharing Equilibrium*.

Observe that the Verified Sharing Equilibrium and the Verified Killing Equilibrium have the same outcome — all news is verified by the seed and only true news is spread within the network. These two equilibria are often "adjacent" to each other, differing only by the seeds' strategies to share or kill unverified news. In Section 3.3, we explore how changes in the parameters could shift the equilibrium from one to the other.

3.3 Comparative Static

In this section, we examine the effects of the key variables on the Sharing and Killing Equilibria separately and build the full picture of how their transitions. We first study the effects of the reputational loss R and the information accuracy μ on the probabilities p_d^*, q_d^*, and z^*. Next, we study the network effects and cost effects.

3.3.1 Reputational loss

Proposition 3.2.

(i) Suppose $\mu < \frac{R}{R+1}$. In equilibrium, for all degree d, p_d^*, q_d^* and z^* are invariant to R.

(ii) Suppose $\mu \geq \frac{R}{R+1}$. In equilibrium, there exists a unique $\hat{R} \in \left[\frac{\underline{c}}{(1-\mu)d_{max}}, \frac{\bar{c}}{(1-\mu)d_{min}}\right)$ such that for all degree d:

- *p_d^* is increasing in R;*
- *q_d^* is increasing in R when $R < \hat{R}$ and decreasing in R otherwise;*
- *z^* is increasing in R.*

First, in the Killing Equilibrium, seeds kill unverified news so non-seeds only receive true news. Both of them will never receive punishment. All equilibrium probabilities are independent of R.

Second, when $\mu \geq \frac{R}{R+1}$, unverified news is always shared. A higher reputational loss R increases the ex-ante punishment for sharing unverified news. As a result, both seeds and non-seeds have incentives to verify more, which improves the quality of indirect news z^*. However, a non-seed's incentive to verify diminishes as z^* improves. At high levels of punishment, all seeds will verify so non-seeds will not. We show that q_d^* is initially increasing and then decreasing in R, and that there exists a point $\hat{R}$ where q_d^* is maximised. Overall, the impact of higher seed verification on the quality of indirect news outweighs the impact of lower non-seeds verification. Therefore, z^* is increasing in R. The formal proof is left in Appendix C.4.

The point $\hat{R}$ where q_d^* is maximised is constant for all degree d. The intuition is as follows: As punishment increases, non-seeds only start to verify less when the improvement in the quality of indirect news z^* outpaces the rise in punishment R. So the point at which q_d^* is maximised depends only on z^*. Since z^* is a function of the average verification probabilities $\bar{p}^*$ and $\bar{q}^*$, z^* only depends on the degree distribution, which is common for all agents. Moreover, an agent's own-degree d only impacts the magnitude of q_d^*: as seen in Figure 3.1, a non-seed with higher degree d is more likely to verify than a non-seed with lower degree d_{min}, $q_d^* \geq q_{d_{min}}^*$.

Finally, we place the Sharing Equilibrium together with the Killing Equilibrium. Suppose $\mu d_{min} \geq \bar{c}$, the least connected seed verifies with certainty in the Killing Equilibrium. So as R increases to the point where $(1-\mu)R = \mu$, all seeds verify with certainty (Verified Sharing Equilibrium). Then all seeds continue to verify with certainty past the transition point (Verified Killing Equilibrium). This is shown in Figure 3.1a. If the punishment is sufficiently high, then all news is verified by the seed and no false news spreads. Conversely, suppose $\mu d_{min} < \bar{c}$. As R increases, not all seeds verify in the Sharing Equilibrium ($p_d^* < 1$ for some d) before transitioning to the Killing Equilibrium. When the punishment is high, even though no false news spreads, not all true news is verified and shared by the seeds in the Killing Equilibrium (Figure 3.1b). Overall, the results of (i) p_d^* and z^* being (weakly) increasing in R and (ii) q_d^* being increasing then decreasing in R remain valid across these two scenarios.

3.3.2 Information accuracy

Proposition 3.3.

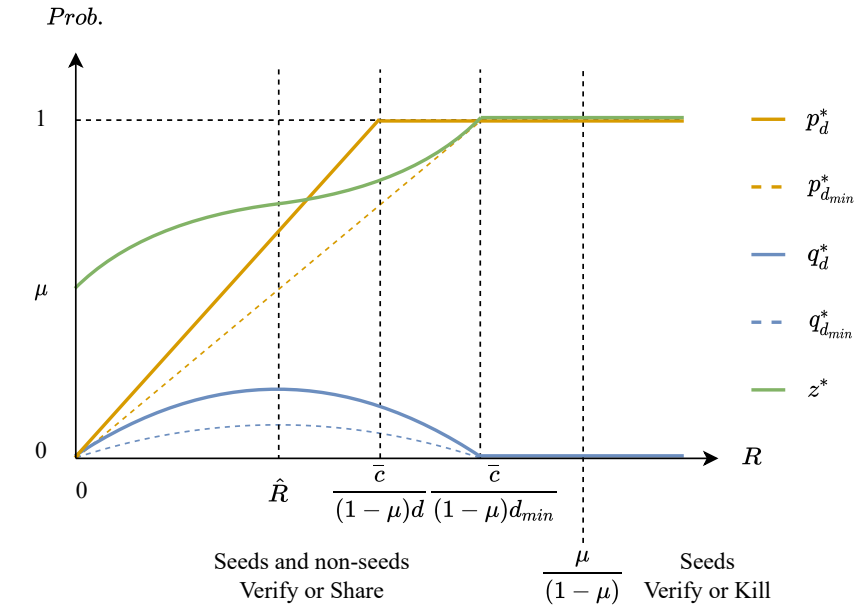
(i) Suppose $\mu < \frac{R}{R+1}$. In equilibrium, for all degree d, p_d^* is increasing in μ, while q_d^* and z^* are invariant to μ.

(ii) Suppose $\mu \geq \frac{R}{R+1}$. In equilibrium, there exists a unique $\hat{\mu} \in (1 - \frac{\bar{c}}{Rd_{min}}, 1)$ such that for all degree d:

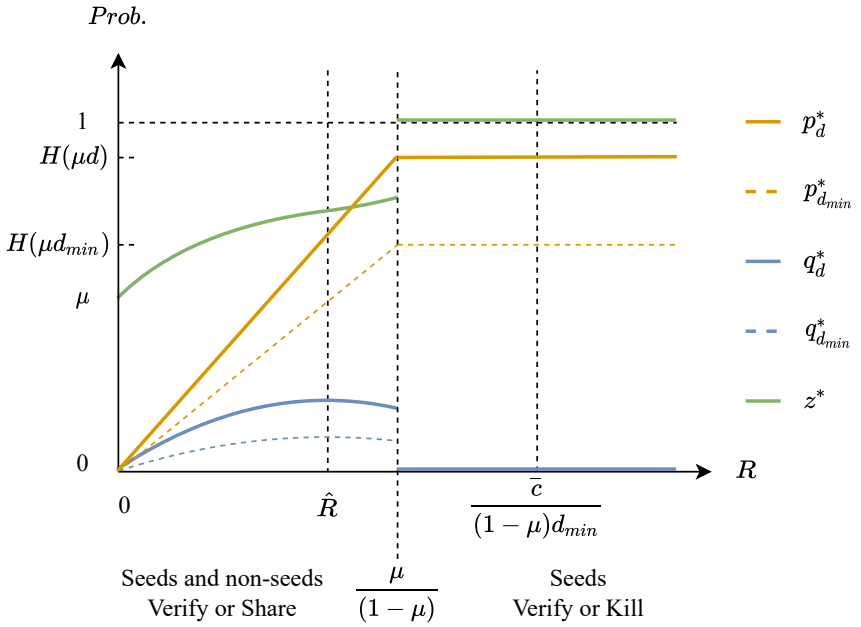
- *p_d^* is decreasing in μ;*
- *q_d^* is increasing in μ when $\mu < \hat{\mu}$ and decreasing in μ otherwise;*
- *z^* is decreasing in μ when $\mu < \hat{\mu}$ and increasing in μ otherwise.*

First, in the Killing Equilibrium, a higher information accuracy from the source encourages seeds to verify and share true news. So p_d^* is increasing in μ. However, $z^* = 1$ and $q_d^* = 0$, both independent of μ.

Second, in the Sharing Equilibrium, an improvement in the information accuracy reduces the need for seeds to verify, so p_d^* decreases. On one hand, the quality of indirect news worsens because seeds verify less as μ increases. On the other hand, the quality of indirect news improves because the information accuracy improves. As μ tends to 1, all indirect



(a) When $\mu d_{min} \geq \bar{c}$.



(b) When $\mu d_{min} < \bar{c}$.

Fig. 3.1 Effects of punishment R on p^*, q^*, z^* for agent of degree d and d_{min}. Assume $\underline{c} = 0$.

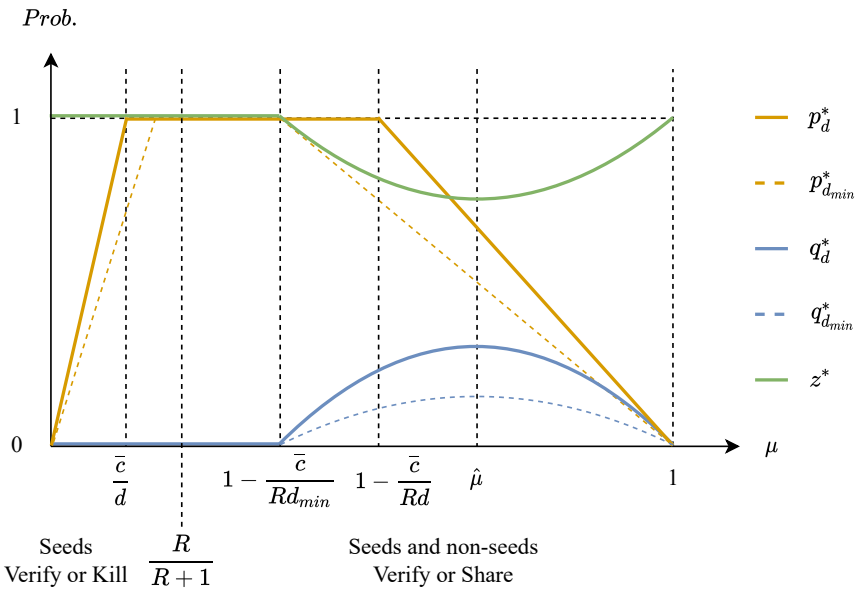
news tends to be true, $z^* = 1$. We show that z^* is initially decreasing and then increasing in μ. There exists a point $\hat{\mu}$ where z^* is minimised. Non-seed verification probability responds inversely to changes in the quality of indirect news: q_d^* is initially increasing and then decreasing in μ and maximised at $\hat{\mu}$. The formal proof is left in Appendix C.4. By the same logic as $\hat{R}$, the point $\hat{\mu}$ is constant across all agents.

Again, we place the Sharing Equilibrium together with the Killing Equilibrium. Initially, when μ is small, seeds kill unverified news. But as μ increases to the point where $\mu = \frac{R}{R+1}$, both seeds and non-seeds start to share unverified news. The transition from the Killing Equilibrium to the Sharing Equilibrium is continuous if and only if $d_{min}\frac{R}{R+1} \geq \overline{c}$ (Figure 3.2a). If $\frac{R}{R+1}d_{min} < \overline{c}$, lowest degree agents have yet to verify with certainty before transitioning into the Sharing Equilibrium. Non-seeds suddenly begin receiving unverified news which cause a spike in their verification (Figure 3.2b). Overall, p_d^* is initially increasing and then decreasing in μ; q_d^* is at first constant at 0, then increasing, and then decreasing in μ; z^* is at first constant at 1, then decreasing, and then increasing in μ. These observations remain valid across the two scenarios.

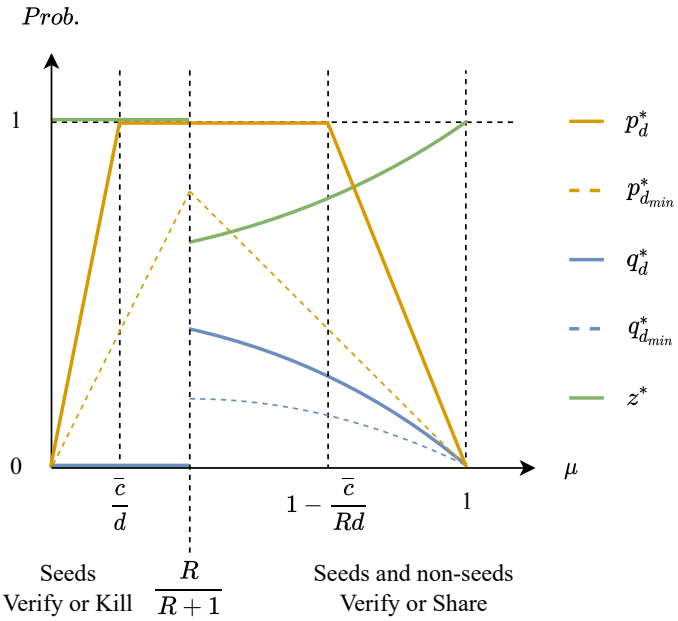
Note that improving information accuracy μ does not always increase the quality of news in circulation z^*. When μ is low, the seed bares the responsibility to verify and ensure good quality news being shared. The seed can afford to shirk when the information becomes more accurate.

3.3.3 Network

Now we examine the effects of degree distribution on the equilibrium. What happens if links are added to the network? What happens if connectivity becomes more concentrated among a selected few? The idea of adding links to make a network denser is captured by the relation of First-Order Stochastic Dominance (FOSD) of degree distribution. The idea of reducing link dispersion or increasing degree concentration is captured by the relation of Second-Order Stochastic Dominance (SOSD). We state the definitions of FOSD and SOSD from Rothschild and Stiglitz (1970) that describe the relationship between the degree distribution rankings



(a) When $\frac{R}{R+1}d_{min} \geq \bar{c}$.



(b) When $\frac{R}{R+1}d_{min} < \bar{c}$.

Fig. 3.2 Effects of prior μ on p^*, q^*, z^* for agent of degree d and d_{min}. Assume $\underline{c} = 0$.

and the expectation of a function. The expectation of the function $u(k)$ over the distribution f is $E_f[u] = \int_{-\infty}^{\infty} f(k)u(k)dk$.

Definition 3.1. *Probability distribution f_1 first-order stochastic dominates f_2 if and only if $E_{f_1}[u] \geq E_{f_2}[u]$ for all non-decreasing functions $u(x)$. Furthermore, f_1 second-order stochastic dominates f_2 if and only if $E_{f_1}[u] \geq E_{f_2}[u]$ for all non-decreasing concave functions $u(x)$. The results hold with strict inequality if $u(x)$ is also a strictly increasing function.*

Next, we explore the effects of dominance shifts in degree distribution on the equilibrium strategies of an agent with degree d. Fixing degree d, suppose there are two conditional degree distributions f_1 and f_2 with respective equilibrium probabilities $p^*_{1,d}, q^*_{1,d}, z^*_1$ and $p^*_{2,d}, q^*_{2,d}, z^*_2$ and average equilibrium probabilities $\bar{p}^*_1, \bar{q}^*_1$ and $\bar{p}^*_2, \bar{q}^*_2$. We show the following results for degree distribution dominance relation:

Proposition 3.4. *Suppose f_1 FOSD f_2, or f_1 SOSD f_2 and $H(c)$ is a concave function in c:*

(i) If $\mu < \frac{R}{R+1}$, for all degree d, p^*_d, q^*_d and z^* remain constant under f_1 and f_2.

(ii) If $\mu \geq \frac{R}{R+1}$, for all degree d, $p^*_{1,d} = p^*_{2,d}$, $q^*_{1,d} \leq q^*_{2,d}$, and $z^*_1 \geq z^*_2$.

If $H(c)$ is also a strictly increasing function in c, then the results hold with strict inequality.

First, the verification probability of a seed with degree d is constant across conditional degree distributions. Observe that the average equilibrium probability of seed verification is:

$$\bar{p}^*_1 = \begin{cases} \sum_k f_1(k)H(\mu k) & \text{if } \mu < \frac{R}{R+1} \\ \sum_k f_1(k)H((1-\mu)Rk) & \text{if } \mu \geq \frac{R}{R+1} \end{cases} \tag{3.10}$$

where $H(.)$ is a cumulative distribution and hence a non-decreasing function. Thus, by Definition 3.1, the average equilibrium probability of seed verification is higher under f_1, $\bar{p}^*_1 \geq \bar{p}^*_2$. The same can be shown in the Killing Equilibrium where $\bar{q}^*_1 \geq \bar{q}^*_2$ given that z^* is invariant to changes in the network.

Second, we argue that when $\mu \geq \frac{R}{R+1}$, the quality of indirect news is higher under the stochastic dominant distribution, $z^*_1 \geq z^*_2$. The formal proof is in Appendix C.4. Instead, we provide a sketch of the proof by contradiction. Suppose the quality of indirect news is poorer

under f_1, $z_1^* < z_2^*$, then all non-seeds will verify more under f_1 than f_2, $q_{1,d}^* > q_{2,d}^*$ for all d. This implies that the average equilibrium probability of non-seed verification $\bar{q}^*$ is higher under f_1 since q_d^* is increasing in d. However, if $\bar{p}^*$ and $\bar{q}^*$ are both higher under f_1, then the probability of receiving truth news indirectly would also be higher under f_1, resulting in $z_1^* > z_2^*$ and reaching a contradiction.

Third, since non-seeds verify less under higher z^*, for all degree d, the equilibrium probability of non-seed verification is lower under f_1, $q_{1,d}^* \leq q_{2,d}^*$. However, the ranking between $\bar{q}_1^*$ and $\bar{q}_2^*$ is ambiguous. On one hand, non-seeds have a lower probability of verification for all degrees under f_1. On the other hand, under FOSD, the conditional degree distribution f_1 has a higher density mass on agents with higher degrees than f_2 (and under SOSD, a higher density mass on agents with intermediate degrees). Given that q_d^* is increasing in d, a denser network (and a more egalitarian network while assuming concavity in $H(.)$) implies a higher average likelihood of verification. Overall, the effect of the network on the average probability of non-seed verification is unclear.

Concave and Non-concave cost function

To illustrate the result, we look at the numerical solution under different degree distributions and cost functions. Consider a network with $n = 20$, $\mu = 0.9$ and $R = 1$. From the perspective of an agent with degree 5, suppose there are 3 conditional degree distributions: (i) f_1 has 1 agent with degree 5, 19 agents with degree 7, (ii) f_2 has 20 agents with degree 5, and (iii) f_3 has 9 agents with degree 3, 2 agents with degree 5, and 9 agents with degree 7. Note that f_1 FOSD f_2 and f_2 SOSD f_3 (because f_3 is a mean-preserving spread of f_2).

Suppose $H(c) = log(c+1)/log(2)$ for $c \in [0,1]$ which is a concave increasing function in c. In equilibrium, $p_{d=5}^*$ equals 0.585 for all conditional degree distributions; $q_{d=5}^*$ equals 0.070, 0.108, 0.110 under conditional degree distribution f_1, f_2, f_3 respectively, while z^* equals 0.990, 0.985, 0.984. The equilibrium probabilities behave as described in Proposition 3.4(ii).

For completeness, suppose $H(c) = c^2$ for $c \in [0,1]$ which is a non-concave increasing function in c. In equilibrium, $p_{d=5}^*$ equals 0.25 for all conditional degree distributions; $q_{d=5}^*$ equals 0.024, 0.047, 0.042 under conditional degree distribution f_1, f_2, f_3 respectively, while

z^* equals 0.969, 0.957, 0.959. $q^*_{d=5}$ is lower and z^* is higher under f_3 than f_2. Intuitively, under a non-concave $H(c)$, a small increase in degree disproportionately increases seed verification. So as there is more dispersion in network degree, the increase in verification from more high-degree seeds outweighs the decrease in verification from more low-degree seeds. Overall, when there is a mean-preserving spread of degrees, the average seed verification is higher which increases z^* and reduces non-seed verification q^*_d.

3.3.4 Verification cost

Similar to the analysis of degree distribution, we examine the effects of the cost distribution on the equilibrium. Suppose there are two cost distributions h_1 and h_2 with the corresponding equilibrium probabilities.

Proposition 3.5. *Suppose h_1 FOSD h_2, or suppose h_1 SOSD h_2 and $F(k)$ is a concave function in k:*

(i) If $\mu < \frac{R}{R+1}$, then for all degree d, $p^*_{1,d} \leq p^*_{2,d}$, while q^*_d and z^* remain constant.

(ii) If $\mu \geq \frac{R}{R+1}$, then for all degree d, $p^*_{1,d} \leq p^*_{2,d}$, and $z^*_1 \leq z^*_2$.

If $F(k)$ is also a strictly increasing function in c, then the results hold with strict inequality.

When the density of costs shifts higher, it is on average more expensive to verify news for all agents. Seeds will naturally verify less, resulting in a poorer indirect quality of news. Non-seeds face two opposing effects: they need to verify more due to the poorer indirect news quality, but the news is becoming more costly to verify. The overall effect is ambiguous. The formal proof is left in Appendix C.4.

3.4 Platform

In this section, we introduce an external agent — social-media platform (hereinafter *platform*) — who wishes to maximise views to generate ad-revenue through a choice of information accuracy. Since the revenue is independent of news veracity, the platform only cares about how many users the content reaches. Given that seeds always receive news, we define *spread*

as the probability for a non-seed to receive a piece of news. A larger spread then translates to a far-reaching piece of news and the platform earns higher revenue.

Definition 3.2. *Spread is defined as*

$$\mathbf{1}_{\mu<\frac{R}{R+1}}\left[\mu\sum_k f(k)H(\mu k)\right]+\mathbf{1}_{\mu\geq\frac{R}{R+1}}\left[\mu+(1-\mu)\frac{1-\bar{p}^*}{1+(n-2)\bar{q}^*}\right] \tag{3.11}$$

The expression of the spread is obtained as follows: In the Killing Equilibrium, true news reaches everyone if and only if the seed verifies. Meanwhile, false news is never shared. So the expected spread when $\mu < \frac{R}{R+1}$ is

$$\mu\sum_k f(k)H(\mu k). \tag{3.12}$$

In the Sharing Equilibrium, true news is always shared with or without verification (spread $= 1$). Conditional on the news being false, as described in eq. (3.8), the probability of a non-seed receiving news is $Pr(\omega|v=F)$. So the expected spread when $\mu \geq \frac{R}{R+1}$ is

$$\mu+(1-\mu)Pr(\omega|v=F)=\mu+(1-\mu)\frac{1-\bar{p}^*}{1+(n-2)\bar{q}^*}. \tag{3.13}$$

Proposition 3.6. *The spread is increasing in the information accuracy μ.*

Proof. Recall that when $\mu < \frac{R}{R+1}$, increasing the information accuracy increases the fraction of true news μ and increases the probability of seeds verifying $H(\mu k)$. Given that only true verified news are shared, both effects increase the spread. When $\mu \geq \frac{R}{R+1}$, increasing μ has two opposing effects on the spread: (i) there is more true news (which are always shared), and (ii) there is less false news (which are only shared when they are unverified). Since $\frac{1-\bar{p}^*}{1+(n-2)\bar{q}^*} < 1$, an increases in μ reduces the amount of false news shared by a fraction less than 1 while increases the amount of true news shared by 1. Overall, the spread is increasing in μ. □

Now suppose the platform can choose the information accuracy (e.g., by monitoring the content) before it enters the network. We superimpose the decision-making process of the

platform onto the verification and sharing model described above. Assume the platform can choose μ at a quadratic investment cost, $\frac{1}{2}K\mu^2$ where $K > 0$. The platform faces the following objective function:

$$argmax_{\mu} \ \mathbf{1}_{\mu < \frac{R}{R+1}} \left[\mu \sum_k f(k) H(\mu k) \right] + \mathbf{1}_{\mu \geq \frac{R}{R+1}} \left[\mu + (1-\mu) \frac{1-\bar{p}^*}{1+(n-2)\bar{q}^*} \right] - \frac{1}{2} K \mu^2 \tag{3.14}$$

For concreteness, assume uniform cost distribution between 0 and 1. Under a regular network with degree d, the equilibrium information accuracy equals

$$\mu^* = \begin{cases} \in [\frac{R}{R+1}, 1) & \text{if } K \in (0, \frac{R+1}{R}] \\ 1/K & \text{if } K \in (\frac{R+1}{R}, d] \\ 1/d & \text{if } K \in (d, 2d] \\ 0 & \text{if } K > 2d \end{cases}. \tag{3.15}$$

The derivation is left in the Appendix C.4. Figure 3.3 visualizes the computational result of this example when the reputational loss $R = 1$. Together, they demonstrate that the platform provides a fixed level of information accuracy for a range of intermediate costs, and stops investing all together at high costs. The intuition is as follows: When $\mu = \bar{c}/d < \frac{R}{R+1}$, all seeds of degree d will verify and share true news. Lowering information accuracy both reduces seed verification $H(\mu d)$ and reduces the fraction of true news μ, which greatly reduces spread. Therefore, the platform has incentives to sustain a minimal level of accuracy required $\mu^* = \bar{c}/d$ such that these agents continue to verify. But once this accuracy becomes too expensive, the platform would rather not invest at all. This *Minimal accuracy required* to sustain verification is lower for denser networks. Following this insight, we compare the optimal information accuracy between two networks using the FOSD and SOSD relation.

Proposition 3.7. *Suppose there are two network distributions f_1 and f_2 where f_1 FOSD f_2, or where f_1 SOSD f_2 and $H(c)$ is a concave function in c. In equilibrium, there exists two*

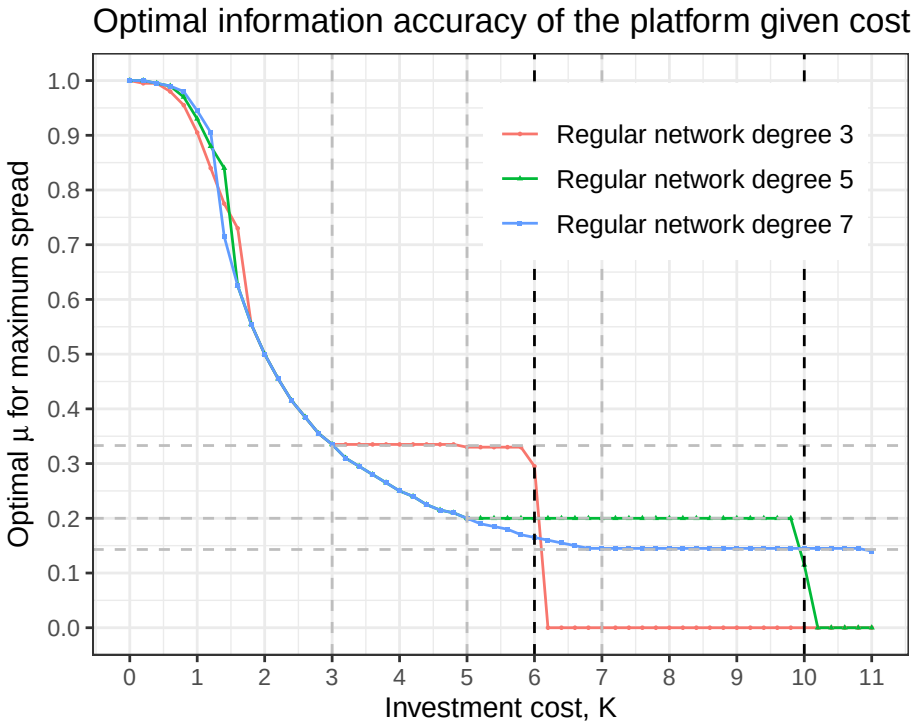


Fig. 3.3 Optimal μ for the platform with investment cost K under a regular network with degree 3 (Green), 5 (Red), and 7 (Blue). Assuming uniform cost distribution between 0 and 1, and $R = 1$.

cost thresholds $\hat{K}$ and $\hat{K}'$ such that

$$\begin{cases} \mu_1^* \leq \mu_2^* & \text{if } K \in (\hat{K}, \hat{K}'), \\ \mu_1^* \geq \mu_2^* & \text{if } K \leq \hat{K} \text{ or } K \geq \hat{K}'. \end{cases} \tag{3.16}$$

The proposition states that when information accuracy is expensive (or cheap), the platform would invest more under a denser (or more egalitarian) network, but otherwise, invest less. We first explain the intuition of the FOSD result and then show how the SOSD result applies with the same logic.

When μ equals 1, all news are true and therefore shared, so the network obtains maximum spread. As discussed before, highly-connected agents verify more, so lowering μ will hurt the spread more in a denser network. The platform has higher incentive to keep μ close to

1 for a denser network. However, as information accuracy becomes more expensive, the platform has less incentive to keep a high μ given that a denser network yields a lower spread.

Next suppose $\mu < \frac{R}{R+1}$. Recall that highly connected agents earn higher benefits from sharing true news, so they verify even under poorer information accuracy. Hence, denser networks can sustain verification at a lower information accuracy. Thus, the platform can afford the Minimal accuracy required under denser networks even when accuracy is expensive. A sparser network, on the other hand, has less incentive to verify, so its Minimal accuracy required is higher. If this level of investment required is expensive, the platform would rather not invest at all. This phenomenon can be seen in Figure 3.3: At $K = 7$, $\mu^* = 1/5$ for a regular network of degree 5 and $\mu^* = 1/7$ for degree 7; At $K = 10$, $\mu^* = 0$ for degree 5 and $\mu^* = 1/7$ for degree 7. In summary, as investment cost increases from 0, μ^* is initially higher for a denser network, but then lower, and then higher.

The intuition of the SOSD relation is analogue to the aforementioned FOSD result. Here, we say the network is "denser" when degrees are less dispersed and the network is more egalitarian. Under the assumption that $H(c)$ is concave, a "denser" network has a higher average probability of verification. The rest of the proof follows as before.

Concave and Non-concave cost function

To illustrate, Figure 3.4a compares the optimal accuracy μ^* under a regular network with degree 5 and a bimodal network with degrees 3 and 7 — half of the agents have degree 3 and the other half have degree 7. This bimodal network distribution is a mean-preserving spread of the regular network. Therefore, by definition, the regular network second-order stochastically dominates the bimodal network. Observe that the optimal accuracy plateaus at $\mu^* = 1/5$ for the regular network, compared to the bimodal network with two plateaus: one at $\mu^* = 1/3$ and the other at $1/7$. Intuitively, low-degree agents are the first to stop verifying as information accuracy decreases. So the platform initially provides the Minimal accuracy required for low-degree agents to continue verifying. But once the information accuracy becomes too expensive, the platform then only provides the Minimal accuracy required for high-degree agents.

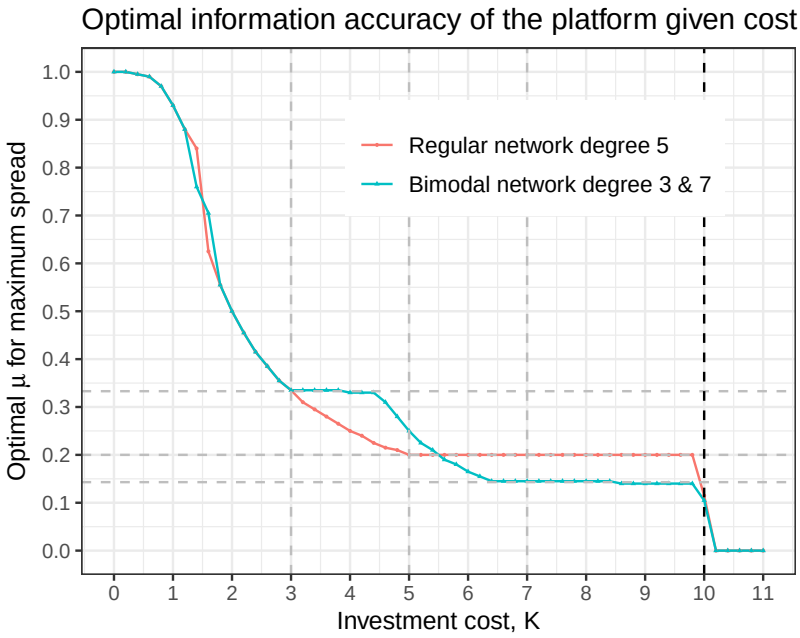
For completeness, now suppose $H(c)$ is a strictly convex function. As seen in Figure 3.4b, the network effect on the Minimal accuracy required persists when information accuracy is expensive.

3.5 Conclusion

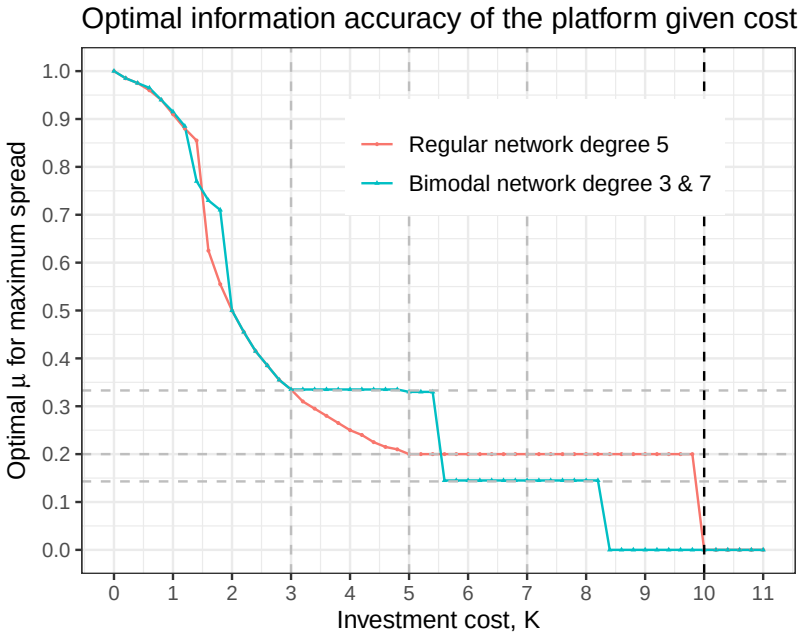
This paper analyzes the impact of the social network on the spread of misinformation. Our model captures the key features of content sharing on social media sites: (i) consumers share content over social networks, (ii) consumers verify independently for sharing benefits while in fear of reputational loss, and (iii) platforms and media producers benefit from consumer viewership. Upon receiving information, an agent can decide to share without verification, kill when not verifying, or verify at a cost to transmit only true news. Not all truthful news is shared when agents kill unverified news, while misinformation spreads when agents share without verification.

We provide several comparative static results. Of particular interest is the effect of the network. Overall, denser networks and more egalitarian networks reduce the amount of falsehood in circulation. Additionally, we show that a views-maximising platform has incentive to invest in higher information accuracy for these networks when monitoring news sources is expensive. In the modern world where fact-checking vast amount of content is expensive, the result suggests better news quality under denser and more egalitarian networks. However, following from the fact that the quality of indirect news may not be increasing in information accuracy, the platform's investment in the quality of news sources may crowd out consumer verification and increase misinformation.

This paper serves as a basis for the study of socially-networked information verification and sharing. Several directions for future work could build on our model. Following Acemoglu et al. (2021), one possible extension is for agents to derive value from getting additional shares or "retweets" on the stories they shared downstream. Under degree independence, all agents face the same conditional degree distribution. So the expected benefit from subsequent shares is simply a lump sum constant across agents, thus increases incentives to



(a) When $H(c) = c$ for $c \in [0,1]$.



(b) When $H(c) = c^2$ for $c \in [0,1]$.

Fig. 3.4 Optimal μ for the platform with investment cost K under a regular network with degree 5 (Blue), and a bimodal network with degrees 3 and 7 (Red). Assume $R = 1$.

verify. Another extension is to allow network users to punish their senders. Verification from agents downstream will reveal that their sender has shared false news and impose punishment on their sender. Then the game features both strategy substitutes and strategic complements. We briefly discuss this extension in Appendix C.3.

There are a few limitations of our model. First, we assume the sharing network is a tree which suppresses interesting network features such as clustering and homophily. Second, we assume degrees are private knowledge. It is possible to allow degrees of neighbours to be observable. Then agents' posterior beliefs would depend on the degree of their sender. Agents will likely share content from well-connected sender without verification and kill content from others. Platforms would then target influencers not only because of their reach, but also because they have good incentives to verify and hence are trusted. This can be further extended to a model of network formation where there are strong incentives to link to high degree agents who are more trusted. This would complement work like the law of the few (Galeotti and Goyal, 2010).

Bibliography

Acemoglu, D., Dahleh, M. A., Lobel, I. and Ozdaglar, A. (2011), 'Bayesian learning in social networks', *The Review of Economic Studies* **78**(4), 1201–1236.

Acemoglu, D., Ozdaglar, A. and ParandehGheibi, A. (2010), 'Spread of (mis) information in social networks', *Games and Economic Behavior* **70**(2), 194–227.

Acemoglu, D., Ozdaglar, A. and Siderius, J. (2021), 'Misinformation: Strategic sharing, homophily, and endogenous echo chambers', *Working paper* .

Agranov, M., Gillen, B. and Persitz, D. (2020), 'Information aggregation on networks: an experimental study', *Working paper* **.**(.), .

Ali, S. N. and Miller, D. A. (2016), 'Ostracism and forgiveness', *American Economic Review* **106**(8), 2329–48.

Allcott, H. and Gentzkow, M. (2017), 'Social media and fake news in the 2016 election', *Journal of economic perspectives* **31**(2), 211–36.

Altay, S., Hacquin, A.-S. and Mercier, H. (2020), 'Why do so few people share fake news? it hurts their reputation', *new media & society* p. 1461444820969893.

Ambrus, A., Mobius, M. and Szeidl, A. (2014), 'Consumption risk-sharing in social networks', *American Economic Review* **104**(1), 149–82.

Bala, V. and Goyal, S. (1998), 'Learning from neighbours', *The review of economic studies* **65**(3), 595–621.

Bala, V. and Goyal, S. (2001), 'Conformism and diversity under social learning', *Economic theory* **17**(1), 101–120.

Bandiera, O., Barankay, I. and Rasul, I. (2009), 'Social connections and incentives in the workplace: Evidence from personnel data', *Econometrica* **77**(4), 1047–1094.

Barabási, A.-L. and Albert, R. (1999), 'Emergence of scaling in random networks', *Science* **286**(5439), 509–512.

Barr, A. and Oduro, A. (2002), 'Ethnic fractionalization in an african labour market', *Journal of Development Economics* **68**(2), 355–379.

Becker, J., Brackbill, D. and Centola, D. (2017), 'Network dynamics of social influence in the wisdom of crowds', *Proceedings of the national academy of sciences* **114**(26), E5070–E5076.

Becker, J., Porter, E. and Centola, D. (2019), 'The wisdom of partisan crowds', *Proceedings of the National Academy of Sciences* **116**(22), 10717–10722.

Benjamini, I., Chan, S.-O., O'Donnell, R., Tamuz, O. and Tan, L.-Y. (2016), 'Convergence, unanimity and disagreement in majority dynamics on unimodular graphs and random graphs', *Stochastic Processes and their Applications* **126**(9), 2719–2733.

Bernheim, B. D., Peleg, B. and Whinston, M. D. (1987), 'Coalition-proof nash equilibria i. concepts', *Journal of economic theory* **42**(1), 1–12.

Besley, T. and Prat, A. (2006), 'Handcuffs for the grabbing hand? media capture and government accountability', *American economic review* **96**(3), 720–736.

Bloch, F., Genicot, G. and Ray, D. (2008), 'Informal insurance in social networks', *Journal of Economic Theory* **143**(1), 36–58.

Bramoullé, Y. and Goyal, S. (2016), 'Favoritism', *Journal of Development Economics* **122**, 16–27.

Bramoullé, Y. and Huremović, K. (2017), 'Promotion through connections: Favors or information?', *arXiv preprint arXiv:1708.07723* .

Bramoullé, Y. and Kranton, R. (2007), 'Risk-sharing networks', *Journal of Economic Behavior & Organization* **64**(3-4), 275–294.

Breza, E. (2016), Field experiments, social networks and development, *in* Y. Bramoulle, A. Galeotti and B. Rogers, eds, 'Oxford Handbook of the Economics of Networks', Oxford University Press.

Brogaard, J., Engelberg, J. and Parsons, C. A. (2014), 'Networks and productivity: Causal evidence from editor rotations', *Journal of Financial Economics* **111**(1), 251–270.

Calvó-Armengol, A. (2004), 'Job contact networks', *Journal of economic Theory* **115**(1), 191–206.

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

Chandrasekhar, A. G., Larreguy, H. and Xandri, J. P. (2020), 'Testing models of social learning on networks: Evidence from two experiments', *Econometrica* **88**(1), 1–32.

Charlson, G. (2022), 'In platforms we trust: misinformation on social networks in the presence of social mistrust'.

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

Choi, S., Gale, D. and Kariv, S. (2005), *Behavioral aspects of learning in social networks: an experimental study*, Emerald Group Publishing Limited.

Choi, S., Gallo, E. and Kariv, S. (2016), Networks in the laboratory, *in* Y. Bramoulle, A. Galeotti and B. Rogers, eds, 'Oxford Handbook of the Economics of Networks', Oxford University Press.

Coleman, J. S., Katz, E. and Menzel, H. (1966), *Medical innovation: A diffusion study*, Indianapolis: Bobbs-Merrill Company.

Currarini, S., Jackson, M. O. and Pin, P. (2009), 'An economic model of friendship: Homophily, minorities, and segregation', *Econometrica* **77**(4), 1003–1045.

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

DeMarzo, P. M., Vayanos, D. and Zwiebel, J. (2003), 'Persuasion bias, social influence, and unidimensional opinions', *The Quarterly journal of economics* **118**(3), 909–968.

Ellman, M. and Germano, F. (2009), 'What do the papers sell? a model of advertising and media bias', *The Economic Journal* **119**(537), 680–704.

Fréchette, G. R. (2012), 'Session-effects in the laboratory', *Experimental Economics* **15**(3), 485–498.

Gagnon, J. and Goyal, S. (2017), 'Networks, markets, and inequality', *American Economic Review* **107**(1), 1–30.

Gale, D. and Kariv, S. (2003), 'Bayesian learning in social networks', *Games and economic behavior* **45**(2), 329–346.

Galeotti, A. and Goyal, S. (2010), 'The law of the few', *American Economic Review* **100**(4), 1468–92.

Galeotti, A., Goyal, S., Jackson, M. O., Vega-Redondo, F. and Yariv, L. (2010), 'Network games', *Review of Economic Studies* .

Galeotti, A. and Merlino, L. P. (2014), 'Endogenous job contact networks', *International economic review* **55**(4), 1201–1226.

Galton, F. (1907), 'Vox populi (the wisdom of crowds)', *Nature* **75**(7), 450–451.

Gentzkow, M., Glaeser, E. L. and Goldin, C. (2006), The rise of the fourth estate. how newspapers became informative and why it mattered, *in* 'Corruption and reform: Lessons from America's economic history', University of Chicago Press, pp. 187–230.

Gentzkow, M. and Shapiro, J. M. (2006), 'Media bias and reputation', *Journal of political Economy* **114**(2), 280–316.

Golub, B. and Jackson, M. O. (2010), 'Naive learning in social networks and the wisdom of crowds', *American Economic Journal: Microeconomics* **2**(1), 112–149.

Golub, B. and Sadler, E. (2016), Learning in social networks, *in* Y. Bramoulle, A. Galeotti and B. Rogers, eds, 'Oxford Handbook of the Economics of Networks', Oxford University Press.

Goyal, S. (forthcoming), *Networks: An Economics Approach*, MIT Press.

Grimm, V. and Mengel, F. (2020), 'Experiments on belief formation in networks', *Journal of the European Economic Association* **18**(1), 49–82.

Haag, M. and Lagunoff, R. (2006), 'Social norms, local interaction, and neighborhood planning', *International Economic Review* **47**(1), 265–296.

Hauser, C., Hopenhayn, H. et al. (2008), 'Trading favors: Optimal exchange and forgiveness', *Carlo Alberto Notebooks* **88**, 864.

Jackson, M. O., Rodriguez-Barraquer, T. and Tan, X. (2012), 'Social capital and social quilts: Network patterns of favor exchange', *American Economic Review* **102**(5), 1857–97.

Kahn, C. M. and Mookherjee, D. (1992), 'The good, the bad, and the ugly: Coalition proof equilibrium in infinite games', *Games and Economic Behavior* **4**(1), 101–121.

Karhunen, P., Kosonen, R., McCarthy, D. J. and Puffer, S. M. (2018), 'The darker side of social networks in transforming economies: Corrupt exchange in chinese guanxi and russian blat/svyazi', *Management and Organization Review* **14**(2), 395–419.

Karlan, D., Mobius, M., Rosenblat, T. and Szeidl, A. (2009), 'Trust and social collateral', *The Quarterly Journal of Economics* **124**(3), 1307–1361.

Katz, E. and Lazarsfeld, P. F. (1966), *Personal Influence, The part played by people in the flow of mass communications*, Transaction Publishers.

Kearns, M., Judd, S. and Vorobeychik, Y. (2012), Behavioral experiments on a network formation game, *in* 'Proceedings of the 13th ACM Conference on Electronic Commerce', pp. 690–704.

Keppo, J., Kim, M. J. and Zhang, X. (2022), 'Learning manipulation through information dissemination', *Operations Research* .

Kets, W., Iyengar, G., Sethi, R. and Bowles, S. (2011), 'Inequality and network structure', *Games and Economic Behavior* **73**(1), 215–226.

Kranton, R. and McAdams, D. (2022), 'Social connectedness and the market for information'.

Lazarsfeld, P. F. and Merton, R. K. (1954), 'Friendship as a social process: A substantive and methodological analysis', *Freedom and control in modern society* **18**(1), 18–66.

Ledeneva, A. C. and Ledeneva, A. V. (1998), *Russia's economy of favours: Blat, networking and informal exchange*, Vol. 102, Cambridge University Press.

Levy, R. (2021), 'Social media, news consumption, and polarization: Evidence from a field experiment', *American economic review* **111**(3), 831–70.

Lippert, S. and Spagnolo, G. (2011), 'Networks of relations and word-of-mouth communication', *Games and Economic Behavior* **72**(1), 202–217.

McDonald, S. (2011), 'What's in the "old boys" network? accessing social capital in gendered and racialized networks', *Social networks* **33**(4), 317–330.

McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001), 'Birds of a feather: Homophily in social networks', *Annual review of sociology* **27**(1), 415–444.

Möbius, M. (2001), 'Trading favors'.

Mobius, M., Phan, T. and Szeidl, A. (2015), 'Treasure hunt: Social learning in the field', *Working Paper Harvard and CEU Budapest* .

Mortensen, D. T. and Vishwanath, T. (1994), 'Personal contacts and earnings: It is who you know!', *Labour economics* **1**(2), 187–201.

Mossel, E., Sly, A. and Tamuz, O. (2014), 'Asymptotic learning on bayesian social networks', *Probability Theory and Related Fields* **158**(1-2), 127–157.

Mostagir, M., Ozdaglar, A. and Siderius, J. (2022), 'When is society susceptible to manipulation?', *Management Science* .

Nava, F. (2016), *Repeated games and networks. In The Oxford Handbook of the Economics of Networks.*, Oxford University Press.

Newman, M. (2010), *Networks: An Introduction*, Oxford University Press, Inc., New York, NY, USA.

Newman, M. (2018), *Networks*, Oxford university press.

Nguyen, N. P., Yan, G., Thai, M. T. and Eidenbenz, S. (2012), Containment of misinformation spread in online social networks, *in* 'Proceedings of the 4th Annual ACM Web Science Conference', pp. 213–222.

Papanastasiou, Y. (2020), 'Fake News Propagation and Detection: A Sequential Model', *Management Science* **66**(5), 1826–1846. Publisher: INFORMS.

Rothschild, M. and Stiglitz, J. E. (1970), 'Increasing risk: I. a definition', *Journal of Economic theory* **2**(3), 225–243.

Shearer, E. and Gottfried, J. (2017), 'News use across social media platforms 2017'.

Törnberg, P. (2018), 'Echo chambers and viral misinformation: Modeling fake news as complex contagion', *PloS one* **13**(9), e0203958.

Walker, M. and Gottfried, J. (2019), 'Republicans far more likely than democrats to say fact-checkers tend to favor one side', *Pew Research Center* .

Zinovyeva, N. and Bagues, M. (2015), 'The role of connections in academic promotions', *American Economic Journal: Applied Economics* **7**(2), 264–92.

Appendix A

Omitted proofs of Chapter 1

A.1 Pairwise favouritism

Suppose instead of random favouritism, agents perform pairwise favouritism — when there is no neighbouring expert, player i always favour neighbour j who then returns the favours. The condition for player i to sustain favouritism with neighbour j becomes

$$-x(1+\frac{\delta}{1-\delta}p(n-1-d_i))-\frac{\delta}{1-\delta}p(|S_F|-2)\alpha+\frac{\delta}{1-\delta}p\beta(n-1-d_j)\geq 0. \quad (1.2")$$

By comparing condition (1.2") against condition (1.2), it is easier to sustain pairwise favouritism than random favouritism if and only if:

$$\left[(n-1-d_j)-\sum_{k\in N_i\cap S_F}\frac{n-1-d_k}{|N_k\cap S_F|}\right]\beta+\Big[(|S_F|-2)-|N_i^c\cap S_F|\Big]\alpha\geq 0. \quad (1.2")$$

Player i gains the undivided favours from neighbour j but loses out on other favour-granting neighbours. If player i on average has less favouritism neighbours than the number of favouritism friends these neighbours have, pairwise favouritism earns more favours than random favouritism. However, as discussed in the main text, unequal connections within the

favouritism group is unlikely to be able sustain favouritism in the first place:

$$(n-1-d_j) - \sum_{k \in N_i \cap S_F} \frac{n-1-d_k}{|N_k \cap S_F|} \leq 1.$$

Therefore, this gain from pairwise favouritism is small. Furthermore, other favouritism players would also favour their partners/groups instead of i when i is the expert. This loss in efficient wages α from pairwise favouritism dominates any potential gain in more favours β:

$$(|S_F|-2) - |N_i^c \cap S_F| > 1.$$

Overall, it is easier to sustain random favouritism than pairwise favouritism and players earn more under random favouritism.

A.2 Endogenous linking

Now consider agents with heterogeneous linking costs who can form connections bilaterally. Suppose there are two types of agents: high type agents with low linking costs and low type agents with high linking costs.

The marginal expected benefit of a market player linking with a favouritism player is $p(\alpha+c)$, whereas the marginal benefit for linking with a market player is pc (eq. (1.4)). This implies that market players prioritize linking with favouritism players. Next, the marginal benefit of a favouritism player linking with a market player is $p(1-\alpha-L+\beta)$ (eq. (1.3)). The marginal expected benefit of a favouritism player linking with another favouritism player j is comprised of efficiency benefits $p(1-\alpha-L+\beta)$, wages as an expert $p\alpha$, and favouritism benefits $p\beta\frac{n-1-d_j}{|N_j \cap S_F|}$. But recall that the removal of a link between favouritism players could reduce favouritism incentives for all others and collapse the favouritism group. Therefore, the benefit for favouritism player linking with another favouritism player j is larger than

$$p\left[(1-\alpha-L+\beta)+\alpha+\beta\frac{n-1-d_j}{|N_j \cap S_F|}\right]. \tag{A.1}$$

Assume an increasing marginal linking cost $f'(d) > 0$. First, consider a high type with degree $d' > d$. A high type would prioritise forming links to sustain favouritism collectively. This means prioritising links with other agents with high degrees, namely other high types. As a result, the high types form the hub (such as a regular subgraph). Second, consider a low type with degree d practising market behaviour. Suppose $pc \geq f'(d)$, then a low type would propose to link with a high type practising favouritism since $p(\alpha + c) \geq pc \geq f'(d)$. If $p(1-\alpha-L+\beta) \geq f'(d')$, the high type with degree d' accepts the link and the equilibrium is a hub-spoke network. If $p(1-\alpha-L+\beta) < f'(d')$, the high type rejects the link and the low types instead link with each other. Thus, the equilibrium is a network where nodes outside the hub form connections among themselves. Instead suppose $pc < f'(d)$, then low types are isolated because linking is too expensive.

In the Pure market equilibrium, low types earn lower expected payoffs than high types because they have fewer connections and are more likely to incur the search cost. The heterogeneity in linking cost induces this payoff inequality. In contrast, when favouritism is sustained in equilibrium, favouritism players (high types) form the hub and cooperate to extract a large portion of the reduced aggregate surplus. As a result, favouritism exacerbates the payoff inequality induced from linking cost heterogeneity.

A.3 Proofs

Proposition 1.3. *Suppose Assumption 1.1 and 1.2 hold. For all networks g, $S_F = N$ is not a subgame perfect equilibrium of the repeated game.*

Proof. Suppose $S_F = N$, then $|N_i^c \cap S_F| = n-1-d_i$ and $|N_j \cap S_F| = d_j$. All players must then satisfy the following condition to not deviate from practising favouritism:

$$-x + \frac{\delta}{1-\delta} p \left[-(n-1-d_i)(1-c-L+\beta) + \beta \sum_{j \in N_i} \frac{n-1-d_j}{d_j} \right] \geq 0 \qquad \text{(A.2)}$$

Consider the player i with the lowest degree, where $d_i \leq d_j \ \ \forall j \in N$. It follows that $\forall j \in N$, $\frac{1}{d_j} \leq \frac{1}{d_i}$, implying $\sum_{j \in N_i} \frac{1}{d_j} \leq \sum_{j \in N_i} \frac{1}{d_i} = 1$. Thus, $\sum_{j \in N_i} \frac{n-1-d_j}{d_j} \leq n-1-d_i$. The left-hand

side of the inequality (A.2) equals:

$$-x+\frac{\delta}{1-\delta}p\left[-(n-1-d_i)(1-c-L)\right] \tag{A.3}$$

which is negative. Player i fails to satisfy the condition because favouritism is aggregate surplus reducing, $1-c>L$, which follows from Assumption 1.2. Therefore, if $S_F=N$, all for network g, the player with lowest degree always has the incentive to deviate to market behaviour. □

Proposition 1.4. *Suppose Assumption 1.1 and 1.2 hold. Under strategy profile s^* and network g, consider any two distinct equilibria, S and S', where $S'_F \neq \varnothing$. If $S_F \subsetneq S'_F$, then $S' \succ S$.*

Proof. The coalition of interest is the set of players in S'_F but not in S_F. We show that they all have incentives to deviate to favouritism in equilibrium S'. The expected payoff of player $i \in S'_F \backslash S_F$ when the coalition practise favouritism equals:

$$p\Big[d_i(1-\alpha)+d_i\alpha+(n-1-d_i)(L-\beta)+|N_i^c\cap S'_M|\alpha+\beta\sum_{j\in N_i\cap S'_F}\frac{n-1-d_j}{|N_j\cap S'_F|}\Big]. \tag{A.4}$$

Her expected payoff when the coalition practise market behaviour (while S_F practises favouritism) equals:

$$p\Big[d_i(1-\alpha)+d_i\alpha+(n-1-d_i)(1-\alpha-c)+|N_i^c\cap S_M|\alpha\Big]. \tag{A.5}$$

There exists a profitable coalition deviation from S' to S if and only if the difference in expected payoff is positive for all players $i \in S_F \backslash S'_F$. By simplifying the expression with $x=(1-\alpha-c)-(L-\beta)$, the difference in payoff equals:

$$p\Big[(n-1)(-x)+(|N_i^c\cap S'_M|-|N_i^c\cap S_M|)\alpha+d_ix+\beta\sum_{j\in N_i\cap S'_F}\frac{n-1-d_j}{|N_j\cap S'_F|}\Big] \tag{A.6}$$

All players in S'_F must satisfy condition (1.2) because S' is an equilibrium: for all $i \in S'_F$ and ergo $i \in S'_F \backslash S_F$, player i must satisfy:

$$d_i x + \beta \sum_{j \in N_i \cap S'_F} \frac{n-1-d_j}{|N_j \cap S'_F|} \geq (\frac{n}{\delta} - n + 1)(n-1)x + |N_i^c \cap S'_F|\alpha \tag{A.7}$$

Substituting this into the expression (A.6), the difference in expected payoff becomes:

$$\begin{aligned}
&\geq p[(n-1)(-x) + (|N_i^c \cap S'_M| - |N_i^c \cap S_M|)\alpha + (\frac{n}{\delta} - n + 1)(n-1)x + |N_i^c \cap S'_F|\alpha] \\
&= p[(\frac{n}{\delta} - n)(n-1)x + (|N_i^c \cap S'_M| + |N_i^c \cap S'_F| - |N_i^c \cap S_M|)\alpha] \\
&= p[(\frac{n}{\delta} - n)(n-1)x + |N_i^c \cap S_F|\alpha] \\
&\geq 0
\end{aligned}$$

For all players in $S'_F \backslash S_F$, there exists a profitable coalition deviation from S to S'. Hence, $S' \succ S$. □

Corollary 1.2. *Suppose there is a set of equilibria $\mathbb{S}$ under strategy profile s^* and network g. An equilibrium S is coalition-proof if and only if for all $S' \in \mathbb{S}$, $S_F \nsubseteq S'_F$.*

Proof. Suppose an equilibrium favouritism group S_F is a proper subset of another equilibrium favouritism group S'_F, then S_F is dominated (Proposition 1.4).

Suppose for all $S' \in \mathbb{S}$, S_F is not a proper subset of S'_F. Suppose there exists an equilibrium S' that dominates S, i.e. there exists a profitable coalition deviation from S to S'. The coalition C either comprises of all market players under equilibrium S or at least one favouritism player.

First, suppose all players in coalition C are market players under equilibrium S. If all of them have the incentives to collectively deviate to another equilibrium S', they would either join the original favouritism group S_F or form a competing favouritism group against them. Either way, since S' is an equilibrium and $S'_F = S_F \cup C$, S_F would then be a proper subset of S'_F, forming a contradiction.

Second, suppose all players in coalition C are favouritism players in equilibrium S. By the definition of being in S_F, principal i in C is happy to sustain favouritism despite incurring a loss in the current period. This means her future expected payoff while sustaining favouritism is higher than the expected payoff in the punishment phase, the Pure market equilibrium (Condition 1.2). On top of that, the coalition C earns lower expected payoffs by reverting to market behaviour than when all players revert to market behaviour. This is because the presence of any favouritism player extracts surplus from market players. Therefore, all players in $C \subseteq S_F$ earn higher expected payoffs sustaining favouritism in S than collectively deviating to market behaviour in S', and have no incentive to deviate to S'.

Third, suppose the coalition comprises of some players in S_F and some in S_M. The group of players in $S_F \cap C$ will earn less after the coalition deviation by two reasons: deviating to market behaviour reduces their payoffs, as explained previously; the new favouritism group formed by $S_M \cap C$ reduces payoffs for all market players (including $S_F \cap C$). So the favouritism players in equilibrium S have no collective incentive to deviate regardless of who is in the coalition.

Thus, there is no coalition C that earns higher expected payoffs by collectively deviating from equilibrium S. By proof of contradiction, if S_F is not a proper subset of another equilibrium's favouritism group, S is a coalition-proof equilibrium. □

Appendix B

Supplementary materials of Chapter 2

B.1 Simulation

B.1.1 Network generation & selection

There are three networks of interest: Erdös-Rényi (ER), Stochastic Block (SB), and Royal Family (RF) network. All networks have 40 nodes, $n = 40$. To control for the average information received by each node, the networks have an average outdegree of 4 (excluding self links), following Becker et al. (2017). Our DeGroot simulations show that the hypotheses are robust against different outdegrees.

The generation process of each network type is as follows. The parameter specifications in the network generation process were selected to ensure strong connectedness in the networks generated.

- Erdös-Rényi networks are generated according to the Erdös-Rényi model (using the "erdos.renyi.game" function from the igraph package). We specify the number of nodes as n and the total number of edges as $2n$.
- Stochastic Block networks are generated according to the Trait-based random generation (using "sample_pref" function from the igraph package). We specify the number of nodes as n and the size of each community as 5. So there are $n/5$ communities where the probability of linking within a community is $p_{ii} = 0.85$ and between commu-

nities is $p_{ij} = p_{ii}/60$. (These parameter specifications were selected to ensure strong connectedness in the networks generated.)

- Royal Family networks are created by first placing n players in a directed ring (player n observes player 1 who then observes player 2 and so on). Then players 1,2,3 are selected to be the hub where all players observe them. All players have an outdegree of 4 (except for players 1 and 2 with an outdegree of 2, and players 3 and n with an outdegree of 3).

For each network treatment, we randomly generated 100 networks that are (strongly) connected — every node can be reached through a path from every other node. Then we computed network measures such as outdegree, diameter, average path length, and clustering for each network. The average statistics for each network type are presented in Table B.1.

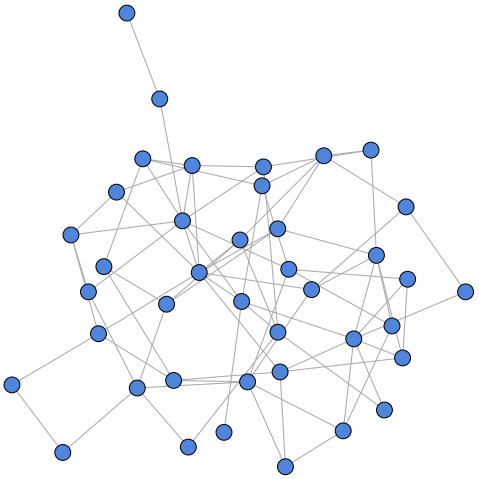
Out of the 100 randomly generated networks, a network with measures closest to the average statistics is then used in the experiment. Table B.2 presents the network statistics of these networks and Figure B.1 presents the network graphs. Note that the Royal Family network is not generated randomly.

Table B.1 Averages network statistics of 100 randomly generated networks

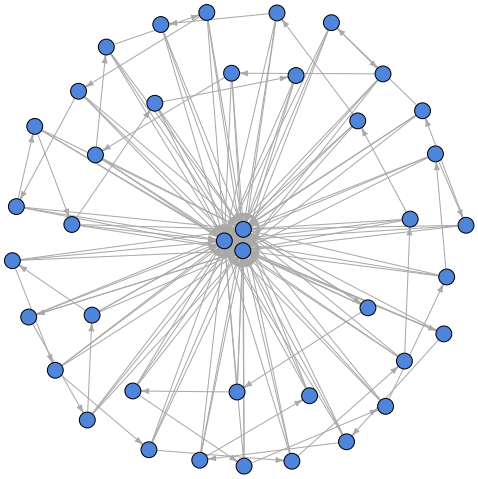
n=40	avg. outdegree	diameter	avg path length	clustering
ER	4.00	5.63	2.73	0.10
SB	3.98	9.15	4.12	0.57
RF	3.85	38.00	12.72	0.26

Table B.2 Network statistics of the networks used in the experiments

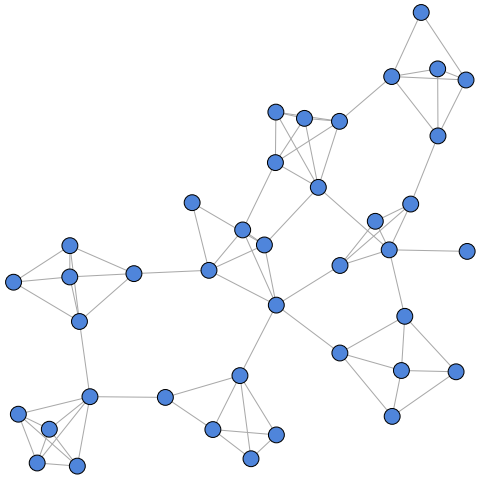
n=40	avg. outdegree	diameter	avg path length	clustering
ER	4.00	5	2.73	0.10
SB	4.00	9	3.85	0.57
RF	3.85	38	12.72	0.26



(a) ER graph



(b) RF graph



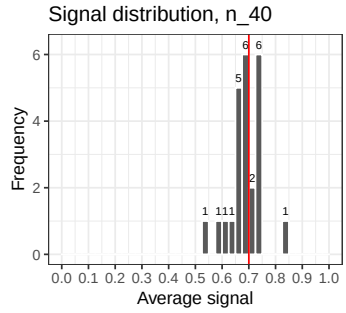
(c) SB graph

Fig. B.1 Network graph of n=40 with average outdegree 4

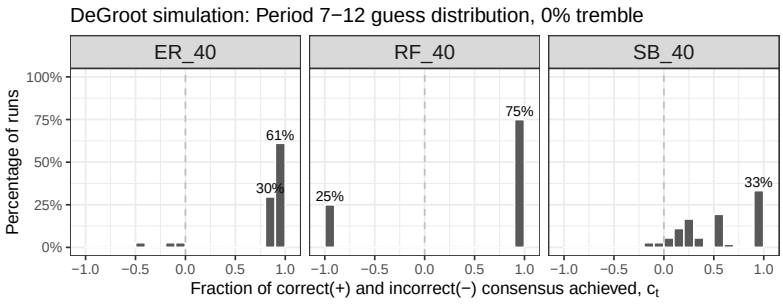
B.1.2 Signal generation & selection

We randomly select 24 sets of signals for the experiment. For each network treatment, there are 4 groups of players each playing 6 rounds. So group 1 in round 1 uses the first set of signals while group 4 in round 6 uses the 24th set. Therefore, the same collection of signals are used across all networks.

We perform two checks to ensure that the 24 sets of signals are representative. First, we note that the distribution of the 24 sets of signals is bell-shaped around the mean 0.7 where 1 represents the correct state (Figure B.2a). Second, we confirm that the simulated guesses following these 24 sets of signals (Figure B.2b) have the same properties as the simulations of the 1000 sets of signals (presented in Figure 2.2c, see main text). The regression on network effects with respect to 'Correct consensus', 'Incorrect consensus' and 'Breakdown of consensus' (as defined in the Consensus Outcomes section in the main text) confirms the main hypotheses (Table B.3).



(a) Distribution of signals



(b) Distribution of c_t simulated from signals

Fig. B.2 Signal distribution and simulation results using the 24 sets of signals for the experiment. (a) Distribution of signals used for all networks in the experiment with mean 0.70, standard deviation 0.06, 1st quartile 0.675, 2nd quartile 0.70 and 3rd quartile 0.75. (n=24) (b) Distribution of c_t under DeGroot simulation using experiment signals. The hypotheses from the simulation of 1000 runs are confirmed: 1) There is more breakdown of consensus in the Stochastic Block network than in the Erdös-Rényi and Royal Family network; 2) There is more incorrect consensus in the Royal Family network than in the Erdös-Rényi and Stochastic Block network.

Table B.3 OLS regression of simulated data, network size 40, k = 0.3

	OLS - Correct Consensus	OLS - Incorrect Consensus	OLS - Breakdown
(Intercept)	0.88^{***}	0.04	0.08
	(0.09)	(0.05)	(0.07)
typeRF	-0.08	0.17^{**}	-0.08
	(0.12)	(0.08)	(0.10)
typeSB	-0.37^{***}	-0.04	0.42^{***}
	(0.12)	(0.08)	(0.10)
R^2	0.13	0.11	0.31
Adj. R^2	0.10	0.08	0.29
Num. obs.	72	72	72

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

B.1.3 Variations on DeGroot updating rule

All network effects identified in the simulations are robust to alternative variations on DeGroot updating rule.

Deterministic DeGroot. In the case of indifference, suppose an individual persists with her last period's guess. Formally, we say:

$$a_{i,t} = \begin{cases} 1 & \text{if } \mu_{i,t} > \frac{1}{2}, \\ 0 & \text{if } \mu_{i,t} < \frac{1}{2}, \\ a_{i,t-1} & \text{if } \mu_{i,t} = \frac{1}{2} \end{cases} \tag{B.1}$$

Simulations of this variant of the DeGroot are presented in Figure B.3.

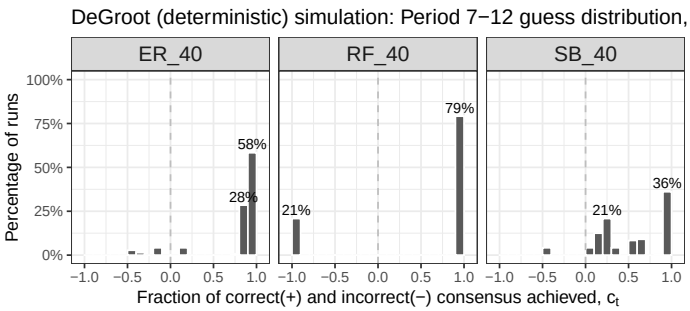
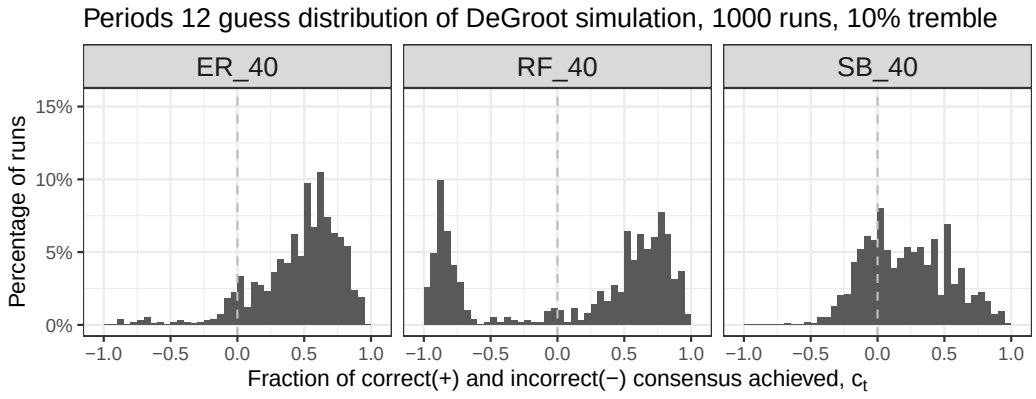
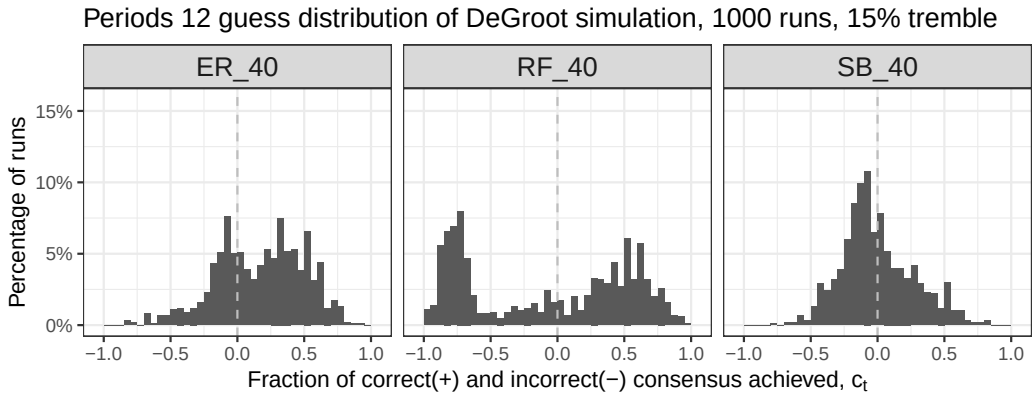


Fig. B.3 Distribution of c_t under Deterministic DeGroot simulation using experiment signals.



(a) 10% trembling



(b) 15% trembling

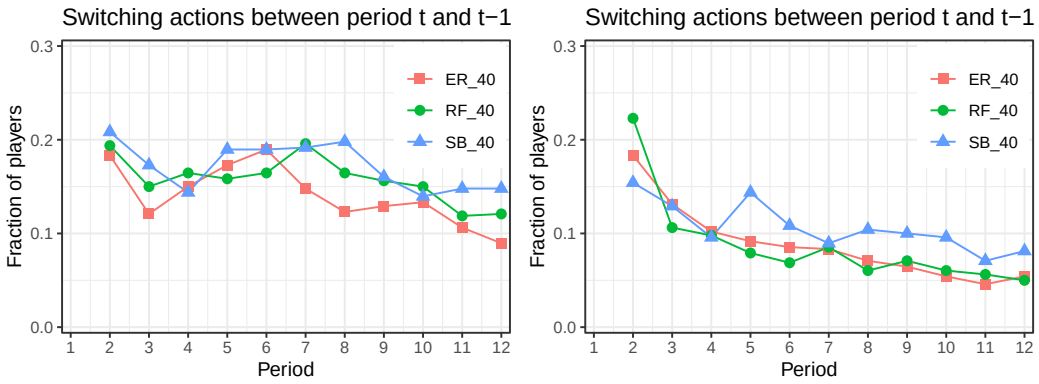
Fig. B.4 Distribution of c_t under simulation with trembling.

DeGroot with Trembling. Suppose an individual observes a majority guess of Red: if we use DeGroot updating rule with 10% trembling, that means she would guess Green 10% of the time and Red 90% of the time. Figure B.4 shows that the networks effects identified with the original DeGroot (as in Figure 2.2c in the main text) are robust.

B.2 Findings

B.2.1 Convergence

The rapid convergence of guesses in the experiment is supported by evidence on switching frequency: 20% of individuals switched their guesses at period 2 after observing the first period guesses of their neighbors, this switching frequency falls to 10% toward the end of the experiment in period 12. The switching probability falls significantly as subjects learn across rounds: as a result, it is only 5% in the last three rounds (Figure B.5). We argue that the residual switching in guesses in the final periods are not due to further learning by subjects, but due to random guessing.



(a) n=40, round 1-3 (b) n=40, round 4-6

Fig. B.5 Percentage of subjects switching guesses per period. (a) In the first three rounds, the percentage of switching falls from 20% in period 2 to 10∼15% in period 12. (b) In the last three rounds, the percentage of switching falls from around 20% in period 2 to 5∼8% in period 12. Therefore, adjusting for learning across rounds, there are less than 8% of subjects switching guesses by period 12.

We estimate that 10% of the guesses are random in the experiment, using the following technique: Irrespective of whether a myopic player follows Bayesian or DeGroot learning rule, in period 1, it is optimal to guess her initial signal. In period 2, both (myopic) Bayesian and DeGroot learning rules predict that player should follow the majority guess in her neighbourhood in period 1. Table B.4 shows that about 10% of guesses do not follow

subjects' initial signals in period 1 and contradict both learning rules in period 2. This suggests that about 10% of guesses ignore information.

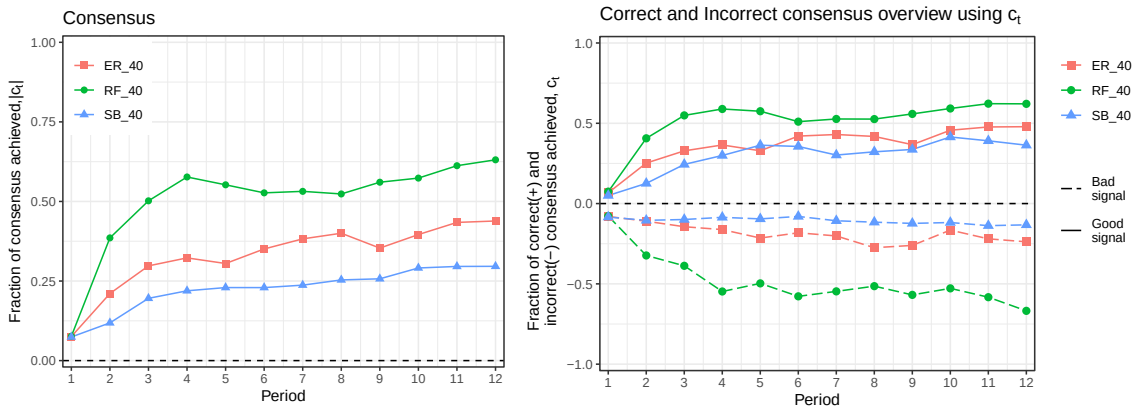
Table B.4 Fraction of guesses against Bayesian and DeGroot prediction, network size 40

	Guess against majority in period 1,2	
	OLS (Bayesian, DeGroot predicts 0)	Logit
(Intercept)	0.10^{***}	-2.24^{***}
	(0.01)	(0.08)
typesizeRF_40	0.02^{*}	0.24^{*}
	(0.01)	(0.13)
typesizeSB_40	0.02^{***}	0.25^{***}
	(0.01)	(0.08)
R^2	0.00	
Adj. R^2	0.00	
Num. obs.	5760	5760
AIC		4029.57
BIC		4049.54
Log Likelihood		−2011.78
Deviance		4023.57

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

B.2.2 Consensus

The simulations lead us to propose two hypotheses: One, the breakdown of consensus is most likely in the Stochastic Block network, followed by the Erdös-Rényi network and lastly the Royal Family network; Two, the Royal Family network leads to the wrong consensus more often than the Erdös-Rényi network. Figure B.6a presents the evolution of consensus across periods across all networks, while Figure B.6b presents the evolution of c_t partitioned by 'good' and 'bad' signals. Under DeGroot updating simulation, the set of 'good' signals would lead to $c_t \geq 0$ (correct consensus), while the 'bad' signals would lead to $c_t < 0$ (incorrect consensus). They show that the rankings in the hypotheses are maintained across all periods. The regression Table B.7 shows the statistical significance of the estimates (presented in Figure 2.4 in the main text), supporting our hypotheses. The estimate of 'incorrect consensus' on 'typeRF' represents the difference in fraction of incorrect consensus achieved between the Royal Family network and the Erdös-Rényi network.



(a) Evolution of consensus $|c_t|$

(b) Evolution of consensus c_t

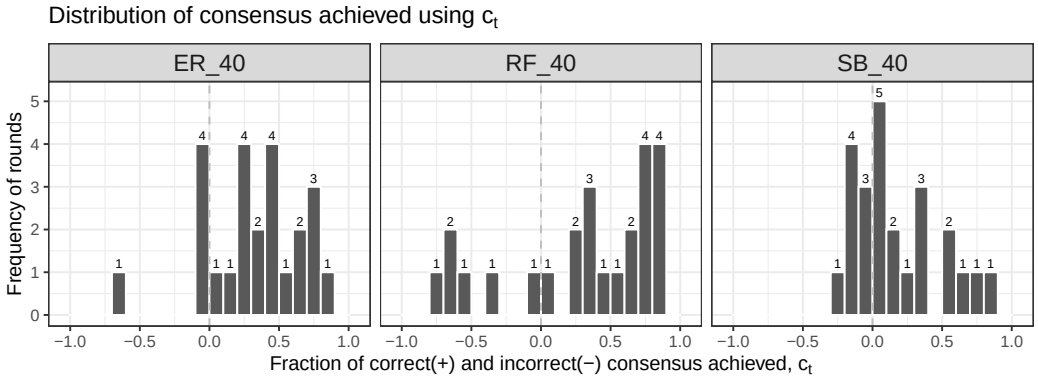
Fig. B.6 Evolution of $|c_t|$ and partitioned c_t. (a) In period 12, RF, ER, SB reach 63%, 44%, 30% of consensus, respectively. (b) We partitioned c_t averaged across all games by 'good' and 'bad' signals. The ranking of correct and incorrect consensus reached is preserved across most periods.

A similar distribution of c_t obtains if we consider fewer periods (periods 10-12) or rounds (rounds 4-6) (Figure B.7).

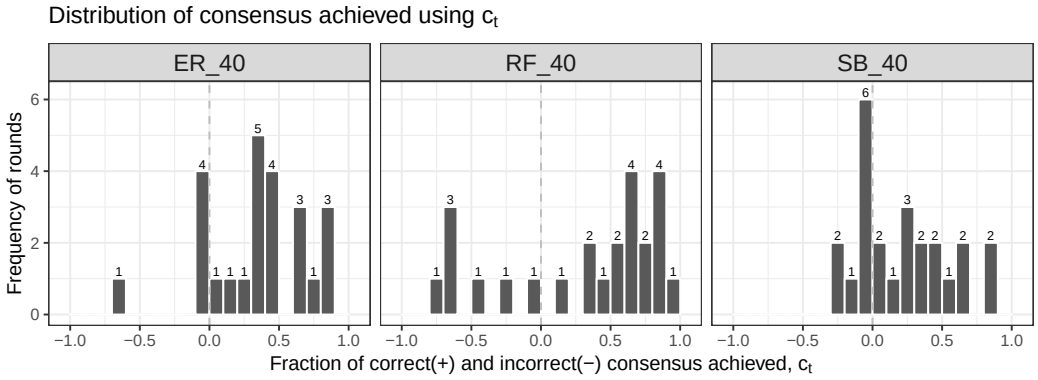
Recall, we defined binary variables of correct consensus (if $c_t > k$), incorrect consensus (if $c_t < -k$), and breakdown of consensus (if $-k \leq c_t \leq k$) based on the value of c_t. Our main findings are robust to 1) different widths k (Tables B.6 to B.8), 2) an alternative model specification such as the logit model (Table B.9), and 3) a continuous definition of consensus outcomes (Table B.10).

A continuous variation on the definition of consensus would be as follows: Consensus is defined as the absolute value of c_t, $|c_t|$; correct consensus is defined as censoring negative values of c_t to 0; incorrect consensus censors positive values of c_t to 0; breakdown is defined as the negative of consensus, $-|c_t|$.

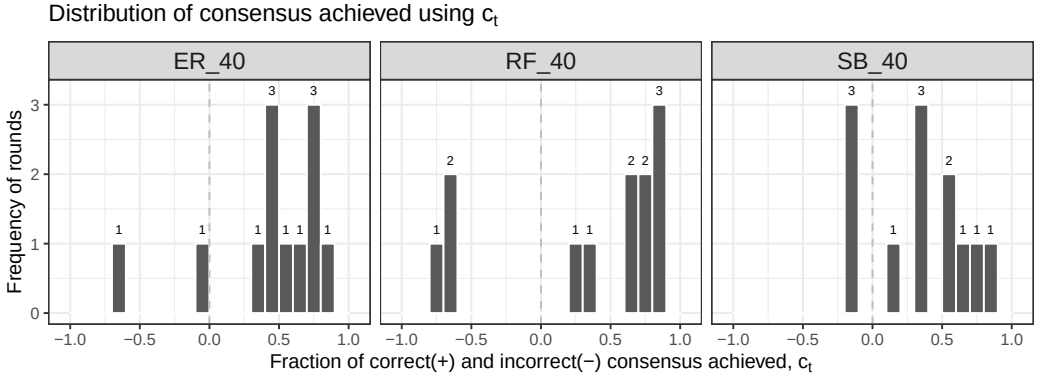
For the Stochastic Block network, we show that communities are more likely to reach consensus compared to Erdös-Rényi network despite the networks being less likely to reach consensus. The subset of subjects in each block of the Stochastic Block model may be seen as constituting a 'community'. Given the network generation methods in the Stochastic Block model, subjects with location id $1-5$ is a community while id $6-10$ is another community, and so on. So in all three networks, we define a community by the same location ids. We



(a) Distribution of averaged c_t, between period 7-12, round 1-6 (n=24 per network)



(b) Distribution of averaged c_t, between period 10-12, round 1-6 (n=24 per network)



(c) Distribution of averaged c_t, between period 7-12, round 4-6 (n=12 per network)

Fig. B.7 Distribution of averaged c_t robust over period and round selections.

define *community consensus* as 1 when all 5 subjects in a community reaches complete consensus, and 0 otherwise.

Table B.5 shows that 52% of communities reach consensus in the Stochastic Block network which is 14% point higher than in Erdös-Rényi network. We next look closer at the dispersion of average guesses of communities. The maximum difference in average guesses of communities is equal to 1: when there exists one community with correct consensus and one with incorrect consensus. Figure B.8 shows that 75% of rounds in the Stochastic Block network have large dispersion in community guesses (greater than 0.7) while only 50% in Erdös-Rényi network and 46% in Royal Family network. This implies that disagreements between communities are the principal source of the consensus breakdown in the Stochastic Block network.

B.2.3 Updating rule

On average, 88% of guesses match with the DeGroot rule. This is higher than the baseline of how well *guessing randomly* matches with DeGroot predictions: simulations show that on average 60% pseudo subjects' random guesses match with DeGroot. This is also higher than the baseline of how well *guessing signal* matches with DeGroot predictions: simulations show that on average 75% guesses of pseudo subjects (if guessing only signal) match with DeGroot (Figure B.11a).

Suppose that 10% of guesses are randomly made. We show that the level of consensus attained in the experiment is comparable the simulation under 10% trembling for Erdös-Rényi (Figure B.10a) and 15% trembling for Stochastic Block and Royal Family network (Figure B.10b).

We next delve deeper by looking at subject level match with DeGroot. Because each subject plays a total of 6 rounds and 12 periods per round and their guesses are not statistically independent, we treat each subject as a data point. Figure B.11 presents the histogram of how well a subject's guesses match with DeGroot predictions. For all networks, there are significantly more subjects whose guesses match with DeGroot than pseudo subjects who guess their signals or randomly.

Table B.5 Regression of community consensus on network treatment

	OLS - Community Consensus	Logit - Community Consensus
(Intercept)	0.38***	−0.49***
	(0.03)	(0.15)
typeRF	0.16*	0.66*
	(0.09)	(0.37)
typeSB	0.14***	0.55***
	(0.04)	(0.16)
R^2	0.02	
Adj. R^2	0.02	
Num. obs.	576	576
AIC		791.85
BIC		804.92
Log Likelihood		−392.93
Deviance		785.85

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

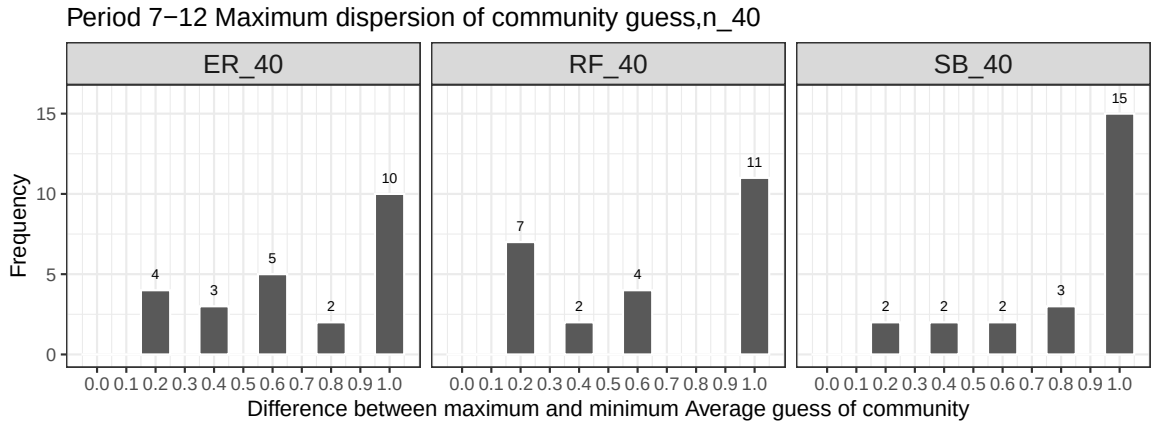


Fig. B.8 Distribution of the maximum dispersion in average guesses between communities for each network. 75% of rounds in the SB have more than 0.7 dispersion in average guesses between communities, 50% in ER and 46% in RF (n=24 per network).

Table B.6 OLS regression c_t, k=0.2, n=40

	OLS - Correct Consensus	OLS - Incorrect Consensus	OLS - Breakdown
(Intercept)	0.71^{***}	0.04	0.25^{***}
	(0.07)	(0.04)	(0.04)
typeRF	-0.00	0.17^{**}	-0.17^{*}
	(0.15)	(0.08)	(0.09)
typeSB	-0.33^{***}	-0.00	0.33^{***}
	(0.08)	(0.05)	(0.06)
R^2	0.10	0.07	0.20
Adj. R^2	0.08	0.04	0.18
Num. obs.	72	72	72

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table B.7 OLS regression c_t, k=0.3, n=40

	OLS - Correct Consensus	OLS - Incorrect Consensus	OLS - Breakdown
(Intercept)	0.54^{***}	0.04	0.42^{***}
	(0.07)	(0.04)	(0.04)
typeRF	0.08	0.17^{**}	-0.25^{**}
	(0.16)	(0.08)	(0.11)
typeSB	-0.21^{***}	-0.04	0.25^{***}
	(0.07)	(0.04)	(0.04)
R^2	0.06	0.11	0.17
Adj. R^2	0.03	0.08	0.15
Num. obs.	72	72	72

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table B.8 OLS regression c_t, k=0.4, n=40

	OLS - Correct Consensus	OLS - Incorrect Consensus	OLS - Breakdown
(Intercept)	0.46^{***}	0.04	0.50^{***}
	(0.09)	(0.04)	(0.06)
typeRF	0.04	0.12^{*}	-0.17
	(0.18)	(0.07)	(0.12)
typeSB	-0.25^{**}	-0.04	0.29^{***}
	(0.12)	(0.04)	(0.09)
R^2	0.07	0.08	0.14
Adj. R^2	0.04	0.05	0.12
Num. obs.	72	72	72

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table B.9 Logistic regression c_t, k=0.3, n=40

	Logit - Correct Consensus	Logit - Incorrect Consensus	Logit - Breakdown
(Intercept)	0.17	-3.14^{***}	-0.34^{*}
	(0.28)	(0.92)	(0.17)
typeRF	0.34	1.80^{*}	-1.27^{*}
	(0.66)	(1.01)	(0.77)
typeSB	-0.86^{***}	-16.43^{***}	1.03^{***}
	(0.28)	(1.05)	(0.17)
AIC	101.41	38.88	90.78
BIC	108.24	45.71	97.61
Log Likelihood	−47.71	−16.44	−42.39
Deviance	95.41	32.88	84.78
Num. obs.	72	72	72

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table B.10 OLS regression c_t censored, n=40

	OLS - Correct Consensus	OLS - Incorrect Consensus	OLS - Breakdown
(Intercept)	0.36^{***}	0.02	0.62^{***}
	(0.03)	(0.02)	(0.01)
typeRF	0.07	0.08^{**}	-0.15^{*}
	(0.10)	(0.04)	(0.08)
typeSB	-0.14^{***}	−0.02	0.16^{***}
	(0.04)	(0.02)	(0.02)
R^2	0.08	0.09	0.19
Adj. R^2	0.06	0.06	0.17
Num. obs.	72	72	72

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

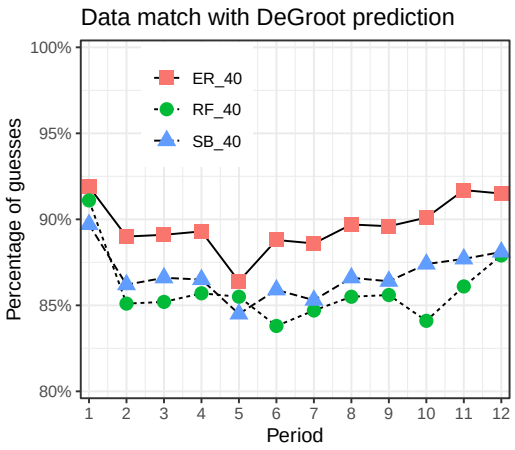
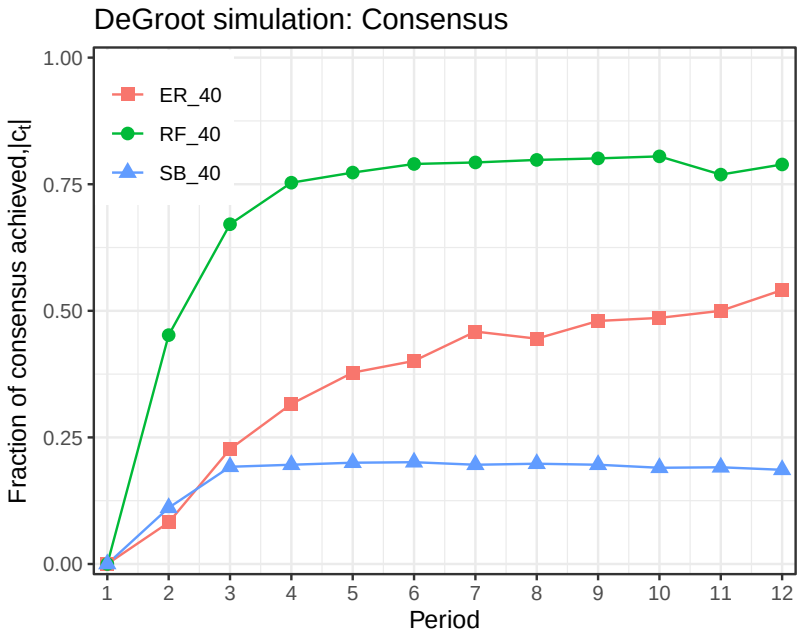


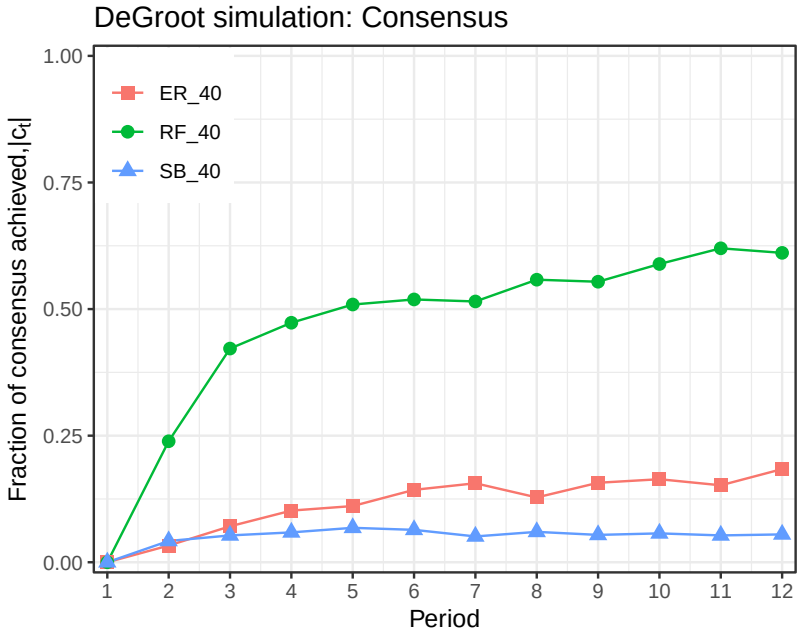
Fig. B.9 The percentage of guesses matching DeGroot prediction across periods. Across all networks at least 90% of first guesses matched with the DeGroot prediction (i.e., guess follows the signal). This percentage falls to 80%∼85% in the second period and then steadily increases until it reaches 85%∼90% in later periods.

Bayesian learning: Information Leader. When DeGroot prediction contradicts with information leader's guess, a Bayesian player should follow their information leader while a DeGroot player should follow the majority of their neighbours. Table B.11 show that when the two are in conflict, only around 10% of subjects follow Bayesian prediction (ER:10%, RF:4%, SB:14%), while the rest follow DeGroot.

No learning: Stubborn players that only follow their signal. Similarly, when DeGroot prediction contradicts goes against initial signal received, a stubborn player should only follow their own signal. Table B.12 show that around 25% of subjects follow initial signal (ER:25%, RF:29%, SB:29%) while the rest follow DeGroot. As before, we show that there is significant learning across rounds.

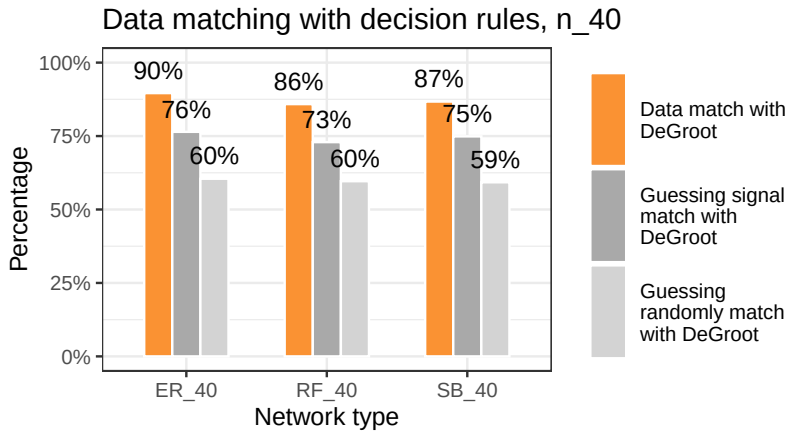


(a) DeGroot simulation with 10% trembling

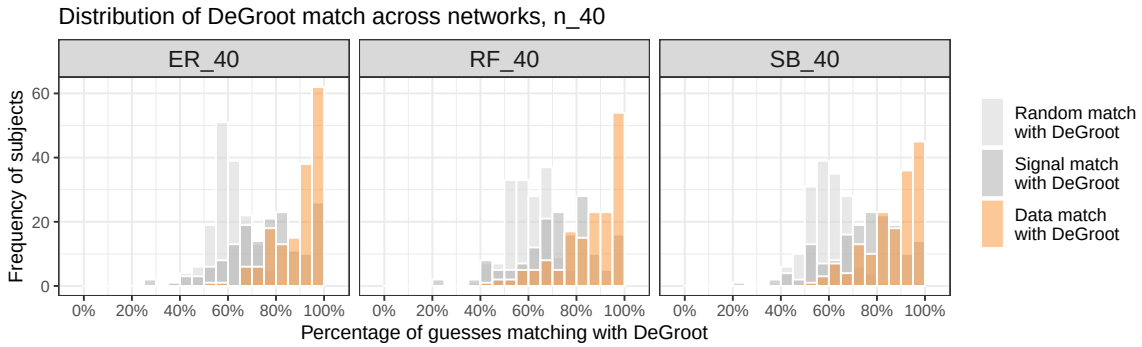


(b) DeGroot simulation with 15% trembling

Fig. B.10 Consensus achieved under DeGroot simulation with trembling.



(a) Fraction of guesses that match with DeGroot, compared to baselines.



(b) Fraction of subjects that match with DeGroot, compared to baselines

Fig. B.11 Percentage of guesses/subjects match with the DeGroot rule. Guesses matching DeGroot prediction are in orange; (Simulation) Guessing signal matching DeGroot prediction are in dark grey; (Simulation) Guessing randomly matching DeGroot prediction are in light grey. (a) Roughly 88% of guesses match with DeGroot predictions, significantly higher than the other two baselines of 75% and 60% respectively. (n=46,080: 11,520 per network) (b) 80% of subjects in ER match with DeGroot predictions at least 80% of the time; these fractions are 72% in the RF and 76% in the SB. This is again compared to the baseline of how well guessing signal matches with DeGroot predictions: Only 44% of pseudo subjects' guesses (if guessing only signal) in ER match with DeGroot predictions at least 80% of the time (37% in RF, and 41% in SB); A negligible fraction of pseudo subjects' guesses (if guessing randomly) match with DeGroot predictions at least 80% of the time. (n=960: 240 per network)

Table B.11 Fraction of guesses imitate leader against DeGroot prediction

	Correctly follow leader			
	OLS (Bayesian predicts 1)	Logit	OLS	OLS
(Intercept)	0.10^{***}	-2.20^{***}	0.18^{***}	0.18^{***}
	(0.02)	(0.18)	(0.03)	(0.02)
RF_40	-0.06^{***}	-0.91^{**}		
	(0.02)	(0.39)		
SB_40	0.04^{**}	0.41^{*}		
	(0.02)	(0.21)		
period			-0.01^{***}	
			(0.00)	
round				-0.02^{***}
				(0.01)
R^2	0.01		0.01	0.01
Adj. R^2	0.01		0.00	0.01
Num. obs.	1870	1870	1870	1870
AIC		1388.23		
BIC		1404.83		
Log Likelihood		−691.12		

$^{***}p < 0.01$; $^{**}p < 0.05$;$^{*}p < 0.1$

Table B.12 Fraction of guesses following signal against DeGroot prediction

	Always follow signal			
	OLS (Stubbornness predicts 1)	Logit	OLS	OLS
(Intercept)	0.25^{***}	-1.07^{***}	0.35^{***}	0.40^{***}
	(0.01)	(0.07)	(0.02)	(0.02)
RF_40	0.04	0.21		
	(0.04)	(0.21)		
SB_40	0.04^{*}	0.22^{*}		
	(0.02)	(0.12)		
period			-0.01^{***}	
			(0.00)	
round				-0.03^{***}
				(0.01)
R^2	0.00		0.00	0.01
Adj. R^2	0.00		0.00	0.01
Num. obs.	9366	9366	9366	9366
AIC		11185.38		
BIC		11206.81		
Log Likelihood		−5589.69		

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

B.3 Related experiments

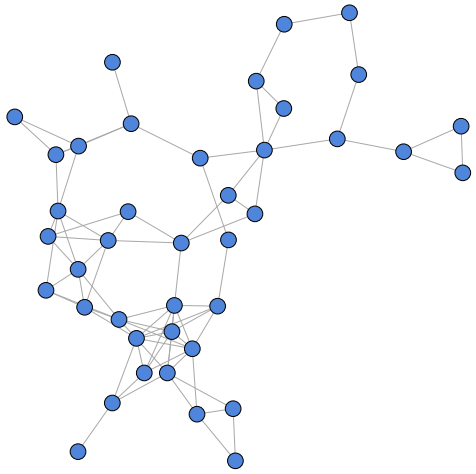


Fig. B.12 RGG graph of n=40 with average outdegree 4

In a recent paper, Chandrasekhar et al. (2020) looked at the mixture model of Random Geometric Graphs and Erdös-Rényi Graphs. We denote it as the RGG network from this point forward. This model captures the idea of sparse and clustered networks from the real world where the share of 'clans' — a set of nodes that are more connected among themselves than to those outside — is non-vanishing as n grows. This feature of inward-looking clans is also present in the 5-player communities within the Stochastic Block network. Under DeGroot updating rule, 'clans' being inward-looking facilitates the breakdown of consensus.

The network generation process is as follows: These exists a Poisson point process on the latent space $\Omega = [0,1]^2 \subset \mathbb{R}^2$, which determines the latent location of n nodes, with uniform intensity $\lambda > 0$. For any subset $A \subset \Omega$, $n_A \sim Poisson(\nu_a)$, where $\nu_a := \lambda \int_A dy$. If the Euclidean distance between two nodes i and j are at most $r = 0.2$, then i and j are linked with probability $\alpha = 0.95$. Otherwise, they are linked with probability $\beta = \alpha/(3n) < \alpha$. These parameter specifications were selected to ensure strong connectedness in the networks generated. Figure B.12 presents an example of the RGG network which is also used in the experiment.

Figure B.13a presents the simulation results of DeGroot updating rule on the RGG network and compares it with the Erdös-Rényi, Royal Family and Stochastic Block network.

Table B.13 Quartile and Mean of c_t under DeGroot simulation.

type	1st quartile	2nd quartile	3rd quartile	mean
RF	1	1	1	0.792
ER	0.95	1	1	0.953
RGG	0.825	0.925	1	0.882
SB	0.75	0.875	1	0.864

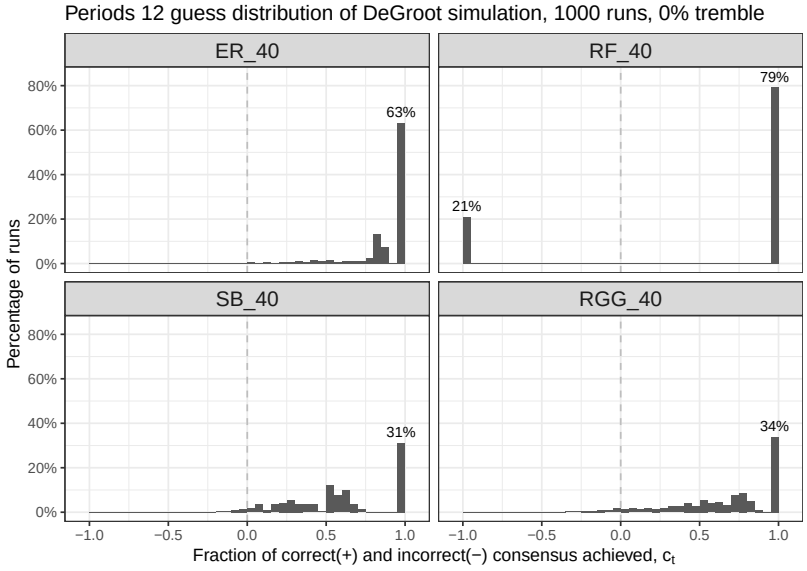
These simulations suggest that the RGG "lies between" the Erdös-Rényi and Stochastic Block network. The quartiles and the mean of the distribution of simulated c_t confirm this (Table B.13). We observe the same results in the experiment (Figure B.13b).

The Erdös-Rényi and Stochastic Block networks are canonical networks. Given the simulations and the experimental findings noted above, for expositional reasons, we felt it was best to present the Erdös-Rényi and Stochastic Block networks in the main text and move the RGG network to this Appendix.

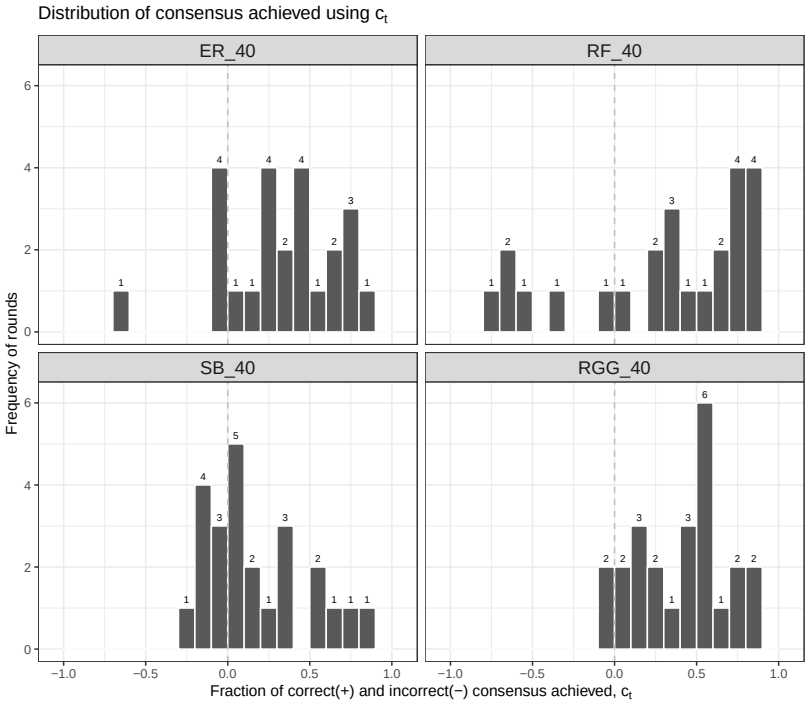
B.4 Experimental Design

The experiment took place at the Laboratory for Research in Experimental and Behavioral Economics (LINEEX) at the University of Valencia. Subjects were recruited through the online recruitment system of LINEEX. All subjects who participated in this study provided informed consent at the LINEEX laboratory, and the procedure of this study was approved by the Institutional Review Board of the University of Valencia. In the experiment, subjects interacted through computer terminals in the LINEEX laboratory, and the experimental software was programmed in HTML, PHP, Javascript, and SQL.

Upon starting an experimental session, subjects read the paper-based instructions, which were also read aloud by an experimenter to guarantee that everyone received the same information (Supplementary Materials). The subjects were then provided with a step-by-step interactive tutorial on their computer screen, which allowed them to get familiarized with the software interface and the game (Figure B.14). To clarify possible consequences of guesses in different periods of a round, subjects were shown a sample network (with only 10 players but with similar features as the network used in the actual game, depending on



(a) Distribution of c_t under Simulation (with RGG)



(b) Distribution of averaged c_t from experimental data (with RGG)

Fig. B.13 Distribution of averaged c_t. (a) The simulation of 1000 sets of signals shows that the distribution of consensus achieved by RGG lies between ER and SB. (b) Our experiment confirms the results from the simulation. (n=24 per network)

the experimental condition) highlighting what guesses would be observed by subjects as a decision-maker from their neighbours, and their neighbours' neighbours.

Details about the decision screens were also provided to subjects: during any period of the game, each subject was shown the colour of the ball initially drawn, and guesses made by neighbours in the network during the previous period (Figure B.16). Subjects also could view guesses made by those individuals (and themselves) in earlier periods of the game through a slider button. At the end of a round, a feedback screen revealed information about the payoff effective period that has been randomly selected, the guesses made by the subject and all others in this period, the bag actually selected, and consequently the payoffs received by the subject in this round (0 or 3 euros depending on whether the guess matches the bag) (Figure B.17). Prior to starting the first round of the game, all subjects also filled up a short questionnaire (4 questions) about their comprehension of the decision screens (Figure B.15). Correct answers were shown after each guess made by the subjects.

To prevent long inactivity during the game, subjects were asked to make all guesses within 30 seconds (in any period of any round). If no guess was made before this time limit, a guess was made automatically, replicating the most recent guess or choosing at random in the first period. Throughout the experiment, all guesses, with no exception, were made by subjects within this time limit.

B.5 Dataset

Figures B.18 to B.20 present the evolution of the average guesses of each network treatment (ER, SB, RF), group (1-4), and round (1-6) from the experiment. Original datasets are available upon request.

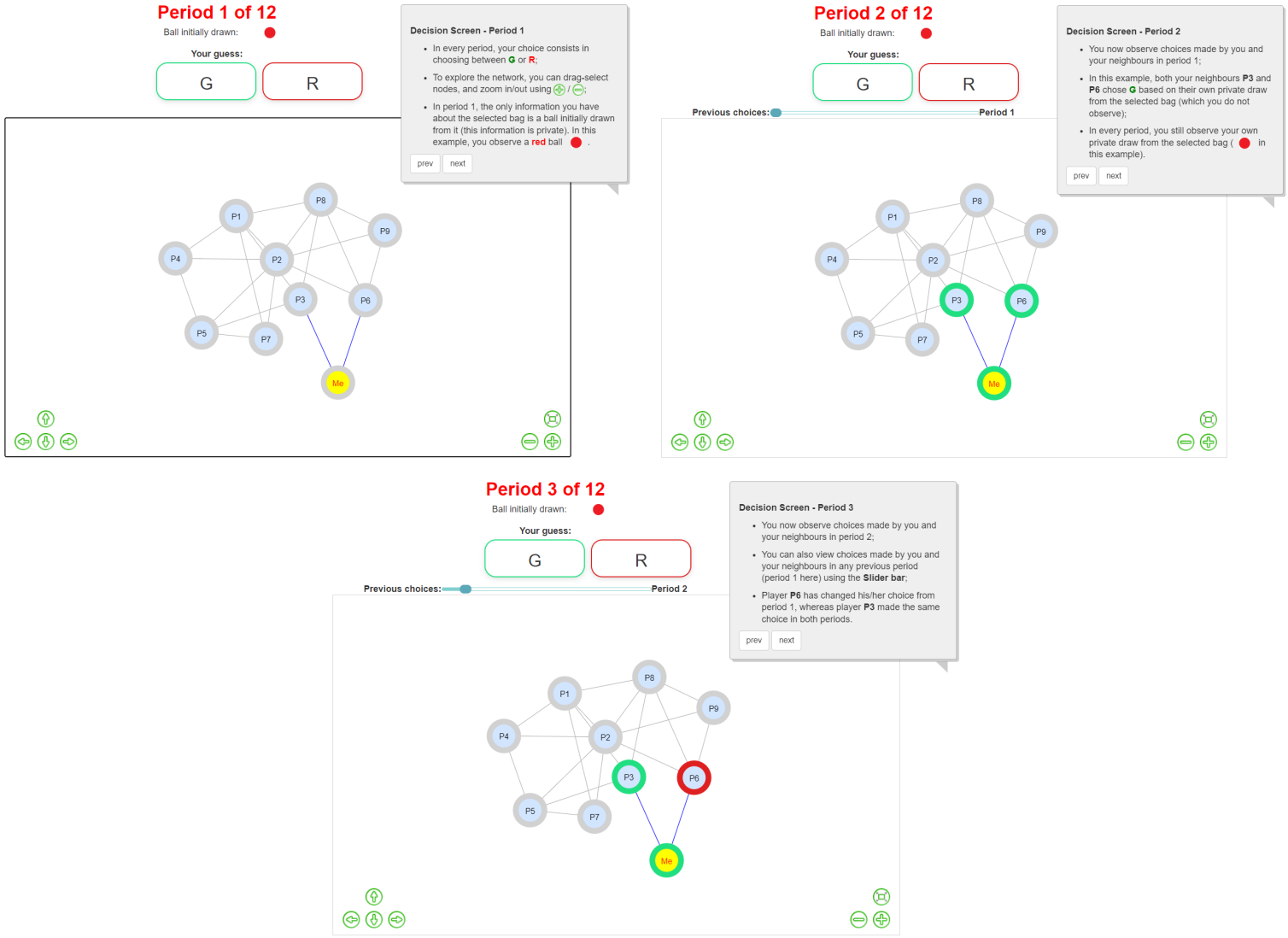


Fig. B.14 Tutorials from the experiment

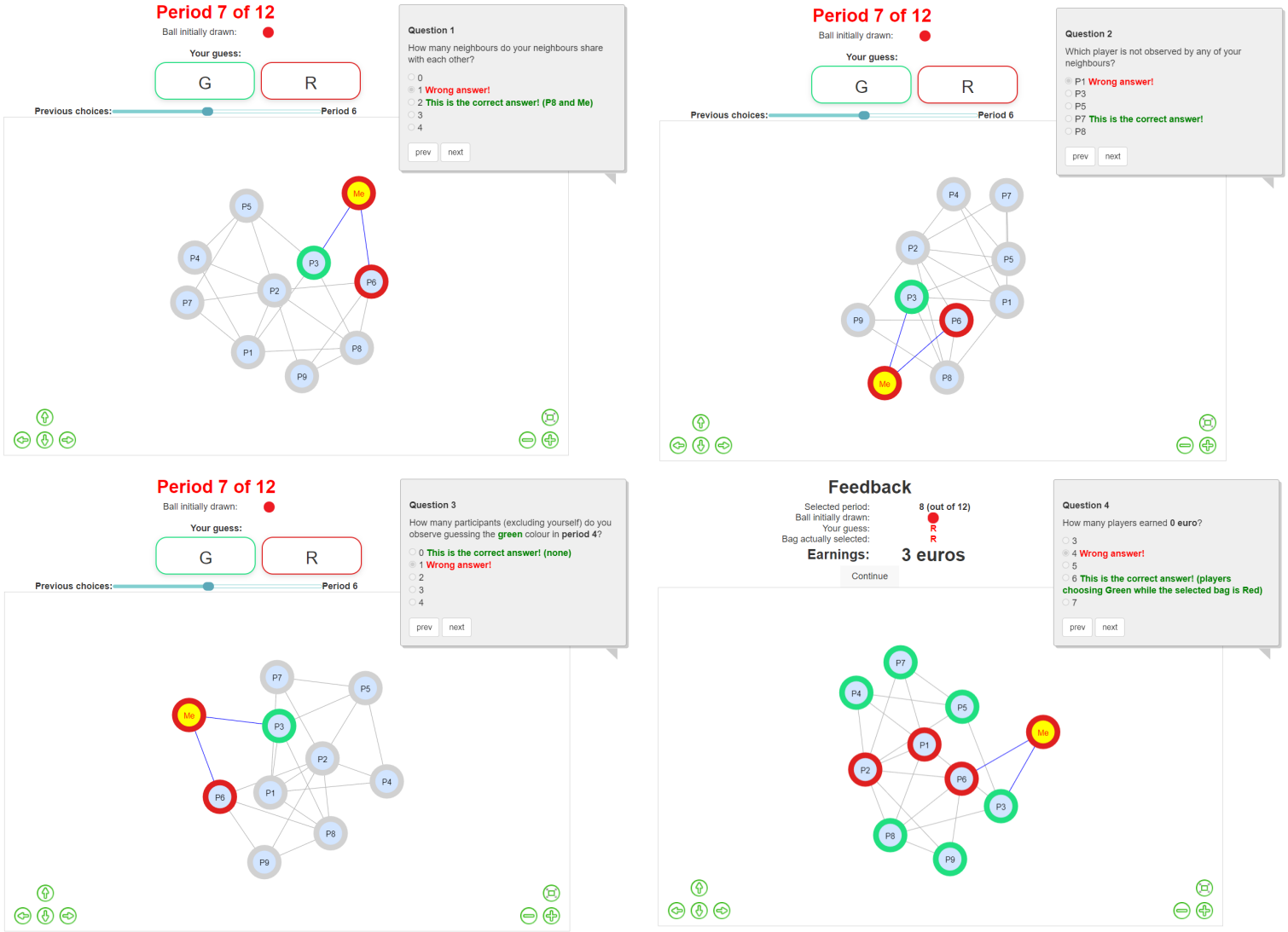


Fig. B.15 Questionnaires from the experiment

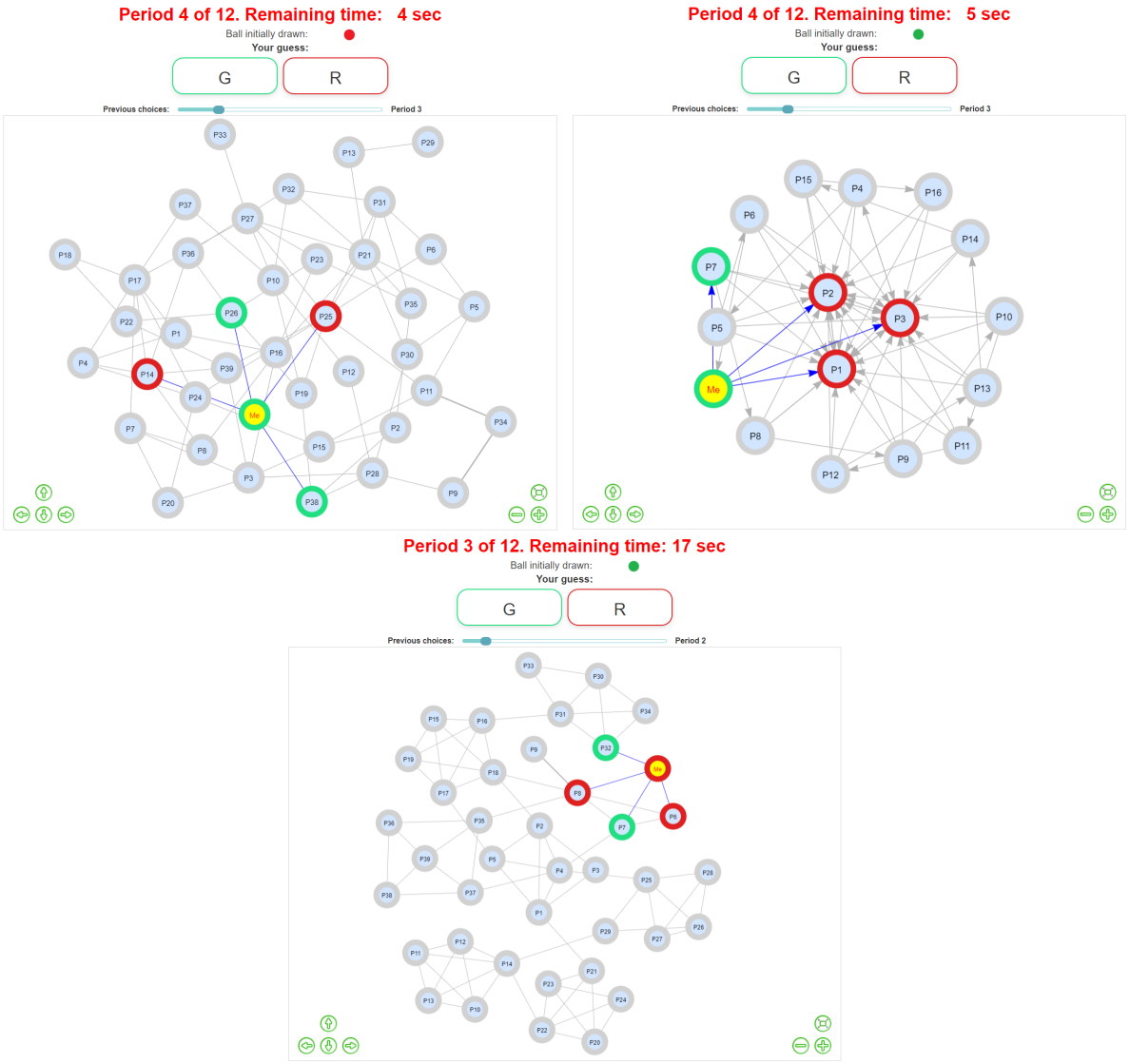


Fig. B.16 Screenshots from the experiment during the game

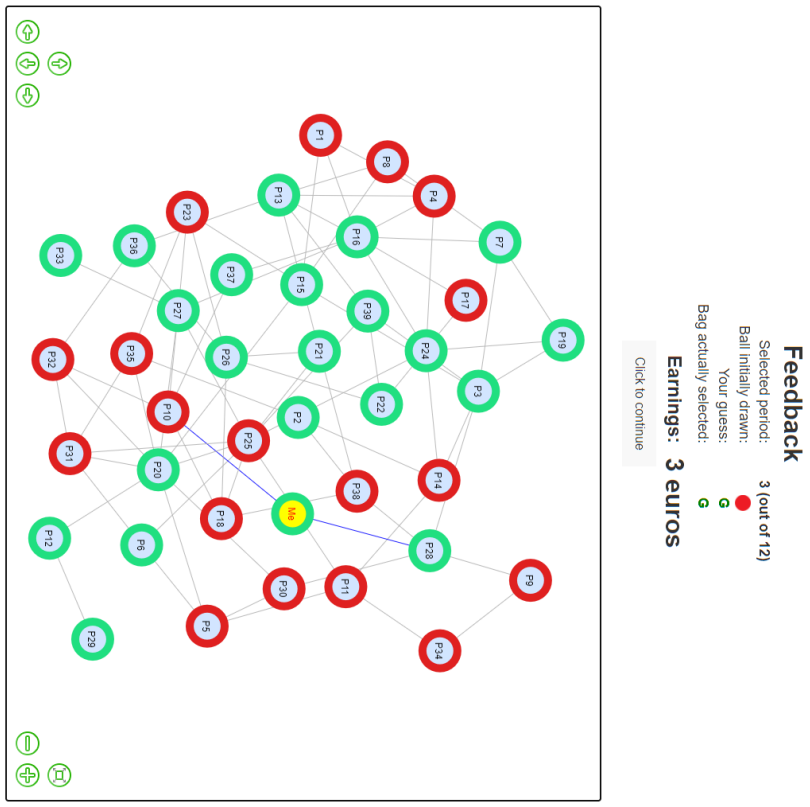


Fig. B.17 Feedback screen from the experiment during the game

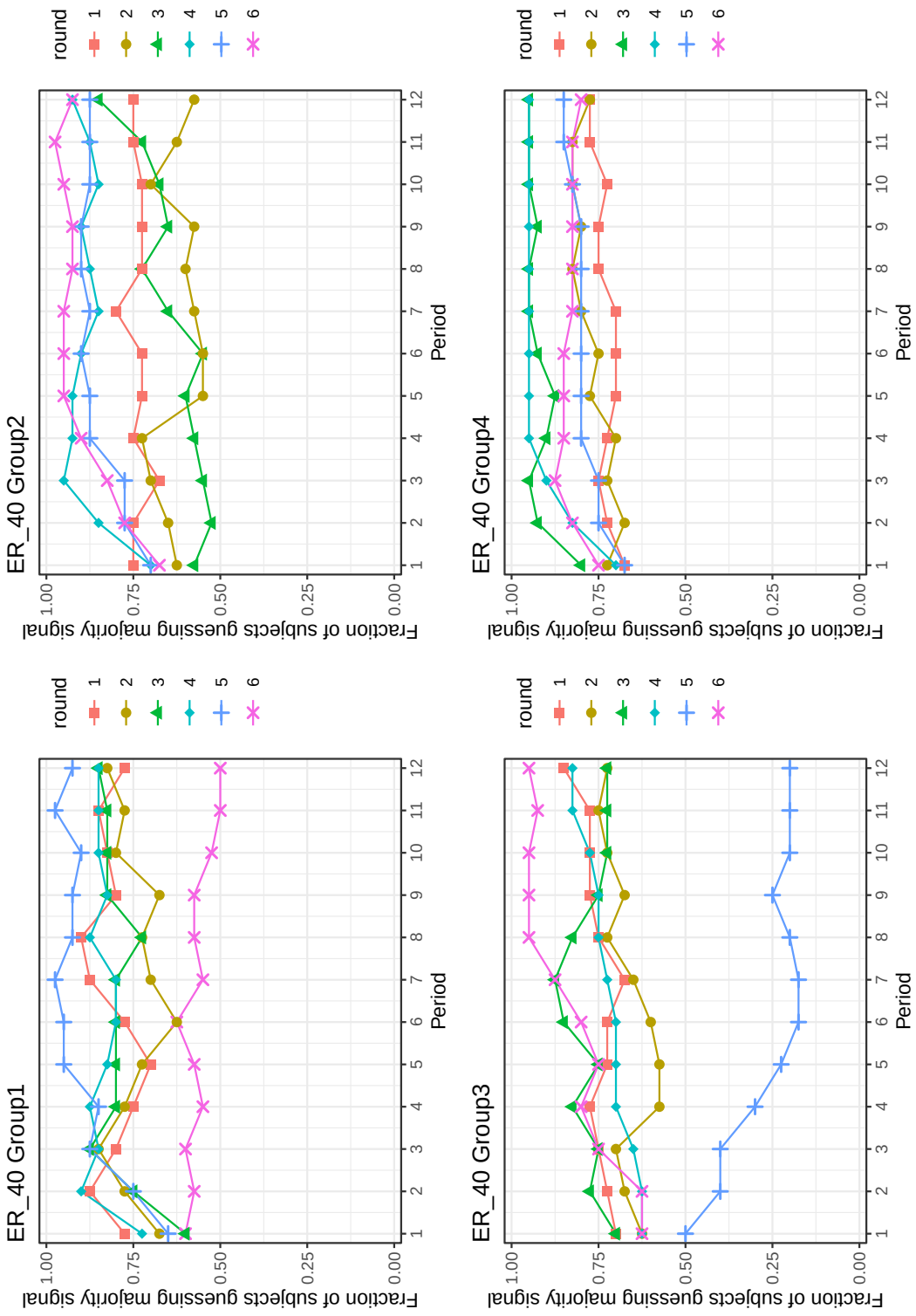


Fig. B.18 Experimental results — Development of guesses n=40

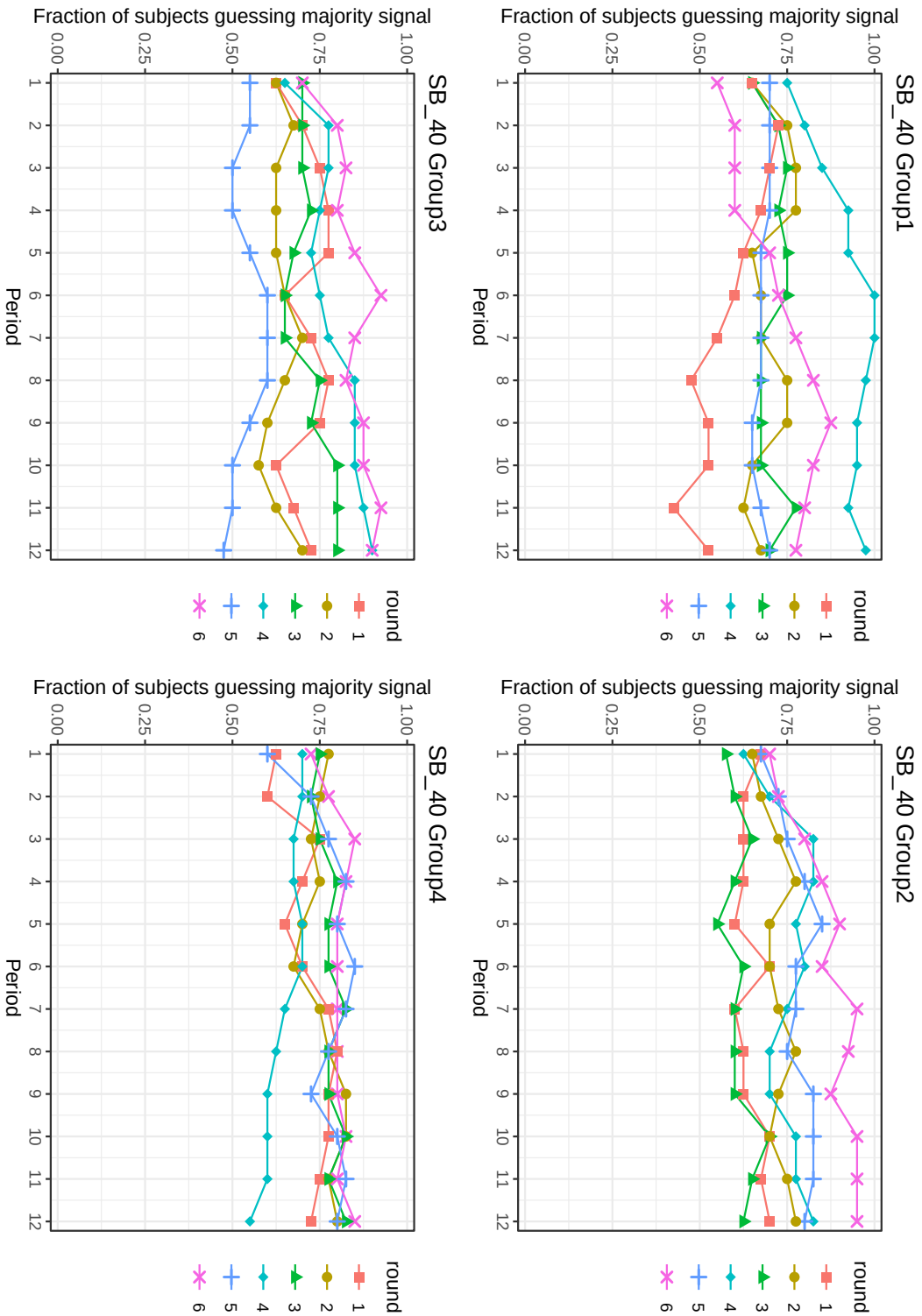


Fig. B.19 Experimental results — Development of guesses n=40

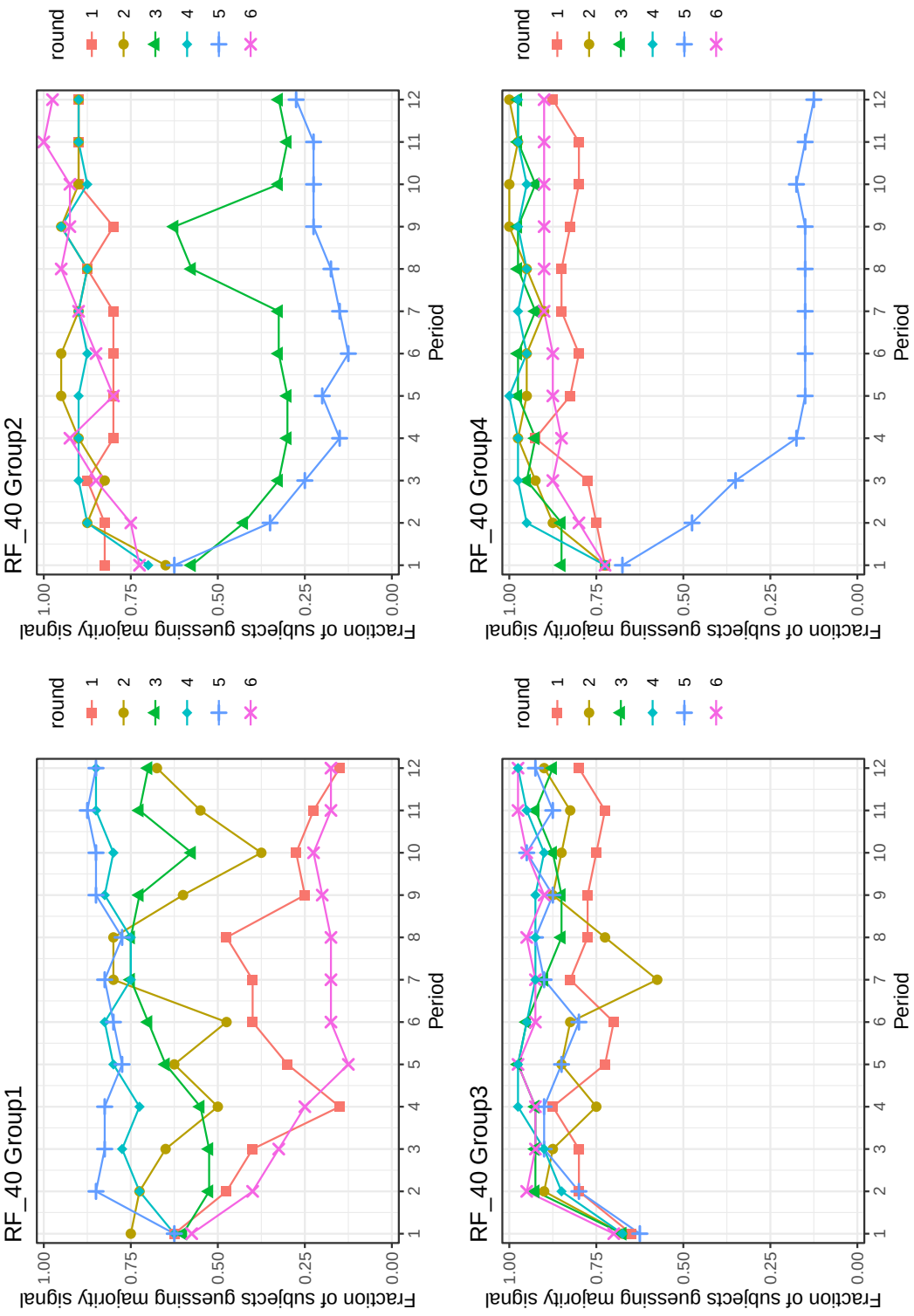


Fig. B.20 Experimental results — Development of guesses n=40

Appendix C

Omitted proofs of Chapter 3

C.1 Alternate sharing benefit

Consider that the sharing benefit is independent of news veracity, i.e., individuals earn a sharing benefit of 1 for each neighbour. Additionally, if external examination reveals news to be false, then the agent suffers a reputational loss of $R' \geq 0$ per neighbour. We show that this payoff structure is effectively identical to the model discussed in the main text.

Given that inspection reveals veracity of news perfectly, agents only share news verified to be false if and only if $d - R'd - c \geq 0$ or equivalently when $R' \leq 1$. In equilibrium, there is incentive to share both true and false news hence no one verifies. But this scenario is not interesting for our analysis. Instead, we focus on the case where news that has been verified to be false is not shared, $R' > 1$.

The seed's ex-ante payoff equals

$$U_{d,c}^{seed} = \begin{cases} \mu - (1-\mu)Rd & \text{if } a_{d,c}^{seed} = \mathscr{S} \\ 0 & \text{if } a_{d,c}^{seed} = \mathscr{K} \\ \mu d - c & \text{if } a_{d,c}^{seed} = \mathscr{V} \end{cases}. \tag{C.1}$$

Similarly for the non-seed (replacing μ with z). This yields us the following observation: The seed (and non-seed) shares unverified news instead of killing it so long as:

$$\mu \geq (R' - 1)/R' \tag{C.2}$$

In equilibrium, when $\mu < (R' - 1)/R'$, the seed will verify if and only if $c \leq \mu d$ and only share true news; non-seeds will not verify and pass on all news received. When $\mu \geq (R' - 1)/R'$, the seed will verify if and only if $c \leq (1 - \mu)(R' - 1)d$ and share unverified news; the non-seed will verify if and only if $c \leq (1 - z)(R' - 1)d$ and share unverified news. By defining $R' - 1$ to be R, we reach the identical equilibrium strategies described in our baseline model.

C.2 Alternate reputational loss

Consider that the total reputational loss from sharing false news is invariant to degree. The seed's ex-ante payoff equals

$$U_{d,c}^{seed} = \begin{cases} \mu d - (1 - \mu)R & \text{if } a_{d,c}^{seed} = \mathscr{S} \\ 0 & \text{if } a_{d,c}^{seed} = \mathscr{K} \\ \mu d - c & \text{if } a_{d,c}^{seed} = \mathscr{V} \end{cases}. \tag{C.3}$$

Similarly for the non-seed (replacing μ with z). In this case the average reputational loss per neighbour is decreasing in degree d — there is diminishing marginal punishment for sharing unverified news to an additional neighbour. An individual with high degree faces lower average punishment per neighbour, hence, she is more likely to share without verification. This yields us the following observation: The seed (and non-seed) shares unverified news instead of killing it if and only if:

$$d \geq \frac{(1 - \mu)R}{\mu} \quad \left(\text{and } d \geq \frac{(1 - z)R}{z}\right) \tag{C.4}$$

There exists two threshold $\hat{d}$ and $\tilde{d}(z)$ such that $\hat{d} = \frac{(1-\mu)R}{\mu}$ and $\tilde{d}(z) = \frac{(1-z)R}{z}$. Since $z \geq \mu$, $\frac{1-z}{z} \geq \frac{1-\mu}{\mu}$, $\hat{d} \geq \tilde{d}(z)$ for all z.

A non-seed must form expectations on whether their sender would share unverified news based on the degree distribution. The probability of receiving news from a sender with degree k conditional on news being false equals $Pr(\omega_k|v=F)$: a seed (non-seed) with degree below $\hat{d}$ ($\tilde{d}$) will kill unverified news, therefore an agent never receives false news; a seed (non-seed) with degree above $\hat{d}$ ($\tilde{d}$) will share unverified news, therefore an agent receives false news when it is not verified. Together, the ex-ante probability of receiving false news equals

$$\begin{aligned} Pr(\omega|v=F) &= \sum_k f(k) Pr(\omega_k|v=F) \\ &= \frac{1}{n-1} \sum_{k \geq \hat{d}} f(k)(1-p_k) + \frac{n-2}{n-1} Pr(\omega|v=F) \sum_{k \geq \tilde{d}} f(k)(1-q_k). \end{aligned} \tag{C.5}$$

Solving for $Pr(\omega|v=F)$

$$Pr(\omega|v=F) = \frac{\sum_{k \geq \hat{d}} f(k)(1-p_k)}{(n-1) - (n-2)\sum_{k \geq \tilde{d}} f(k)(1-q_k)} \tag{C.6}$$

where $p_k = H((1-\mu)R)$ and $q_k = H((1-z)R)$.

The probability of receiving news from a sender with degree k conditional on news being true equals $Pr(\omega_k|v=T)$: a seed (non-seed) with degree below $\hat{d}$ ($\tilde{d}$) will kill unverified news, therefore an agent receives true news when it is verified; a seed (non-seed) with degree above $\hat{d}$ ($\tilde{d}$) will share unverified news, therefore an agent always receives true news. Together, the ex-ante probability of receiving true news equals

$$\begin{aligned} Pr(\omega|v=T) &= \sum_k f(k) Pr(\omega_k|v=F) \\ &= \frac{1}{n-1} \sum_{k \geq \hat{d}} f(k) + \frac{1}{n-1} \sum_{k < \hat{d}} f(k) H(\mu k) \\ &+ \frac{n-2}{n-1} Pr(\omega|v=F) \sum_{k \geq \tilde{d}} f(k) + \frac{n-2}{n-1} Pr(\omega|v=F) \sum_{k < \tilde{d}} f(k) H(zk). \end{aligned} \tag{C.7}$$

Solving for $Pr(\omega|v=T)$

$$Pr(\omega|v=T)=\frac{\sum_{k\geq\hat{d}}f(k)+\sum_{k<\hat{d}}f(k)H(\mu k)}{(n-1)-(n-2)\left(\sum_{k\geq\tilde{d}}f(k)+\sum_{k<\tilde{d}}f(k)H(zk)\right)}. \tag{C.8}$$

Substituting it all into the definition of z gives us

$$z=Pr(v=T|\omega)=\frac{Pr(\omega|v=T)\mu}{Pr(\omega|v=T)\mu+Pr(\omega|v=F)(1-\mu)}. \tag{C.9}$$

Apart from the baseline effects on the verification probabilities, now changes in key parameters will also affect the proportion of agents sharing and killing unverified news. In the baseline model, when the quality of news μ improves, at first the quality of indirect news z worsens as agents verify less, but eventually z improves with the higher quality of direct news (Proposition 3.3). In this extension, as μ improves, there is an additional effect where a low-degree seed would shift from killing to sharing unverified news which further worsens the quality of indirect news z.

A FOSD shift in the conditional degree distribution will increase the fraction of agents with degree above $\hat{d}$ who would share unverified news. Hence, it reduces the improvement in quality of news received z^* from the baseline model.

C.3 Strategic complementarity in the verification

Another extension is to allow network users to punish their senders. Verification from an agent downstream will reveal that her sender has shared false news and impose punishment on her sender. Agents who do not verify and share false news now receive a reputational loss of R for each neighbour who verifies in the next period. Given that degrees are private information, the ex-ante probability of a neighbour verifying is $\sum_k q_k=\bar{q}$. For an agent with degree d, the expected number of neighbours verifying equals $\bar{q}d$. Therefore, the payoffs are

as follows:

$$U_{d,c}^{seed} = \begin{cases} \mu d - (1-\mu)R\bar{q}d & \text{if } a_{d,c}^{seed} = \mathscr{S} \\ 0 & \text{if } a_{d,c}^{seed} = \mathscr{K} \\ \mu d - c & \text{if } a_{d,c}^{seed} = \mathscr{V} \end{cases} \tag{C.10}$$

Payoffs for non-seeds are similar (replacing μ with z). As in the baseline model, by Brouwer's fixed-point theorem, there exists an equilibrium.

The game now features both strategy substitutes and strategic complements: Previously, verification upstream reduces the need to verify downstream. But now, verification downstream also increases the risk of punishment and therefore leads to more verification upstream. A high-degree sender faces higher risks of her receivers verifying, so she has additional incentives to verify.

By the same analysis as in the main text, both seeds and non-seeds share unverified news if $\mu \geq \frac{R\bar{q}}{R\bar{q}+1}$, or equivalently if $\frac{\mu}{(1-\mu)R} \geq \bar{q}$. Agents only share unverified news when downstream agents do not verify very often. When $\mu \geq \frac{R\bar{q}}{R\bar{q}+1}$, equilibrium verification probability equals $p_d^* = H((1-\mu)Rd\bar{q}^*)$ and $q_d^* = H((1-z^*)Rd\bar{q}^*)$ where z^* is as defined in Equation (3.5). When $\mu < \frac{R\bar{q}}{R\bar{q}+1}$, $p_d^* = H(\mu d)$, $q_d^* = 0$, and $z^* = 1$.

Note that no one verifying is always an equilibrium. If non-seeds do not verify $\bar{q}^* = 0$, no one receives punishment, so both seeds and non-seeds have no incentives to verify. All news is shared without verification since $\mu \geq \frac{R\bar{q}}{R\bar{q}+1} = 0$, leaving $z^* = \mu$. This is the scenario with *extremely viral misinformation*.

Proposition C.1. *No one verifying is always an equilibrium. (Unverified Sharing Eqm.)*

Next, we show that under the Sharing Equilibrium indirect news will never be true with certainty. Assume $z^* = 1$, then non-seeds will never verify. But if $\bar{q}^* = 0$, seeds will never verify. This implies that $z^* = \mu$, thus reaching a contradiction. As a result, seeds or non-seeds verifying with certainty is never an equilibrium.

Proposition C.2. *For all Sharing Equilibrium, $z^* < 1$.*

To illustrate, assume the verification cost function is uniform between 0 and 1, and consider a d-regular network. Since $p^* = 1$ or $q^* = 1$ is not an equilibrium, we can treat

$H(x) = x$. Ignoring the equilibrium where $q^* = 0$, we solve for a closed-form solution of q^*:

$$q^* = \frac{(1-\mu)Rd - 1}{(1-\mu)^2(Rd-1)Rd + (n-2)\mu} \tag{C.11}$$

which determines p^* and hence z^*. The comparative statics with respect R, μ, and network density d are presented in Figure C.1. Observe that Propositions 3.2 to 3.4 still hold.

C.4 Proofs

Proposition 3.2.

(i) Suppose $\mu < \frac{R}{R+1}$. In equilibrium, for all degree d, p_d^*, q_d^* and z^* are constant in R.

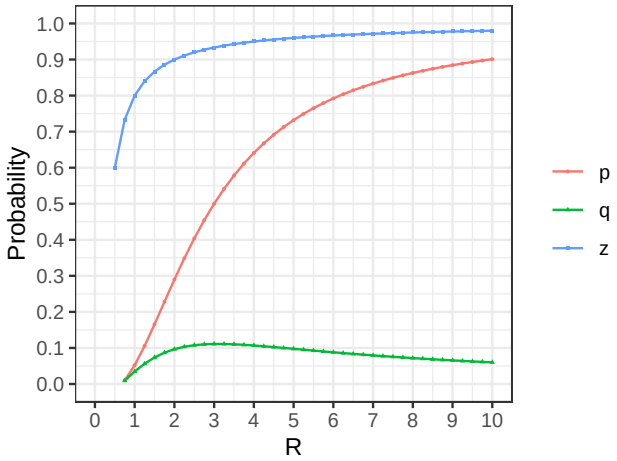
(ii) Suppose $\mu \geq \frac{R}{R+1}$. In equilibrium, there exists a unique $\hat{R} \in \left[\frac{\underline{c}}{(1-\mu)d_{max}}, \frac{\overline{c}}{(1-\mu)d_{min}}\right)$ such that for all degree d:

- *p_d^* is increasing in R;*
- *q_d^* is increasing in R when $R < \hat{R}$ and decreasing in R when $R > \hat{R}$;*
- *z^* is increasing in R.*

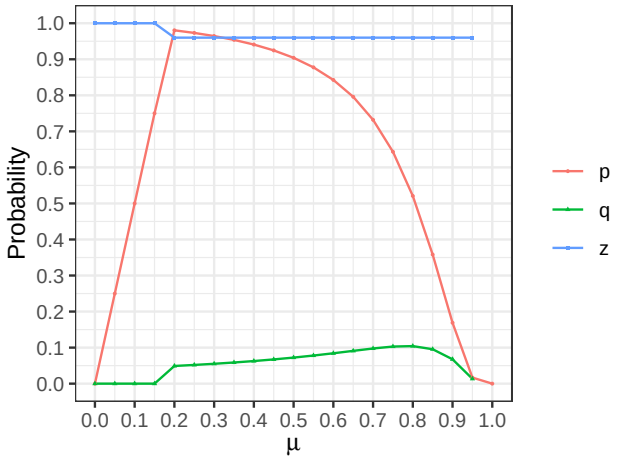
Proof. We prove the effects of the parameters R on the equilibrium strategies of verification p_d^* and q_d^* and the equilibrium probability of receiving true news z^*.

Recall that $p_k^* = 0$ when $k < \frac{\underline{c}}{(1-\mu)R}$ and $p_k^* = 1$ when $k \geq \frac{\overline{c}}{(1-\mu)R}$. The average equilibrium probability of seed verifying is

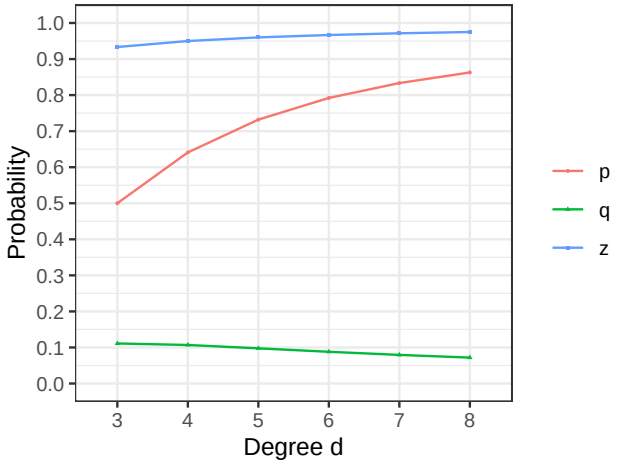
$$\begin{aligned}
\bar{p}^* &= \int_k f(k)p_k^* dk \\
&= \int_{-\infty}^{\frac{\underline{c}}{(1-\mu)R}} f(k)0dk + \int_{\frac{\underline{c}}{(1-\mu)R}}^{\frac{\overline{c}}{(1-\mu)R}} f(k)H((1-\mu)Rk)dk + \int_{\frac{\overline{c}}{(1-\mu)R}}^{\infty} f(k)1dk \\
&= F\left(\frac{\underline{c}}{(1-\mu)R}\right)0 + \int_{\frac{\underline{c}}{(1-\mu)R}}^{\frac{\overline{c}}{(1-\mu)R}} f(k)H((1-\mu)Rk)dk + \left(1 - F\left(\frac{\overline{c}}{(1-\mu)R}\right)\right)1 \\
&= \int_{\frac{\underline{c}}{(1-\mu)R}}^{\frac{\overline{c}}{(1-\mu)R}} f(k)H((1-\mu)Rk)dk + 1 - F\left(\frac{\overline{c}}{(1-\mu)R}\right)
\end{aligned} \tag{C.12}$$



(a) Reputational loss R



(b) News quality μ



(c) Network density d

Fig. C.1 Comparative statics on p^*, q^*, z^*. $n = 20, R = 5, \mu = 0.7, d = 5$

Using Leibniz integral rule, the derivative of $\bar{p}^*$ over R is:

$$\frac{d\bar{p}^*}{dR} = f\left(\frac{\overline{c}}{(1-\mu)R}\right)H(\overline{c})\frac{d}{dR}\left(\frac{\overline{c}}{(1-\mu)R}\right) - f\left(\frac{\underline{c}}{(1-\mu)R}\right)H(\underline{c})\frac{d}{dR}\left(\frac{\underline{c}}{(1-\mu)R}\right) + \int_{\frac{\underline{c}}{(1-\mu)R}}^{\frac{\overline{c}}{(1-\mu)R}} \frac{d}{dR}\left(f(k)H((1-\mu)Rk)\right)\ dk - f\left(\frac{\overline{c}}{(1-\mu)R}\right)\frac{d}{dR}\left(\frac{\overline{c}}{(1-\mu)R}\right) \quad \text{(C.13)}$$

Since $H(\overline{c}) = 1$ and $H(\underline{c}) = 0$, the derivative can be simplified to

$$\frac{d\bar{p}^*}{dR} = (1-\mu)\int_{\frac{\underline{c}}{(1-\mu)R}}^{\frac{\overline{c}}{(1-\mu)R}} f(k)h((1-\mu)Rk)k\ dk \quad \text{(C.14)}$$

By the same logic, the derivative of $\bar{q}^*$ over R is:

$$\frac{d\bar{q}^*}{dR} = (1-z^*-\frac{dz^*}{dR}R)\int_{\frac{\underline{c}}{(1-z^*)R}}^{\frac{\overline{c}}{(1-z^*)R}} f(k)h((1-z^*)Rk)k\ dk \quad \text{(C.15)}$$

Using expression of z^* in eq. (3.9), the derivative of z^* over R is:

$$\frac{dz^*}{dR} = \frac{(1-\mu)\mu\left[(1+(n-2)\bar{q}^*)\frac{d\bar{p}^*}{dR} + (1-\bar{p}^*)(n-2)\frac{d\bar{q}^*}{dR}\right]}{\left[(1-\mu)(1-\bar{p}^*)+\mu(1+(n-2)\bar{q}^*)\right]^2} \quad \text{(C.16)}$$

To simplify the algebra, define the following expressions (which are all positive):

$$\int_{\bar{p}} := \int_{\frac{\underline{c}}{(1-\mu)R}}^{\frac{\overline{c}}{(1-\mu)R}} f(k)h((1-\mu)Rk)k\ dk \quad \text{(C.17)}$$

$$\int_{\bar{q}} := \int_{\frac{\underline{c}}{(1-z^*)R}}^{\frac{\overline{c}}{(1-z^*)R}} f(k)h((1-z^*)Rk)k\ dk \quad \text{(C.18)}$$

$$a_p := (1-\mu)(1-\bar{p}^*) \quad \text{(C.19)}$$

$$a_q := \mu(1+(n-2)\bar{q}^*) \quad \text{(C.20)}$$

By solving the simultaneous equations (C.14), (C.15), and (C.16), we get:

$$\frac{d\bar{p}^*}{dR} = (1-\mu)\int_{\bar{p}} \tag{C.21}$$

$$\frac{d\bar{q}^*}{dR} = \frac{(1-z^*)(a_p+a_q)^2-(1-\mu)^2 R a_q \int_{\bar{p}}}{\int_{\bar{q}}^{-1}(a_p+a_q)^2+\mu R(n-2)a_p} \tag{C.22}$$

$$\frac{dz^*}{dR} = \frac{(1-\mu)^2 a_q \int_{\bar{p}} + (n-2)(1-z^*)\mu a_p \int_{\bar{q}}}{(a_p+a_q)^2+(n-2)\mu R a_p \int_{\bar{q}}} \tag{C.23}$$

Furthermore, the derivatives of p_d^* and q_d^* are analogues to the derivatives of $\bar{p}^*$ and $\bar{q}^*$ in expression (C.14) and (C.15).

$$\frac{dp_d^*}{dR} = (1-\mu)\cdot h((1-\mu)Rd)d \tag{C.24}$$

$$\frac{dq_d^*}{dR} = (1-z^*-\frac{dz^*}{dR}R)\cdot h((1-z^*)Rd)d$$

$$\implies \frac{dq_d^*}{dR} = \frac{(1-z^*)(a_p+a_q)^2-(1-\mu)^2 R a_q \int_{\bar{p}}}{\left[h((1-z^*)Rd)d\right]^{-1}(a_p+a_q)^2+\mu R(n-2)a_p} \tag{C.25}$$

It is clear that $\frac{d\bar{p}^*}{dR}$, $\frac{dp_d^*}{dR}$, and $\frac{dz^*}{dR}$ are non-negative. This implies that $\bar{p}^*$, p_d^* and z^* are (weakly) increasing in R.

The sign of $\frac{d\bar{q}^*}{dR}$ and $\frac{dq_d^*}{dR}$ are equivalent to the sign of the following expression:

$$(1-z^*)((1-\mu)(1-\bar{p}^*)+\mu(1+(n-2)\bar{q}^*))^2-(1-\mu)^2 R\mu(1+(n-2)\bar{q}^*)\int_{\bar{p}}. \tag{C.26}$$

If $(1-\mu)Rd_{min} > \bar{c}$, then Verified Sharing Equilibrium is reached where $q_d^* = 0$; If $(1-\mu)Rd < \underline{c}$, implying $(1-z^*)Rd < \underline{c}$, then $q_d^* = 0$. q_d^* is strictly concave in R because the second order derivative is negative. Therefore, there exists a unique global maximum point $\hat{R}$ between $\frac{\underline{c}}{(1-\mu)d}$ and $\frac{\bar{c}}{(1-\mu)d_{min}}$ where q_d^* is increasing in R when $R < \hat{R}$ and decreasing in R when $R > \hat{R}$. □

Proposition 3.3.

(i) Suppose $\mu < \frac{R}{R+1}$. In equilibrium, for all degree d, p_d^* is increasing in μ, while q_d^* and z^* are constant in μ.

(ii) Suppose $\mu \geq \frac{R}{R+1}$. *In equilibrium, there exists a unique* $\hat{\mu} \in (1 - \frac{\bar{c}}{Rd_{min}}, 1)$ *such that for all degree d:*

- p_d^* *is decreasing in* μ*;*
- q_d^* *is increasing and* z^* *is decreasing in* μ *when* $\mu < \hat{\mu}$*;*
- q_d^* *is decreasing and* z^* *is increasing in* μ *when* $\mu \geq \hat{\mu}$.

Proof. Using the method and definitions of expressions as above, the implicit derivatives of $\bar{p}^*$, $\bar{q}^*$ and z^* over μ are:

$$\frac{d\bar{p}^*}{d\mu} = -R\int_{\bar{p}} \tag{C.27}$$

$$\frac{d\bar{q}^*}{d\mu} = -R\frac{dz^*}{d\mu}\int_{\bar{q}} \tag{C.28}$$

$$\frac{dz^*}{d\mu} = \frac{(1-\bar{p}^*)(1+(n-2)\bar{q}^*) + (1-\mu)\frac{d\bar{p}^*}{d\mu}a_q + \mu(n-2)\frac{d\bar{q}^*}{d\mu}a_p}{(a_p+a_q)^2} \tag{C.29}$$

We use these equations above to solve for the explicit derivatives.

$$\frac{d\bar{p}^*}{d\mu} = -R\int_{\bar{p}} \tag{C.30}$$

$$\frac{d\bar{q}^*}{d\mu} = -\frac{(1-\bar{p}^*)(1+(n-2)\bar{q}^*) - (1-\mu)Ra_q\int_{\bar{p}}}{(a_p+a_q)^2 + (n-2)\mu Ra_p\int_{\bar{q}}}R\int_{\bar{q}} \tag{C.31}$$

$$\frac{dz^*}{d\mu} = \frac{(1-\bar{p}^*)(1+(n-2)\bar{q}^*) - (1-\mu)Ra_q\int_{\bar{p}}}{(a_p+a_q)^2 + (n-2)\mu Ra_p\int_{\bar{q}}} \tag{C.32}$$

Furthermore, the derivatives of p_d^* and q_d^* are analogues to the derivatives of $\bar{p}^*$ and $\bar{q}^*$ in eqs. (C.30) and (C.31). So we can solve for simultaneous equations as before.

$$\frac{dp_d^*}{d\mu} = -Rh((1-\mu)Rd)d \tag{C.33}$$

$$\frac{dq_d^*}{d\mu} = -\frac{(1-\bar{p}^*)(1+(n-2)\bar{q}^*) - (1-\mu)Ra_q\int_{\bar{p}}}{(a_p+a_q)^2 + (n-2)\mu Ra_p\int_{\bar{q}}}h((1-z^*)Rd)Rd \tag{C.34}$$

It is clear that $\frac{d\bar{p}^*}{d\mu}$ and $\frac{dp_d^*}{d\mu}$ are non-positive. This implies that $\bar{p}^*$ and p_d^* are (weakly) decreasing in μ.

The sign of $\frac{dz^*}{d\mu}$ and the inverse sign of $\frac{d\bar{q}^*}{d\mu}$ and $\frac{dq_d^*}{d\mu}$ are equivalent to the sign of the following expression:

$$(1-\bar{p}^*)(1+(n-2)\bar{q}^*)-(1-\mu)R\mu(1+(n-2)\bar{q}^*)\int_{\bar{p}}. \tag{C.35}$$

Since $(1+(n-2)\bar{q}^*)>0$, the expression is positive if and only if $1-\bar{p}^*-\mu(1-\mu)R\int_{\bar{p}}>0$. When $\mu=1$, the expression is positive — z^* is increasing in μ while q_d^* is decreasing in μ. When $\mu<1-\frac{\bar{c}}{Rd_{min}}$, all seeds verify $\bar{p}^*=1$, so the expression is negative — z^* is decreasing in μ while q_d^* is increasing in μ. z^* is strictly convex in R because the second order derivative is positive. Therefore, there exists a unique global maximum point $\hat{\mu}$ between $1-\frac{\bar{c}}{Rd_{min}}$ and 1 where z in decreasing (q_d^* is increasing) in μ when $\mu<\hat{\mu}$ and z in increasing (q_d^* is decreasing) in μ when $\mu>\hat{\mu}$.

□

Proposition 3.4. *Suppose f_1 FOSD f_2, or f_1 SOSD f_2 and $H(c)$ is a concave function in c:*

(i) If $\mu<\frac{R}{R+1}$, for all degree d, p_d^*, q_d^* and z^* remain constant under f_1 and f_2.

(ii) If $\mu\geq\frac{R}{R+1}$, for all degree d, $p_{1,d}^*=p_{2,d}^*$, $q_{1,d}^*\leq q_{2,d}^*$, $z_1^*\geq z_2^*$, and $\bar{p}_1^*\geq\bar{p}_2^*$.

If $H(c)$ is also a strictly increasing function in c, then the results hold with strict inequality.

Proof. Suppose there are two degree distributions f_1 and f_2 with corresponding average equilibrium probabilities $\bar{p}_1^*,\bar{q}_1^*,z_1^*$ and $\bar{p}_2^*,\bar{q}_2^*,z_2^*$.

We first compare $\bar{p}_1^*$ and $\bar{p}_2^*$ under the FOSD relation. Definition 3.1 states that if f_1 FOSD f_2 then $E_{f_1}[u]\geq E_{f_2}[u]$ for all non-decreasing function u. $\bar{p}_1^*$ is the probability of seed verifying averaged over the degree distribution f_1, i.e., $E_{f_1}[p_k^*]$. Since $p_k^*=H((1-\mu)Rk)$, the cumulative distribution of costs, p_k^* is a non-decreasing function of k for all degrees $k\in\{d_{min},d_{max}\}$. Therefore, if f_1 FOSD f_2 then $E_{f_1}[p_k^*]\geq E_{f_2}[p_k^*]$ implying $\bar{p}_1^*\geq\bar{p}_2^*$.

Next, we evaluate the difference between $\bar{q}_1^*$ and $\bar{q}_2^*$:

$$\begin{aligned}
\bar{q}_1^* - \bar{q}_2^* &= \int f_1(k)H((1-z_1^*)Rk)dk - \int f_2(k)H((1-z_2^*)Rk)dk \\
\bar{q}_1^* - \bar{q}_2^* &= \int [f_1(k) - f_2(k)]H((1-z_1^*)Rk)dk \\
&\quad + \int f_2(k)[H((1-z_1^*)Rk) - H((1-z_2^*)Rk)]dk
\end{aligned} \tag{C.36}$$

The first integral of eq. (C.36) is equivalent to $E_{f_1}[u(k)] - E_{f_2}[u(k)]$, where $u(k) = H((1-z_1^*)Rk)$. Given the fact that z_1^* is constant in k and H is a cumulative distribution, $u(k)$ is a non-decreasing function of k. Therefore, by Definition 3.1, if f_1 FOSD f_2 then the first integral of eq. (C.36) is greater than or equals to 0.

The second integral is positive when $(1-z_1^*) - (1-z_2^*)$ is positive, negative when $(1-z_1^*) - (1-z_2^*)$ is negative and 0 otherwise. In order to evaluate the second term, we look at the sign of $(1-z_1^*) - (1-z_2^*)$:

$$\begin{aligned}
&(1-z_1^*) - (1-z_2^*) \\
&= \frac{(1-\mu)(1-\bar{p}_1^*)}{(1-\mu)(1-\bar{p}_1^*) + \mu(1+(n-2)\bar{q}_1^*)} - \frac{(1-\mu)(1-\bar{p}_2^*)}{(1-\mu)(1-\bar{p}_2^*) + \mu(1+(n-2)\bar{q}_2^*)} \\
&\propto (1-\mu)(1-\bar{p}_1^*)\mu(1+(n-2)\bar{q}_2^*) - (1-\mu)(1-\bar{p}_2^*)\mu(1+(n-2)\bar{q}_1^*) \\
&= \mu(1-\mu)\left[(1-\bar{p}_1^*) - (1-\bar{p}_2^*) + (1-\bar{p}_1^*)(n-2)\bar{q}_2^* - (1-\bar{p}_2^*)(n-2)\bar{q}_1^*\right] \\
&\propto -(\bar{p}_1^* - \bar{p}_2^*) + (n-2)\left[(1-\bar{p}_1^*)\bar{q}_2^* - (1-\bar{p}_2^*)\bar{q}_1^*\right] \\
&= -(\bar{p}_1^* - \bar{p}_2^*) - (n-2)(1-\bar{p}_1^*)(\bar{q}_1^* - \bar{q}_2^*) - (n-2)(\bar{p}_1^* - \bar{p}_2^*)\bar{q}_1^* \\
&= -(\bar{p}_1^* - \bar{p}_2^*)(1+(n-2)\bar{q}_1^*) - (n-2)(1-\bar{p}_1^*)(\bar{q}_1^* - \bar{q}_2^*)
\end{aligned}$$

This gives us the following relation:

$$(1-z_1^*) - (1-z_2^*) \propto -(\bar{p}_1^* - \bar{p}_2^*)(1+(n-2)\bar{q}_1^*) - (n-2)(1-\bar{p}_1^*)(\bar{q}_1^* - \bar{q}_2^*) \tag{C.37}$$

If f_1 FOSD f_2, then we have shown that $\bar{p}_1^* \geq \bar{p}_2^*$ and the first term in eq. (C.36) is non-negative. Suppose $(1-z_1^*) - (1-z_2^*)$ is positive, then the second term in eq. (C.36)

is also positive, implying $\bar{q}_1^* - \bar{q}_2^*$ is positive. This means both terms on the right-hand side of eq. (C.37) are negative and $(1-z_1^*)-(1-z_2^*)$ should be negative, thus reaching a contradiction. Therefore, if f_1 FOSD f_2, then $(1-z_1^*)-(1-z_2^*)$ is less than or equals to 0, meaning $z_1^* \geq z_2^*$. This implies that $q_{1,d}^* \leq q_{2,d}^*$.

Knowing that the second term of eq. (C.36) is non-positive while the first term is non-negative, $\bar{q}_1^* - \bar{q}_2^*$ is greater than or equals to 0 if and only if the following inequality holds:

$$\int [f_1(k) - f_2(k)] H((1-z_1^*)Rk)dk + \int f_2(k)[H((1-z_1^*)Rk) - H((1-z_2^*)Rk)]dk \geq 0. \tag{C.38}$$

In summary, if f_1 FOSD f_2, then $\bar{p}_1^* \geq \bar{p}_2^*$ and $z_1^* \geq z_2^*$. Additionally, if inequality (C.38) is satisfied, $\bar{q}_1^* \geq \bar{q}_2^*$. Otherwise, $\bar{q}_1^* < \bar{q}_2^*$.

We can use the exact same method for SOSD by using Definition 3.1. If f_1 SOSD f_2 and $H(c)$ is non-decreasing and **concave** function in c, then we reach the same result as the FOSD relation. □

Proposition 3.5. *Suppose h_1 FOSD h_2, or suppose h_1 SOSD h_2 and $F(k)$ is a concave function in k:*

(i) If $\mu < \frac{R}{R+1}$, then for all degree d, $p_{1,d}^* \leq p_{2,d}^*$, while q_d^* and z^* remain constant.

(ii) If $\mu \geq \frac{R}{R+1}$, then for all degree d, $p_{1,d}^* \leq p_{2,d}^*$, $z_1^* \leq z_2^*$, and $\bar{p}_1^* \leq \bar{p}_2^*$.

If $F(k)$ is also a strictly increasing function in c, then the results hold with strict inequality.

Proof. Suppose there are two cost distributions h_1 and h_2 with corresponding average equilibrium probabilities $p_{1,d}^*, q_{1,d}^*, \bar{p}_1^*, \bar{q}_1^*, z_1^*$ and $p_{2,d}^*, q_{2,d}^*, \bar{p}_2^*, \bar{q}_2^*, z_2^*$.

Firstly, we evaluate the difference between $\bar{p}_1^*$ and $\bar{p}_2^*$ when $\mu \geq \frac{R}{R+1}$:

$$\bar{p}_1^* - \bar{p}_2^* = \int f(k)\big[H_1((1-\mu)Rk) - H_2((1-\mu)Rk)\big]dk \tag{C.39}$$

using integration by parts

$$\bar{p}_1^* - \bar{p}_2^* = \Big[F(k)\big[H_1(.) - H_2(.)\big]\Big]_0^\infty - \int_{k=0}^{\infty} F(k)\big[h_1(.) - h_2(.)\big] \cdot (1-\mu)R \ \ dk \tag{C.40}$$

evaluating the first term and substituting $b = (1-\mu)Rk$

$$\bar{p}_1^* - \bar{p}_2^* = 0 - \int_{b=0}^{\infty} \big[h_1(b) - h_2(b)\big] F(b/(1-\mu)R)db \tag{C.41}$$

$$\bar{p}_1^* - \bar{p}_2^* = -\Big[E_{h_1}(F(c/(1-\mu)R)) - E_{h_2}(F(c/(1-\mu)R))\Big] \tag{C.42}$$

Since $F(c/(1-\mu)R)$ is the cumulative distribution of degree, it is a non-decreasing function of c for all costs $c \in \{\underline{c}, \overline{c}\}$. Therefore, by Definition 3.1, if h_1 FOSD h_2 then $\bar{p}_1^* \leq \bar{p}_2^*$. The proof for when $\mu < \frac{R}{R+1}$ is similar.

Next, we evaluate the difference between $\bar{q}_1^*$ and $\bar{q}_2^*$ when $\mu \geq \frac{R}{R+1}$:

$$\begin{aligned}
\bar{q}_1^* - \bar{q}_2^* &= \int f(k)\big[H_1((1-z_1^*)Rk) - H_2((1-z_2^*)Rk)\big]dk \\
&= \int f(k)\big[H_1((1-z_1^*)Rk) - H_1((1-z_2^*)Rk)\big]dk \\
&\quad + \int f(k)\big[H_1((1-z_2^*)Rk) - H_2((1-z_2^*)Rk)\big]dk
\end{aligned} \tag{C.43}$$

Using the same method as evaluating $\bar{p}_1^* - \bar{p}_2^*$, the second integral of Equation (C.43) is equivalent to $E_{h_1}[u(c)] - E_{h_2}[u(c)]$, where $u(c) = F(\frac{c}{(1-z_2^*)R})$. Given the fact that z_2^* is constant in c and F is a cumulative distribution, $F(.)$ is non-decreasing functions in c. By Definition 3.1, if h_1 FOSD h_2, the second integral is non-positive. The first integral of Equation (C.43) is positive when $(1-z_1^*) - (1-z_2^*)$ is positive, negative when $(1-z_1^*) - (1-z_2^*)$ is negative, and 0 otherwise. We can evaluate the sign of the first integral using the same relation as before (Equation (C.37)). If h_1 FOSD h_2, then $\bar{p}_1^* - \bar{p}_2^*$ and the second integral in Equation (C.43) are both negative. Suppose $(1-z_1^*) - (1-z_2^*)$ is negative, the first integral of Equation (C.43) is also negative. By Equation (C.43), $\bar{q}_1^* - \bar{q}_2^*$ is negative which means both terms of Equation (C.37) are positive. This implies that $(1-z_1^*) - (1-z_2^*)$ is positive, thus reaching a contradiction. Therefore, $(1-z_1^*) - (1-z_2^*)$ must be non-negative and $z_1^* \leq z_2^*$.

The proof for the relationship between $\bar{q}_1^*, \bar{q}_2^*$ and z_1^*, z_2^* follows as above when h_1 SOSD h_2 and $F(.)$ is a concave cumulative degree distribution.

□

Optimal information accuracy

The platform faces the following objective function:

$$argmax_{\mu} \ \mathbf{1}_{\mu<\frac{R}{R+1}}\left[\mu\sum_{k} f(k)H(\mu k)\right]+\mathbf{1}_{\mu\geq\frac{R}{R+1}}\left[\mu+(1-\mu)\frac{1-\bar{p}^{*}}{1+(n-2)\bar{q}^{*}}\right]-\frac{1}{2}K\mu^{2} \tag{3.14}$$

First consider when $\mu \geq \frac{R}{R+1}$. The first-order condition (FOC) equals

$$1-Pr(\omega|v=F)+(1-\mu)\frac{d}{d\mu}Pr(\omega|v=F)-K\mu. \tag{C.44}$$

When $\mu = 1$, $Pr(\omega|v=F)=1$, and the first-order condition equals $-K$. So as long as investment cost is positive, monitoring all sources is too costly and $\mu = 1$ is not an equilibrium.

Next, consider when $\mu < \frac{R}{R+1}$. The FOC equals

$$\left(\frac{1}{\mu}\sum_{k} f(k)H(\mu k)+\frac{d}{d\mu}\sum_{k} f(k)H(\mu k)-K\right)\mu=0. \tag{C.45}$$

When $\mu \geq \frac{\bar{c}}{d_{min}}$, $H(\mu k)$ equals 1 and the FOC becomes $1/\mu - K = 0$, so the equilibrium investment $\mu^{*} = 1/K$. Therefore, for investment cost $K \in (\frac{R+1}{R}, \frac{d_{min}}{\bar{c}}]$, μ^{*} equals $1/K$. Assume $\mu > 0$, the payoff equals $\left[\frac{1}{\mu}\sum_{k} f(k)H(\mu k)-\frac{1}{2}K\right]\mu^{2}$. If $K > \frac{2}{\mu}\sum_{k} f(k)H(\mu k)$, any positive μ earns negative payoffs, so it is optimal for the platform to lower μ. When the investment cost is large, it is optimal to reduce μ^{*} to 0.

Under a regular network of degree d, it is optimal to lower μ if $K > \frac{2}{\mu}H(\mu d)$. Suppose $K \leq \frac{2}{\mu}H(\mu d)$, then the platform has incentive to increase μ until $\mu = 1/d$. Any further increases in μ will not improve the amount of verification, but only increase the costs (more than the spread induced through more true news). Therefore, for some intermediate range of investment cost, it is optimal to set $\mu^{*} = 1/d$ for a regular network.