Belief diffusion in social networks

Patrick Girard, Valery Pavlov, Mark C. Wilson

June 12, 2015

Abstract

We report the results of a laboratory experiment investigating propagation of beliefs in a social network. During the experiment, participants faced several questions having objectively correct answers and could update their answers several times based on aggregate information about answers chosen by their neighbours. One of the novelties of the experiment was that the binary choices faced by participants were augmented to include an "I do not know" option and incentives to choose it when they felt indifferent between the actual options.

We observe that the dynamics of decisions in the network strongly depends on the question type, logical or factual. The results also indicate that propagation of beliefs can be more accurately described by a threshold model rather than models of probabilistic contagion. However, in contrast with assumptions underlying standard threshold models, our results suggest that it is not the larger proportion of neighbours that is driving participants' choices but the difference between the proportions of neighbours opting for the competing options.

Keywords: social networks, influence, advice taking, threshold model, propagation

1 Introduction

Social learning is one of the most important phenomena affecting behaviours of individuals and shaping behaviours of societies. People can learn from others about facts ("Yesterday's fog caused more than 10 car accidents"), about their actions and preferences ("I did not buy it, no one wears this 1960s style nowadays"), and about their beliefs ("I am sure I can learn to ride a unicycle in just a couple of days"). The present study specifically focuses on the spread of beliefs, a type of information that cannot be verified by a person receiving within a reasonable time. Even if a person sharing the information actually knows the information is correct, it is not possible to demonstrate to the recipient that it is correct. For example, if someone simply claims that the Great Wall of China is the only man-made object visible from the Moon it is impossible to verify the claim without resorting to other sources. Our interest in the propagation of beliefs is driven mainly by two factors. First, from the practical standpoint, it is arguably a more common situation than the spread/formation of social norms or propagation of verifiable information. Indeed, most often we cannot easily verify the information we receive by email or read on the internet. Second, from the academic standpoint, propagation of beliefs is another "pure" case, but, unlike the propagation of actions (e.g., herding), appears to be much less understood.

The main part of this study is a laboratory experiment. Participants, connected via a network, aim to select the correct answer to a multiple choice question with a single objectively correct answer. They perform several iterations of this process, simultaneously, at each iteration being given information on the distribution of answers of their neighbours at the last iteration. We are interested in whether social influence and learning occurs — that is, how close do the agents come to a unanimously correct selection, and how quickly. Furthermore, we are interested in whether there are significant differences between types of questions (for example, purely logical questions versus those that require knowledge of the world), and whether allowing agents to admit ignorance has an impact on learning.

In order to explain the positioning of our study in more detail we review in Section 1.1–1.4 some of the streams of research most closely related to our study. We give an overview of our specific hypotheses and results in Section 1.5. Section 2 describes our experimental setup. In Section 3 we present data and analysis, and Section 4 gives discussion, conclusions and comparison with other work.

1.1 Models of group behaviour

Social sciences as a whole, from history, philosophy and anthropology to political science and finance, aim at improving our understanding of human societies — a problem mankind has been challenged with since its own inception. The problem is complicated by the complexity of the society, owing to multiple levels of interactions and a variety of roles played by every member of the society. For example, a typical person belongs to a family, acting as a parent for her/his children, and, at the same time, as a sibling and a child her/himself. The same person may have a circle of friends (possibly having some set of unwritten rules and a hierarchy), belong to a religious group, be employed at a company (governed by its constitution and by laws), have responsibilities as a citizen (e.g. voting), etc. This (typical) person is often caught between conflicting goals, principles, disconnects between desires and capabilities, and has to face multidimensional trade-offs. The interests, beliefs, capabilities, and roles of individuals, companies, unions, etc., interact in incredibly complex ways, on multiple levels of legal, cultural and emotional ties.

Different fields approach the problem from the different perspectives and with a variety of methodologies. A common feature of most of these methodologies is individualism — although "emergent phenomena" may occur in groups, the individual is considered as the basic unit of analysis.

Tasks performed by individuals in groups are often modelled by game theory. The key concept is that of *best response*, a choice of actions for each situation such that each action optimizes the agent's payoff in that situation. When all agents simultaneously adopt such a strategy, this leads to *Nash equilibrium.* Furthermore, in some situations, iterated best response by players in some order will converge to an equilibrium. Of course, there are several problems with this approach: informational and computational requirements for agents to compute their equilibrium strategy may be enormous even in relatively small games; there may be more than one equilibrium, and no "obvious" one to coordinate on; convergence is not always achieved.

Experiments such as ours are often modelled by a *graphical game*, a strategic game in which each player's payoff depends only on the actions of her neighbours in the network. More specifically, since information about specific neighbours is not known, we are dealing with a *semi-anonymous graphical game* (Jackson 2010, Chapter 9). Many authors have studied *coordination games*, games that exhibit *strategic complementarity*, where the payoff increases with the number of neighbours sharing the same action as a given agent. Such games have been well studied, for example to model the adoption of new technologies, and they often lead to *herd behaviour* where players converge on a unanimous action.

However, game-theoretic models are of limited relevance for us. This is because the game is degenerate — each agent's payoffs are not affected by the actions of any other player. This is because our experimental design aims to incentivise participants to find the correct answer, rather than conform to the majority answer. This latter problem, namely that of emergence of social norms, has been widely studied in the past. Our main subject of interest in the present paper, however, is the propagation of knowledge and beliefs.

Of course, even though we designed the experiment to avoid explicit complementarities (or other externalities), we cannot rule out subjects deriving utility from being in the majority opinion even when it is wrong, for example. However, even if such were the case (which we doubt), we do not know how to estimate this utility quantitatively and compare it with the utility derived (via monetary payment) from correctly answering. Thus we assume that as stated above, strategic considerations do not really enter agents' analyses.

Our experimental situation is better modelled as an instance of *social learning*, which is closely related to *diffusion*, *influence*, and *contagion*. The answers of its neighbours may influence the answers of an agent, even though the payoffs are not directly influenced. A network diffusion model is specified by a set of *nodes* (agents) connected in a some way (*topology*), each node characterized by a *state*, and some *transition function* describing an evolution of states over time. Typical examples include models of infectious disease (where the state of an agent may be "infected", "susceptible", "recovered", etc), information flow ("informed", "uninformed"), product adoption ("has adopted", "has not yet adopted").

The special case where each agent is directly connected to each other (the complete graph) typically has much simpler dynamics, so the topology (the structure of the underlying graph of the social network) is usually very important in such models. However, in some models even the case of a complete graph is challenging, owing to the relative complexity of the transition function. Studies on diffusion in network models are mostly separated into those studying the effects of network topology and those studying the effects of local interactions between agents.

The earliest models of contagion come from the study of infectious diseases, and do not discuss networks at all. In effect, they use a complete graph, probabilistic transitions and "mean-field" approximations looking only at population means. They are also typically in continuous time and hence use the machinery of differential equations. By analogy, such models can also be used to describe the spread of rumours or information.

For example, the SIR model (Kermack & McKendrick 1927), apparently adopted from earlier models describing the rate of a chemical reaction, describes the spread of a contagious disease in a population. In our terminology, each node in a complete graph can be in one of three states: S (susceptible), I (infected), or R (recovered). The incubation period is zero and any infected person is contagious. At any point in time any infected person has a chance γ to heal and a susceptible person gets infected with some probability β. The dynamics of the system is fully described by the following system of ordinary non-linear differential equations

$$\frac{dS}{dt} = -\beta SI$$
$$\frac{dI}{dt} = \beta SI - \gamma I$$
$$\frac{dR}{dt} = \gamma I.$$

Note that although contagion and recovery take place probabilistically the transition equations are deterministic due to the assumption of a large population.

Variants of this model include the SI, SIRS, SIER models.

Epidemic models of the above type in which network structure is explicitly considered have even more interesting behaviour. Their long-run behaviour and time evolution can be studied using percolation theory and other mathematical tools. See Newman (2010) for background. Much is known about how the network topology controls the rate and extent of contagion in such models. These findings can be summarized roughly as: contagion spreads faster when there are "long ties", and in "small world" networks with low diameter. Such models have been used to study information flow, in addition to the spread of diseases.

More recently, researchers have realized that there is a difference between simple contagions of this type and *complex contagions* (Centola & Macy 2007). For the latter, simple exposure to an infected neighbour is not usually sufficient for a node to change state, and multiple exposures are required. For example, this appears to better model behavioral changes such as smoking cessation. The reason, apparently, is that, in contrast with the case of catching an infection, adoption of a behavior requires reinforcement, multiple exposures to the behavior adopted by close neighbours. Changing belief about a factual question, as in our study, is likely (we think) to require substantial reinforcement. Overall, complex contagions exhibit much different behaviour from simple contagions. Centola & Macy (2007) give four reasons why complex contagion may occur: strategic complementarity, credibility, legitimacy, emotional contagion. The second of these seems especially relevant to our study.

The role of topology in complex contagions was clarified by (Centola 2010), who used an experiment to study adoption of healthy behaviors in online communities, Manipulating the network topology showed that, contrary to the intuition derived from epidemic models, a small-world network is less effective than a highly clustered network in promoting diffusion of such behaviours, strongly suggesting that influence in such situations is a complex contagion.

For understanding complex contagions, the class of (linear) *threshold models* is often used. These were introduced in Schelling (1969)), in the context of segregation. In such a model, each node has an individual threshold (a real number in the interval $[0, 1]$) for changing its state, and will change state to s if and only if the fraction of its neighbours having state s exceeds the threshold. This is stark contrast to epidemic models in which a very small initial contagion can spread throughout the network much more easily. Schelling's model was the first that demonstrated the importance of allowing for both key components of network models, the micro interactions and the topology.

Threshold models are deterministic, although obvious probabilistic analogues exist. Granovetter (1973) argues that threshold models are well adapted to describe collective behaviour like riots, strikes, and diffusion of some types of innovations. The last situation has been heavily studied, sometimes under the name of "product adoption", first studied in Bass (1969) via a network-free simple contagion model, a special case of the SIR model.

A topic that has not been much studied is the node states and their role. Historically, the choice of a node state space has been motivated either empirically or by tractability. The most immediate example is the SIR model. The three possible states of a node match the actual conditions that a person may go through during an epidemic. However, extending the model to include the latent period (condition E) resulted in a model more closely matching the actual situation. In another example, Schelling's model of segregation, a node can be in one of a finite number of states, corresponding to the number of races. In such a model there is simply no room for interracial households and, if their proportion in the population of interest is large, one might consider extending the nodes' states to allow for mixed cases.

However, in situations when a node's evolution is a result of the node's decision-making, choosing an appropriate state space is much less straightforward. In particular, if the node state represents a belief (say, a subjective probability), then a continuous state is appropriate, in contrast to the discrete-state models described above.

In order to better align our experiment with reality, we allowed participants to report one of three states, one of which is "I don't know".

1.2 Decision-making in groups

Studies of group decision-making specifically focusing on social influence date back to the middle of the 20th century (see Costanzo et al. (1968) and references therein). This literature mostly deals with experiments on small groups in which information can flow easily between all members. Our study uses topologies and information restrictions in which this is not the case. Nevertheless, this literature contains some insights relevant to our study.

Costanzo et al. (1968) report that adoption by a group member of the same behavior as the majority of the group (conformity) is directly related to the perceived expertise of the majority of the group and inversely proportional to the (self)-perceived expertise of the person making the choice. In our experiment, we aim to implicitly manipulate the perceived expertise by choosing different types of questions. One of the questions was similar to a task used in Costanzo et al. (1968), involving estimating the area of figures.

Cox & Hayne (2006) note that most of research in decision-making focuses on individual-decision makers whereas most important decisions in real life are made by groups and little systematic work has been done. They also find, in the context of the common-value auction (known for the "winner's curse") that the performance does not improve with the size of a bidding unit (in the experiment they used 1, 3, 5 and 7 people in a unit). In the contrast, Slembeck & Tyran (2004) find that in a "Three Door" task, famous in part because less than 20% of participants choose to switch the door, introduction of groups results in nearly 100% correct decisions. Kugler et al. (2012) review the literature on group decision-making published over 25 years (including "groupthink") and conclude that in interactive settings (games) groups tend to make decisions closer to the standard game-theoretic predictions but in non-interactive settings, groups tend to mitigate some of the biases typical for individuals while perhaps exacerbating others. Overall, groups may perform better, worse or on par with individuals. The point, as Davis (1992) states it having reviewed the consensus group research published over 1950-1990 period, is that "group-level phenomena are often counterintuitive".

In the above studies, the group members typically reach a consensus on the group's answer, via an explicit or implicit *social decision scheme*. The most commonly studied are: "truth wins", where a single participant with the correct answer suffices for the entire group to adopt that answer; simple majority voting over two alternatives; choosing an answer uniformly at random; "truth supported wins", where two correct answers suffice for the entire group. Each of these explains observed behaviour best for certain types of questions. For example, "truth wins" works well for mathematical questions, while "truth supported wins" better explains performance on general world knowledge. An explanation for this in terms of *demonstrability* is given in Laughlin & Ellis (1986). Unsurprisingly, question types and experimental setups in which a participant knowing the correct answer can convince a skeptic of its correctness require fewer group members to support that answer initially in order for it to be adopted by the group. In our study, we chose different types of questions in order to better detect such differences. We did not go to the extreme of questions involving preferences with no objective justification, but we did include both "logical" (essentially mathematical) questions and "factual" questions involving general world knowledge. Our experiment did not allow for an communication between agents. Thus even though we chose question types that would have had high demonstrability if in-group discussion were allowed, our setup promotes low demonstrability.

A substantial body of research on advice-giving and advice-taking suggests that whether people take advice or discount it, and whether advising improves the quality of decisions, depends on factors such as the expertise of the advisor, the perceived expertise of the advisor, the confidence of the advisor, the expertise of the judge (a person taking advice) in the problem at hand, the value of rewards for high performance, etc. (see Bonaccio & Dalal (2006) for a literature review). The interplay between these factors can be important for the overall performance but not straightforward to predict. For example, Bonner et al. (2002) report that groups are more willing to follow advice of experts when the latter are legitimate, i.e. have a proven record of expertise. However, the overall performance may actually suffer because the non-expert group members are not scrutinizing the opinions of the legitimate experts, and when the latter make mistakes other members are unable to spot them.

From the network research perspective, it is interesting that most of the studies investigating the effect of group size do not generally find a substantial effect. There are exceptions, e.g. Laughlin et al. (2006) find that teams of 3, 4 or 5 people do better in the Letter-to-Number task, but the most prominent one is in coordination games — a robust finding here is that coordination typically fails once the team size goes over 4 people (Feri et al. 2010).

1.3 Social learning

The main focus of the present article is on how information about beliefs of other group members can, over time, lead to more accurate beliefs by individuals. If a single group belief is to be reported (for example in a jury trial) this can be computed via some aggregation method such as a fixed voting rule. Note that the idea that aggregation can substantially improve group decisions made by imperfect agents, without any interaction between agents, is at least as old as the Condorcet Jury Theorem, and is commonly referred to as "the wisdom of crowds". The mathematics behind such results often expresses the fact that independent random errors essentially cancel out under aggregation, at least when the number of agents is large. However, if there is bias in the errors, aggregation can exacerbate this bias. In our study, by using questions known from the psychology literature to elicit an intuitively obvious but factually incorrect answer, we introduce bias.

Different mechanisms have been proposed for how individuals learn from others. Where sufficient time and opportunity for communications between agents is available, improvements in skill (the traditional meaning of "learning") are possible. However in many situations of interest (for example, anonymous online networks), agents can only (repeatedly) observe actions of other agents (*observational learning* as opposed to *communicational learning*). This does not rule out (by provoking further introspection) agents improving their skill, but can also lead to simple imitation without an increase in understanding. This distinction between *process contagion* and *output contagion* is made in Rahwan et al. (2014). Our study is designed to be purely observational.

The literature discussed in Section 1.2 is all concerned with communicational learning, and in the case where the topology is trivial (a complete graph). From now on we focus on learning on networks. This closely related to diffusion — the node state is typically called a *belief*, and we expect beliefs to diffuse somehow through the network. However typically there is a notion of objective truth which we hope that beliefs will converge to. The basic questions are: is convergence to an equilibrium belief reached? is the equilibrium belief unanimous (a consensus)? is the crowd eventually wise? Of course, beliefs cannot usually be measured directly, and are typically inferred from actions.

Lazer & Friedman (2007) derive a model for collaborative problem-solving in which the underlying topology is a variable and agents imitate others with better solutions. The model predicts that for difficult tasks involving a "rugged" landscape of solutions, networks that optimize the spread of information lead to worse overall solutions in the long run than less efficient networks, although they do better on shorter time scales. The explanation is that there is a tradeoff between "exploration" (generating candidate solutions) and "exploitation", and too much communication early in the process reduces diversity and thus curtails exploration, leading to locally optimal but globally suboptimal solutions. Mason & Watts (2012) describe a laboratory experiment along similar lines which finds, in contrast, that performance is better overall for efficient networks. One explanation is that inefficient networks, by allowing complex contagion, promote imitation and act against exploration. Mason et al. (2008) also describes a laboratory experiment which refines the above analysis. They find that very different network topologies work best for different types of problems — it matters whether there is a single good solution, many acceptable ones, etc. The problem described in this paragraph differs markedly from ours, in that participants in our study do not receive any information on the quality of their proposed answer until after the experiment.

Closer to our setup, Lorenz et al. (2011) describe an experiment (on a complete network) involving observational learning, in which social aggregation of information is relatively poor, and give three main reasons for this. Two of these are related to social influence — a reduction in diversity as in Lazer & Friedman (2007), and increased confidence of agents in their answers because of herd effects. Rahwan et al. (2014) describe an experimental study of observational learning whose main finding is that higher connectivity leads to better convergence to the correct answer.

The main models of observational learning on networks involve subjective probability. Each agent maintains a number between 0 and 1, which is updated based on the beliefs of other agents. *Bayesian learning* (Gale & Kariv 2003) requires (rather unrealistically) each agent to consider the full history of actions of each neighbour, and typically leads to convergence to a consensus belief in the long run. A simpler myopic update rule was proposed by DeGroot (1974), in which the state of a node is updated by forming a weighted average of all neighbouring states (the weights are fixed, so that node do not revise their beliefs about reliability of their neighbours). This typically leads to convergence (guaranteed by theorems about Markov chains) to a common consensus belief. Furthermore, crowds are eventually wise under many natural sufficient conditions. More detailed information can be derived about convergence rates, situations where convergence fails, and the relationship between convergence speed and wisdom (Golub & Jackson 2010).

Note that both of the above models assume knowledge by each agent of each neighbour's state. More typically agents can only observe actions of others, and actions are discrete. In situations where anonymity is preserved, such as large online networks, typically only summary statistics are available. This means that individual neighbours cannot be distinguished. A discrete analogue of the DeGroot model in the case where there are 2 actions is described in Chandrasekhar et al. (2012). Each agent has a subjective probability, and a threshold t with $0 \leq t \leq 1$. At each time step, each agent reports a discrete opinion 0 or 1. Each agent updates its state by simple averaging (possibly with a different weight given to its own opinion) and reports 1 if and only if this average exceeds t.

Our study is designed so that participants receive only fully anonymized summary feedback about their neighbours. Furthermore there are only 3 signals observed.

Some experimental work on observational social learning in networks has been conducted and brought a number of important insights (Berninghaus et al. 2002, Cassar 2007, Centola 2010, Chandrasekhar et al. 2012, Choi & Lee 2014, Corbae & Duffy 2008, Jia et al. 2014, Watts & Dodds 2007). For example, Choi et al. (2005) have the agents play a coordination game in which they receive a payment if they all choose the same action, and the nodes whose action was taken receive more than the rest. They find that more connected network structures as well as longer communication improve the efficiency and the equity of the outcomes. Chandrasekhar et al. (2012) conduct a unique lab experiment in rural India to test the performance of Bayesian and DeGroot models and find that a simple DeGroot averaging model with 50% threshold better models the learning process occurring on the network.

1.4 Belief revision

Our study concentrates on beliefs about matters of objective fact, where the observed actions of agents are "believe P is true", "believe P is false", or "don't know". A theoretical motivation for focusing on a three-state model of belief stems from the extensive literature in belief revision (an active area of research by computer scientists, logicians and philosophers since the seminal work (Alchourran et al. 1985) (see also Gärdenfors (1988)). An important distinction in belief revision is that between *revision* and *contraction*. A revision is a belief change in which an agent adopts a new belief, and a contraction is one in which the agent drops a belief without endorsing a new one. An example of a belief contraction is someone who becomes agnostic (refrains from belief or disbelieving in God) without becoming atheist (disbelieving in God). Another instance is to stop condemning something without condoning it. Finite state models make more sense in this framework. It is a notorious problem to draw a distinction between revision and contraction in a probabilistic framework.

1.5 Our contribution

We performed a controlled laboratory experiment with human participants, designed to detect propagation of beliefs among them, and analysed the dynamics of subjects' answers. We used two "logical" questions (Q1, Q2) for which the correct answer can be deduced from the information given in the question, two "factual" questions (Q3,Q4) that required knowledge of famous but difficult to remember facts about the world that could not be deduced from the question, and one factual question (Q5) chosen so that no one could possibly know the correct answer with certainty except for the experimenters. In the latter case we provided the correct answer to some participants and made it common knowledge that some participants know the answer. For each question participants were provided two possible answers, of which one was correct, and a possibility to admit they do not know the answer by choosing the "I do not know" option. We considered two questions to be "fair" (Q1, Q3), two to be "tricky" (Q2, Q4), and one to be "impossible" (Q5).

1.5.1 Research hypotheses

(i) (topology)

- (a) Convergence will be faster in the complete graph than in the other topology.
- (b) There will be a larger degree of unanimity with the complete graph than with the other topology.

(ii) (rationality)

- (a) The group will learn the correct answer on Q3 and Q5.
- (b) The number of incorrect responses will be less on Q3 and Q5 than on the other questions.

(iii) (logical versus factual questions)

- (a) The level of susceptibility of participants to influence from neighbours is lower for the logical questions than for the factual questions.
- (b) The degree of unanimity among participants is lower for the logical questions than for the factual questions.
- (c) The likelihood of answering "I don't know" is lower for the logical questions than for the factual questions.
- (d) The group is more likely to learn the correct answer on factual than on logical questions.

(iv) (fair versus tricky questions)

- (a) For fair questions, aggregation by plurality voting will eventually yield the correct answer, whereas this will not occur for tricky questions.
- (b) The group is more likely to learn the correct answer on fair than on tricky questions.
- (c) For tricky questions, better eventual aggregation of the correct answer will occur in the other topology, whereas for fair questions, it will be the same for both topologies.
- (d) The speed of convergence for tricky questions will be greater than for the fair questions.

(v) (belief revision)

- (a) Participants require more reinforcement from neighbours before deciding to switch from Answer 1 or "don't know" to Answer 2 than to switch from Answer 1 to "don't know" (here 1 and 2 denote the two definite answers, either of which could be correct).

1.5.2 Discussion of research hypotheses

For (i), note that whether the contagion is simple or complex, all standard models predict faster convergence on complete graphs. The idea of lack of diversity begin caused by communication efficiency leads to the second assertion.

For (ii), note that given sufficient confidence in the rationality of others, it is rational to answer Q5 correctly (if the subject has been given the answer), or answer "I do not know" until some neighbour gives either answer 1 or 2, in which case (assuming there is only one such answer given by neighbours) the subject should copy that answer. For Q3, a similar argument holds: we expect that participants will know immediately that they do not know the answer, and expect that at least someone else in their network does know (for Q5, this fact was common knowledge).

The reasoning behind (iii) is as follows. For logical questions, each participant will consider her expertise to be fairly high relative to the group's expertise, and in any case will know that in principle she can answer the question correctly without external help. For factual questions, she is more likely to realise her own ignorance (and for the impossible question, ignorance of most participants is essentially common knowledge). Thus participants will be more susceptible to influence from neighbours in the factual questions than in the logical questions. Furthermore fewer wrong answers will be given, since answering "I don't know" gives better payoff than guessing. Finally, the lack of diversity of answers should lead to a higher degree of unanimity caused by herding.

For (iv), note that we chose Q2 and Q4 to have commonly chosen false answers. Our hypothesis is that participants giving these answers (more than half of them) are less likely to be influenced than on other questions, and in any case even if they are influenced, most of their neighbours will also be wrong. This explains the fourth assertion. It is known that the expected success rate is about 10–20% for the Wason task. It is hard to imagine a convincing learning model that will lead to the majority of the group converging to the correct answer, given the purely observational nature of the study (low demonstrability). The second assertion is motivated by the idea that herding will occur owing to a lack of diversity of answers. The third assertion is based on the findings of Lazer & Friedman (2007).

As for (v), our understanding of belief revision implies that giving up a belief ("becoming agnostic") requires more reinforcement than adopting the opposite belief ("becoming atheist").

1.5.3 Summary of results

Our study shows that under the experimental conditions described, subjects' answers are influenced by the answers of others, even for purely logical questions and where almost no information can be deduced about the correctness of others' answers during the experiment.

We broadly confirm the research hypotheses (i) – (v) above, with some exceptions (see Section 3 for details). In particular, the effect on convergence rate of topology and question type was not as expected.

Refuting a model with many unknown parameters is beyond the power of our study. Our results are broadly consistent with threshold models. However, we can refute models that predict that the fraction of participants giving the correct answer should increase monotonically with iteration number.

Also, models predicting eventual convergence to unanimity in all cases are refuted by our data.

This includes many standard infection models, which gives more justification for us to consider threshold models.

Interestingly, we find that standard threshold-type models, in which the likelihood of switching opinion to a given answer is an increasing function of the absolute fraction of neighbours exhibiting that opinion, are not well suited to our data. Much more compelling are models in which the relevant variable is the *net* fraction of neighbours holding an opinion.

2 Experimental setup

Considering the exploratory focus of the study, a major requirement for the experiment design was to enable us to obtain a rich data set. To this end, two key design parameters, the decision-making tasks and the network topology, were chosen to ensure a wide range of distinct decision-making conditions.

2.1 Participants

Participants were students of a large public university in Australasia from a variety of majors, mostly business school undergraduates. Overall, 52 people took part in the study. To recruit participants we used ORSEE (Greiner n.d.). All sessions were conducted in a decision-making research computer laboratory designed for running decision-making experiments. Participants were paid around $5 of "show-up" fee and an additional amount of money proportional to the profit (in experimental tokens) earned as a result of the decisions they made during the experiment. The average payment, including the show-up fee, was around $20. Prior the experiment, participants did not know the exact nature of the experiment, only that it is about belief propagation in social networks. Participants may or may not have had prior information about each other. However the experimental setup ensured that all information about a given participant was anonymised, so that no participant could know anything about the answers given by any other.

2.2 Incentives

Money was the only pecuniary incentive provided. During the experiment participants were presented with five different questions (see Appendix C). On each question they could answer 10 times by choosing among three options that were the same every time. One of the options was the correct answer (each question had an objectively correct answer), one was an incorrect answer and the third option was "I do not know". After a participant answered the question 10 times, the correct answer was revealed.

Participants were paid as follows (see the experiment instructions in Appendix A):

- out of ten answers that each participant can give for each question, only two contributed to the profit; the very first answer and one other chosen at random;
- the correct answer results in 10 tokens;
- the incorrect answer, or no answer, results in 0 (zero) tokens;
- answering "I do not know" results in 6 tokens.

We chose these parameters in order to induce participants to report "don't know" rather than not answering, or guessing an answer uniformly at random.

2.3 Decision-making tasks (Questions)

Our selection was subordinate to the goal of inducing different degrees of confidence in the correct answer. Two cognitive questions (Questions 1 and 2 in Appendix C), one of the questions used in Frederick's Cognitive Reflection Test (Frederick 2005) and a standard variant of the Wason selection task (Wason 1968), are well-known to be challenging, producing substantial proportions of incorrect answers. These questions are self-contained and no extra information is needed to answer them correctly. In contrast, the experience-based questions testing factual knowledge have the property that the correct answer may be not known to everyone. Therefore, receiving information about answers given by other people is the only possibility for those who do not know the fact underlying the question to give a correct answer. Two of the factual questions (Q3 & Q4) are based on publicly known facts. One question (Q5) is based on a fact privately communicated to some participants, while all other participants were assured that some people know the correct answer.

To collect the participants' responses we used zTree (Fischbacher 2007). The very first screen displayed the experiment instructions (Appendix A). Then participants were presented with each question 10 times. At each of the 10 iterations for a given question, they were provided the information about their last answer and the distribution of answers given by their neighbours. They were then given the opportunity to change their answer if desired. They could not change any of the past answers. Participants were told how many neighbours they had, but nothing that would identify who they were.

2.4 Topology

We used two quite different network topologies in our experiment. One of these was a complete (undirected) graph in which each pair of distinct nodes is linked. We also devised another network topology having several interesting properties. First, the links are directed. The influence can only propagate in the direction of the arrows. Simply put, a node only "see" those nodes that have arrows going to that node. One goal we pursued with this design feature was to avoid endogeneity in one part on the network (horizontal nodes at the bottom of Figure 1). Another goal was to make it possible to create a more uniform distribution of the degrees on the network, and this is the second main property of the topology.

For each of the 5 questions, the node numbers of the physical machines were permuted, so that subjects usually occupied different positions in the network. For each iteration of a fixed question, the node numbers were fixed.

3 Experimental results

We captured a complete record of subjects' answers at every iteration of each question. We also have complete information about the answers from the previous iteration fed to each node.

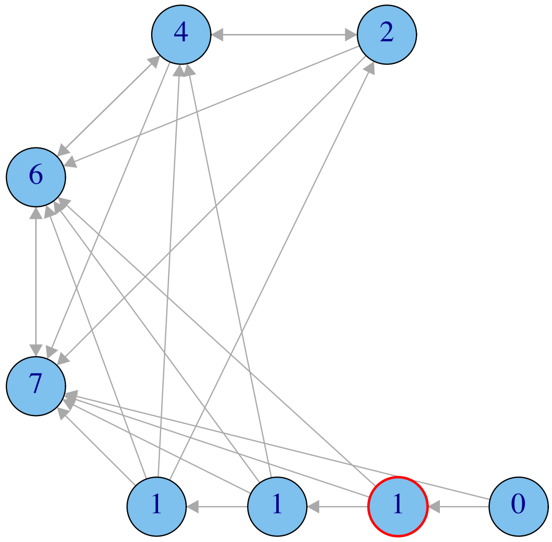


Figure 1: The principle behind the network topology used in the experiment. Labels on the nodes indicate the number of incoming links. The nodes have been divided in two groups. Each of the bottom nodes has only one feeding node, and those closer to the left feed increasingly more nodes above. The bottom nodes are not influenced by the nodes above.

Table 1: Parameters of the different treatments used

Treatment ID	Treatment date	Topology	Number of participants
A	140822_1152	directed	14
B	140910_1337	directed	8
C	141001_1308	complete	18
D	141002_1255	complete	12

3.1 Descriptive statistics

We first present some graphs. A label such as A, 1 refers to Treatment A (see Table 1), Question 1. For each treatment, we use several indicator variables, indexed by subject s, question q and iteration i. These are shown in Table 2. For each such variable, omitting an index simply averages over all values of the index. Thus, for example, l_s is the mean of the last iteration on which subject s changed, over all questions q.

First, we present Figure 2 which displays for each treatment and each subject s, the value of n_s, the fraction of times s changed answer while answering the same question. Figure 2 shows that subjects were generally engaged in the experiment, and made substantial numbers of changes to their answers. There was a substantial variation between subjects.

Figure 3 shows, for each question and treatment, the fraction $(C_{q,i}, I_{q,i}, U_{q,i}, A_{q,i})$ of subjects giving each possible type of answer at each iteration.

Figure 4 shows the (approximate, since repeated data points were "jittered") empirical cumulative

Figure 2: Changes by treatment and subject

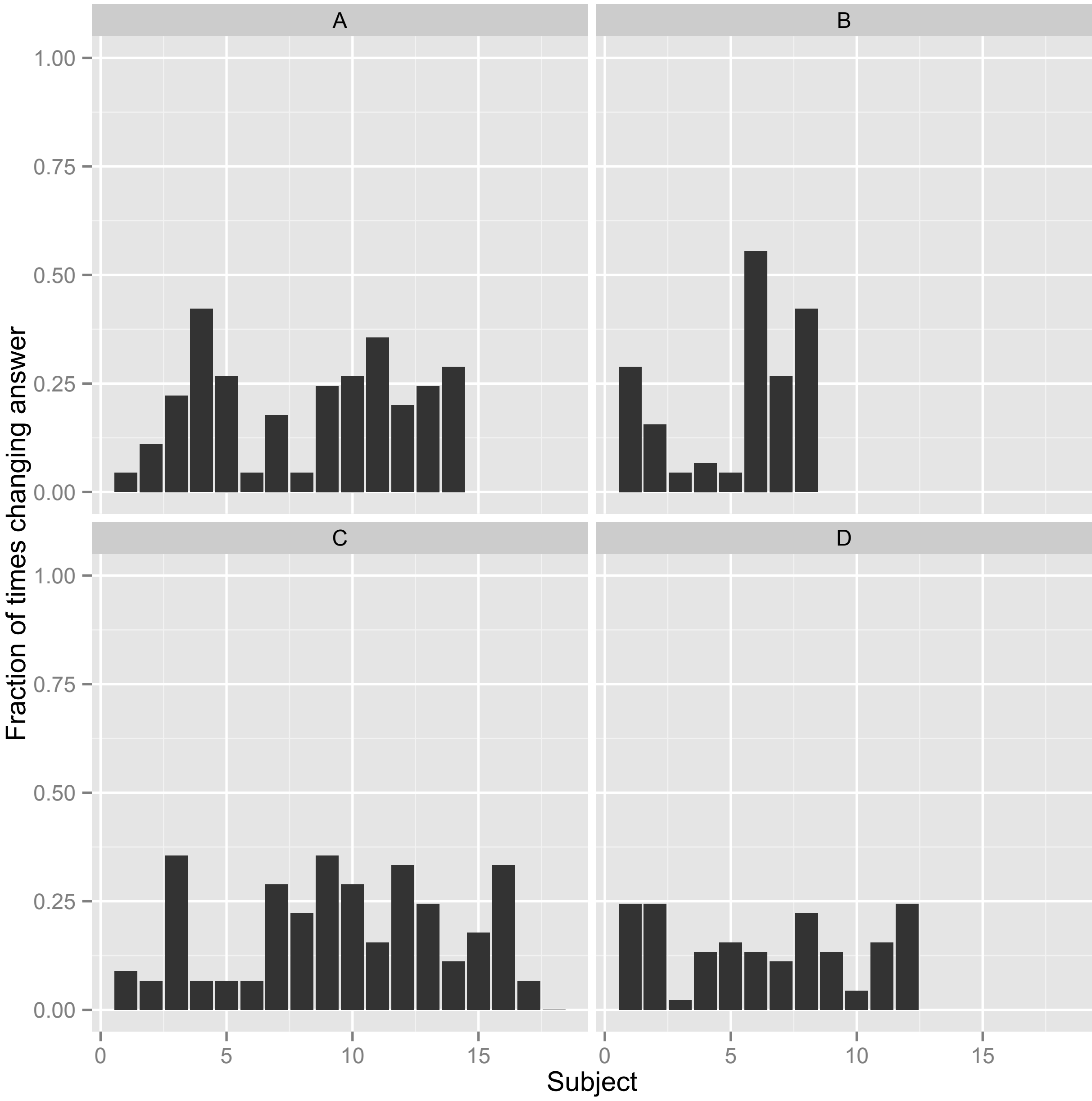


Figure 3: Answer type distribution by (treatment, question)

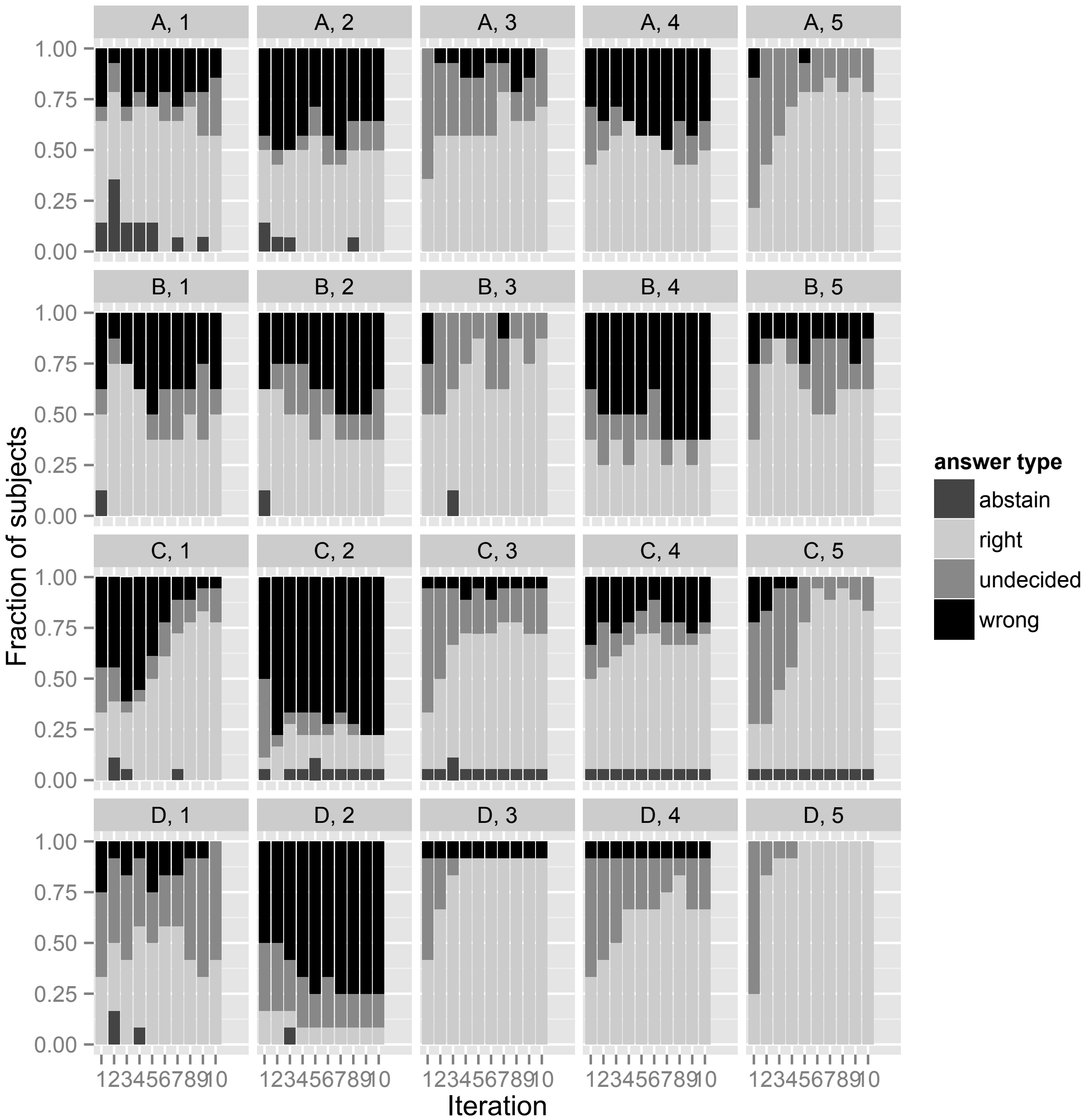


Table 2: Indicator variables used in analysis

Variable	Meaning
n	Change in answer
n^1	Change to answer 1
n^2	Change to answer 2
n^3	Change to answer 3 ("don't know")
L	Last iteration when answer changed
C	Correct answer given
I	Incorrect answer given
U	"Don't know" given
A	No answer given
M	Answer given is the modal answer
logical	Question is "logical" (Q1, Q3)
tricky	Question is "tricky" (Q2, Q4)
topology	Did not use complete graph (Treatment C, D)

distribution functions for all changes to a given answer, the colours red, green, and gold corresponding to the answers 1, 2, 3 (= "don't know") respectively. The horizontal axis measures the fraction of neighbours that had the given colour. We observe that a large number of changes occurred to answer 1 (answer 2) by nodes with small fractions of neighbours having given that answer the the previous iteration. Also, there is a clear difference between the behaviour of answer 3 (the top curve) and that for the other answers, formalized by first-order dominance.

3.2 Statistical tests

We formulate our research hypotheses, listed in Section 1.5.1, in terms of the experimental data as follows.

(i) L will be smaller in treatments C and D than in A and B, while for $i = 10$, M_i will be larger. Similarly, $M_{10} - M_1$ will be larger for treatments C and D.

(ii) When $i = 10$, $C_{q,i} > 0.5$ for $q \in \{3, 5\}$. Also, I_q will be smaller for $q \in \{3, 5\}$ than for $q \in \{1, 2, 4\}$.

(iii) There will be a difference in the patterns of answers associated with "logical" versus "factual" questions. Specifically:

- For both $i = 1$ and $i = 10$, $U_{q,i}$ will be smaller in the logical questions ($q \in \{1, 2\}$) than in the factual ones ($q \in \{3, 4, 5\}$).
- L_q and n_q will be larger in the "factual" questions than in the "logical" ones.
- When $i = 10$, $M_{q,i}$ will be larger for $q \in \{3, 4, 5\}$ than for $q \in \{1, 2\}$.
- $C_{q,10} - C_{q,1}$ will be smaller for $q \in \{1, 2\}$ than for $q \in \{3, 4, 5\}$.

(iv) There will be a difference in the patterns of answers associated with "fair" versus "tricky" questions. Specifically:

- When $i = 10$ $I_{q,i} \geq 0.5$ for $q \in \{2, 4\}$, while when $i = 10$ $C_{q,i} > 0.5$ for $q \in \{1, 3, 5\}$).

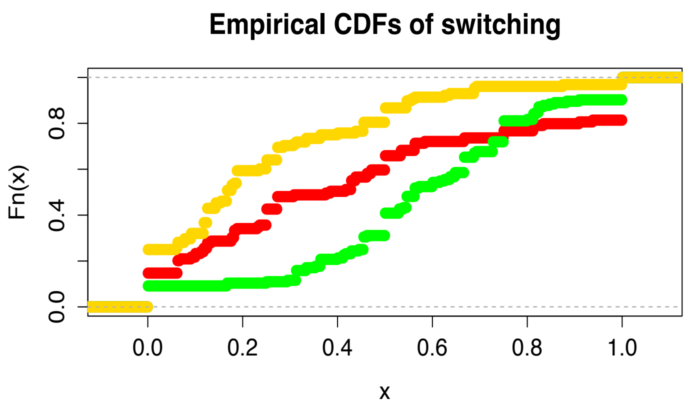


Figure 4: Empirical distributions of switching depending on the proportion of neighbours choosing the given answer. For example, about 70% of switches to answer 3 ("don't know") occurred with at most 40% of neighbours reporting that answer in the previous iteration.

- $C_{q,10} - C_{q,1}$ will be smaller for $q \in \{2, 4\}$ than for $q \in \{1, 3, 5\}$.
- For tricky questions, when $i = 10$, $C_{q,i}$ will be larger for treatments A and B than for treatments C and D. For fair questions there will be no difference.
- L_q and n_q will be larger in the "fair" questions than in the "tricky" ones.

(v) There will be a difference between the empirical distribution functions of "don't know" and the other two answers.

We present the results of several regression analyses. Throughout, we excluded data points in which a subject did not make any choice (Figure 3 shows the number of such points). In each case, the regressors include the indicators of whether the question is logical or factual, and whether it is "fair" or "tricky", as described above, and there is one data point for each subject. Tables 3–5 present models fitted to the data and our conclusions are based on the significance results reported in these tables. Note that R-squared tend to be very low, and, therefore, our assessment of the detected effects is, correspondingly, that they are very moderate despite being statistically significant.

We observe a clear difference between types of questions. Table 3 shows regression results for the case where the dependent variable is respectively $U_q, U_{q,1}, U_{q,10}$, while the independent variables are indicators for logical and tricky questions.

At the first iteration, "I do not know" is less likely on the logical questions than the factual ones. However, at the 10th iteration the difference is no longer statistically significant. Similarly, there is a clear difference at the first iteration between tricky and fair questions, which persists (albeit rather weakly) until the 10th iteration.

Regression analysis for correctness is shown in Table 5. We observe a clear difference between question types in terms of fraction of correct answers.

We are interested to see whether "crowds became wiser" and if the type of question mattered. To this end, we regress the difference between two indicators on the question type. Model (1) in Table

5 presents the results. The variable diffCorrect is simply $C_{s,q,10} - C_{s,q,1}$, measuring the coincidence with the correct answer at the last and at the first iteration. In the same table, model (2) tests whether subjects become closer to the modal answer. The variable diffModal is $M_{s,q,10} - M_{s,q,1}$.

We fit a regression model to test whether the convergence depends on the question type. The results shown in Table 4. In the first model, the dependent variable is $L_{s,q}$. In the second model, the dependent variable is $n_{s,q}$.

Table 3: Factors affecting "I do not know" choice (logit model)

	Dependent variable:		
	idk		
	(1)	(2)	(3)
logical1	−0.21	−1.02	0.49
	$p = 0.05^{**}$	$p = 0.0005^{***}$	$p = 0.15$
tricky1	−0.44	−0.63	−0.76
	$p = 0.0001^{***}$	$p = 0.03^{**}$	$p = 0.05^{**}$
Constant	−1.17	0.14	−1.52
	$p = 0.00^{***}$	$p = 0.45$	$p = 0.00^{***}$
Akaike Inf. Crit.	2,487.12	319.13	235.20
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 4: The pattern of changes (OLS model)

	Dependent variable:	
	lastChanged	subjChanged
	(1)	(2)
logical1	0.63	0.12
	$p = 0.20$	$p = 0.62$
tricky1	−1.17	−0.77
	$p = 0.02^{**}$	$p = 0.002^{***}$
topology	0.42	
	$p = 0.38$	
Constant	3.61	2.35
	$p = 0.00^{***}$	$p = 0.00^{***}$
Observations	256	280
Adjusted R^2	0.02	0.03
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

We find a clear asymmetry between dynamics of the "don't know" option and the other options. This is evident from Figure 4, and confirmed by a Kolmogorov-Smirnov test. The p-values of the KS tests all for three pairwise comparisons are less than 0.001. More changes are made to "undecided" than to other answers. This is consistent with a 3-state threshold model incorporating

Table 5: Correct and modal answers

	Dependent variable:			
	diffCorrect	diffModal	correct	isModal
	OLS	*OLS*	*logit*	*logit*
	(1)	(2)	(3)	(4)
logical1	−0.26	−0.09	−1.16	−0.54
	p = 0.0001***	p = 0.32	p = 0.0001***	p = 0.06*
tricky1	−0.24	−0.02	−1.19	−0.38
	p = 0.0002***	p = 0.79	p = 0.0001***	p = 0.19
topology	−0.18	−0.15	−0.20	−0.73
	p = 0.003***	p = 0.07*	p = 0.49	p = 0.01***
Constant	0.54	0.27	1.56	1.55
	p = 0.00***	p = 0.0002***	p = 0.00***	p = 0.00***
Observations	256	256		
Adjusted R^2	0.15	0.01		
Akaike Inf. Crit.			307.97	309.69
Note:			*p<0.1; **p<0.05; ***p<0.01	

both revision and contraction.

4 Discussion and conclusions

4.1 Research hypotheses

Of our 13 research hypotheses, 8 were supported by our data and analysis. The following were not confirmed.

- We did not observe a statistically significant increase in convergence rate when using the complete topology as compared to the other topology.
- We did not observe a statistically significant increase in convergence rate when using logical questions as compared to factual questions (in Table 4), there is no significant coefficient for the indicator *logical*).
- Although the fraction of subjects answering “don’t know” was significantly smaller for logical than factual questions over all iterations and for the first iteration, this effect was no longer significant when we looked only at the last iteration.
- We did not observe that for tricky questions, social learning of the correct answer was greater when using the incomplete topology (the signs of the coefficients in the first column of Table 5 should be different, but they are the same; performance on fair questions appears better in the complete topology).

Overall, we find that participants are more influenced by neighbours' answers to logical questions than we had expected.

4.2 Implications for modelling

Non-monotonicity

We note that the fraction of correct answers is mostly non-monotonic as a function of iteration number. The crowds may "eventually become wise", but their correctness is not monotonically improving over time. Crowds may also become unwise. The (treatment, question) pair labelled (B,2) is an instance where half of the participants had the correct answer to begin with, but fewer than half had the correct answer by the end. Even with the complete graph (D,1), well over 50% were correct at iteration 6 and 7, but well under 50% were correct in the following 3 iterations. This data refutes any model that predicts monotonicity. Also, models predicting eventual convergence to unanimity in all cases are refuted by our data. This includes many standard infection models.

Rationality

We have observed some violations of rationality. For example, for Q5 all participants were informed that the correct answer had been given to some participants. For a participant to whom the correct answer was given, answering correctly at every iteration dominates every other option (because of the payoff structure). For the other participants, answering "don't know" until another answer is detected among the feeds, then switching to that answer, dominates any other strategy (assuming belief in rationality of other players). However in Q5 we observed some subjects answering incorrectly.

Such violations were relatively few, however. We designed Q3 so that subjects would be led to reason as for Q5 - it seems reasonable to assume that even if I don't know the answer, at least one other participant will know it. The number of incorrect answers was fairly small for this question, and the correct answer was eventually chosen by a large majority in both of these questions, with the initially rather large number of undecided subjects rapidly becoming small.

Threshold models

Table 6 presents models (1)-(5), one for each question. In each model the dependent variable is $a1$, an indicator equal to 1 whenever a subject chooses the first option among the three answer choices. The independent variables $p1$ and $p2$ are the proportions of neighbours that chose at the previous iteration answers 1 and 2 respectively. Note that $p3$ is not included because $p1 + p2 + p3 = 1$).

Note that we are using linear models in a situation when the dependent variable is binary while a logit model would be normally used.

However, these models illustrate two important points. First, the coefficients (regardless of the fact that the models are not particularly meaningful) are very different across models. That is, participants' behavior was very different in different questions. Second, the values of R^2 are vastly different too, ranging from 7% to 73%. In other words, a model perfectly fitting the data in one question may be unable to describe what happens in another question, let alone to predict. This is consistent with our analyses above showing substantial sensitivity of dynamics to the type of the question.

Table 6: The explanatory power of neighbours' influence

	Dependent variable:				
	a1				
	(1)	(2)	(3)	(4)	(5)
p1	0.51	0.68	0.15	0.56	0.94
	p = 0.00***	p = 0.00***	p = 0.05**	p = 0.00***	p = 0.00***
p2	0.16	0.19	0.08	0.14	0.72
	p = 0.0002***	p = 0.0000***	p = 0.0001***	p = 0.0004***	p = 0.0000***
Observations	448	455	457	459	459
Adjusted R^2	0.25	0.35	0.06	0.26	0.73
Note:					*p<0.1; **p<0.05; ***p<0.01

Next, Table 7 presents three logit models explaining the role of the neighbours on the choice of the first answer. The dependent variable a1 is an indicator that equals to 1 when a subject chooses the first answer. Model (1) follows, roughly, the general "threshold" idea that the likelihood of choosing the first answer depends only on the proportion of neighbours choosing the same answer. Model (2) is more in spirit of preferential attachment models stating that a node may imitate any of its neighbours but the chances depend on the proportions of neighbours of each type. The proportion of neighbours choosing "I do not know" option is not included into the model because $p1 + p2 + p3 = 1$ and so introducing $p3$ cannot add any new information. As important result following from model (2) is that the coefficients of $p1$ and $p2$ are nearly equal in absolute value but have different signs (this would be a trivial result if there were only two options because then $p1 = 1-p2$ would hold but there are three options in our experiment). In effect, this is equivalent to saying that the likelihood of a subject choosing answer 1 depends on the difference, $(p1-p2)$. To test this data-driven hypothesis, in model (3) we excluded $p1$ and $p2$ but introduced diff12 $= (p1 - p2)$ and also $p3$. Judging by AIC, model (2) is better than model (1) but (3) is better than (2). The interpretation of model (3) is somewhat counter-intuitive. For example, compare two situations in which the proportions of neighbours who have chosen options 1, 2 and 3, respectively, are 0.55, 0.45, 0.0 and 0.1, 0.0, 0.9. The difference between the proportions of the first and the second choices is the same 0.1 but the proportion of option 3 in the second scenario is much larger. According to the model, the likelihood of the node choosing option 1 is much smaller than in the first case. The intuition, apparently, seems to suggest the opposite because 0.55 is not very different from 0.45 and so a person should be almost indifferent between the two options while 0.1 is infinitely larger than 0.0 and so the first option seems clearly dominant. Yet, apparently, larger proportions of people abstaining from taking either side may serve as a kind of a warning signal that other people may be seeing some "red flag". Similar observations hold for the second answer, only less pronounced.

Turning to the third, "I do not know" answer, Table 8 presents several models we fit to uncover how much this choice is affected by the neighbours' decisions. The most simple Model (1) turns out to have an excellent fit, better than Model (2) in which we introduced the absolute difference between the proportions of neighbours choosing options 1 and 2, and only slightly worse than the complete Model (3) in which we included not the proportions of neighbours choosing each option. However,

Table 7: Compensatory effects of neighbours' choices

	Dependent variable:		
	a1		
	(1)	(2)	(3)
p1	2.92	1.85	
	p = 0.00***	p = 0.00***	
p2		−1.70	
		p = 0.00***	
diff12			1.74
			p = 0.00***
p3			−0.99
			p = 0.0001***
Constant	−1.73	−0.66	−0.42
	p = 0.00***	p = 0.0000***	p = 0.00***
Akaike Inf. Crit.	2,432.69	2,364.35	2,345.41
Note:			*p<0.1; **p<0.05; ***p<0.01

Model (3) has redundancy due to $p1 + p2 + p3 = 1$ and, therefore, coefficients cannot interpreted due to the issue with identifiability. Interestingly, excluding $p3$ from the predictors results in a very poorly fit Model (4) despite the added absolute difference. Therefore, our conclusion is simply that $p3$ is the only reliably informative predictor.

Table 8: How neighbours affect the "I do not know" choice

	Dependent variable:			
	a3			
	(1)	(2)	(3)	(4)
diff12abs		−0.09		−0.42
		p = 0.65		p = 0.07*
p1			0.68	−0.56
			p = 0.11	p = 0.05**
p2			0.74	−0.48
			p = 0.07*	p = 0.07*
p3	1.77	1.71	2.41	
	p = 0.00***	p = 0.00***	p = 0.0000***	
Constant	−1.92	−1.86	−2.60	−0.94
	p = 0.00***	p = 0.00***	p = 0.00***	p = 0.00***
Akaike Inf. Crit.	2,058.59	2,060.38	2,058.62	2,101.36
Note:				*p<0.1; **p<0.05; ***p<0.01

A pure linear threshold model, where the agent places zero weight on her own current belief, would predict that changes to option 1 (respectively 2) would be impossible if the feeds were all of the opposite opinion. We observe no violations of this in our data. Another basic prediction of such models is that if 100% of feeds agree on 1 or 2 being the correct answer, any switch would be to that answer. We see only one violation of this, in (A, 3) where 100% of feeds choosing answer 2 led to a switch to "undecided".

Although the logical questions were able to be solved without any input from other participants, there are many instances where participants changed their answer. For example, in Q2 several switches to the wrong answer occurred. This is consistent with the idea that some participants have low thresholds for influence.

Question 2 (the Wason task) displayed some interesting behaviour. Not only did the crowd not become wise in any of the 4 treatments, there were some violations of standard threshold models. For example, in (C,2) there are several changes to the correct answer when 80% of feeds supplied the wrong answer. This behaviour did not occur in the factual questions and indicates that a larger weight on each agent's own opinion is required when modelling logical questions (as seems intuitively reasonable).

Our data allow for more testing of threshold models. For example, we can find upper and lower bounds on subjects' thresholds by observing when they change or did not change their answer. Thus, the upper bound of the threshold for switching to option 1 of a given subject for a given question is defined as the proportion of neighbours choosing 1 observed when the focal subject switched to option 1. The reasoning is that since the subject switched then it must be the case that the proportion of neighbours exceeded the threshold. The threshold is still unknown but smaller than the observed proportion of neighbours. The lower bound is defined as the proportion of neighbours when the subject did not switch to 1. That is, since the subject did not change it must be the case that the proportion of neighbours is not below the threshold. An inconsistency occurs whenever the lower bound exceeds the upper bound. In order to have enough data, we pooled all answers for each subject. We observe substantial inconsistency of choice. For option 1 alone, 2/3 of participants made decisions such that their lower bound strictly exceeded the upper bound. However, the data in Table 9 suggests that choice inconsistency is highly dependent on the question type. Although one might expect that people would be more likely to drastically change their opinions on logical questions, supposedly because the questions are self-contained, most violations come from the factual question 5.

Table 9: Violations of the upper and lower bounds

	Session	Period	Subject	violations
1	A	1	10	1
2	A	2	9	2
3	A	3	9	2
4	A	4	5	4
5	A	5	4	4
6	B	3	6	1
7	C	1	3	1
8	C	1	9	2
9	C	4	16	2
10	C	5	9	7
11	C	5	16	7
12	D	1	12	2

4.3 Relation to literature

The closest previous work to ours is Rahwan et al. (2014). The authors found that on “logical” questions from the Cognitive Reflection Test, there is strong evidence of output contagion but no evidence of process contagion. This would suggest that under the conditions of their experiments, “System 2” (analytical reasoning) is not engaged. Our study was not designed to differentiate between output contagion and process contagion. Furthermore we have some doubts about the conclusions of Rahwan et al. (2014). The results on random and clustered topologies in Rahwan et al. (2014) show that random graphs propagate the correct answer much better, consistent with standard infection/information models but not with complex contagions (Centola 2010). Since we believe that in an observational study of this type, simple contagion should not cause output contagion, this is puzzling and deserves further study.

Note that one possible weakness in both studies is the amount of time given to participants to revise their answers. In our experiment, participants were allowed 30 seconds in iterations after the first, while in Rahwan et al. (2014) 15 seconds were allowed. It seems likely that for very small allowed times, only output contagion can occur, because there is not enough time for further thinking. This is another reason for our skepticism above. A further study to investigate the role of time constraints is desirable.

Our study is different from Rahwan et al. (2014) in three key aspects. First, we study how the dynamics of belief propagation depends on the type of question, by using factual questions that are not “logical”. Second, by letting participants not only guess the answer but also admit ignorance we are able to obtain a cleaner view on the propagation of beliefs. Third, we require participants to choose between two options (or admit ignorance), whereas Rahwan et al. (2014) allows free-form answers. Allowing free-form answers can result in many answers, very few of which have substantial support, which may reduce influence from neighbours. On the other hand, for certain types of questions (those where a proposed solution can be verified much more easily than it can be derived, as for the “widgets” question common to our study and Rahwan et al. (2014)), it may be much easier to spread the correct answer via free-form rather than via multiple choice answers (a simple rather than a complex contagion). This may be the reason behind the behaviour on random versus clustered topologies mentioned above. Note that on the one question (widgets) and topology (complete) where the two studies overlap, our study finds a higher proportion of correct answers, and also exhibits nonmonotonicity in this proportion over iterations, showing quite different behaviour from Rahwan et al. (2014). This strengthens our belief in the sensitivity of results to experimental design features.

Another related work is Lorenz et al. (2011) which uses the complete graph and a problem-solving task that asks participants to estimate the value of factual numbers, such as the length of the border between two countries. The study also varied the information given to participants (none, arithmetic mean of neighbours’ guesses, full information about neighbours’ guesses), and found little difference between the last two, but an improvement by both of them in collective accuracy over the first regime.

4.4 Future work

We conjecture that allowing the “don’t know” option allows for better overall social learning. This seems plausible, because in the 2-state model the wrong answer can gain majority support early, and convergence to that answer occurs quickly. However, we have not yet performed a control experiment in which only two answers were allowed, and the presence of "tricky" questions complicates things considerably. This is an obvious area for future work.

Various studies have used different answer formats (multiple choice, free-form) and different question types (single correct answer, optimisation problem). We suspect that these choices have a substantial effect on which type of network topology is best for social learning and aggregation. A study concentrating on the interaction of topology with question/answer format looks promising.

The exploration/exploitation literature deals with optimisation problems and the overall state of knowledge is unclear. Focusing on learning of unique correct answers as in the present paper might be helpful. For example, the idea of what is a "difficult" question in the optimisation framework may have some, or no, relation to what we call "tricky" questions. It may be that we can unearth an underlying difference in types of questions, which may incorporate both global properties of the solution space and psychological and cognitive information about the agents, which explains both types of experiments.

Figure 4 displays for answers 1 and 2 three main small intervals in the x-axis where relatively large jumps in the cdf occur. These are near 0, near 0.5 and near 1. We hypothesize that this indicates three main types of subjects: those with low and high thresholds for influence by neighbours, and those who tend to adopt majority opinion. Figure 2 supports this to some extent, but the data is not clear. Further experimentation to determine whether these types really exist in the population would be very interesting. Such an experiment would probably use the complete graph.

Using experimental data to compare threshold models with other models of diffusion and learning, and even to fit them to real data, is another obvious area to explore. Our results showing that threshold models are not always appropriate in the case of more than two states deserve more scrutiny.

Appendices

A Experiment instructions

Information about the experimental setup.
Together with several other people in the room you are part of a connected network. Some of the links between people are "one-way" while others are "two-way".

Some of you are connected to more people than others (the number of links ranges from 1 to 17). Participation is anonymous so that you will not know the identities of people you are connected to at any moment.

You will be asked several questions in turn. Each question will be asked simultaneously to everybody in the network. You will be asked the same question several times. After each iteration you will receive a summary of answers supplied by your "feeds" (people connected to you in the network). You will also be feeding your answer to the people to whom you are connected (they may be different from people feeding their answers to you because some links are one-way). At each iteration you will have an opportunity to update your answer.

Before every question the positions of people on the network will be changed randomly. Therefore, you may be connected to a different number of people, and to different people, than before.

Your decisions and how you will be paid.
You will receive $5 for participating in this experiment. In addition you can earn money based on your answers. Each question has 2 possible answers, plus a third option "I am not sure".

An incorrect answer is worth 0 (zero) tokens. A correct answer is worth 10 tokens. Choosing "I am not sure" will give you 6 tokens. Not choosing anything will give you 0 (zero) tokens.

For each question, you will receive a payment for your very first answer and for your answer in another randomly chosen iteration. Note that not answering is guaranteed the lowest payment, and choosing an answer randomly has an expected payment of 5 tokens, which is lower than the 6 token payment for "I am not sure".
At the end of the experiment tokens will be converted to [redacted to conceal the national currency] paid in cash privately.

B Decision-making interface

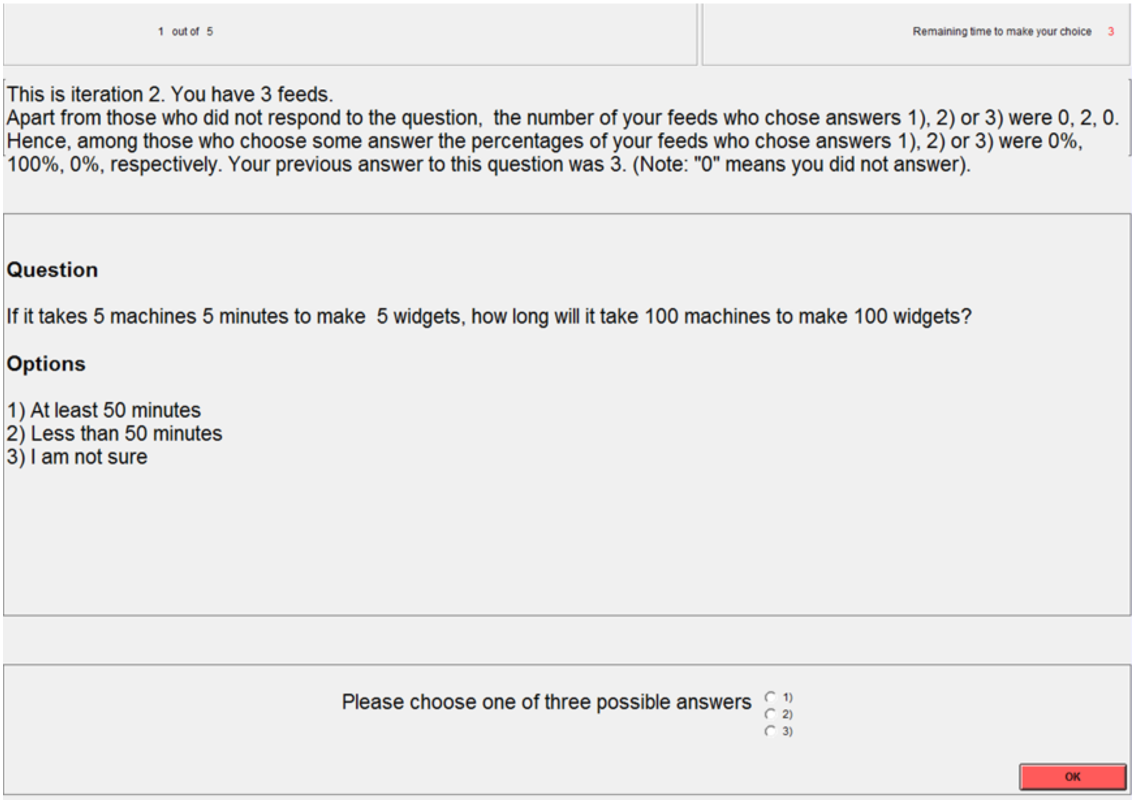


Figure 5: Information provided to every participant included the question, available answers, the participant's previous answer as well as the distribution of answers among the feeding nodes.

C Questions used in the experiment

The specific questions and answers have been chosen to simulate different degrees of knowledge among the participants.

Question 1. If it takes 5 machines 5 minutes to make 5 widgets, how long will it take 100 machines to make 100 widgets?

1. At least 50 minutes
2. Less than 50 minutes
3. I am not sure

Question 2. Suppose you have a set of four cards placed on a table, each of which has a number on one side and a coloured patch on the other side. The visible faces of the cards show 3, 8, red and brown. Which card(s) must you turn over in order to test the truth of the following claim: “if a card shows an even number on one face, then its opposite face is red” ?

1. 8 and brown
2. 8 and red
3. I am not sure

Question 3. The name of the character played by Paul Walker in the “Fast and Furious” movies is:

1. Dominic
2. Brian
3. I am not sure

Question 4. True or false: the Great Wall of China is the only manmade object visible from the Moon.

1. True
2. False
3. I am not sure

Question 5. Does the picture below contain more white or black dots?

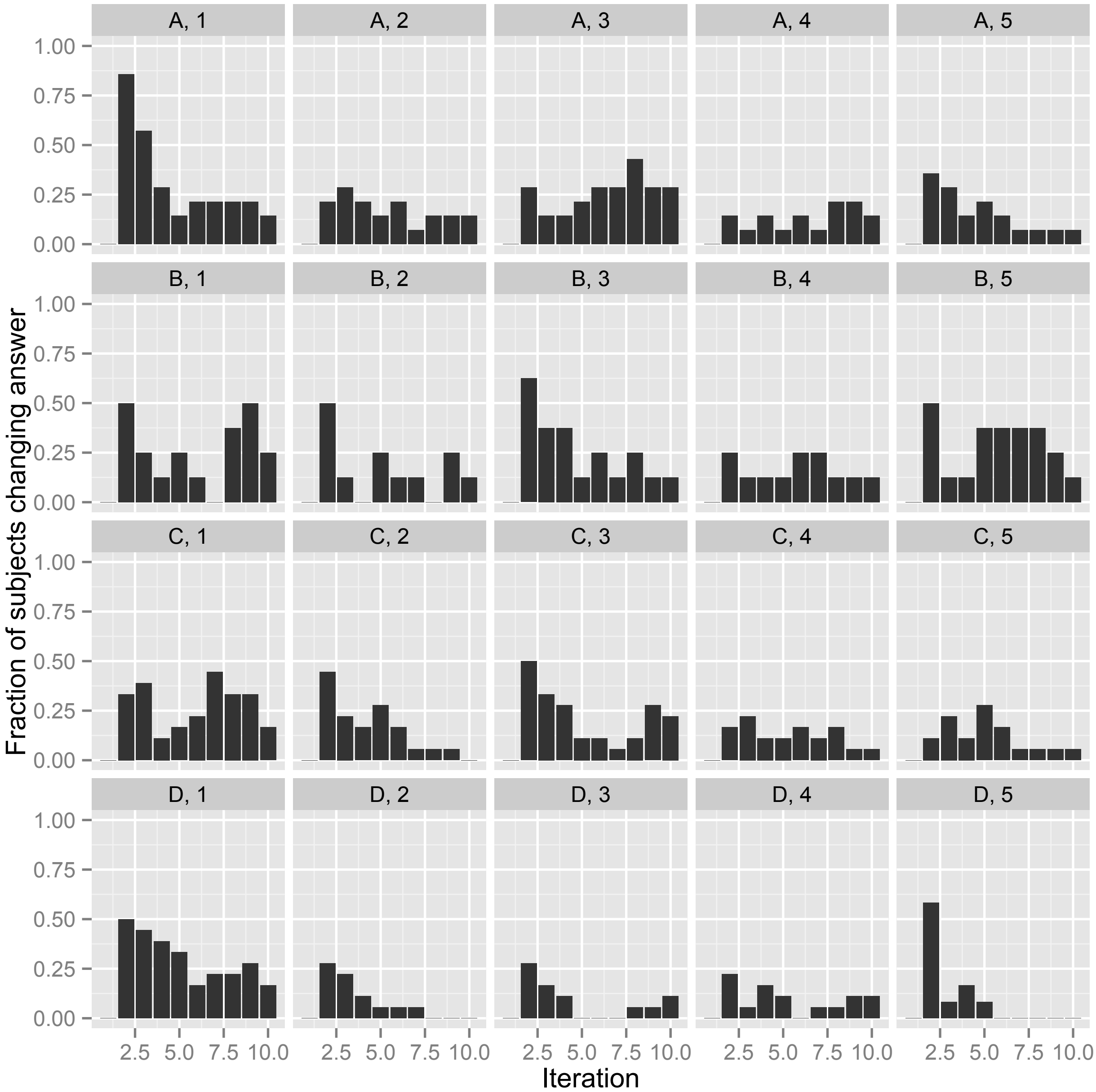
1. More white dots
2. More black dots
3. I am not sure

For Question 5 a picture has been converted to black and white format and adjusted such that the experimenters thought it was impossible to tell whether it had more black or white dots.

D More detailed data analysis

Figure 6 shows for each treatment and question the fraction of subjects who changed their answer, at each iteration. We observe that in most of these cases, convergence (defined as, say, no changes for at least 2 consecutive iterations) was not achieved by our 10 iteration cutoff. However, in many cases by iteration 10 convergence appears to be almost reached. This apparent convergence is much faster for the complete graph than for the other topology, and for the factual questions (especially Q5) than for the logical ones.

Figure 6: Changes by question and treatment



References

Alchourran, C. E., Gärdenfors, P. & Makinson, D. (1985), 'On the logic of theory change: Partial meet contraction and revision functions', *The Journal of Symbolic Logic* **50**(2), 510–530.

Bass, F. M. (1969), 'A new product growth for model consumer durables', *Management Science* **15**(5), pp. 215–227.

Berninghaus, S. K., Ehrhart, K.-M. & Keser, C. (2002), 'Conventions and local interaction structures: Experimental evidence', *Games and Economic Behavior* **39**(2), 177–205.

Bonaccio, S. & Dalal, R. S. (2006), 'Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences', *Organizational Behavior and Human Decision Processes* **101**(2), 127–151.

Bonner, B. L., Baumann, M. R. & Dalal, R. S. (2002), 'The effects of member expertise on group decision-making and performance', *Organizational Behavior and Human Decision Processes* **88**(2), 719–736.

Cassar, A. (2007), 'Coordination and cooperation in local, random and small world networks: Experimental evidence', *Games and Economic Behavior* **58**(2), 209–230.

Centola, D. (2010), 'The spread of behavior in an online social network experiment', *Science* **329**(5996), 1194–1197.

Centola, D. & Macy, M. (2007), 'Complex contagions and the weakness of long ties', *American Journal of Sociology* **113**(3), 702–734.

Chandrasekhar, A., Larreguy, H. & Xandri, J. (2012), 'Testing models of social learning on networks: Evidence from a framed field experiment.', *Working paper.* .

Choi, S., Gale, D. & Kariv, S. (2005), 'Behavioral aspects of learning in social networks: an experimental study', *Advances in Applied Microeconomics* **13**, 25–61.

Choi, S. & Lee, J. (2014), 'Communication, coordination, and networks', *Journal of the European Economic Association* **12**(1), 223–247.

Corbae, D. & Duffy, J. (2008), 'Experiments with network formation', *Games and Economic Behavior* **64**(1), 81–120.

Costanzo, P. R., Reitan, H. T. & Shaw, M. E. (1968), 'Conformity as a function of experimentally induced minority and majority competence', *Psychonomic Science* **10**(10), 329–330.

Cox, J. C. & Hayne, S. C. (2006), 'Barking up the right tree: Are small groups rational agents?', *Experimental Economics* **9**(3), 209–222.

Davis, J. H. (1992), 'Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples 1950äì1990', *Organizational Behavior and Human Decision Processes* **52**(1), 3–38.

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

Feri, F., Irlenbusch, B. & Sutter, M. (2010), 'Efficiency gains from team-based coordinationâîlarge-scale experimental evidence', *American Economic Review* **100**(4), 1892–1912.

Fischbacher, U. (2007), 'z-tree: Zurich toolbox for ready-made economic experiments', *Experimental Economics* **10**(2), 171–178.

Frederick, S. (2005), 'Cognitive reflection and decision making', *The Journal of Economic Perspectives* **19**(4), pp. 25–42.

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

Gärdenfors, P. (1988), *Knowledge in Flux: Modeling the Dynamics of the Epistemic States. Bradford Book*, MIT Press, Cambridge, MA.

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

Granovetter, M. S. (1973), 'The strength of weak ties', *American Journal of Sociology* **78**(6), 1360–1380.

Greiner, B. (n.d.), 'Online recruitment system for economic experiments'. http://www.orsee.org/web/.

Jackson, M. O. (2010), *Social and economic networks*, Princeton University Press.

Jia, P., Mirtabatabaei, A., Friedkin, N. & Bullo, F. (2014), 'Opinion dynamics and the evolution of social power in influence networks', *Working paper.* .

Kermack, W. O. & McKendrick, A. G. (1927), 'A contribution to the mathematical theory of epidemics', *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **115**(772), 700–721.

Kugler, T., Kausel, E. E. & Kocher, M. G. (2012), 'Are groups more rational than individuals? a review of interactive decision making in groups', *Wiley Interdisciplinary Reviews: Cognitive Science* **3**(4), 471–482.

Laughlin, P. R. & Ellis, A. L. (1986), 'Demonstrability and social combination processes on mathematical intellective tasks', *Journal of Experimental Social Psychology* **22**, 177–189.

Laughlin, P. R., Hatch, E. C., Silver, J. S. & Boh, L. (2006), 'Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. [miscellaneous article]', *Journal of Personality* **90**(4), 644–651.

Lazer, D. & Friedman, A. (2007), 'The network structure of exploration and exploitation', *Administrative Science Quarterly* **52**(4), 667–694.

Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. (2011), 'How social influence can undermine the wisdom of crowd effect', *Proceedings of the National Academy of Sciences* **108**(22), 9020–9025.

Mason, W. A., Jones, A. & Goldstone, R. L. (2008), 'Propagation of innovations in networked groups.', *Journal of Experimental Psychology: General* **137**(3), 422.

Mason, W. & Watts, D. J. (2012), ‘Collaborative learning in networks’, *Proceedings of the National Academy of Sciences* **109**(3), 764–769.

Newman, M. (2010), *Networks: an introduction*, Oxford University Press.

Rahwan, I., Krasnoshtan, D., Shariff, A. & Bonnefon, J.-F. (2014), ‘Analytical reasoning task reveals limits of social learning in networks’, *Journal of The Royal Society Interface* **11**(93), 20131211.

Schelling, T. C. (1969), ‘Models of segregation’, *The American Economic Review* **59**(2), 488–493.

Slembeck, T. & Tyran, J.-R. (2004), ‘Do institutions promote rationality?: An experimental study of the three-door anomaly’, *Journal of Economic Behavior & Organization* **54**(3), 337–350.

Wason, P. C. (1968), ‘Reasoning about a rule’, *The Quarterly Journal of Experimental Psychology* **20**(3), 273–281.

Watts, D. & Dodds, P. (2007), ‘Influentials, networks, and public opinion formation’, *Journal of Consumer Research* **34**(4), 441–458.