Networked crowds answer tricky questions poorly

Patrick Girard[a], **Valery Pavlov**[a], **and Mark C. Wilson**[a,1]

[a]University of Auckland, Auckland, New Zealand

This manuscript was compiled on September 28, 2016

We focus on the following basic group decision situation, which we call *iterative distributed jury* (IDJ), a variant of the *Delphi technique*. Group members seek to answer truthfully a question having a well-defined objectively correct answer; they revise answers iteratively; only summary feedback on group members' answers is available at each iteration; individual estimates are aggregated to form a group answer.

Experimental studies of the effectiveness of Delphi-like methods have yielded mixed results. To investigate further, we designed a laboratory multiple choice IDJ experiment having some novel features. One novelty was that we incentivized participants to reveal their ignorance; another is the use of both *logical* and *factual* questions.

We find that, perhaps surprisingly, substantial social influence occurs even in this highly anonymized and information-restricted setting, and even for purely logical questions. Eventual group accuracy is strongly dependent on the *trickiness* (likelihood of being answered confidently but wrongly, a concept distinct from *difficulty*) of the question. Also, the bulk of learning occurs by those who were willing to admit to being undecided. We find that question factors are more important than participant characteristics.

In addition to consequences for the practical use of this group decision method, our quantitative results suggest specific new models of opinion dynamics that deserve detailed study.

social networks | influence | wisdom of crowds | advice taking

1. Introduction

We focus in this paper on situations where opinions held by individuals about objective facts are shared repeatedly and updated. The one-iteration special case has been widely studied under the name "wisdom of crowds", since the description by Galton [1] of estimating the weight of an ox. Its discrete choice version, often modelled by social choice theory, has an even longer history dating back to Condorcet [2]. Positive theoretical results on the accuracy of such group judgments rely on independence or negative correlation of estimates and a good aggregation rule (mean, median, plurality voting, etc) although some conditions can be relaxed [3, 4].

When iteration and feedback is allowed, it is clear theoretically that crowd estimates can converge to very low quality answers (for example, when herding toward a wrong answer occurs). The literature on group decision-making shows that group accuracy may suffer when group members deliberate in an unstructured fashion, for reasons such as groupthink [5] and excessive reliance on common information [6].

More structured iterative methods for improving group estimates include prediction markets and the *Delphi technique* [7]. The advent of the Internet has made it possible to assemble rather large and diverse groups for decision-making. Structured distributed decision-making techniques with limited discussion, such as we study here, seem likely to have more applications in the near future, and hence deserve serious study.

Iterative distributed jury judgments. We focus on the *iterative distributed jury* situation (our term):

- each participant aims to find the true answer to each question asked;
- anonymity of participants is preserved;
- participants iteratively and simultaneously revise answers;
- feedback to participants is controlled by a central agent (in particular, open discussion is not allowed);
- at each iteration, each participant is given statistical feedback about the answers of other participants.

The last four conditions are often used to define the Delphi technique, which has been widely studied since its introduction. The first condition is often implicit, but unless participants are incentivized for individual correctness, for example if they are explicitly rewarded for group correctness, strategic behaviour may lead to unexpected dynamics of answers. The original Delphi techniques implicitly assume that each participant is connected to each other. However, we may also consider situations in which information flow is controlled by a nontrivial network topology (such as an existing social network).

In experimental tests of this group decision-making technique, group members independently answer a question. After receiving feedback on the distribution of answers of the other group members, they have the opportunity to revise their

Significance Statement

We carry out laboratory experiments to test a version of the classic *Delphi method* of group decision-making that appears to be increasingly important in online situations. We find that substantial social influence occurs even for questions with a single correct answer that can be logically deduced. Group decisions can easily be wrong when the questions involved are *tricky* — that is, they suggest a single wrong but compelling answer. Motivated by the prevalence of online *lurkers*, our study incentivized participants to truthfully report "I am not sure". Those who did so initially exhibited more learning than those who confidently gave the wrong answer initially. Confidently given wrong answers were given frequently, social influence exacerbated this, and the wrongness was caused more by group interaction and individual confidence than by individual competence. We advise caution in using such methods for important group decisions.

PG, VP and MW devised the research program containing this article. VP and MW planned the experiments. VP carried out the experiments with assistance from MW and PG. VP and MW carried out the analysis. VP and MW wrote the paper with help from PG. Author names are listed alphabetically.

No conflict of interest.

[1] Corresponding author; E-mail: mc.wilson@auckland.ac.nz

answer. This continues until a pre-specified number of rounds in our experiment; other possible stopping criteria include reaching a pre-specified level of approximation to unanimity, or reaching a point at which agents no longer change their answers.

A few recent experimental studies [8, 9] have used the general setup above. We compare our results with theirs in the Discussion section.

2. Our experiment

We carried out an experimental study of IDJ type. We focus on little-studied details such as the type of question and the answer format, and the confidence of the participants.

Answer format. Unlike the above-mentioned experiments [8, 9], which used free-form answer format, we focused on a multiple choice situation. This was in order to give us more control over the answers supplied. For example, in the worst case using free-form answers, every participant may give a different answer. For non-numerical questions in which no obvious statistical summary other than the mode is useful, feedback would potentially be too time-consuming for participants to read. Another reason is to control communication — if the text of participant answers is readable by other participants, a form of discussion could ensue, and we wished to prevent any discussion.

Most papers using multiple choice in the experimental choice literature allow only two choices. We introduced a third option, rather than forcing participants to choose between two alternatives. We incentivized participants to report truthfully "I am not sure" (rather than "tossing a coin") if they were indifferent between the two alternatives but to answer correctly if they strongly believed that they knew the answer. The major reason for including the third answer option was to allow us to estimate the confidence of participants.

Question type. We designed the suite of questions to contain two types of questions. *Logical* questions are self-contained and can be solved by analytical reasoning. *Factual* questions rely on information gained from experience which may not be available to all participants. Lorenz *et al.* [8] used only factual questions, whereas Rahwan *et al.* [9] used only logical questions.

Our selection of questions was subordinate to the goal of inducing different degrees of confidence in the correct answer (exact wording of the questions is in the Appendix). We used the informal idea of "tricky" and "difficult" questions to guide our selection (we operationalize these concepts in the Results section). By "tricky" we mean that some questions suggest to many people a wrong answer that seems obviously correct — this is distinct from mere hardness and has been heavily studied by psychologists in the framework of *dual process reasoning*. Our two logical questions, the arithmetic "widget" question from Frederick's Cognitive Reflection Test [10] and a standard card variant of the Wason selection task [11] dealing with logical implication, are well-known to fall in this category, the latter being considerably more tricky than the former. One of our factual questions has an answer choice that is a known common misconception, also making it somewhat tricky. By contrast, the other two factual questions we considered to be rather straightforward (difficult, but not tricky). One of them deals with a fact about the world that we expected would be known to some, but not all, of the participants. Furthermore, we expected that those who did not know this fact would realize that many of their peers (who were not explicitly known to them, but obviously shared common demographic characteristics) would know the answer. The remaining factual question was chosen so that no one could know the answer with certainty, and we made it common knowledge that we had communicated the correct answer to someone in the group.

Given the payment structure and the beliefs about other participants we expected participants to form, it is rational for someone not knowing the answer to either of the two factual questions above to answer "I am not sure" until receiving a signal for one or the other answer option, and to then imitate that answer. No wrong answers should be observed under common knowledge of rationality.

Confidence. It seems obvious that if there is any improvement in group accuracy, this occurs through those who do not know the answer learning in some way from feedback (it is not important now whether this learning is simply imitation or the product of better analysis as a result of being confronted with different opinions to one's own). Note that it is possible that by having several iterations worth of time in which to think about the question, participants may improve their answer. Our discussion in Section 6 of the totally disconnected participants shows this to be unlikely.

Those who do know the answer clearly contribute to group accuracy. However, those who do not know the correct answer may either answer incorrectly or answer "I am not sure". We hypothesize that participants who are confident enough in their answers (or perhaps sufficiently risk-loving) that they are willing to give a definite wrong answer before receiving any feedback, when a reasonable payoff is obtained by answering "I am not sure", learn worse than others. Our results as shown in Figure 2 and Table 4 indeed show that most of the contribution to group learning is achieved by those who are willing to admit their own ignorance initially.

3. Details of experimental design

Participants were students of a large public university in Australasia from a variety of majors. Overall, 52 people took part in the study. To recruit participants we used ORSEE [12].

All sessions were conducted in a dedicated computer laboratory designed for running decision-making experiments. Participants were paid a "show-up" fee of $5 and an additional amount of money determined by the number of correct answers and answers "I am not sure" made during the experiment. The average payment, including the show-up fee, was around $20. Prior to 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 completely anonymised, so that no participant could know which answer had been provided by which participant.

To collect the participants' responses we used zTree [13]. The very first screen displayed the experiment instructions (Appendix A). During the experiment participants were presented with five different questions (see Appendix B). On each question they could answer 10 times by choosing (within 90 seconds on the first iteration and 30 seconds on each subsequent one) 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 am not sure”. We randomized the option number, 1 or 2, assigned to the correct answer on each question. Option “I do not know” was always the last choice on the list. At each of the 10 iterations for a given question, subjects were provided the information about their last answer and the distribution of answers given by their neighbors (but no information identifying any specific neighbor). 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 neighbors they had, but nothing that would identify who they were. Each participant was shown the correct answer after answering the question 10 times.

Incentives. Money was the only incentive provided to participants. Participants were paid as follows:

- 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 yields 10 tokens;
- the incorrect answer, or no answer, yields 0 tokens;
- answering “I am not sure’ yields 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. An alternative procedure used by some researchers is to pay for every correct answer, rather than a randomly chosen one. We chose random payments in order to avoid “portfolio effects” [14].

4. Analysis of experimental results

Each data point consists of a participant, an experimental treatment, a question, an iteration number, and an answer. We used the following basic statistical measures. For a given question $q \in \{1, \ldots, 5\}$ and iteration $i \in \{1, \ldots, 10\}$, we denote the fraction of participants answering at iteration i for question q correctly C_{qi}, answering “I do not know” U_{qi}, and answering incorrectly I_{qi}. Dropping a subscript means we are averaging over all values of that subscript, so that, for example, C is the fraction of correct answers given in the entire experiment. We define the *group learning* $L_q := C_{q,10} - C_{q,1}$ on a question q to be the difference between final and initial group correctness.

Similarly, for a given participant s we can define the *participant correctness* C^s as the fraction of correct answers given by s over all iterations of all questions. The *participant changeability* Δ^s is the fraction of available iterations in which s changed their previous answer. Note that the first iteration is never available, so each participant had 45 chances to change opinion (9 iterations for each of 5 questions).

For the data analysis we used the standard software R. The key descriptive statistics and different model estimations are provided in Appendix C. With one exception, we use multinomial logit, fixed effects models.

Basic properties of participants. We find participants to have been generally engaged in the task, as described in Table 2. More than half of participants never abstained. Only one participant had abstention rate higher than 8% (this rate was 78%). In all, only 72 of our 2600 data points involved abstention; we eliminated these data points in the analysis below. Participants changed their previous answer (not including abstentions) on average about one-quarter of the time.

We find substantial variation in participant answer scores. However the data does not support the hypothesis that we have substantial groups of either extremely high skill or extremely low skill participants. Only 4 of 52 participants (8%) never gave a wrong answer. Furthermore all but 4 participants (92%) answered “I am not sure” at least once.

Difficulty and trickiness of questions. While the difference between logical and factual questions is obvious, it is not clear how to define *a priori* whether a question is “tricky” or “difficult”. We decided to define these as continuous variables, and in terms of the answers the question received. Specifically, we define the **trickiness** of a question to be the fraction of participants answering *incorrectly at the first iteration*, $T_q := I_{q1}$. We also define the **difficulty** of a question as the fraction of participants *not answering correctly at the first iteration*, $D_q := I_{q1} + U_{q1}$. By definition, difficulty is always at least as large as trickiness. However Table 3 shows clearly that the two measure very different things. Note that in the two-choice case where “I am not sure” is not an option, trickiness and difficulty cannot be distinguished simply by looking at answers.

We also include a slight reformulation which may be helpful. Define **perceived difficulty** to be U_{q1}. The trickiness is then the difference between difficulty and perceived difficulty.

Our predictions that the factual questions Q3 and Q5 would have very low trickiness but not low difficulty were correct. Also, we predicted that the Wason task Q2 would have the highest trickiness score, as it did. Note that Q5 and Q2 have very similar difficulty but very different trickiness. Trickiness and difficulty as measured with aggregated data did not vary much when we broke the data down by experimental session.

Dynamics of answers. Figure 1 shows the dynamics of group correctness C_{qi}. The analogous figures for I_{qi} and U_{qi} also show substantial difference between questions consistent with this.

We separated the data into subsets corresponding to subjects who answered correctly, incorrectly, or “I am not sure” on the first iteration of the question, and computed analogues of the results the previous subsection for each of these three subgroups. The trajectories for correctness varied strongly between these types as shown in Figure 2. The group correctness for the subgroup of participants answering the first iteration of a given question correctly was about 90% for all subsequent iterations of that question. In contrast the group correctness for the subgroups who answered “I am not sure” (respectively incorrectly) at the first iteration increased to around 50% (respectively 25%) by the final iteration.

Clearly, the members of the initially undecided group exhibited the most learning overall. Table 4 presents an OLS model used to estimate the impact of the initial answer on the final correctness, and this shows similar effect sizes.

Fig. 1. Group correctness by question and iteration

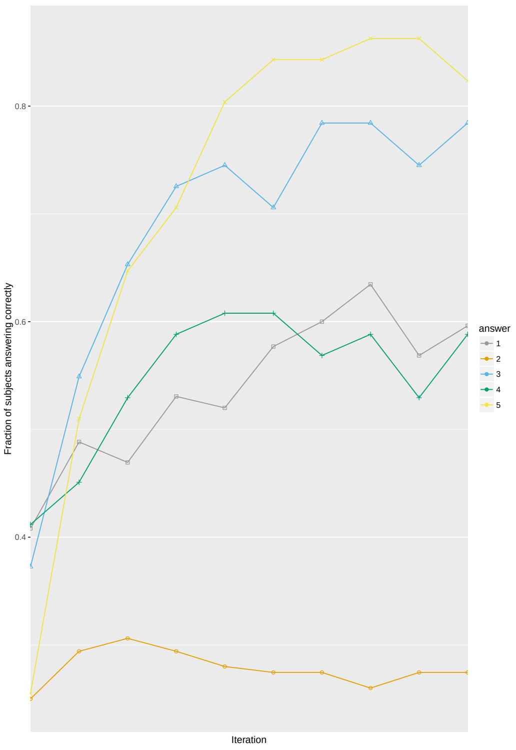
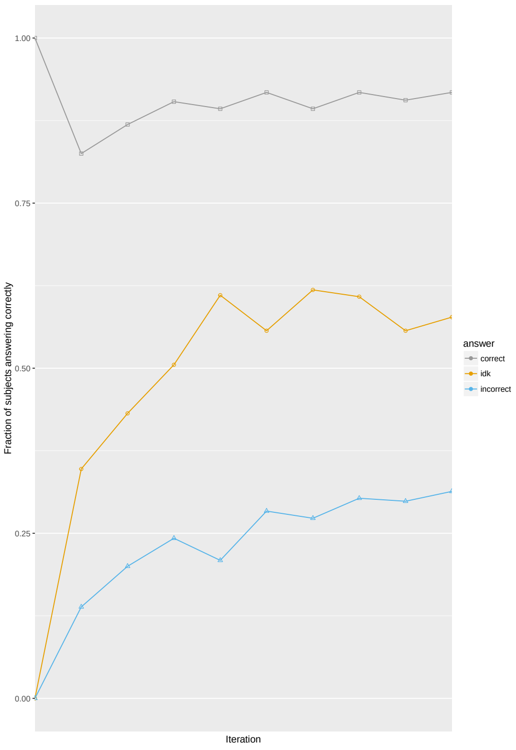


Fig. 2. Group correctness by initial answer type and iteration



5. Further discussion

Robustness checks. We segmented the data according to various parameters. For example, we performed the analyses listed above for each network topology separately, and for each experimental session separately. Results were broadly in line with the overall results reported here, with more noise, as expected. Overall, our data appears to have a "healthy" degree of variability showing that there are neither any influential outliers nor artifacts due to some peculiarity in the experiment design.

Related work. The Delphi method, developed at RAND Corporation in the 1950s and first discussed in the research literature in the 1960s [15], incorporated aspects of what we have termed IDJ methodology. The main difference is that in early implementations of the method, participants were allowed to request information from the central controller. The method has been used extensively for questions for which the correct answer cannot be known (for example, forecasting), and retains popularity more than 50 years after its first use. Several studies in the 1960s by Dalkey and collaborators [7, 16–18] investigated the effect on group accuracy of various adjustments to the procedure (such as varying the presentation of statistical feedback).

More recently, a few studies have been carried out in the "wisdom of crowds" literature. Lorenz, Rauhut, Schweitzer and Helbing [8] carried out an experiment of the IDJ type involving factual questions (such as the number of murders in Switzerland in a given year) having numerical answers. They report that providing more information to group members about their fellow members' answers, and allowing them to revise their answer, led to a reduced diversity of answers, and higher confidence by those answering, but less overall accuracy. This interpretation has been criticized [19] and a similar study using different statistical tests showed that increasing feedback led to better group decisions [20].

Rahwan *et al.* [9], with a similar experiment involving logical questions having numerical answers, reported that group correctness is enhanced by higher connectivity (which allows for more information sharing). They used 5 groups of 20 participants, each group having a different network topology. They used 7 questions in total, but reported in detail only on the 3 questions comprising the CRT. They also used 4 numerical questions from the Berlin Numeracy Test [21], which they judged to be either too easy or too hard to allow comparison between the different topologies they studied.

The more general issue of exploration of possibilities versus exploitation of group knowledge, and its effect on solution quality, has been addressed in a stream of work, both experimental and theoretical [22, 23], with similarly conflicting results. Lazer and Friedman [22] showed via simulation that for hard optimization tasks, topologies promoting rapid information flow may, by promoting premature convergence, lead to suboptimal group solutions. However, there is no such difficulty for easy tasks. Mason & Watts [23] performed online experiments which led to the exact opposite conclusion: efficient networks produced better collective solutions.

Trickiness. In our multiple choice format, questions with low difficulty (and hence low trickiness) present little problem — very little social learning is required for the group to make a correct judgment. Questions with high difficulty, but low trickiness, also present little problem — substantial group learning occurs, leading to eventual group correctness. However, questions with high enough trickiness (which still may have relatively low difficulty) may present serious problems for group performance.

This has implications not only in our multiple-choice setup but also for free-form answer situations. For example, there are many wrong answers to most questions, and if the question is not tricky we expect a relatively large set of wrong answers to be given, each with low frequency. In this case it is relatively easy for the correct answer to stand out, and convergence to it by the group is more likely. However for tricky questions, a single compelling but wrong answer can presumably skew the group answer dynamics.

A new operationalization of trickiness is required for free-form answers to logical questions; an obvious choice is the frequency of the most common wrong answer. The trickiness by our above definition of the three CRT questions used by Rahwan *et al.* was respectively 0.31, 0.57, 0.40 (note that the frequency of the correct answer was respectively 0.25, 0.20, 0.28). Our Q1 (the “widgets” question) was also used by Rahwan *et al.* The focal wrong answer (namely 100) is also clearly the modal answer in applications of the CRT by Frederick and coauthors (S. Frederick, personal communication). Thus for free-form questions where we have reason to believe that a common reasoning error will occur, the concept of trickiness makes sense.

For free-form answers where the range of plausible answers is much larger, for example almanac-type questions as used by Lorenz *et al.*, the concept of trickiness will require further refinement, because almost every answer will occur only once at the first iteration. We believe that questions where there is a common misconception of the order of magnitude of the answer, for example, should lead to similar phenomena. In general, however, we believe that questions of this type will not have a high trickiness score, however measured.

Questions versus participants. Tricky questions are by definition tricky because a large fraction of people confidently give a wrong answer. An obvious hypothesis about our experimental data is that the same people are being tricked on each question. This is related to the well-known *Dunning-Kruger effect* [24] which describes the general phenomenon of overconfidence of low skill participants and underconfidence of high skill participants.

In order to test this, we define a participant to be *tricked* if they gave an incorrect answer on the first iteration of a question, and define their *trickability* to be the fraction of times this occurred. Note that no subject had trickability of more than 0.6, while only 10 of the 52 subjects had trickability 0. Defining “low skill” to mean “tricked 2 times on the two logical questions”, “medium skill” to mean “tricked 1 time on the two logical questions”, and “high skill” to mean “tricked 0 times on the two logical questions”, we reanalysed the data for these three subgroups of participants (having respectively 7,24,14 members — 7 participants abstained). The mean group correctness for the low skill group was slightly higher than that for the high skill group, and various other analyses all failed to find any effect. We conclude that for this type of participant and experiment, at least, different participants are tricked by different questions, and then find it relatively hard to learn thereafter, even if their overall unaided performance is good and the correct answer is available among their neighbours.

Demonstrability. Group decision-making research has discovered that for certain tasks (such as mathematical and logical reasoning), a single correct answer from a group member rapidly propagates to the group (“truth wins"), whereas for other tasks (such as world knowledge) this is generally not so. This is usually explained in terms of *demonstrability* [25]. Our experimental setup promoted low demonstrability, because there was no direct communication allowed and the multiple choice answer format allowed for less precise answers in some questions.

Logical questions have higher demonstrability than factual ones in our setting (although both have rather low demonstrability). When confronted with the fact that others have supplied a different answer, a participant may be able to corroborate that answer by reasoning if the question is logical, whereas there is no way to corroborate it (other than searching one’s memory more thoroughly) if the question is factual. This gives a reason to expect more influence, rather than less, on logical questions than on factual ones. Thus we have two conflicting forces: logical questions are more demonstrable, but there is no compelling reason to use others’ information.

Our results clearly show (see Table 6 that there is no major difference in how much influence operates for logical and factual questions, and that influence is substantial.

Topology. While topology was not the main focus of our study, we did use two very different topologies which allows for comparison. We used the usual complete undirected graph in two sessions accounting for 30 participants, and for the other two sessions (accounting for 22 participants) a novel directed (“spiral”) topology designed to create heterogeneity in node degrees. When using the spiral topology on the trickiest (Wason) question, eventual group accuracy ($C_{10} =$) is lower than for the complete topology ($C_{10} =$), but higher for all other questions. More importantly, group learning was higher in all questions on the complete topology. This aggregate picture persists when individual experimental sessions are investigated. We interpret this as supporting the general conclusion of Watts & Mason rather than that of Lazer & Friedmanin our setting, but the question definitely deserves further investigation. It is not completely clear whether trickiness or difficulty is more relevant when classifying a problem as “hard” in this setup.

6. Conclusions

Implications for use of Delphi-like techniques. The clear negative effect of question trickiness on group performance shows that the Delphi approach may fail drastically in some situations. Although preventing premature convergence to low quality solutions by means of restricting information flow is a prominent theme in the literature, we hypothesize that for tricky enough questions, lack of discussion reduces demonstrability and increases the chance of the group converging to a bad answer. How to determine *a priori* how to balance these competing considerations is not clear to us. Of course, we did not perform an experiment in which we varied the level of feedback, and this is an obvious candidate for future work. Note that when discussion among group members is allowed, the performance of teams has been shown to be superior to individuals in the Wason task (see [26] and references therein).

The efficacy of the Delphi method has been criticized from several directions [27], including anonymity and iteration leading to low engagement and low quality answers, and any increase in accuracy being caused by greater reflection and pressure to conformity rather than improved reasoning caused by feedback. To our knowledge, our work here is the first to point out that certain types of (tricky) questions having definite answers may be inappropriate for this method.

Implications for modeling. 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 (or worsening) over time. Also, any models predicting rapid convergence to unanimity, such as standard infection models, are clearly inconsistent with our data (note that most experimental work of this kind uses 5 iterations per question, while we used 10).

Our results suggest the following basic prediction which is particularly appealing for logical questions: those who are correct at the first iteration remain correct; those who are incorrect remain incorrect; those who are undecided eventually follow whichever of the above two groups is larger. This would predict, for example, that the eventual group correctness scores in Table 3 would be respectively 0.65, 0.25, 0.92, 0.73, 0.74.

Rahwan *et al.* discuss two theories of how social influence may operate in such situations, which they call *processing contagion* (improved reasoning) and *output contagion* (simple imitation of answers). They conclude that output contagion explains their data much better than processing contagion. Our results seem consistent with this, although because of our study design we cannot make a definitive statement. Although Rahwan *et al.* claim this as a key result of their paper, it seems completely obvious. A time of 15–30 seconds for participants to read the summary of responses of neighbors on the previous iteration, and then revise their answers, is unlikely to allow for detailed reflection. Experiments in the psychological literature confirm this expectation [28]. Experience with teaching mathematical topics to undergraduates leads us to believe that processing contagion can only take place over a period of days rather than minutes. Note that Maciejovsky *et al.* [26] report lasting improvements in performance by team members for Wason-like tasks in a setup where team members confer.

A simple hypothesis is that there is no influence on participants' answers from information about their neighbours' answers. For example, two of the questions were purely logical and self-contained, so there was no need to take account of the answers of other group members. However, this hypothesis does not explain the data well. Table 5 shows substantial correlation of answers with answers given by neighbors at the previous iteration, no matter what the initial answer of the participant. Similar analysis holds when we restrict only to logical questions. Thus models of social influence should be highly relevant to the further study of learning in such situations, and we intend to study these in detail in future work. Preliminary analysis shows that in our experiment the dynamics of correctness are controlled largely by the difference $C_i - I_i$ between group correctness and incorrectness, and also the group undecidedness U_i. The latter is inversely related to the rate of convergence. Table 6 supports this. The sign of the former quantity determines whether eventual group correctness is reached, while its magnitude controls convergence. This suggests several models worthy of analysis. For example, models of the threshold type, or probabilistic models based on Markov chains, seem very natural here. We are aware of very little theoretical work on models of belief change with more than two states (see [29]), and none with asymmetry between the states, whereas the "I am not sure" state is clearly different from the other two in our setup.

Following the basic principle of explicitly providing participants with as much information as possible subject to the requirements of the experiment, we supplied complete information on how many neighbours gave each of answer option 1, 2, or 3 ("I am not sure"). In real networks this third option is usually unreported, and those who only view discussion but do not provide their opinion ("lurkers") may be very numerous. Further experiments to determine whether or not reporting the numbers of undecided neighbors makes a difference to dynamics (and how different it is to the case where everyone must vote and option 3 is not allowed) are desirable.

Methodological issues. The controlled laboratory experiment with undergraduate subjects has been widely used in social sciences. For the kinds of experiments discussed here, we may need to develop different methodology. Each network experiment is in some sense a single data point, and generating enough data points for strong statistical results in this framework is costly. Results are very variable and sensitive to initial conditions and apparently insignificant experimental design choices. As an example, consider the difference in results obtained by our experiment and that of Rahwan *et al.* on the CRT "widgets" question. In each of our experimental sessions, whichever of the two definite answers was more frequent initially became the group decision eventually. By contrast, in the other experiment, in each session the modal wrong answer was initially several times more common than the correct answer, and yet convergence to the correct answer occurred rapidly in 3 of the 4 connected topologies used. One possible explanation is the answer format which causes differences in demonstrability. Questions such as this one are presumably easier to solve when given an exact numerical answer, rather than a choice of two intervals as in our experiment, because verifying an equality is much easier than verifying an inequality. Of course our multiple choice format could be used with a larger number of options, corresponding to commonly given wrong answers, and this would make for an interesting experiment.

Looking at the results for control groups in which there is no feedback information (the totally disconnected topology) we see a large difference between the results of Rahwan *et al.* [9] and Lorenz *et al.* [8]. In the former case, participants typically made few changes to their first answer over the 5 iterations, but in the latter case large numbers of changes, involving large percentage changes in the supplied answer, were typical. We did not include a control group. Although our spiral topology yielded participants receiving no information from neighbors, there were too few of them to allow for meaningful analysis.

1. Galton F (1907) Vox populi (the wisdom of crowds). *Nature* 75:450–451.
2. de Condorcet N (1995) Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. 1785. Facsimile reprint of original published in Paris, 1972, by the Imprimerie Royale. English translation appears in I. McLean and A. Urken. *Classics of Social Choice* pp. 91–112.
3. Kanazawa S (1999) Using laboratory experiments to test theories of corporate behavior. *Rationality and Society* 11(4):443–461.
4. Davis-Stober CP, Budescu DV, Dana J, Broomell SB (2014) When is a crowd wise? *Decision* 1(2):79.
5. Janis IL (1972) Victims of groupthink: a psychological study of foreign-policy decisions and fiascoes.
6. Stasser G, Titus W (1985) Pooling of unshared information in group decision making: Bi-

ased information sampling during discussion. *Journal of personality and social psychology* 48(6):1467.

7. Dalkey NC, Brown BB, Cochran S (1969) The Delphi method: An experimental study of group opinion, Technical report.
8. 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.
9. Rahwan I, Krasnoshtan D, Shariff A, Bonnefon JF (2014) Analytical reasoning task reveals limits of social learning in networks. *Journal of The Royal Society Interface* 11(93):20131211.
10. Frederick S (2005) Cognitive reflection and decision making. *The Journal of Economic Perspectives* 19(4):pp. 25–42.
11. Wason PC (1968) Reasoning about a rule. *The Quarterly Journal of Experimental Psychology* 20(3):273–281.
12. Greiner B (2015) Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association* 1(1):114–125.
13. Fischbacher U (2007) z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2):171–178.
14. Cox JC (2010) Some issues of methods, theories, and experimental designs. *Journal of economic behavior & organization* 73(1):24–28.
15. Dalkey N, Helmer O (1963) An experimental application of the Delphi method to the use of experts. *Management science* 9(3):458–467.
16. Brown B, Cochran SW, Dalkey NC (1969) The Delphi Method, II, Technical report.
17. Dalkey N, Brown B, Cochran S (1969) The Delphi method, III: Use of self-ratings to improve group estimates, (DTIC Document), Technical report.
18. Dalkey NC, Brown BB, Cochran SW (1970) The Delphi Method, IV, Technical report.
19. Farrell S (2011) Social influence benefits the wisdom of individuals in the crowd. *Proceedings of the National Academy of Sciences* 108(36):E625–E625.
20. Gũrçay B, Mellers BA, Baron J (2015) The power of social influence on estimation accuracy. *Journal of Behavioral Decision Making* 28(3):250–261.
21. Cokely ET, Galesic M, Schulz E, Ghazal S, Garcia-Retamero R (2012) Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making* 7(1):25.
22. Lazer D, Friedman A (2007) The network structure of exploration and exploitation. *Administrative Science Quarterly* 52(4):667–694.
23. Mason W, Watts DJ (2012) Collaborative learning in networks. *Proceedings of the National Academy of Sciences* 109(3):764–769.
24. Dunning D (2015) On Identifying Human Capital: Flawed Knowledge Leads to Faulty Judgments of Expertise by Individuals and Groups in *Advances in Group Processes.* (Emerald Group Publishing Limited), pp. 149–176.
25. Laughlin PR, Ellis AL (1986) Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology* 22(3):177–189.
26. Maciejovsky B, Sutter M, Budescu DV, Bernau P (2013) Teams Make You Smarter: How Exposure to Teams Improves Individual Decisions in Probability and Reasoning Tasks. *Management Science* 59(6):1255–1270.
27. Woudenberg F (1991) An evaluation of Delphi. *Technological forecasting and social change* 40(2):131–150.
28. Evans JSB, Curtis-Holmes J (2005) Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning* 11(4):382–389.
29. Mossel E, Neeman J, Tamuz O (2014) Majority dynamics and aggregation of information in social networks. *Autonomous Agents and Multi-Agent Systems* 28(3):408–429.

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. 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*

Table 1. Treatment parameters

ID	Treatment date	Topology	Participants	Edges
A	140822_1152	directed	14	61*
B	140910_1337	directed	8	22*
C	141001_1308	complete	18	153**
D	141002_1255	complete	12	66**
Note:	* - unidirectional; ** - bidirectional			

Table 2. Participant characteristics — overall statistics

Quantity	min	median	mean	sd	max
Abstention	0%	0%	3%	11%	78%
Correctness	18%	55%	54%	17%	94%
Incorrectness	0%	23%	24%	13%	48%
Undecidedness	0%	20%	19%	13%	50%
Changeability	7%	27%	29%	12%	64%
Trickability	0%	20%	26%	18%	60%

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.

C. Statistics, models and results

Table 3. Question characteristics — overall statistics

Question	Type	T	D	C_{10}	L
1	logical	0.35	0.59	0.60	0.19
2	logical	0.46	0.75	0.27	0.02
3	factual	0.08	0.63	0.78	0.41
4	factual	0.27	0.59	0.59	0.18
5	factual	0.16	0.75	0.82	0.57

Table 4. Those who were IDK improved more than incorrect ones

	Dependent variable:
	correct (at the last iteration)
firstAnsweridk	−0.34***
firstAnswerincorrect	−0.61***
Constant	0.92***
Observations	256
Adjusted R^2	0.24
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5. Factors affecting correctness, by participant's first answer

	right (1), wrong (2) and IDK (3)		
	correct		
	(1)	(2)	(3)
trickiness	−1.83**	−1.91***	−2.38**
propFeedsCorrect	0.35	3.87***	5.60***
propFeedsWrong	−3.79***	−0.32	0.44
propFeedsIDK	−1.98***	1.55***	2.08***
Constant	5.28***	−2.66***	−3.32***
Observations	2,366	1,880	928
Log Likelihood	−529.82	−742.87	−415.10
Akaike Inf. Crit.	1,071.64	1,497.73	842.19
Bayesian Inf. Crit.	1,106.25	1,530.97	871.19
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 6. The difference $C - I$ and the previous answer predict a participant's next answer fairly well; whether the question is logical or factual is irrelevant.

	Dependent variable:	
	a1	a2
	(1)	(2)
diff12	0.91***	−1.01***
p3	−0.74*	0.07
previousAnswer1	1.69***	−2.62***
previousAnswer2	−2.73***	1.64***
previousAnswer3	−1.40***	−1.62***
logical	−0.04	−0.07
Akaike Inf. Crit.	581.44	608.16
Note:	*p<0.1; **p<0.05; ***p<0.01	