2740.

Trust and Manipulation in Social Networks

MANUEL FÖRSTER, ANA MAULEON AND VINCENT VANNETELBOSCH

Network Science, 4(2), 216-243, 2016

50 YEARS OF CORE REPRINTS

CORE
Voie du Roman Pays 34, L1.03.01
Tel (32 10) 47 43 04
Fax (32 10) 47 43 01
Email: immaq-library@uclouvain.be
http://www.uclouvain.be/en-44509.html

Network Science

http://journals.cambridge.org/NWS

Additional services for ***Network Science:***

Email alerts: Click here
Subscriptions: Click here
Commercial reprints: Click here
Terms of use : Click here



Trust and manipulation in social networks

MANUEL FÖRSTER, ANA MAULEON and VINCENT J. VANNETELBOSCH

Network Science / Volume 4 / Issue 02 / June 2016, pp 216 - 243
DOI: 10.1017/nws.2015.34, Published online: 11 February 2016

Link to this article: http://journals.cambridge.org/abstract_S205012421500034X

How to cite this article:
MANUEL FÖRSTER, ANA MAULEON and VINCENT J. VANNETELBOSCH (2016). Trust and manipulation in social networks. Network Science, 4, pp 216-243 doi:10.1017/nws.2015.34

Request Permissions : Click here

CAMBRIDGE JOURNALS

Downloaded from http://journals.cambridge.org/NWS, IP address: 91.177.13.245 on 26 May 2016

Network Science **4** (2): 216–243, 2016. © Cambridge University Press 2016
doi:10.1017/nws.2015.34

Trust and manipulation in social networks

MANUEL FÖRSTER
Department of Economics, University of Hamburg, Germany
(e-mail: manuel.foerster@wiso.uni-hamburg.de*)*

ANA MAULEON and VINCENT J. VANNETELBOSCH
CEREC, Saint-Louis University – Brussels;
CORE, University of Louvain, Louvain-la-Neuve, Belgium
(e-mail: ana.mauleon@usaintlouis.be, vincent.vannetelbosch@uclouvain.be*)*

Abstract

We investigate the role of manipulation in boundedly rational opinion dynamics. Agents are subject to persuasion bias and repeatedly communicate with their neighbors in a social network. They can exert effort to manipulate trust in the opinions of others in their favor and update their opinions about some issue of common interest by taking weighted averages of neighbors' opinions. We show that manipulation can connect a segregated society and thus lead to mutual consensus. Second, we show that manipulation fosters opinion leadership; and surprisingly agents with low trust in their own opinion might get more influential even by being manipulated. Finally, comparative simulations reveal that manipulation is beneficial to information aggregation when preferences and abilities for manipulation are homogeneous, but detrimental in case abilities are concentrated at few powerful agents.

Keywords: *social networks, trust, manipulation, persuasion bias, opinion leadership, consensus, misinformation*

1 Introduction

Individuals often rely on social connections (friends, neighbors, and coworkers, as well as political actors and news sources) to form beliefs or opinions on various economic, political, or social issues. Every day individuals make decisions on the basis of these beliefs and moreover, it is often desirable that others hold similar beliefs as this would make them taking similar decisions. The latter is important if individuals need the support of others to enforce their interests. In politics, majorities are needed to pass laws, and in organizations, decisions might be taken by a hierarchical superior or in a decentralized fashion. Hence, it may be advantageous for individuals to increase their influence on others by manipulating the way they form their beliefs.

This paper studies a model of opinion formation where agents are subject to persuasion bias, i.e. fail to adjust for possible repetitions of information they receive and instead treat all information as new, and can exert effort to manipulate trust in the opinions of others in their favor.[1] We can see persuasion bias "as a simple,

[1] In this light, we also model the incentives for manipulation in a boundedly rational way, meaning that agents have a limited understanding of its consequences.

boundedly rational heuristic for dealing with a very complicated inference problem." (DeMarzo et al., 2003) Correctly adjusting for repetitions of information would become extremely complicated when individuals interact repeatedly on complex social networks.[2]

To give some concrete examples for this setup, consider an organization where decision making is decentralized, but decisions may have spillovers on other divisions. If this is the case, then opinions of other division leaders on issues related to the decisions are of importance as they will influence their decisions. And thus, leaders have incentives to manipulate the way the others form these opinions in their favor. Second, consider a network of policy-makers and lobbyists that hold opinions on some policy issue and assume that the social network reflects the (subjective) precisions of each agent's initial information as in DeMarzo et al. (2003).[3] Lobbyists then have incentives to manipulate the social network of policy-makers to make the latter implement policies in their favor, e.g. by convincing them of their (alleged) expertise or by providing (alleged) credible arguments for their opinion; in this context, we can view manipulation as persuasive or informational lobbying.[4] And more generally, we model manipulation as a persuasive activity that changes the persuasion bias of the manipulated agent in favor of the manipulator.[5]

Our setup is based on the model developed by DeGroot (1974). Agents hold *opinions* about some issue of common interest, are organized in a *weighted social network* and repeatedly meet and communicate with their neighbors in the network. We refer to the weighted and directed links from some agent to her neighbors as her *social trust*. At each discrete time instance, first one pair of agents is selected by a stochastic process such that one of them has the opportunity to *manipulate* the other one. The former can then decide whether or not to exert costly effort to manipulate the social trust of the other agent. If she decides to do so, this increases the social trust of the manipulated agent in the manipulator, with the magnitude of the increase depending on the manipulator's *ability to manipulate*. The manipulator's decision depends on her preferences for manipulation. She faces a trade-off between her personal gain from manipulation and its cost. Our model covers a large class of possible *reward functions* to measure the personal gain from manipulation of an agent, ranging from functions that depend only on current opinions and the agent's effort to functions where agents can foresee how opinions will change in the short-term. These are two possible extensions of persuasion bias to manipulation: one where agents do not need to know the network (apart from their own social trust), and one where agents do know the network, but are nevertheless unable to adjust for repetitions of information.[6] Agents might also have

[2] We refer to DeMarzo et al. (2003) for a detailed discussion of this assumption, including references on psychological evidence; for experimental evidence, see Choi et al. (2012); Chandrasekhar et al. (2012).

[3] A lobbyist would be an agent that has high ability to manipulate policy-makers, while these would have low (or no) ability to manipulate others.

[4] Potters & Van Winden (1992) define persuasive or informational lobbying as "the use by interest groups of their (alleged) expertise or private information on matters of importance for policy-makers in an attempt to persuade them to implement particular policies."

[5] Notice that agents are not compensated for being manipulated. This means that, for instance, the effort of a lobbyist is understood as being related to providing, e.g. (alleged) credible arguments for his opinion.

[6] Notice that indeed, inferring where current opinions come from requires much more computational ability than simply computing how an opinion will change in the short-term.

more sophisticated reward functions if they are less subject to persuasion bias and not manipulable themselves. Second, all agents communicate with their neighbors and update their opinions according to their (possibly manipulated) social trust, i.e. an agent's updated opinion is a weighted average of her neighbors' opinions (and possibly her own opinion) from the previous time instance. Without incentives to manipulate our model reverts to the standard DeGroot model of opinion formation.

We first study segregated societies and show that manipulation can connect segregated groups and thus make them reaching a consensus. In particular, an agent outside these groups can serve as a mediator by manipulating members of both groups, meaning that a mediator needs to have the ability to convince both groups of his expertise or credibility on the issue. Second, we show that manipulation fosters opinion leadership in the sense that the manipulating agent always increases her influence on the long-run opinions and the emerging (partial) consensus. For the other agents, the effect of manipulation is ambiguous and, depending on the social network, they might either gain or lose. In particular, we find that, surprisingly, agents with low trust in their own opinion might get more influential even by being manipulated. To round off this part of the analysis, we establish that under a weak minimum rationality requirement on the agents' gain from manipulation, manipulation eventually comes to an end. Finally, we investigate the tension between information aggregation and spread of misinformation. We rely on comparative simulations to obtain insights on the expected impact of manipulation on information aggregation. The simulations reveal that manipulation is beneficial to information aggregation when preferences and abilities for manipulation are homogeneous. This result turns out to be very robust with respect to changes in the initial opinions and the social network; and it even holds when agents manipulate without knowing the social network, basing their decision only on current opinions and their effort. The intuition for why this holds is that agents benefit more from manipulation when the manipulated agent is influential as in this case they indirectly get more influential on many other agents. And thus, as the average influence of the other agents in the society decreases in the agent's own influence, agents benefit more from manipulation the lower their own influence. However, we also find that manipulation is detrimental to information aggregation in case abilities are concentrated at few powerful agents; such agents, e.g. lobby organizations, might severely harm information aggregation and spread their (potentially) misleading information. Agents with higher ability benefit more from manipulation and hence increase their influence even if they were already rather influential before.

There is a large and growing literature on opinion formation in social networks, either using a Bayesian perspective or some boundedly rational framework where agents are subject to persuasion bias.[7] We study a model of boundedly rational opinion formation on a social network based on the seminal paper by DeGroot (1974), see also DeMarzo et al. (2003), and Golub & Jackson (2010) for related works. DeMarzo et al. (2003) show that persuasion bias implies the phenomena of social influence and unidimensional opinions. Golub & Jackson (2010) study social

[7] Acemoglu et al. (2011) develop a model of Bayesian learning on general social networks, and Acemoglu & Ozdaglar (2011) provide an overview of recent research on opinion dynamics and learning in social networks.

learning under persuasion bias and find that all opinions in a large society converge to the truth if and only if the influence of the most influential agent vanishes as the society grows.

Buechel et al. (2015) develop a model of opinion formation under persuasion bias where agents may state an opinion that differs from their true opinion due to their preferences for conformity. They find that lower conformity fosters opinion leadership. In addition, the society becomes wiser if agents who are well informed are less conform, while uninformed agents conform more with their neighbors. Friedkin & Johnsen (1990) propose a variation where agents are subject to persuasion bias but can adhere to their initial beliefs to some degree. This leads to persistent disagreement among the agents in the long-run. And similar results hold when assuming some kind of homophily, see, e.g. Axelrod (1997); Hegselmann & Krause (2002); Deffuant et al. (2000). Acemoglu et al. (2010) is related to Section 5 of our work as they investigate the tension between information aggregation and spread of misinformation in a related model. They characterize how the presence of forceful agents affects information aggregation. Forceful agents influence the beliefs of the other agents they meet, but only change very slowly their own opinions. In our framework, these would be agents with very high trust in their own opinion. They show that all beliefs converge to a stochastic consensus and quantify the extent of misinformation by providing bounds on the gap between the consensus value and the benchmark without forceful agents where there is efficient information aggregation. Friedkin (1991) studies measures to identify opinion leaders under persuasion bias.

Furthermore, Watts (2014) studies the influence of social networks on correct voting. Agents have beliefs about each candidate's favorite policy and update their beliefs based on the favorite policies of their neighbors and on whom the latter support. She finds that political agreement in an agent's neighborhood facilitates correct voting, i.e. voting for the candidate whose favorite policy is closest to his own favorite policy. Our paper is also related to the literature on lobbying as costly signaling, e.g. Austen-Smith & Wright (1994); Esteban & Ray (2006). These papers do not consider networks and model lobbying as providing one-shot costly signals to decision-makers in order to influence a policy decision.[8] To the best of our knowledge, we are the first allowing for the possibility to manipulate the social network of other agents under persuasion bias.

The paper is organized as follows. In Section 2, we introduce the model of opinion formation. In Section 3, we study segregated societies. Section 4 looks at the long-run effects of manipulation. In Section 5, we investigate how manipulation affects the extent of misinformation in society. Section 6 concludes. The proofs are presented in the appendix.

2 Model and notation

We consider a set $\mathcal{N} = \{1, 2, \ldots, n\}$, $n \geqslant 2$, of agents who repeatedly communicate with their neighbors in a social network. Each agent $i \in \mathcal{N}$ has an initial *opinion*

[8] Notice that we study how (repeated) manipulation and lobbying affect public opinion (and potentially single decision-makers) in the long-run, but do not model explicitly any decision-making process.

or *belief* $x_i(0) \in \mathbb{R}$ about some issue of common interest and an initial vector of *social trust* $m_i(0) = (m_{i1}(0), m_{i2}(0), \ldots, m_{in}(0))$, with $0 \leqslant m_{ij}(0) \leqslant 1$ for all $j \in \mathcal{N}$ and $\sum_{j \in \mathcal{N}} m_{ij}(0) = 1$, that captures how much attention agent i initially pays to each of the other agents. More precisely, $m_{ij}(0)$ is the initial weight or trust that agent i places on the current opinion of agent j in forming her updated belief. We assume that $m_{ii}(0) > 0$ for all $i \in \mathcal{N}$, which can be interpreted as each agent having some trust in her own initial opinion.[9] At time instance $t = 0, 1, 2, \ldots$, the agents' beliefs are represented by the vector $x(t) = (x_i(t))_{i \in \mathcal{N}}$ and their social trust by the matrix $M(t) = (m_{ij}(t))_{i,j \in \mathcal{N}}$. Furthermore, let $\alpha_{ij} \geqslant 0$ denote agent i's *ability to manipulate* with respect to agent j.[10]

First, one agent is chosen randomly (probability $1/n$ for each agent) to meet and to have the opportunity to manipulate another agent. If agent $i \in \mathcal{N}$ is chosen at time t, she meets agent $j \in \mathcal{N}$ with probability $\alpha_{ij}/(\sum_{k \neq i} \alpha_{ik})$ and can decide on the effort $\alpha \in \{0, \alpha_{ij}\}$ she would like to exert on j, and no meeting takes place if $\sum_{k \neq i} \alpha_{ik} = 0$.[11] We write $\Gamma(t) = (i, j; \alpha)$ when agent i is chosen to manipulate agent j at time t and decides to exert effort α on j. The decision of agent i leads to the following updated trust weights of agent j:

$$m_{jk}(t+1) = \begin{cases} m_{jk}(t)/(1+\alpha) & \text{if } k \neq i \\ \big(m_{jk}(t) + \alpha\big)/(1+\alpha) & \text{if } k = i \end{cases}.$$

The trust of j in i is increasing in the effort of agent i and all trust weights of j are normalized.[12] Notice that this normalization implies that the trust of j in an agent other than i decreases by the factor $1/(1+\alpha)$, i.e. the absolute decrease in trust is proportional to its level. If i decides not to manipulate ($\alpha = 0$), the trust matrix does not change. To emphasize that the resulting updated trust matrix $M(t+1)$ is derived by applying the manipulation decision $\Gamma(t) = (i, j; \alpha)$ to $M(t)$, we also write $[M(t)](i, j; \alpha)$ instead of $M(t+1)$. Agent i decides on whether or not to manipulate agent j according to her utility function

$$u_i\big(\alpha \mid x(t), M(t), j\big) = v_i\big(\alpha \mid x(t), M(t), j\big) - c_i(\alpha \mid j),$$

where $v_i\big(\alpha \mid x(t), M(t), j\big)$ is the *(relative) reward* for her effort $\alpha \in \{0, \alpha_{ij}\}$ and $c_i(\alpha \mid j)$ is its *cost*. We assume that for all $j \neq i$, $c_i(\alpha_{ij} \mid j) > c_i(0 \mid j) = 0$ such that effort is costly. Notice that $v_i\big(\alpha_{ij} \mid x(t), M(t), j\big) - v_i\big(0 \mid x(t), M(t), j\big)$ measures agent i's *gain from manipulation*. Second, all agents communicate with their neighbors and update their opinions using the updated trust weights, reflecting that they are subject to persuasion bias and fail to adjust for possible repetitions of information:

$$x(t+1) = M(t+1)x(t) = [M(t)](i, j; \alpha)x(t).$$

[9] We impose this weak assumption as it guarantees convergence to consensus if the social network is strongly connected, i.e. if there is a directed path between any two agents. Notice, however, that it would suffice to assume aperiodicity of the initial trust matrix.

[10] Notice that we can fit the model to a situation of lobbying: take α_{ij} large for some i (the lobbyist) and all $j \neq i$ (policy-makers), while $\alpha_{jk} \approx 0$ for $j \neq i$ and all $k \neq j$.

[11] To keep things simple, we assume that the meeting probabilities are equal to the relative abilities to manipulate of the agents. Notice that we do not assume that agents choose an optimal effort level $\alpha^* \in [0, \alpha_{ij}]$ as they are boundedly rational and as this would require high computational abilities involving calculating derivatives.

[12] This reflects that manipulation is a persuasive activity that changes the persuasion bias of the manipulated agent in favor of the manipulator.

We can rewrite this equation as $x(t+1) = \overline{M}(t+1)x(0)$, where $\overline{M}(t) := M(t)M(t-1)\cdots M(1)$ denotes the *overall trust matrix*. Now, let us give some examples of reward functions.

Example 1 (*Reward functions*)
Consider the function

i.
$$v_i\big(\alpha \mid x(t), M(t), j\big) = -\frac{1}{n-1}\sum_{k\neq i}\Big(x_i(t) - x_k(t+1)\Big)^2,$$
where $x_k(t+1) = \big([M(t)](i,j;\alpha)x(t)\big)_k$. Agent i's objective is to bring the other agents' updated opinions as close as possible to her current opinion, weighting equally the other agents and disregarding that her own opinion will change as well.[13] The latter represents that she is subject to persuasion bias.

ii.
$$v_i\big(\alpha \mid x(t), M(t), j\big) = -\sum_{k\neq i} m_{ik}(t)\Big(x_i(t) - x_k(t+1)\Big)^2,$$
where $x_k(t+1) = \big([M(t)](i,j;\alpha)x(t)\big)_k$. Here, agent i's objective is to bring the updated opinions of agents she trusts as close as possible to her current opinion, the importance of each agent being given by her trust weights.

iii.
$$v_i\big(\alpha \mid x(t), M(t), j\big) = \alpha\big(x_i(t) - x_j(t)\big)^2,$$
a simple reward function that is independent of the trust matrix $M(t)$ and thus builds on the persuasion bias assumption. The product of the exerted effort and the squared distance between the two agents' current opinions serves as a boundedly rational proxy for the reward from manipulation. It seems natural that, without knowledge of the network, agents manipulate those agents whose opinions are far from their own opinion.

We will frequently use the first reward function in examples, together with a cost function $c_i(\alpha_{ij} \mid j) = c$ for some constant $c > 0$. Our model reverts to the standard DeGroot model without manipulation.

Remark 1
If we choose either constant reward functions $v_i \equiv v$ for all $i \in \mathcal{N}$ or manipulation abilities $\alpha_{ij} = 0$ for all $i, j \in \mathcal{N}$, then agents have no incentives to or cannot manipulate, respectively, and our model reverts to the standard model of DeGroot (1974).

3 Segregated societies

In this section, we study segregated societies. We investigate how manipulation can change segregated groups and, in particular, under which conditions it might lead to a connected society. We first shortly recall some graph-theoretic terminology. We call a group of agents $C \subseteq \mathcal{N}$ *minimal closed* at time t if these agents only trust

[13] For instance, consider an issue on which there will be a vote at some point. Then, the equal weights represent that each opinion is equally important as each agent has one vote.

agents inside the group, i.e. $\sum_{j\in C} m_{ij}(t) = 1$ for all $i \in C$, and if this property does not hold for a proper subset $C' \subsetneq C$. The set of minimal closed groups at time t is denoted by $\mathscr{C}(t)$ and is called the *trust structure*. The society is *segregated* at time t if $\mathscr{C}(t)$ contains more than one group. Notice that agents in a minimal closed group reach a consensus in the long-run conditional on that the trust matrix does not change any more. A walk at time t of length K (from agent i_1 to agent i_{K+1}) is a sequence of agents $i_1, i_2, \ldots, i_{K+1}$ such that $m_{i_k,i_{k+1}}(t) > 0$ for all $k = 1, 2, \ldots, K$. A walk is a *path* if all agents are distinct. At each time t, we can decompose the set of agents $\mathscr{N}$ into minimal closed groups and agents outside these groups, the *rest of the world*, $R(t)$:

$$\mathscr{N} = \bigcup_{C\in\mathscr{C}(t)} C \cup R(t).$$

Within minimal closed groups, all agents interact indirectly with each other, i.e. there is a path between any two agents.[14] Moreover, notice that agent $i \in \mathscr{N}$ is part of the rest of the world $R(t)$ if and only if there is a path at time t from her to some agent in a minimal closed group $C \not\ni i$. We say that a manipulation at time t does not change the trust structure if $\mathscr{C}(t+1) = \mathscr{C}(t)$, which implies that $R(t+1) = R(t)$.

We find that manipulation changes the trust structure when the manipulated agent belongs to a minimal closed group and additionally the manipulating agent does not belong to the same group. The group of the manipulated agent is either disbanded or the manipulating agent (and possibly others) join the group.

Proposition 1
Suppose that at time t, $\Gamma(t) = (i, j; \alpha_{ij})$, $\alpha_{ij} > 0$.

i. Let $i \in \mathscr{N}$, $j \in R(t)$ or $i, j \in C \in \mathscr{C}(t)$. Then, the trust structure does not change.
ii. Let $i \in C \in \mathscr{C}(t)$ and $j \in C' \in \mathscr{C}(t)\backslash\{C\}$. Then, C' is disbanded, i.e. $\mathscr{C}(t+1) = \mathscr{C}(t)\backslash\{C'\}$.
iii. Let $i \in R(t)$ and $j \in C \in \mathscr{C}(t)$.
    a. Suppose that there exists no path from i to k for any $k \in \cup_{C'\in\mathscr{C}(t)\backslash\{C\}} C'$. Then, $R' \cup \{i\}$ joins C, i.e.
    $$\mathscr{C}(t+1) = \mathscr{C}(t)\backslash\{C\} \cup \{C \cup R' \cup \{i\}\},$$
    where $R' = \{l \in R(t)\backslash\{i\} \mid \text{there is a path from } i \text{ to } l\}$.
    b. Suppose that there exists $C' \in \mathscr{C}(t)\backslash\{C\}$ such that there exists a path from i to some $k \in C'$. Then, C is disbanded.

The proofs of this and other results can be found in Appendix A. Manipulation of an agent in the rest of the world or within a minimal closed group (part (i)) does not change the trust structure as all groups remain closed. Second, if both agents belong to different groups, the one of the manipulated agent is disbanded and its agents join the rest of the world as it is no longer closed after manipulation. And third, if only the manipulated agent belongs to a minimal closed group, then the effect on the group of the manipulated agent depends on the trust matrix. If the manipulating agent does not (indirectly) trust anyone in another minimal

[14] Notice that minimal closed groups are also called strongly connected and closed groups, see Golub & Jackson (2010).

closed group, then she and possibly others join the group of the manipulated agent (part (iii,a)), while otherwise her group is disbanded (part (iii,b)).[15] The next remark follows immediately from the fact that the society reaches a consensus if and only if it is not segregated, i.e. there is a single minimal closed group.[16] In particular, it shows that an agent outside two segregated groups can serve as a mediator by manipulating members of both groups.

Remark 2
Suppose that at time t, $\mathscr{C}(t) = \{C, C'\}$ and $R(t) = \{k\}$. Society reaches a consensus if either

i. agent $i \in C(C')$ manipulates agent $j \in C'(C)$ at some time instance $t' \geqslant t$, or
ii. agent k manipulates agent $k' \in C$ at some time instance $t' \geqslant t$ and agent $k'' \in C'$ at some time instance $t'' \geqslant t$.

The first part follows as one of the two minimal closed groups is disbanded (Proposition 1 part (ii)) and thus there is only one group left. Similarly, if agent k manipulates an agent in one of the groups, then she either joins the group (Proposition 1 part (iii,a)) or it is disbanded (Proposition 1 part (iii,a)). In the first case, the other group will be disbanded after she has manipulated as well an agent therein. Thus, one of the groups will be disbanded in any case as in the first part. The following example illustrates how manipulation can enable a society to reach a consensus when there are two segregated groups and no rest of the world.

Example 2 (*Consensus due to manipulation*)
Consider $\mathscr{N} = \{1, 2, 3\}$ and

$$u_i\big(\alpha \mid x(t), M(t), j\big) = -\frac{1}{2}\sum_{k \neq i}\big(x_i(t) - x_k(t+1)\big)^2 - \frac{\alpha}{10}$$

for all $i \in \mathscr{N}$. Notice that the first part of the utility is the reward function from Example 1 part (i). Furthermore, the initial opinions and trust matrix are

$$x(0) = \begin{pmatrix} 0.9 \\ 0.6 \\ 0.05 \end{pmatrix} \text{ and } M(0) = \begin{pmatrix} 0.9 & 0.1 & 0 \\ 0.05 & 0.95 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and $\alpha_{ij} = 0.2$ is the ability of i to manipulate j, for all $i \neq j$. Notice that the agents form two minimal closed groups: $\mathscr{C}(0) = \{\{1, 2\}, \{3\}\}$. Without manipulation, agents 1 and 2 would thus reach a consensus, while the stubborn agent 3 would stick to her initial opinion. However, with manipulation, the agents reach a consensus with probability 1 (at least one of the agents 1 and 2 manipulates agent 3 or vice versa, implying that one of the groups is disbanded). Figure 1 depicts one simulation run

[15] We say that agent i indirectly trusts agent j (at time t) if there is a path (at time t) from i to j.
[16] There is always at least one minimal closed group; agents in the rest of the world adopt the consensus of this group.

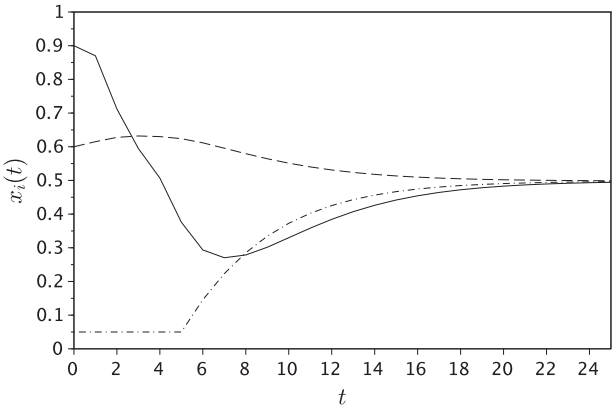


Fig. 1. Opinion dynamics in Example 2. The solid line represents agent 1, the dashed line agent 2, and the dashed-dotted line agent 3.

of the long-run opinion dynamics, where[17]

$$\Gamma(t) = \begin{cases} (3,1;0.2), & t = 1,4 \\ (2,3;0.2), & t = 5 \\ (\cdot,\cdot;0), & \text{otherwise} \end{cases}.$$

The group of agent 1 is disbanded at time instance 1 due to the manipulation of agent 3 (see part (ii) of Proposition 1). The last manipulation takes place at time 5, where agent 2 manipulates agent 3, meaning that agents 1 and 2 join the group of agent 3 (see part (iii,a) of Proposition 1) such that they eventually form one minimal closed group, $\mathscr{C}(6) = \{\mathcal{N}\}$, and reach a mutual consensus. This results in the following final trust matrix:

$$M(t) \simeq \begin{pmatrix} 0.625 & 0.069 & 0.306 \\ 0.05 & 0.95 & 0 \\ 0 & 0.167 & 0.833 \end{pmatrix}, t \geqslant 6.$$

4 Long-run opinion dynamics

In this section, we investigate the long-run effects of manipulation. First, we study the consequences of a single manipulation on the long-run opinions of minimal closed groups. In this context, we are interested in the role of manipulation in opinion leadership. Second, we study the outcome of the influence process and illustrate our results by means of an example.

4.1 Opinion leadership

Typically, an agent is called *opinion leader* if she has substantial influence on the long-run beliefs of a group. That is, if she is among the most influential agents in the

[17] The trajectories of the opinions can differ substantially between runs as they depend on the order of the agents' meetings.

group. Intuitively, manipulating others should increase influence on long-run beliefs and thus foster opinion leadership. In what follows, we investigate this issue.

We denote by $\pi(C;t) = (\pi_i(C;t))_{i\in C}$ the probability vector of the agents' influence on the final consensus of their group $C \in \mathscr{C}(t)$ at time t, conditional on that the trust matrix does not change any more.[18] In this case, the group converges to the consensus

$$x(\infty) = \pi(C;t)\, x(t)|_C = \sum_{i\in C} \pi_i(C;t) x_i(t),$$

where $x(t)|_C = (x_i(t))_{i\in C}$ is the restriction of $x(t)$ to agents in C. In other words, $\pi_i(C;t)$, $i \in C$, is the influence weight of agent i's opinion at time t, $x_i(t)$, on the consensus of C. Notice that the influence vector $\pi(C;t)$ depends on the trust matrix $M(t)$ and therefore changes with manipulation. A higher value of $\pi_i(C;t)$ corresponds to more influence of agent i on the consensus. Each agent in a minimal closed group has at least some influence on the consensus: $\pi_i(C;t) > 0$ for all $i \in C$.[19]

We restrict our analysis to the case where both the manipulating and the manipulated agent are in the same minimal closed group. Since in this case the trust structure is preserved (Proposition 1 part (i)), we can compare the influence on the consensus of the group before and after manipulation. We show that the manipulating agent always gains influence, which confirms our intuition that manipulation fosters opinion leadership. For any other agent, the change can be positive or negative, implying that even the manipulated agent might gain influence.

Proposition 2
Suppose that at time t, $\Gamma(t) = (i, j; \alpha)$, $i, j \in C$. If $\alpha = \alpha_{ij} > 0$, then

i. agent i strictly increases her long-run influence, $\pi_i(C;t+1) > \pi_i(C;t)$,
ii. any other agent $k \neq i$ of the group can either gain or lose influence, depending on the trust matrix,
iii. agent $k \neq i, j$ loses influence for sure if j trusts solely her, i.e. $m_{jk}(t) = 1$.

We refer to Appendix B for technical details and the proof of Proposition 2. For agent i, the manipulator, the direct gain of influence (due to an increase of trust from j) always dominates her indirect loss of influence (due to a decrease of trust from j faced by agents that (indirectly) trust i). For agents $k \neq i, j$, there is a trade-off between an indirect gain of influence (due to the increase of trust that i obtains from j, but only in case i (indirectly) trusts k), on the one hand, and a direct loss of influence (due to a decrease of trust from j) as well as an indirect loss of influence (due to a decrease of trust from j faced by agents that (indirectly) trust k), on the other hand. In the extreme case where j only trusts k, the direct loss of influence dominates the indirect gain for sure.

Similarly, in case of the manipulated agent j, the trade-off is between an indirect gain of influence (due to the increase of trust that i obtains from her, but only if i (indirectly) trusts j), and an indirect loss of influence (due to a decrease of trust from her faced by agents that (indirectly) trust her). In particular,

[18] In the language of Markov chains, $\pi(C;t)$ is known as the unique stationary distribution of the communication class C.
[19] See Golub & Jackson (2010).

this means that, somehow surprisingly, even the manipulated agent j might gain influence. In the following, we investigate this observation more closely. We show that the manipulated agent gains from being manipulated in situations where the manipulating agent trusts her significantly and at the same time she does not have much trust in her own opinion, i.e. if $m_{ij}(t)$ is large and $m_{jj}(t)$ small.

Corollary 1
Suppose that at time t, $\Gamma(t) = (i, j; \alpha)$, $i, j \in C$, $|C| \geqslant 3$ and $\alpha = \alpha_{ij} > 0$. Then, there exists $q(\eta) < 1$ such that agent j gains influence from being manipulated if

i. $m_{ij}(t) > q(\eta)$ and
ii. $m_{jj}(t) < (1-\eta)/(3-\eta) \cdot \sum_{k \in C \setminus \{i,j\}} m_{jk}(t)$,

where $\eta := \max_{k \in C \setminus \{i,j\}} m_{kj}(t)$.

The intuition behind this result is that an agent j that does not have much trust in her own opinion (condition (ii)) basically gives it up immediately and thus relies only on intermediaries to spread her opinion.[20] In these circumstances, agent j gains influence from being manipulated by an intermediary agent i that trusts her significantly (condition (i)). In a sense, the manipulated agent has trusted the "wrong" agents before, i.e. agents $k \neq i, j$ that did not trust her as much as the manipulating agent i does. In particular, condition (ii) implies two necessary conditions for that agent j gains influence from being manipulated: first, j needs to trust at least some agent $k \neq i, j$ (as otherwise the right-hand side vanishes), and second, $m_{jj}(t) < 1/4$ (as $\eta \geqslant 0$ and $\sum_{k \in C \setminus \{i,j\}} m_{jk}(t) \leqslant 1 - m_{jj}(t)$).

4.2 Convergence

We have already seen to which consensus value the agents converge to if no more manipulation takes place. We now investigate this issue from an ex-ante point of view. To ensure that manipulation decisions satisfy a minimum rationality requirement, we consider the following assumptions on the reward functions of the agents.

Assumption 1 (*Minimum reward rationality*)

A1. Agent i's gain from manipulating agent j vanishes if the effect of manipulation on the updated opinions tends to zero, i.e.

$$v_i\big(\alpha_{ij} \mid x, M, j\big) - v_i\big(0 \mid x, M, j\big) \to 0 \text{ if } \|[M](i, j; \alpha_{ij})x - [M](i, j; 0)x\| \to 0,$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^n$.

A2. Agent i's gain from manipulating agent j vanishes if their opinion difference tends to zero, i.e.

$$v_i\big(\alpha_{ij} \mid x, M, j\big) - v_i\big(0 \mid x, M, j\big) \to 0 \text{ if } |x_i - x_j| \to 0.$$

[20] Notice that for agents with high trust in their own opinion, their influence is partially due to the fact that they spread it over time.

The idea behind (A1) is that there should be no gain from manipulation without manipulation affecting the updated opinions. For (A2), the idea is that there should be no gain from manipulation without the current opinions of the manipulator and the manipulated agent being different. Notice that the reward functions in Example 1 (i) and (ii) satisfy (A1) but not (A2) and vice versa for that in Example 1 (iii).

We first show that the trust structure eventually settles down. Second, if all reward functions satisfy either (A1) or (A2), manipulation within the minimal closed groups that have finally been formed comes to an end. We also determine the final consensus opinion of each group.

Proposition 3

i. There exists an almost surely finite stopping time τ such that under the event $\{\tau = t\}$, we have $\mathscr{C}(t') = \mathscr{C}(t)$ for all $t' \geqslant t$.
ii. Take $C \in \mathscr{C}(t)$. If for all $i \in C$ either (A1) or (A2) holds, then there exists an almost surely finite stopping time $\widehat{\tau}$ such that almost surely $\widehat{\tau} \geqslant t$ and under the event $\{\widehat{\tau} = t''\}$ agents in C are not manipulated from time t'' on. Moreover, they converge to the random variable

$$x(\infty) = \pi(C;\widehat{\tau})\, M(\widehat{\tau}-1)|_C\, M(\widehat{\tau}-2)|_C \cdots M(1)|_C\, x(0)|_C\,.$$

Part (i) holds as the number of possible changes in the trust structure is bounded. To understand why part (ii) holds, notice that when the trust structure has settled down, the agents continue to approach a mutual consensus after each manipulation. At some time t'' (under the event $\{\widehat{\tau} = t''\}$), the effect of manipulation on the updated opinions becomes small enough (the current opinions become close enough) such that $v_i\big(\alpha_{ij} \mid x(t''), M(t''), j\big) - v_i\big(0 \mid x(t''), M(t''), j\big) < c_i(\alpha_{ij} \mid j)$, i.e. the gain from manipulation is too small to outweigh its cost and thus agent i decides not to manipulate j from time t on. And as this holds for any pair of agents (i, j) in C, manipulation comes to an end.

Denote by $\overline{\pi}_i(C;t)$, the *overall influence* of agent i's initial opinion on the consensus of group C at time t conditional on that no more manipulation affecting C takes place. The formula for the overall influence is implicitly given by Proposition 3. In particular, it turns out that an agent outside a minimal closed group that has finally been formed can never have any influence on its consensus opinion.

Corollary 2

Under the event $\{\tau = t\}$, the overall influence of the initial opinion of agent $i \in \mathscr{N}$ on the consensus of the group $C \in \mathscr{C}(t)$ is given by the random variable

$$\overline{\pi}_i(C;\widehat{\tau}) = \begin{cases} \big(\pi(C;\widehat{\tau})\, M(\widehat{\tau}-1)|_C\, M(\widehat{\tau}-2)|_C \cdots M(1)|_C\big)_i & \text{if } i \in C \\ 0 & \text{if } i \notin C\text{,} \end{cases}$$

where τ and $\widehat{\tau}$ are as defined in Proposition 3.

4.3 Three-agents example

Finally, let us consider an example to illustrate the results of this section. We use a utility model based on the reward function in Example 1 (i).

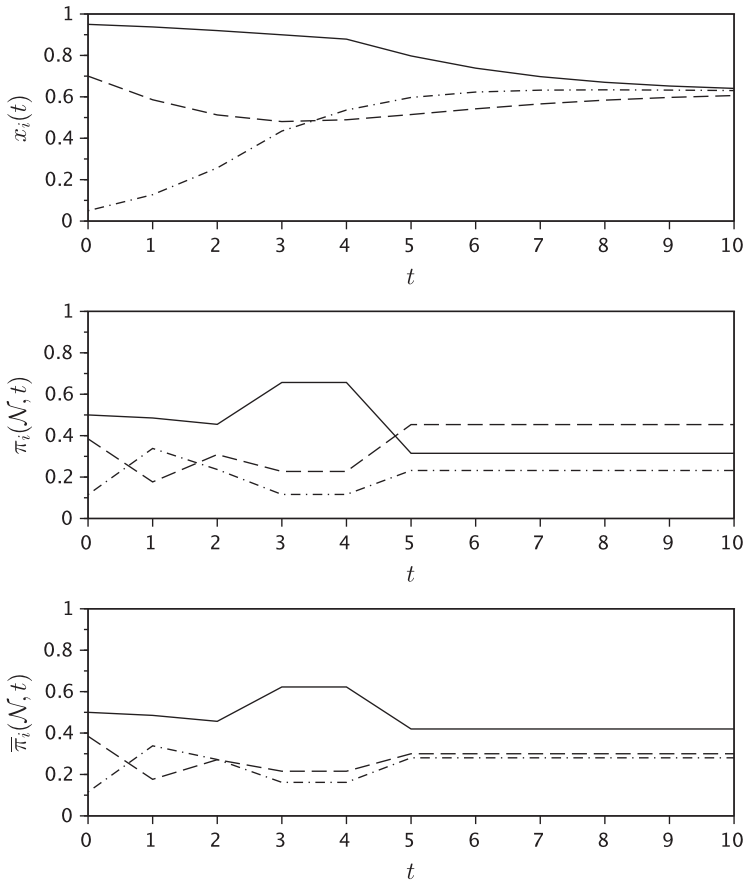


Fig. 2. Opinion dynamics (top), and the corresponding influence (middle), and overall influence (bottom) of the agents in Example 3. The solid line represents agent 1, the dashed line agent 2, and the dashed-dotted line agent 3.

Example 3 (*Three-agents society*)
Consider $\mathcal{N} = \{1, 2, 3\}$ and

$$u_i\big(\alpha \mid x(t), M(t), j\big) = -\frac{1}{2}\sum_{k \neq i}\big(x_i(t) - x_k(t+1)\big)^2 - \frac{\alpha}{20}$$

for all $i \in \mathcal{N}$. Furthermore, the initial opinions and trust matrix are

$$x(0) = \begin{pmatrix} 0.95 \\ 0.7 \\ 0.05 \end{pmatrix} \text{ and } M(0) = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.05 & 0.92 & 0.03 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}$$

and $\alpha_{ij} = 0.2$ is the ability of i to manipulate j, for all $i \neq j$. Notice that the agents form one minimal closed group. The vector of initial influence and overall influence, which is also the vector of influence in the DeGroot model without manipulation (denoted by π^{DG}), is $\pi(\mathcal{N}; 0) = \overline{\pi}(\mathcal{N}; 0) = \pi^{\mathrm{DG}} \simeq (0.5, 0.385, 0.115)$. Thus, without manipulation, the agents would reach the consensus $x^{\mathrm{DG}}(\infty) = \pi^{\mathrm{DG}} \cdot x(0) \simeq 0.75$. The influence weights and hence the consensus the agents reach might change significantly with manipulation. Figure 2 depicts one simulation run of the long-run

opinion dynamics and the corresponding influence and overall influence of the agents, where

$$\Gamma(t) = \begin{cases} (3,2;0.2), & t=0 \\ (2,3;0.2), & t=1 \\ (1,3;0.2), & t=2 \\ (2,1;0.2), & t=4 \\ (\cdot,\cdot;0), & \text{otherwise} \end{cases}.$$

Notice that although the last manipulation makes agent 2 the most influential agent from time 5 on, agent 1 stays the most influential agent in terms of the initial opinions. By the time agent 2 manipulates agent 1, the initial opinion of the latter agent has already spread in the network. This reflects the intuition that manipulation has the most bite in the beginning, before potentially misleading opinions (in this case agent 1's opinion) have spread. This results in the following final trust matrix:

$$M(t) \simeq \begin{pmatrix} 0.792 & 0.208 & 0 \\ 0.042 & 0.767 & 0.192 \\ 0.201 & 0.174 & 0.625 \end{pmatrix}, t \geqslant 5.$$

The agents reach the consensus $x(\infty) \simeq 0.623$, which reflects the gain of overall influence of agent 3 (and the loss of the other agents) as the consensus value is closer to agent 3's initial opinion compared to the case without manipulation.

5 Spread of (mis)information

In this section, we investigate how manipulation affects the aggregation of information and the spread of misinformation in society. First, we study conditions under which a single manipulation is beneficial for information aggregation. Second, we rely on comparative simulations to obtain insights on the long-run consequences of manipulation on information aggregation and the spread of misinformation.

We assume that the society forms one minimal closed group as segregated societies clearly fail to aggregate dispersed information.[21] We use an approach similar to Acemoglu et al. (2010) and assume that there is an *underlying state* $\mu = 1/n \cdot \sum_{i \in \mathcal{N}} x_i(0)$ that corresponds to the average of the initial opinions of the n agents. Information about the underlying state is dispersed, but can easily be aggregated by the agents: uniform overall influence on the long-run beliefs leads to perfect aggregation of information.[22] We measure the *spread of misinformation* in society at time t by the gap between the consensus reached conditional on that no more manipulation takes place and the underlying state:

$$\overline{\pi}(\mathcal{N};t)x(0) - \mu = \sum_{i \in \mathcal{N}} \left(\overline{\pi}_i(\mathcal{N};t) - \frac{1}{n} \right) x_i(0).$$

Thus, we can quantify the extent of the spread of misinformation by looking at the deviations of the overall influences from the average influence in society.

[21] Nevertheless, manipulation that leads to a connected society as in Example 2 can be interpreted as improving information aggregation.

[22] We could also think of the initial opinions as being drawn independently from some distribution with mean μ. Then, uniform overall influence would lead as well to optimal aggregation, the difference being that it would not be perfect due to the finite number of samples drawn by the agents. Furthermore, notice that the agents do not have any objectives regarding the underlying state.

Definition 1 (*Extent of misinformation*)

i. The *extent of misinformation* in society at time t is

$$\|\overline{\pi}(\mathcal{N};t) - 1/n \cdot \mathbb{I}\|_2,$$

where $\mathbb{I} = (1,1,\ldots,1) \in \mathbb{R}^n$ is a vector of 1s and $\|x\|_2 = \sqrt{\sum_{k\in\mathcal{N}} |x_k|^2}$ the standard Euclidean norm of $x \in \mathbb{R}^n$,

ii. manipulation at time t *decreases* (*increases*) the extent of misinformation in society if

$$\|\overline{\pi}(\mathcal{N};t+1) - 1/n \cdot \mathbb{I}\|_2 < (>)\|\overline{\pi}(\mathcal{N};t) - 1/n \cdot \mathbb{I}\|_2, \text{ and}$$

iii. manipulation at time t *strongly decreases* (*strongly increases*) the extent of misinformation in society if

$$|\overline{\pi}_i(\mathcal{N};t+1) - 1/n| < (>)|\overline{\pi}_i(\mathcal{N};t) - 1/n| \text{ for all } i \in \mathcal{N}.$$

Notice that the above measure is also useful in other situations. If the issue at stake is, for instance, normative, we can think of the extent of misinformation as a measure for how egalitarian the society is in terms of influence; in this sense, the underlying state reflects the consensus a fully egalitarian society would attain. The next result provides conditions under which manipulation strongly decreases the extent of misinformation in society. We show that first, the ability of the manipulating agent should be limited and second, agents with excess overall influence ($\overline{\pi}_k(\mathcal{N};t) > 1/n$) should lose and others gain overall influence.

Proposition 4

Suppose that $\overline{\pi}_k(\mathcal{N};t) \neq 1/n$ for all $k \in \mathcal{N}$. Then, there exists $\overline{\alpha} > 0$ such that the manipulation $\Gamma(t) = (i,j;\alpha_{ij})$, $\alpha_{ij} > 0$, strongly decreases the extent of misinformation if

i. $\alpha_{ij} \leqslant \overline{\alpha}$, and

ii. $\sum_{l=1}^{n} \overline{m}_{lk}(t)\big(\pi_l(\mathcal{N};t+1) - \pi_l(\mathcal{N};t)\big) < (>)\ 0$ for $k \in \mathcal{N}$ such that $\overline{\pi}_k(\mathcal{N};t) > (<)\ 1/n$.

Intuitively, condition (ii) says that agents whose current overall influence is above the average influence in society ($\overline{\pi}_k(\mathcal{N};t) > 1/n$) should lose overall influence due to the manipulation and agents whose current overall influence is below the average influence ($\overline{\pi}_k(\mathcal{N};t) < 1/n$) should gain. Then, this leads to a strong reduction of the extent of misinformation if i's ability to manipulate j is below some threshold (condition (i)). Otherwise, manipulation makes some agents too influential, in particular the manipulating agent, and might lead to an increase of the extent of misinformation.

5.1 Comparative simulations

Example 3 indicates that when preferences for manipulation and abilities to manipulate are homogeneous, then manipulation might overall—i.e. in terms of the overall influences after the last manipulation has taken place—decrease the extent of misinformation.[23] In the following, we conduct comparative simulations to further

[23] Figure 2 depicts a simulation run where manipulation overall has strongly decreased the extent of misinformation.

investigate this observation. Doing comparative simulations is a risky business as the resulting opinion dynamics may highly depend on the order in which the agents manipulate in a particular simulation run and even more on the particular initial opinion vectors and trust matrices we use. We therefore consider many different combinations of initial opinion vectors and trust matrices and also take the average values over a large number of simulation runs to ensure that our results are robust and precise approximations of the expected impact of manipulation. We investigate environments with (1) homogeneous preferences and abilities, and (2) homogeneous preferences but heterogeneous abilities. We conduct the simulations for $\mathcal{N} = \{1, 2, 3\}$ and do $R = 100,000$ simulation runs for each combination of preferences and parameter values, which allows us to precisely approximate the expected final overall influences and the corresponding extent of misinformation. These quantities indicate the tendency of the impact of manipulation in each of the settings.

We start with homogeneous preferences and abilities for manipulation, i.e.

$$u_i\big(\alpha \mid x(t), M(t), j\big) = -\frac{1}{2}\sum_{k\neq i}\big(x_i(t) - x_k(t+1)\big)^2 - \chi_{\{\alpha\neq 0\}}(\alpha)\cdot\frac{1}{100}$$

for all $i \in \mathcal{N}$ and $\alpha_{ij} = 0.2$ for all $i \neq j$.[24] We consider all permutations of the initial opinion vectors

$$x(0) = \begin{pmatrix}0.9\\0.7\\0.1\end{pmatrix} \text{ and } x(0) = \begin{pmatrix}0.9\\0.5\\0.1\end{pmatrix}.^{25}$$

The initial trust matrix in the basic setting (1) is

$$M(0) = M^{(1)} = \begin{pmatrix}0.95 & 0.05 & 0\\0.05 & 0.92 & 0.03\\0.05 & 0.05 & 0.9\end{pmatrix}.$$

To test how the impact of manipulation depends on the initial trust matrix, we consider modifications of $M^{(1)}$. We increase one (or two) trust weight(s) $m_{ij}(0)$, $i \neq j$, by 0.1 (0.05 each), and decrease the corresponding weight $m_{ii}(0)$ accordingly. Thus, we run simulations with the initial trust matrix $M^{(1)}$ and the following modifications:

$$M^{(2)} = \begin{pmatrix}0.85 & 0.15 & 0\\0.05 & 0.92 & 0.03\\0.05 & 0.05 & 0.9\end{pmatrix}, M^{(3)} = \begin{pmatrix}0.85 & 0.05 & 0.1\\0.05 & 0.92 & 0.03\\0.05 & 0.05 & 0.9\end{pmatrix},$$

[24] $\chi_{\{\alpha\neq 0\}}(\alpha) = \begin{cases}1, & \text{if } \alpha \neq 0\\ 0, & \text{otherwise}\end{cases}$ denotes the characteristic function of the set $\{\alpha \neq 0\}$.

[25] Notice that only the absolute differences of each pair of opinions matter for the agents' preferences. Thus, we implicitly also cover permutations of the vector $\begin{pmatrix}0.9\\0.3\\0.1\end{pmatrix}$ as, e.g. the opinion dynamics with this vector of initial opinions is equivalent in terms of manipulations to the one with initial opinions $\begin{pmatrix}0.1\\0.7\\0.9\end{pmatrix}$. And we also cover permutations of vectors $\begin{pmatrix}0.9+h\\0.7+h\\0.1+h\end{pmatrix}, h \in \mathbb{R}$.

$$M^{(4)} = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.15 & 0.82 & 0.03 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}, M^{(5)} = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.05 & 0.82 & 0.13 \\ 0.05 & 0.05 & 0.9 \end{pmatrix},$$

$$M^{(6)} = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.05 & 0.92 & 0.03 \\ 0.15 & 0.05 & 0.8 \end{pmatrix}, M^{(7)} = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.05 & 0.92 & 0.03 \\ 0.05 & 0.15 & 0.8 \end{pmatrix},$$

$$M^{(8)} = \begin{pmatrix} 0.85 & 0.1 & 0.05 \\ 0.05 & 0.92 & 0.03 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}, M^{(9)} = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.1 & 0.82 & 0.08 \\ 0.05 & 0.05 & 0.9 \end{pmatrix},$$

$$M^{(10)} = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.05 & 0.92 & 0.03 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}, M^{(11)} = \begin{pmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.87 & 0.08 \\ 0.05 & 0.05 & 0.9 \end{pmatrix},$$

$$M^{(12)} = \begin{pmatrix} 0.9 & 0.1 & 0 \\ 0.05 & 0.92 & 0.03 \\ 0.05 & 0.1 & 0.85 \end{pmatrix}, M^{(13)} = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.1 & 0.87 & 0.03 \\ 0.1 & 0.05 & 0.85 \end{pmatrix}.$$

For a given combination of initial opinion vector and trust matrix defined above, the extent of misinformation corresponding to the mean final overall influences is given by $\|\overline{\pi}^{*,\text{mean}} - 1/3 \cdot \mathbb{1}\|_2$, where $\overline{\pi}_i^{*,\text{mean}} = 1/R \cdot \sum_{r=1}^{R} \overline{\pi}_i^{*,r}$ denotes the mean final overall influence of agent i and $\overline{\pi}_i^{*,r}$ the final overall influence of agent i in run r, i.e. the overall influence of agent i in run r after the last manipulation has taken place. The results of the simulations in terms of the extents of misinformation corresponding to the mean final overall influences are presented in Table 1.

The results show that, in expectation, manipulation decreases the final extent of misinformation for all combinations of initial opinions and trust matrices; and moreover, the decrease is strong for most combinations.[26] In particular, it is striking that this is even the case in settings (3) and (11), where the extent of misinformation is already rather small ex-ante. We have also conducted a robustness check with the reward function from Example 1 (iii) to see whether these results depend on the fact that the agents need to know the social network. The results were almost identical,[27] which confirms that manipulation, in expectation, decreases the extent of misinformation when preferences and abilities for manipulation are homogeneous. And in particular, this even holds when agents only take into account their ability and current opinions for their decisions and do not know the social network.

The intuition for this finding is that given a certain ability to manipulate, the benefit from manipulation depends mainly on the influence of the manipulated agent. If the latter is influential, then the manipulator benefits a lot and gains substantially in influence, while she does not benefit much otherwise. The reason is that manipulating an influential agent makes the manipulator indirectly more influential on many other agents. Furthermore, notice that the higher an agent's influence, the lower the average influence of the other agents in the society. Therefore,

[26] Notice that $\overline{\pi}_1^{(11)}(\mathcal{N};0) = \overline{\pi}_1^{(12)}(\mathcal{N};0) = 1/3$ and thus a strong decrease of the extent of misinformation is not possible in settings (11) and (12).

[27] The final extent of misinformation decreased for all combinations of initial opinions and trust matrices, and strongly decreased for most combinations.

Table 1. Simulation results with homogeneous preferences and abilities in terms of the extents of misinformation corresponding to the mean final overall influences. * indicates a decrease and ** a strong decrease compared to the DeGroot model without manipulation.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
0.9, 0.7, 0.1	0.11**	0.19**	0.07**	0.24**	0.17*	0.23**	0.2**	0.12**	0.18**	0.21**	0.06*	0.19*	0.25*
0.9, 0.5, 0.1	0.1**	0.16**	0.06**	0.19**	0.11**	0.21**	0.2*	0.1**	0.14**	0.2**	0.05*	0.18*	0.2**
0.1, 0.7, 0.9	0.11**	0.2**	0.06**	0.18**	0.12**	0.23*	0.24*	0.12**	0.14**	0.23*	0.07*	0.23*	0.21**
0.9, 0.1, 0.7	0.2**	0.16**	0.08*	0.34**	0.14**	0.3**	0.28**	0.09**	0.23**	0.28**	0.05*	0.21*	0.32**
0.9, 0.1, 0.5	0.13**	0.18**	0.06**	0.23**	0.1**	0.24**	0.22**	0.1**	0.15**	0.22**	0.04*	0.2*	0.24**
0.1, 0.9, 0.7	0.14*	0.23**	0.07**	0.2**	0.11**	0.26*	0.25*	0.13**	0.14**	0.25*	0.06*	0.25*	0.23**
0.7, 0.9, 0.1	0.13**	0.19*	0.08**	0.26**	0.14**	0.25**	0.21**	0.13**	0.2**	0.22**	0.04*	0.19*	0.26*
0.5, 0.9, 0.1	0.12**	0.14**	0.07**	0.24**	0.13**	0.23**	0.21**	0.1**	0.17**	0.22**	0.04*	0.17*	0.26**
0.7, 0.1, 0.9	0.18**	0.14*	0.09*	0.32**	0.14**	0.3**	0.26**	0.1**	0.23**	0.28**	0.05*	0.16*	0.32**
DeGroot	0.28	0.3	0.1	0.48	0.21	0.36	0.33	0.18	0.34	0.34	0.08	0.31	0.44

Table 2. Simulation results with homogeneous preferences and heterogeneous abilities in terms of the mean final overall influences and the corresponding extents of misinformation.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	DeGroot
$\overline{\pi}_1^{*,\text{mean}}$	0.337	0.375	0.374	0.41	0.438	0.264	0.299	0.287	1/3
$\overline{\pi}_2^{*,\text{mean}}$	0.325	0.263	0.362	0.291	0.282	0.37	0.402	0.425	1/3
$\overline{\pi}_3^{*,\text{mean}}$	0.338	0.362	0.264	0.299	0.281	0.365	0.299	0.288	1/3
$\lVert\overline{\pi}^{*,\text{mean}} - 1/3 \cdot \mathbb{1}\rVert_2$	0.01	0.087	0.085	0.094	0.128	0.084	0.084	0.112	0

in expectation, agents benefit more from manipulation the lower their own influence and thus manipulation decreases the extent of misinformation.

Second, we conduct simulations with heterogeneous abilities for manipulation. Therefore, we take an otherwise symmetric environment where preferences are

$$u_i\big(\alpha \mid x(t), M(t), j\big) = -\frac{1}{2}\sum_{k\neq i}\big(x_i(t) - x_k(t+1)\big)^2 - \chi_{\{\alpha\neq 0\}}(\alpha)\cdot\frac{1}{100}$$

for all $i \in \mathcal{N}$ and initial opinions and initial trust matrix are given by

$$x(0) = \begin{pmatrix}0.9\\0.5\\0.1\end{pmatrix} \text{ and } M(0) = \begin{pmatrix}0.9 & 0.05 & 0.05\\0.05 & 0.9 & 0.05\\0.05 & 0.05 & 0.9\end{pmatrix}.$$

The vector of initial overall influence is $\overline{\pi}(\mathcal{N};0) = (1/3, 1/3, 1/3)$. Notice that the agents are not perfectly symmetric in the distance between their initial opinions and thus we can only treat agents 1 and 3 as equivalent. The abilities to manipulate in the basic setting (1) are given by $\alpha_{ij}^{(1)} = 0.2$ for all $i \neq j$. To test how the impact of manipulation depends on the abilities, we increase first one ability of agent 1 by 0.4, and then both abilities by 0.4 as well as by 1; and the same is done for agent 2. Thus, we run simulations with the abilities $\alpha_{ij}^{(1)}$ and the following modifications (abilities that do not change are omitted):[28]

$$\alpha_{12}^{(2)} = 0.6; \alpha_{13}^{(3)} = 0.6; \alpha_{12}^{(4)} = \alpha_{13}^{(4)} = 0.6; \alpha_{12}^{(5)} = \alpha_{13}^{(5)} = 1.2;$$
$$\alpha_{21}^{(6)} = 0.6; \alpha_{21}^{(7)} = \alpha_{23}^{(7)} = 0.6; \alpha_{21}^{(8)} = \alpha_{23}^{(8)} = 1.2.$$

The results of the simulations in terms of the mean final overall influences are presented in Table 2.[29]

Settings (2) and (3) show that increasing one ability by 0.4 already benefits agent 1 significantly in expectation, while the agent more susceptible to manipulation loses influence. Notice that the agent not concerned by the change in ability also significantly gains influence in expectation. Increasing both abilities by 0.4 (setting (4)) and by 1 (setting (5)) further increases the gain of influence of agent 1, while the other agents lose equally. Moreover, we also find that the extent of misinformation significantly increases in expectation when some agent gets more ability to manipulate. The increase of both abilities from 0.2 to 1.2 has increased

[28] We omit the case $\alpha_{23}^{(6')} = 0.6$ as it is equivalent to $\alpha_{21}^{(6)} = 0.6$.

[29] Notice that the different mean final overall influences in the basic setting (1) between agent 1 (agent 3) and agent 2 are due to their asymmetry with respect to the initial opinions; and the slight difference between agent 1 and agent 3 reflects the approximation error of our simulations.

agent 1's mean final overall influence by around 30% and furthermore, it has increased more than tenfold the extent of misinformation corresponding to the mean final overall influences. The tests with agent 2 (settings (6)–(8)) confirm these observations, the only difference being that the increase of influence is slightly lower due to her weaker position in terms of initial opinions. Thus, our results confirm the intuition that agents that are more powerful than others in terms of the abilities to manipulate, e.g. lobbyists or lobby organizations, might severely harm the aggregation of information and spread their (potentially) misleading information.

Here, the tendency of the average benefit from manipulation to decrease with the agent's own influence is dominated by the effect coming from the asymmetry in abilities. Agents with higher ability to manipulate benefit more from manipulation and hence, in expectation, increase their influence even if they were already rather influential before.

In total, our simulations suggest that manipulation is beneficial to information aggregation when preferences and abilities for manipulation are homogeneous, but detrimental in case abilities are concentrated at few powerful agents.

6 Conclusion

This paper investigates a model of opinion formation where agents are subject to persuasion bias, i.e. fail to adjust for possible repetitions of information they receive and instead treat all information as new, and can exert effort to manipulate trust in the opinions of others in their favor. Our analysis is motivated by various examples where it is desirable for individuals that others hold similar beliefs as this would make them taking similar decisions.

Agents hold opinions about some issue of common interest, are organized in a weighted social network and repeatedly meet and communicate with their neighbors in the network. At each time instance, first a pair of agents is selected by a stochastic process such that one of them has the opportunity to manipulate the other, which is modeled as changing the persuasion bias of the manipulated agent in favor of the manipulator. Second, all agents communicate with their neighbors and update their opinions according to their (possibly manipulated) social trust.

We first study segregated societies and show that manipulation can connect segregated groups and thus make them reaching a consensus. In particular, an agent outside these groups can serve as a mediator by manipulating members of both groups. Second, we investigate the long-run effects of manipulation and find that it fosters opinion leadership; and surprisingly agents with low trust in their own opinion might get more influential even by being manipulated. Finally, we investigate the tension between information aggregation and spread of misinformation. Our comparative simulations reveal that manipulation is beneficial to information aggregation when preferences and abilities for manipulation are homogeneous. This result turns out to be very robust and even holds when agents manipulate without knowing the social network. However, the simulations also show that manipulation is detrimental to information aggregation in case abilities are concentrated at few powerful agents; such agents, e.g. lobby organizations, might severely harm information aggregation and spread their (potentially) misleading information.

We should notice at this point that manipulation how we model it has no bite in large societies as studied by Golub & Jackson (2010). They study social learning under persuasion bias and find that all opinions in a large society converge to the truth if and only if the influence of the most influential agent vanishes as the society grows. In large societies, manipulation does not change convergence to the underlying state since its consequences are negligible compared to the size of society. However, our intuition is that an agent being able to manipulate a substantial proportion of the society could spread misinformation also in large societies.

We view our paper as a first attempt in studying manipulation and spread of (mis)information in social networks. Our approach incorporates boundedly rational decision making in a model where agents are subject to persuasion bias. We made several simplifying assumptions and derived results that apply to general societies. One line of future investigation is to relax the restriction to manipulation of a single agent. Extending manipulation to groups is interesting in particular in the context of large societies as studied by Golub & Jackson (2010). Second, it is important to study manipulation when agents are, at least partly, Bayesian. While our boundedly rational approach is a natural starting point as the notion of manipulation is more difficult to introduce into a Bayesian setup, it would be interesting to study manipulation with rational agents.

Acknowledgments

We thank the editor, two anonymous referees, Michel Grabisch, Jean-Jacques Herings, Dunia López-Pintado, Agnieszka Rusinowska and numerous seminar participants for helpful comments. Most of the work was carried out while Manuel Förster was affiliated to CES, Université Paris 1 Panthéon-Sorbonne, France, and CORE, University of Louvain, Louvain-la-Neuve, Belgium. Vincent Vannetelbosch and Ana Mauleon are Senior Research Associates of the National Fund for Scientific Research (FNRS). Financial support from the Doctoral Program EDE-EM (European Doctorate in Economics - Erasmus Mundus) of the European Commission, from the Spanish Ministry of Economy and Competition under the project ECO2012-35820, from the Fonds de la Recherche Scientifique – FNRS research grant J.007315 and from the Belgian French speaking community ARC project 15/20-072 of Saint-Louis University—Brussels are gratefully acknowledged.

References

Acemoglu, D., Dahleh, M., Lobel, I., & Ozdaglar, A. (2011). Bayesian learning in social networks. *The Review of Economic Studies*, **78**(4), 1201–36.

Acemoglu, D., & Ozdaglar, A. (2011). Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, **1**(1), 3–49.

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

Austen-Smith, D., & Wright, J. R. (1994). Counteractive lobbying. *American Journal of Political Science*, **38**(1), 25–44.

Axelrod, R. (1997). The dissemination of culture a model with local convergence and global polarization. *Journal of Conflict Resolution*, **41**(2), 203–26.

Buechel, B., Hellmann, T., & Klößner, S. (2015). Opinion dynamics and wisdom under conformity. *Journal of Economic Dynamics and Control*, **52**, 240–57.

Chandrasekhar, A., Larreguy, H., & Xandri, J. (2012). Testing models of social learning on networks: Evidence from a framed field experiment. Mimeo, Massachusetts Institute of Technology.

Choi, S., Gale, D., & Kariv, S. (2012). Social learning in networks: A quantal response equilibrium analysis of experimental data. *Review of Economic Design*, **16**(2–3), 135–57.

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, **3**(01n04), 87–98.

DeGroot, M. (1974). Reaching a consensus. *Journal of the American Statistical Association*, **69**(345), 118–21.

DeMarzo, P., Vayanos, D., & Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *Quarterly Journal of Economics*, **118**(3), 909–68.

Esteban, J., & Ray, D. (2006). Inequality, lobbying, and resource allocation. *The American Economic Review*, **96**(1), 257–79.

Friedkin, N. E. (1991). Theoretical foundations for centrality measures. *American journal of Sociology*, **96**(6), 1478–504.

Friedkin, N. E., & Johnsen, E. C. (1990). Social influence and opinions. *Journal of Mathematical Sociology*, **15**(3–4), 193–206.

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

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence – models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, **5**(3), 1–33.

Hunter, J. J. (2005). Stationary distributions and mean first passage times of perturbed Markov chains. *Linear Algebra and its Applications*, **410**, 217–43.

Potters, J., & Van Winden, F. (1992). Lobbying and asymmetric information. *Public choice*, **74**(3), 269–92.

Watts, A. (2014). The influence of social networks and homophily on correct voting. *Network Science*, **2**(01), 90–106.

A Appendix

A.1 Proof of Proposition 1

i. Follows immediately since all minimal closed groups remain unchanged.

ii. If agent i manipulates agent j, then $m_{ji}(t+1) > 0$ and thus, since $C' \ni j$ is minimal closed at time t, there exists a path at $t+1$ from l to i for all $l \in C'$. Since C is still minimal closed, it follows that $R(t+1) = R(t) \cup C'$, i.e. $\mathscr{C}(t+1) = \mathscr{C}(t) \backslash \{C'\}$.

iii. a. If agent i manipulates agent j, then it follows that $\sum_{l \in C \cup \{i\}} m_{kl}(t+1) = 1$ for all $k \in C$ since C is closed at time t. Furthermore, since by assumption there is no path from i to k for any $k \in \cup_{C' \in \mathscr{C}(t) \backslash \{C\}} C'$ and by definition of R', $\sum_{l \in C \cup R' \cup \{i\}} m_{kl}(t+1) = 1$ for all $k \in R' \cup \{i\}$. Hence, it follows that $\sum_{l \in C \cup R' \cup \{i\}} m_{kl}(t+1) = 1$ for all $k \in C \cup R' \cup \{i\}$, i.e. $C \cup R' \cup \{i\}$ is closed. Moreover, since by assumption there is no path from i to k for any $k \in \cup_{C' \in \mathscr{C}(t) \backslash \{C\}} C'$, there is a path from i to j (otherwise $R' \cup \{i\}$ was closed at t). Thus, since C is minimal closed and i manipulates j, there is a path from k to l for all $k, l \in C \cup \{i\}$ at $t+1$. Then, by definition of R', there is also a path from k to l for all $k \in C \cup \{i\}$ and $l \in R'$. Moreover, again by assumption and definition of R', there exists a path from k to l for all $k \in R'$ and all $l \in C$ (otherwise a subset of R' was closed at t).

Combined, this implies that the same holds for all $k, l \in C \cup R' \cup \{i\}$. Hence, $C \cup R' \cup \{i\}$ is minimal closed, i.e. $\mathscr{C}(t+1) = \mathscr{C}(t) \backslash \{C\} \cup \{C \cup R' \cup \{i\}\}$.

b. If agent i manipulates agent j, then $m_{ji}(t+1) > 0$ and thus, since $C \ni j$ is minimal closed at time t, there exists a path at $t+1$ from l to i for all $l \in C$. Hence, by assumption there exists a path from agent j to k, but not vice versa since $C' \ni k$ is minimal closed. Thus, $R(t+1) = R(t) \cup C$, which finishes the proof.

A.2 Proof of Corollary 1

First, it follows from Remarks 3 and 4 in Appendix B that for all $k \in C \setminus \{i, j\}$,

$$\begin{aligned} r_{kj}(t) = 1 + \sum_{l \in C \setminus \{j\}} m_{kl}(t) r_{lj}(t) \geqslant 1 + \sum_{l \in C \setminus \{j\}} m_{kl}(t) &= 2 - m_{kj}(t) \\ &\geqslant 2 - \max_{u \in C \setminus \{i,j\}} m_{uj}(t) \\ &= 2 - \eta > 1 \end{aligned}$$

since $m_{uu}(t) > 0$ for all $u \in C \setminus \{i, j\}$. And as $m_{ij}(t) \to 1$ implies $r_{ij}(t) \to 1$, there exists $q(\eta) < 1$ such that $r_{ij}(t) \leqslant (3-\eta)/2$ if $m_{ij}(t) > q(\eta)$.

Next, suppose that $m_{ij}(t) > q(\eta)$ and $m_{jj}(t) < (1-\eta)/(3-\eta) \cdot \sum_{k \in C \setminus \{i,j\}} m_{jk}(t)$. It follows that

$$\begin{aligned} \sum_{k \in C \setminus \{i,j\}} m_{jk}(t) \underbrace{\big(r_{kj}(t) - r_{ij}(t)\big)}_{\geqslant 2-\eta-(3-\eta)/2=(1-\eta)/2} &\geqslant \frac{1-\eta}{2} \sum_{k \in C \setminus \{i,j\}} m_{jk}(t) > \frac{3-\eta}{2} m_{jj}(t) \\ &\geqslant m_{jj}(t) r_{ij}(t), \end{aligned}$$

which by Proposition 2 part (ii) (the extended version in Appendix B) implies that agent j gains influence and thus finishes the proof.

A.3 Proof of Proposition 3

Suppose that the sequence $(\tau_k)_{k=1}^{\infty}$ of stopping times denotes the time instances where the trust structure changes, i.e. under the event $\{\tau_k = t\}$ the trust structure changes the kth time at time t. Notice that the event $\{\tau_k = +\infty\}$ means that the kth change never happens. By Proposition 1, it follows that under the event $\{\tau_k = t < +\infty\}$, either

a. $1 \leqslant |\mathscr{C}(t+1)| < |\mathscr{C}(t)|$ and $|R(t+1)| > |R(t)|$, or
b. $|\mathscr{C}(t+1)| = |\mathscr{C}(t)|$ and $0 \leqslant |R(t+1)| < |R(t)|$

holds. This implies that the maximal number of changes in the trust structure is finite, i.e. there exists a positive integer K such that there are at most K changes in the structure and thus, almost surely $\tau_{K+1} = +\infty$. Hence, $\tau := \max\{\tau_k + 1 \mid \tau_k < +\infty\}$, where $\tau_0 \equiv 0$, is the desired almost surely finite stopping time, which finishes part (i). Part (ii) follows from the following lemma.

Lemma 1
Suppose that $\mathscr{C}(0) = \{\mathscr{N}\}$. If for all $i \in \mathscr{N}$ either (A1) or (A2) holds, then there exists an almost surely finite stopping time τ such that under the event $\{\tau = t\}$ no more manipulation takes place from time t on. The society converges to the random

variable

$$x(\infty) = \pi(\mathcal{N};\tau)\overline{M}(\tau-1)x(0).^{30}$$

Proof

By Proposition 1, we know that $\mathscr{C}(t) = \{\mathcal{N}\}$ for all $t \geqslant 0$. First, we show that the opinions converge almost surely to a consensus $x(\infty)$. Therefore, suppose to the contrary that the opinions do not converge almost surely, i.e. there exists a value $d > 0$ such that $\max_{i,j\in\mathcal{N}}|x_i(t) - x_j(t)| \to d$ for $t \to \infty$ with positive probability. In this case, there exists $t' > 0$ such that $d \leqslant \max_{i,j\in\mathcal{N}}|x_i(t) - x_j(t)| < d + \varepsilon$ for all $t \geqslant t'$ and some $\varepsilon > 0$. Furthermore, this implies that there exists a periodic trust matrix $M^* \in \mathbb{R}^{n\times n}$ such that $M(t) \to M^*$ for $t \to \infty$.[31] In particular, there are without loss of generality $2 \leqslant k \leqslant n$ agents $i_1, i_2, \ldots, i_k$ such that $m^*_{i_l i_{l+1}} = 1$ for all $l = 1, 2, \ldots, k$, with $k + 1 \equiv 1$, and

$$d \leqslant \max_{l,l'\in\{1,2,\ldots,k\}}|x_{i_l}(t) - x_{i_{l'}}(t)| < d + \varepsilon$$

for all $t \geqslant t'$. Next, take any agent i_l. As $|m_{i_l i_{l+1}}(t') - m^*_{i_l i_{l+1}}| \geqslant m_{i_l i_l}(t') > 0$, there exists $0 < \delta(m_{i_l}(t')) < 1$ such that if $m_{i_l}(t') = m_{i_l}(t' + k)$, i.e. i_l is not manipulated during k consecutive time instances, then

$$\max_{l',l''}|x_{i_{l'}}(t' + k) - x_{i_{l''}}(t' + k)| \leqslant \delta(m_{i_l}(t')) \cdot \max_{l',l''}|x_{i_{l'}}(t') - x_{i_{l''}}(t')|,$$

which implies

$$\max_{l',l''}|x_{i_{l'}}(t' + k) - x_{i_{l''}}(t' + k)| < \delta(m_{i_l}(t')) \cdot (d + \varepsilon).^{32}$$

Moreover, there exists an integer $r^* > 0$ such that $\delta(m_{i_l}(t'))^{r^*} \cdot (d + \varepsilon) < d$, which implies that

$$\max_{l',l''}|x_{i_{l'}}(t' + r^* \cdot k) - x_{i_{l''}}(t' + r^* \cdot k)| < d$$

if i_l is not manipulated during $r^* \cdot k$ consecutive time instances. Hence, as the same argument holds for all $t'' > t'$, it follows that there exists an almost surely finite stopping time τ' such that under the event $\{\tau' = t\}$ $\max_{i,j\in\mathcal{N}}|x_i(t) - x_j(t)| < d$, which is a contradiction.

Having established the convergence of opinions, it follows immediately that $\|[M(t)](i,j;\alpha_{ij})x(t) - [M(t)](i,j;0)x(t)\| \to 0$ and $|x_i(t) - x_j(t)| \to 0$, for $t \to \infty$ and any pair of agents (i,j). Hence, as by assumption either (A1) or (A2) holds, we get $v_i\big(\alpha_{ij} \mid x(t), M(t), j\big) - v_i\big(0 \mid x(t), M(t), j\big) \to 0 < c_i(\alpha_{ij} \mid j)$ for $t \to \infty$ and any pair of agents (i,j), which shows that there exists an almost surely finite stopping time τ such that under the event $\{\tau = t\}$, there is no more manipulation from time

[30] We define $\overline{M}(t) := I_n$ for $t = 0, -1$, where I_n is the $n \times n$ identity matrix.

[31] The reason is that as the agents form a minimal closed group (which will not change due to manipulation), they will fail to converge only in case the trust matrix converges to a periodic matrix. Notice also that $m_{ii}(0) > 0$ implies $m_{ii}(t) > 0$ and thus $M(t) \neq M^*$ for all $t \geqslant 0$.

[32] Notice that $\delta(m_{i_l}(t'))$ can be calculated by using a theoretical benchmark where all trust vectors except that of agent i_l are equal to those in the periodic trust matrix and where no manipulation takes place: set $\tilde{x}(t') := x(t')$, $\widetilde{m_{i_l}} := m_{i_l}(t')$, $\widetilde{m_{i_{l'}}} := m^*_{i_{l'}}$ for all $l' \neq l$ and $\tilde{x}(t' + k) := \widetilde{M}^k\tilde{x}(t')$. Then, $\delta(m_{i_l}(t')) := \max_{l',l''}|\tilde{x}_{i_{l'}}(t'+k) - \tilde{x}_{i_{l''}}(t'+k)| / \max_{l',l''}|\tilde{x}_{i_{l'}}(t') - \tilde{x}_{i_{l''}}(t')| < 1$ as $|\widetilde{m_{i_l i_{l+1}}} - m^*_{i_l i_{l+1}}| \geqslant \widetilde{m_{i_l i_l}} > 0$.

t on. Furthermore, agents reach a stochastic consensus that can be written as

$$\begin{aligned} x(\infty) = \pi(\mathcal{N};\tau)x(\tau) &= \pi(\mathcal{N};\tau)M(\tau)x(\tau-1) \\ &= \pi(\mathcal{N};\tau)M(\tau-1)M(\tau-2)\cdots M(1)x(0) \\ &= \pi(\mathcal{N};\tau)\overline{M}(\tau-1)x(0), \end{aligned}$$

where the second equality follows from the fact that $\pi(\mathcal{N};\tau)$ is a left eigenvector of $M(\tau)$ corresponding to eigenvalue 1, which finishes the proof. □

Finally, notice that the restriction to C of the matrices $M(\cdot)$ in the computation of the consensus belief is possible as C is minimal closed and thus $M(\cdot)|_C$ are stochastic matrices, which finishes the proof.

A.4 Proof of Proposition 4

First, we derive a formula for the absolute change of the overall influence weights due to manipulation.

Lemma 2
For $k \in \mathcal{N}$,

$$\overline{\pi}_k(\mathcal{N};t+1) - \overline{\pi}_k(\mathcal{N};t) = \sum_{l=1}^{n} \overline{m}_{lk}(t)\big(\pi_l(\mathcal{N};t+1) - \pi_l(\mathcal{N};t)\big).$$

Proof
It follows from Corollary 2 that

$$\begin{aligned} \overline{\pi}_k(\mathcal{N};t+1) &= \sum_{l=1}^{n} \overline{m}_{lk}(t)\pi_l(\mathcal{N};t+1) \\ &= \sum_{l=1}^{n} \overline{m}_{lk}(t)\big(\pi_l(\mathcal{N};t+1) - \pi_l(\mathcal{N};t)\big) + \sum_{l=1}^{n} \overline{m}_{lk}(t)\pi_l(\mathcal{N};t) \\ &= \sum_{l=1}^{n} \overline{m}_{lk}(t)\big(\pi_l(\mathcal{N};t+1) - \pi_l(\mathcal{N};t)\big) + \underbrace{\sum_{l=1}^{n} \overline{m}_{lk}(t-1)\pi_l(\mathcal{N};t)}_{=\overline{\pi}_k(\mathcal{N};t)}, \end{aligned}$$

where the last equality follows since $\pi(\mathcal{N};t)$ is a left eigenvector of $M(t)$, which finishes the proof. □

Next, let $N_* := \{k \in \mathcal{N} \mid \overline{\pi}_k(\mathcal{N};t) < 1/n\}$ and $N^* := \{k \in \mathcal{N} \mid \overline{\pi}_k(\mathcal{N};t) > 1/n\}$. Notice that $N_*, N^* \neq \emptyset$. By Lemma 3 in Appendix B, we have $\pi_k(\mathcal{N};t+1) - \pi_k(\mathcal{N};t) \to 0$ for $\alpha_{ij} \to 0$ and all $k \in \mathcal{N}$ and thus, by Lemma 2,

$$\overline{\pi}_k(\mathcal{N};t+1) - \overline{\pi}_k(\mathcal{N};t) \to 0 \text{ for } \alpha_{ij} \to 0 \text{ and all } k \in \mathcal{N}. \tag{A1}$$

Let $k \in N_*$, then by (ii) and Lemma 2, $\overline{\pi}_k(\mathcal{N};t+1) > \overline{\pi}_k(\mathcal{N};t)$. Hence, by Equation (A1), there exists $\overline{\alpha}(k) > 0$ such that

$$1/n \geqslant \overline{\pi}_k(\mathcal{N};t+1) > \overline{\pi}_k(\mathcal{N};t) \text{ for all } 0 < \alpha_{ij} \leqslant \overline{\alpha}(k).$$

Analogously, for $k \in N^*$, there exists $\overline{\alpha}(k) > 0$ such that

$$1/n \leqslant \overline{\pi}_k(\mathcal{N};t+1) < \overline{\pi}_k(\mathcal{N};t) \text{ for all } 0 < \alpha_{ij} \leqslant \overline{\alpha}(k).$$

Therefore, setting $\overline{\alpha} := \min_{k \in \mathcal{N}} \overline{\alpha}(k)$, we have

$$|\overline{\pi}_k(\mathcal{N}; t+1) - 1/n| < |\overline{\pi}_k(\mathcal{N}; t) - 1/n|$$

for all $k \in \mathcal{N}$ and $0 < \alpha_{ij} \leqslant \overline{\alpha}$, which finishes the proof.

B Appendix – Derivation of Proposition 2

The proof of Proposition 2 relies on a measure for how remotely agents are located from each other in the network, i.e. how directly agents trust other agents. This measure is known as the *mean first passage time* in Markov chain theory. Let $(X_s^{(t)})_{s=0}^{\infty}$ denote the homogeneous Markov chain induced by transition matrix $M(t)$. The agents are then interpreted as states of the Markov chain and the trust of i in j, $m_{ij}(t)$, is interpreted as the transition probability from state i to state j.

The *mean first passage time* from state i to state j is defined as $\mathbb{E}[\inf\{s \geqslant 0 \mid X_s^{(t)} = j\} \mid X_0^{(t)} = i]$. Given the current state of the Markov chain is i, the mean first passage time to j is the expected time it takes for the chain to reach state j. In other words, the mean first passage time from i to j corresponds to the average (expected) length of a random walk on the weighted network $M(t)$ from i to j that takes each link with probability equal to the assigned weight.[33] This average length is small if the weights along short paths from i to j are high, i.e. if agent i trusts agent j rather directly. We therefore call this measure *weighted remoteness* of j from i.

Definition 2 (*Weighted remoteness*)
Take $i, j \in \mathcal{N}$, $i \neq j$. The *weighted remoteness* at time t of agent j from agent i is given by

$$r_{ij}(t) = \mathbb{E}[\inf\{s \geqslant 0 \mid X_s^{(t)} = j\} \mid X_0^{(t)} = i],$$

where $(X_s^{(t)})_{s=0}^{\infty}$ is the homogeneous Markov chain induced by $M(t)$.

The following remark shows that the weighted remoteness attains its minimum when i trusts solely j.

Remark 3
Take $i, j \in \mathcal{N}$, $i \neq j$.

i. $r_{ij}(t) \geqslant 1$,
ii. $r_{ij}(t) < +\infty$ if and only if there is a path from i to j, and, in particular, if $i, j \in C \in \mathscr{C}(t)$, and
iii. $r_{ij}(t) = 1$ if and only if $m_{ij}(t) = 1$.

To provide some more intuition, let us look at an alternative (implicit) formula for the weighted remoteness. Suppose that $i, j \in C \in \mathscr{C}(t)$ are two distinct agents in a minimal closed group. By part (ii) of Remark 3, the weighted remoteness is finite for all pairs of agents in that group. The unique walk from i to j with length 1 is assigned weight (or has probability, when interpreted as a random walk) $m_{ij}(t)$. And the average length of walks to j that first pass through $k \in C \backslash \{j\}$ is $r_{kj}(t) + 1$, i.e.

[33] More precisely, it is a random walk on the state space $\mathcal{N}$ that, if currently in state k, travels to state l with probability $m_{kl}(t)$. The length of this random walk to j is the time it takes for it to reach state j.

walks from i to j with average length $r_{kj}(t)+1$ are assigned weight (have probability) $m_{ik}(t)$. Thus,

$$r_{ij}(t) = m_{ij}(t) \cdot 1 + \sum_{k \in C \setminus \{j\}} m_{ik}(t) \cdot (r_{kj}(t) + 1).$$

Finally, applying $\sum_{k \in C} m_{ik}(t) = 1$ leads to the following remark.

Remark 4
Take $i, j \in C \in \mathscr{C}(t)$, $i \neq j$. Then,

$$r_{ij}(t) = 1 + \sum_{k \in C \setminus \{j\}} m_{ik}(t) r_{kj}(t).$$

Note that computing the weighted remoteness using this formula amounts to solving a linear system of $|C|(|C|-1)$ equations, which has a unique solution. The following lemma provides a formula for the absolute change of the influence weights.

Lemma 3
Suppose that at time t, $\Gamma(t) = (i, j; \alpha)$, $i, j \in C$. Then, the influence of agent $k \in C$ on the final consensus of her group changes as follows:

$$\pi_k(C; t+1) - \pi_k(C; t) = \begin{cases} \alpha/(1+\alpha)\pi_i(C;t)\pi_j(C;t+1)\sum_{l \in C \setminus \{i\}} m_{jl}(t) r_{li}(t) & \text{if } k = i \\ \alpha/(1+\alpha)\pi_k(C;t)\pi_j(C;t+1)\left(\sum_{l \in C \setminus \{k\}} m_{jl}(t) r_{lk}(t) - r_{ik}(t)\right) & \text{if } k \neq i \end{cases}.$$

Proof
Suppose without loss of generality that $\mathscr{C}(t) = \{\mathscr{N}\}$. We can write

$$M(t+1) = M(t) + e_j z(t),$$

where e_j is the jth unit vector, and

$$z_k(t) = \begin{cases} \left(m_{ji}(t) + \alpha\right)/(1+\alpha) - m_{ji}(t) & \text{if } k = i \\ \left(m_{jk}(t)\right)/(1+\alpha) - m_{jk}(t) & \text{if } k \neq i \end{cases} = \begin{cases} \alpha(1 - m_{ji}(t))/(1+\alpha) & \text{if } k = i \\ -\alpha m_{jk}(t)/(1+\alpha) & \text{if } k \neq i \end{cases}.$$

From Hunter (2005), we get

$$\pi_k(\mathscr{N}; t+1) - \pi_k(\mathscr{N}; t) = -\pi_k(\mathscr{N}; t)\pi_j(\mathscr{N}; t+1) \sum_{l \neq k} z_l(t) r_{lk}(t)$$

$$= \begin{cases} \alpha/(1+\alpha)\,\pi_i(\mathscr{N};t)\pi_j(\mathscr{N};t+1)\sum_{l \neq i} m_{jl}(t) r_{li}(t) & \text{if } k = i \\ \alpha/(1+\alpha)\,\pi_k(\mathscr{N};t)\pi_j(\mathscr{N};t+1)\left(\sum_{l \neq k} m_{jl}(t) r_{lk}(t) - r_{ik}(t)\right) & \text{if } k \neq i \end{cases},$$

which finishes the proof. □

Notice that the magnitude of the change in long-run influence increases with i's ability to manipulate j, and it is zero if i does not manipulate. We are now in the position to prove the following result.

Proposition 2
Suppose that at time t, $\Gamma(t) = (i, j; \alpha)$, $i, j \in C$. If $\alpha = \alpha_{ij} > 0$, then

i. agent i strictly increases her long-run influence, $\pi_i(C; t+1) > \pi_i(C; t)$,
ii. any other agent $k \neq i$ of the group can either gain or lose influence, depending on the trust matrix. [In particular, she gains if and only if

$$\sum_{l \in C \setminus \{k,i\}} m_{jl}(t)\big(r_{lk}(t) - r_{ik}(t)\big) > m_{jk}(t) r_{ik}(t),]$$

iii. agent $k \neq i, j$ loses influence for sure if j trusts solely her, i.e. $m_{jk}(t) = 1$.

Proof
We know that $\pi_k(C; t), \pi_k(C; t+1) > 0$ for all $k \in C$. Furthermore, $m_{jj}(0) > 0$ implies $m_{jj}(t) > 0$ and thus, by Remark 3, $\sum_{l \in C \setminus \{i\}} m_{jl}(t) r_{li}(t) > 0$. Hence, $\pi_i(\mathcal{N}; t+1) > \pi_i(\mathcal{N}; t)$, which proves part (i). Part (ii) is obvious. Part (iii) follows since $m_{jk}(t) = 1$ implies $\sum_{l \in C \setminus \{k\}} m_{jl}(t) r_{lk}(t) = 0$, which finishes the proof. □