Algorithmically Mediating Communication to Enhance Collective Decision-making in Online Social Networks

Burton, Jason W.; Almaatouq, Abdullah; Rahimian, M. Amin; Hahn, Ulrike

Algorithmically mediating communication to enhance collective decision-making in online social networks

Jason W Burton (ORCID)
Department of Digitalization, Copenhagen Business School, Frederiksberg, Denmark

Abdullah Almaatouq (ORCID)
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA

M Amin Rahimian (ORCID)
Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, PA, USA

Ulrike Hahn
Department of Psychological Sciences, Birkbeck, University of London, London, UK

Abstract
Digitally enabled means for judgment aggregation have renewed interest in "wisdom of the crowd" effects and kick-started collective intelligence design as an emerging field in the cognitive and computational sciences. A keenly debated question here is whether social influence helps or hinders collective accuracy on estimation tasks, with recent results on the role of network structure hinting at a reconciliation of seemingly contradictory past results. Yet, despite a growing body of literature linking social network structure and collective accuracy, strategies for exploiting network structure to harness crowd wisdom are underexplored. We introduce one such strategy: *rewiring algorithms* that dynamically manipulate the structure of communicating social networks. Through agent-based simulations and an online multiplayer experiment, we provide a proof of concept showing how rewiring algorithms can increase the accuracy of collective estimations—even in the absence of knowledge of the ground truth. However, we also find that the algorithms' effects are contingent on the distribution of estimates initially held by individuals before communication occurs.

CCS Concepts: •Human-centered computing → Collaborative and social computing• Applied computing → Psychology.

Keywords
collective intelligence, decision-making, wisdom of crowds, social networks

Corresponding author:
Jason W Burton, Department of Digitalization, Copenhagen Business School, Howitzvej 60, Frederiksberg 2000, Denmark.
Email: jb.digi@cbs.dk

Data Availability Statement included at the end of the article

Significance statement

Many collective decision-making contexts involve communication among individual group members, such as business executives making an investment forecast or a local community choosing whether to endorse proposed legislation. Sometimes, this communication helps the collective reach an accurate decision because it allows individuals to gain otherwise unknown information from their peers; other times, this communication is bad because it gives rise to detrimental social influence or "groupthink." Building on recent results linking communication, social network structure, and collective accuracy, we developed and tested different rewiring algorithms—programmable rules for manipulating who communicates with whom in online social networks—as a way of steering communicating groups towards more accurate decisions when the ground truth is unknown. Our results show that rewiring algorithms affect collective accuracy in different ways depending on a group's distribution of individual beliefs and the decision task it faces, ultimately providing a proof of concept for using rewiring algorithms as tools for collective intelligence design.

Introduction

Researchers have long demonstrated so-called "wisdom of the crowd" effects, wherein the collective judgment of a group is more accurate than the judgments of individual experts or the individual group members themselves (Condorcet, 1785; Galton, 1907; Grofman et al., 1983; Surowiecki, 2005). Yet, with new digitally enabled means for judgment aggregation giving rise to modern applications such as online prediction markets (Arrow et al., 2008; Wolfers and Zitzewitz 2004), crowdsourcing (Howe, 2006), and digital democracy (Morgan, 2014; Simon et al., 2017), the impetus for crowd-wisdom research has been rejuvenated. Thanks to successes of these applications, there is now an emerging field in the cognitive and computational sciences dedicated to *collective intelligence design*, whereby digital tools (e.g., algorithms, artificial intelligence, and collaborative interfaces) are developed so as to effectively extract wisdom from ever-present crowds (Mulgan 2018). In this paper, we draw from existing literature on wisdom of the crowd effects and propose a new tool, familiar from network science, that can enhance the accuracy of collective decisions in the absence of ground truth knowledge: rewiring algorithms.

Social influence, network structure, and collective estimation

The earliest results on wisdom of the crowd effects in collective estimation tasks assumed that individuals' judgments are made independently, meaning that their errors are likely to be uncorrelated and thus cancel out in aggregate (Condorcet, 1785). However, this independence assumption often goes unmet in the real world because people communicate with or otherwise influence one another. Following this line of thought, past research on the effects of social influence in collective estimation tasks has

produced seemingly contradictory findings. On the one hand, there is evidence that social influence indeed undermines crowd wisdom by causing individuals' judgments to become correlated (Hahn et al., 2019; Lorenz et al., 2011; Muchnik et al., 2013), while on the other, there are studies that report an increase in collective accuracy following social influence (Almaatouq et al., 2020; Becker et al., 2017, 2019; Gürçay et al., 2015).

Formal results that incorporate the possibility of non-independence provide a potential explanation of these seeming contradictions (e.g., Ladha, 1992; Page, 2008). That is, social influence is neither inherently beneficial nor inherently detrimental to crowd wisdom; rather, its effects depend on whether the benefits of communication to individual accuracy outweigh the detrimental effects of non-independence on collective accuracy. The logic of this is made clear in the Diversity Prediction Theorem (Page, 2008):

$$(\bar{x} - \theta)^2 = \frac{\sum_{i=1}^n (x_i - \theta)^2}{n} - \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \qquad (1)$$

where n is the number of group members, x_i is the estimate of the group member i, $\bar{x}$ is the group's mean estimate, and θ is the true value of whatever is being estimated. Put simply, the theorem states that collective error squared is the difference between the average individual error squared and the diversity of the individuals' judgments. While providing a mathematical guarantee that the collective estimate will always be more accurate, in terms of error squared, than the average individual's as long as there exists some diversity in the group, this theorem formalizes how social influence can be both good for collective accuracy (if it leads to an increase in average individual accuracy) and bad (if it leads to too much of a decrease in diversity). Whether social influence will increase or decrease collective accuracy for any given group thus depends on which one of these effects is

greater. Although changes in average individual accuracy and diversity are not directly dependent on one another (i.e., it is possible to observe simultaneous increases in average individual accuracy and diversity), "a trade-off does exist in the necessity of these characteristics for an accurate crowd" (Hong and Page, 2012, p. 57). Indeed, this relationship between diversity and judgmental accuracy also emerges in other formulations of crowd wisdom, such as versions of Condorcet's (1785) Jury Theorem that allow for variation in individual competence and dependence between judgments (Ladha, 1992, 1993), or in correlation-based evaluations of individual and collective accuracy (Hahn, 2022; Hogarth 1978).

To provide predictions for when groups will benefit from social influence, recent research has turned towards studying how different social network structures affect collective accuracy (e.g., Almaatouq et al., 2020; Becker et al., 2017; Hahn et al., 2018a, 2019; Jönsson et al., 2015). Since social network structures delineate the paths through which social influence can be exerted in a group, it follows that different structural characteristics will feature in determining whether the net effect of social influence will be beneficial for collective accuracy. For example, high levels of connectivity and free-flowing information can lead to a homogenization of individuals' beliefs that harms collective accuracy (Jönsson et al., 2015), high levels of centralization can lead to certain individuals wielding excessive influence over the network (Becker et al., 2017), and a lack of structural plasticity can prevent networks from effectively responding to feedback about individuals' performance (Almaatouq et al., 2020).

Rewiring algorithms for collective accuracy

A reading of the literature linking network structure and collective accuracy raises the question: can we build optimal social network structures for eliciting the wisdom of the crowd? Despite the abundance of knowledge on the relationship between network structure and collective accuracy, strategies for exploiting network structure to increase collective accuracy remain underexplored. While there may be considerable difficulties in manipulating the structure of social networks in the analog world, the programmability of the digital world provides new opportunities. In fact, the structure of online social networks (e.g., Facebook and X/Twitter) are already being manipulated through opaque recommender systems and promoted content. However, this manipulation has generally been conducted for commercial, revenue-generating interests rather than social good.

We propose that—just as algorithms have already been used to mediate the information presented to online social networks (Lazer, 2015) and to identify influential nodes in social networks (Wei et al., 2018)—it seems plausible that algorithms could be used to rewire the structure of online

social networks to boost the wisdom of crowds. Specifically, we explore the viability of *rewiring algorithms*—programmable rules for manipulating who communicates with whom—as a tool for enhancing the accuracy of collective estimations. We develop and test three candidate algorithms through simulation and experimentally evaluate their effects on the accuracy of collective estimates made by communicating social networks. In addition, we conduct follow-up simulations for the purpose of providing a reconciliation of our initial simulations and our experimental results.

Modeling and simulations

Given the exploratory nature of this work, we first employed agent-based modeling and simulation to operationalize the parameter space and prototype different algorithm designs. Our modeling framework uses networks of 16 simulated agents who are tasked with judging the probability of a single binary hypothesis (i.e., each agent can favor either 0 or 1, with exact beliefs falling between these points). Such judgments readily map to a broad range of real-world scenarios: assessing the truth or falsity of proposition or predicting whether or not a future outcome will occur.

Our model is initiated by first sending "evidence" to each agent, represented as samples from a Bernoulli distribution, which they integrate with a starting prior of 0.5 via Bayes' theorem. This procedure serves to simulate how individuals would have accrued their own independent knowledge on a given topic, rather than entering a discussion with a purely indifferent prior of 0.5. To represent a population of individuals with varying knowledge about the hypothesis at hand, we vary the amount of evidence each agent receives such that some individuals may be more familiar with or knowledgeable about a given hypothesis. We additionally vary the quality of the evidence sent to the agents by introducing two parameters: *sensitivity*, the probability of receiving positive evidence when the hypothesis is true (i.e., the so-called "hit rate" familiar from signal-detection theory), and *specificity*, the probability of receiving negative evidence when the hypothesis is false (i.e., the so-called "correct rejection rate"). These parameters allow us to model "kind" environments where true positive and true negative evidence is prevalent and a majority of the population is already nearly certain of the truth, as well as less favorable environments where true evidence is rare and the beliefs possessed by the population are more widely distributed.

Once their initial estimates are assigned, the agents communicate with one another across a small-world network structure[1] (Watts and Strogatz, 1998) over the course of four discrete time points, $t = 1,2,3,4$. At each time point, each agent i revises their estimate in light of those communicated by their network neighbors according to a DeGroot belief updating rule (Becker et al., 2017):[2]

$$R_{t+1,i} = \alpha_i \times R_{t,i} + (1 - \alpha_i) \times \overline{R}_{t,j \in N_i} \qquad (2)$$

where $R_{t+1,i}$ is the agent's revised estimate following communication, $R_{t,i}$ is the agent's current estimate, $\overline{R}_{t,j \in N_i}$ is the average current estimate of the agent's network neighbors, and α_i and its complement $(1 - \alpha_i)$ represent the weight that agent places on its own estimate versus those of its peers, respectively. Following the empirical analysis of belief revision by Becker et al. (2017), each agent's α at any given time point is determined by the following regression equation:

$$\alpha_i = 0.74 - 0.05\epsilon_i + \mathcal{N} \qquad (3)$$

where ϵ_i is the agent's absolute error and $\mathcal{N}$ is Gaussian noise with $\mu = 0$ and $\sigma = 0.06$. This stochastic process means that there is a modest association ($r \approx 0.21$) between accuracy and resistance to social influence among our agents (Becker et al., 2017). While we use this updating rule because it is grounded in the empirical analysis of Becker et al. (2017), we note that our results are not dependent on the specific correlation between agents' self-weights and accuracy. We observe the same pattern of results, albeit with slightly attenuated effects, when the coefficient in equation (3) is zero, and when the negative sign is revised such that *less* accurate agents are more resistant to social influence (Figure S1).

Network conditions

Of particular interest to the present work is how different network conditions perform in the general modeling framework outlined above. Here we consider collective accuracy in four conditions: static networks (i.e., unchanging network structure) and networks to which we apply one of three candidate rewiring algorithms. For *static* networks (our control condition), the initial small-world network structure does not change and each agent communicates with the same neighbors at each time point. Such static network structures have been the main focus of existing research on social influence and collective intelligence (Becker et al., 2017, 2019; Golub and Jackson, 2010; Hahn et al., 2018a, 2019; Jönsson et al., 2015; Zollman, 2013), and thus provide this exploratory work with a natural starting point of comparison.[3] In our three experimental conditions, we introduce rewiring algorithms that add and/or remove connections between agents at each time point so that certain agents are exposed to the beliefs (estimates) held by certain other agents. We specifically consider three such algorithms: a *mean-extreme* algorithm, a *polarize* algorithm, and a *scheduling* algorithm. Example animations of each network condition can be viewed here: https://tinyurl.com/network-animations.

The mean-extreme algorithm aims to increase the average accuracy of individuals in a network by directing social influence towards individuals with potentially erroneous, outlying estimates. The algorithm first calculates the mean estimate in a network at a given time point and identifies which side of the scale's midpoint (0.5 on a 0–1 probability scale) the network's mean estimate lies. If the network's mean estimate is less than the midpoint, the algorithm identifies the agent with the lowest estimate and adds directed, outgoing ties to the three agents with the highest estimates. If the network's mean estimate is greater than the midpoint, the algorithm identifies the agent with the highest estimate and adds directed, outgoing ties to the three agents with the lowest estimates. This procedure effectively brings the estimates of the outliers closer to the mean.

The polarize algorithm aims to maintain the diversity of estimates in a network and prevent a potentially biasing homogenization. It first identifies the two most extreme agents on either side of the current distribution of estimates (i.e., the agent with the highest estimate and the agent with the lowest estimate) and cuts all incoming ties to these agents so as to preserve their beliefs from social influence (i.e., the agents with the most extreme beliefs cannot observe any other agents). Then, the influence of these extreme agents is increased by granting each of them two directed, outgoing ties to "core" agents. These core agents are the four individuals with the median estimates in the network (e.g., in a 16-agent network, the agent with the lowest estimate receives outgoing ties to the agents with the 7th and 8th lowest estimates, and the agent with the highest estimate receives two outgoing ties to the agents with the 9th and 10th lowest estimates). The net effect of this procedure is that the diversity of beliefs (measured as variance) is increased by ensuring both extreme, "polar" sides of the belief spectrum are heard.

The scheduling algorithm differs from the mean-extreme and polarize algorithms in that it prescribes (or "schedules") a network structure of intermixing dyads, irrespective of individuals' estimates. Specifically, the algorithm pairs agents at each time point such that no agent speaks to the same agent twice, but each individual will, in principle, have the opportunity to be exposed to all the available information in the network by the end of the four rounds of communication. In this way, scheduled networks will have achieved a maximum diversity in interactions—each dyad at each time point will consist of two individuals sharing aggregated information received from individuals in the network that the other has not interacted with; the algorithm prevents any redundant interactions from taking place. However, for this algorithm to function, it assumes that each individual effectively fully communicates all information they possess and fully integrates all information communicated to them by their peer at each time point. This

algorithmic approach offers an alternative for situations where access to individuals' current estimates at each time point are not available.

Simulation results

With the model and rewiring algorithms described above, our simulations proceed along the following steps:

1. Agents' initial estimates are determined according to the Bayesian framework, which produces a distribution of continuous probability estimates between 0 and 1.
2. Agents are then able to observe the estimates of some subset of their peers in the social network.
3. Agents revise their estimates according to the rule given in equation (2) and equation (3).
4. The network is then rewired according to the appropriate algorithm (unless it is a static network).
5. The collective estimate is then calculated as the average probability.
6. The Brier score is then calculated on the collective estimate and used as the metric for error.

Following 500 iterations in nine different information environments (i.e., pairwise combinations of sensitivity = {0.2, 0.4, 0.9} and specificity = {0.2, 0.4, 0.9}) in which four matched networks are simulated (i.e., one of each network condition starting from an identical initial network), we assess collective accuracy by calculating the squared error of the mean estimate post-communication as in the Diversity Prediction Theorem (equation (1)) and consistent with the Brier scoring metric (Brier, 1950; Rufibach, 2010), henceforth referred to as collective error squared (*CES*). The Brier score was invented to assess the performance of meteorological predictions of single events with a binary resolution of 0 or 1, such as "there is a 60% chance of rain tomorrow," just like the task faced by our simulated agents (Brier, 1950). The Brier score is also particularly relevant here because it can be decomposed mathematically into a component that rewards calibration and a component that rewards resolution (i.e., how many different probability levels an agent distinguishes), as shown by Murphy (1973). Nevertheless, further investigation of rewiring algorithms' effects on calibration—the degree to which objective probabilities match subjective ones (Fischhoff et al., 1977; Lichtenstein et al., 1977)—is a potential avenue for continued research.

In addition to *CES*, we also calculate the average individual error squared (*AIES*) and diversity, measured as variance (*VAR*), present in each network as a way of better understanding each algorithm's effects in the context of the Diversity Prediction Theorem (equation 1) (Page, 2008).

Figure 1 displays the results of these simulations by showing the difference between matched static and experimental networks on each measure in each possible information environment. This visualization shows that the algorithms' effects vary across information environments. For example, consider the panels containing the results where sensitivity and specificity are symmetrically high (*sensi* = 0.9, *speci* = 0.9). In such information environments, no algorithm is able to substantially influence collective accuracy because agents in the network are able to form accurate beliefs based on their independently acquired knowledge, leaving little room for communication to improve the collective estimate. However, in each of the other information environments, the mean-extreme and scheduling algorithms improve collective accuracy (displayed here as decreased *CES*), with varying degrees of magnitude. When viewed in conjunction with the impact of the intervention on *AIES*, it can be deduced that these two algorithms succeed by improving the average individual accuracy at the cost of diversity (displayed here as decreased *VAR*). In contrast, the polarize algorithm aims to improve collective accuracy by increasing (or maintaining) the variance of beliefs at the cost of individuals' accuracy. However, this algorithm displays adverse effects on collective accuracy in these simulations. The failure of the polarize algorithm here seems attributable to two aspects in our modeling: the use of unbiased, optimal agents and the failure to sufficiently balance the increase in individual error with an increase in variance. The unbiased, optimal agents simulated have the ability to distinguish "anti-reliable" evidence (Hahn et al., 2018b), meaning that before any communication takes place, the mean belief in the network is favorable and the distribution of beliefs is skewed towards the truth, regardless of the information environment imposed with the sensitivity and specificity parameters. Thus, broadcasting the extreme estimates to the median agents, who would otherwise converge towards the favorable mean estimate, will necessarily steer those receiving the erroneous extreme away from the truth. However, real human groups may possess biases that our simulated agents do not reflect, in which case the effects observed here may differ. Indeed, instilling a preexisting bias in our model by assigning each agent a starting prior of 0.1 when the truth is 1 changes the results such that the mean-extreme algorithm often decreases collective accuracy and the polarize algorithm more frequently increases accuracy, albeit only slightly (Figure 2).

Next, we proceeded to test each of the rewiring algorithms with actual human social networks in an online multiplayer experiment where participants were tasked with predicting the probability that various near future events would occur.
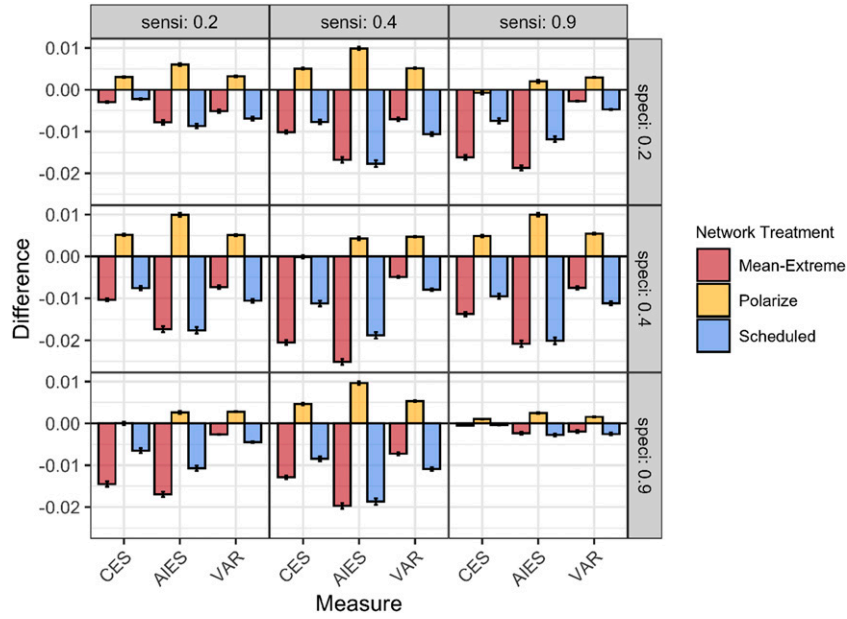
Figure 1. The simulated effects of each algorithm (network condition) on collective error squared *CES*, average individual error squared *AIES*, and belief variance *VAR* averaged across 500 iterations per panel. *Y*-axis values indicate the mean difference on a given measure as compared to a matched static network. A mean falling below zero indicates that the intervention resulted in a decrease of a given measure and vice versa. Black error bars indicate ±1 standard error.
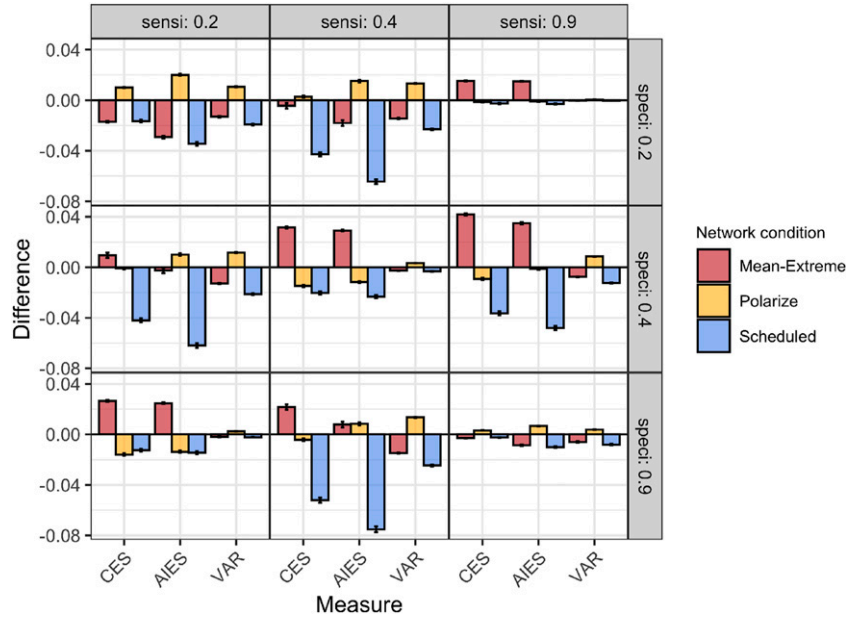
Figure 2. The simulated effects of each algorithm (network condition) on collective error squared *CES*, average individual error squared *AIES*, and belief variance *VAR* averaged across 500 iterations per panel when agents start with priors of 0.1 and the truth is 1. *Y*-axis values indicate the mean difference on a given measure as compared to a matched static network. A mean falling below zero indicates that the intervention resulted in a decrease of a given measure and vice versa. Black error bars indicate ±1 standard error.

Online multiplayer experiment

To move from the simulations towards the real world, we built an online multiplayer experiment with the Empirica software (Almaatouq et al., 2021). This type of "virtual lab" approach allows for flexibility in the design of both a front-end user interface and an experimental back end where we could implement our rewiring algorithms. The preregistration for this study can be accessed here: https://aspredicted.org/9ny8i.pdf.

Table 1. Events predicted by participants in the "collaborative prediction game" experiment. An outcome of 1 indicates the event occurred in reality, and an outcome of 0 indicates the event did not occur in reality.

Event ID	Event Prompt	Outcome
uk_covid	In the UK, the rolling 7-day average of COVID-19 deaths per day will go above 900 between 1 and 14 Feb 2021	0
youtube_subs	There will be at least 10 YouTube channels with more than 63.1 million subscribers on 8 Feb 2021	1
biden_approval	Joe Biden's approval rating with be higher than 55% after 3 weeks as US president	0
us_uk_vax	On 1 Feb 2021, the US will have administered more COVID-19 vaccination doses per 100 people than the UK.	0
Bitcoin	Bitcoin will be valued at less than $30,000 on 8 Feb 2021	0
super_bowl	Both teams in this year's super bowl will score more than 20 points	0
us_climate	The US will rejoin the Paris Climate agreement by 8 Feb 2021	1
sp500	The S&P 500 will close higher on 8 Feb 2021 than it did on 31 December 2020	1
Epl	Liverpool FC will be leading the English premier league on 7 Feb 2021	0
americas_covid	The WHO will report more than 1 million COVID-19 deaths in the Americas by 8 Feb 2021	1

Method

We recruited participants ($N = 704$; 60.94% male, 36.36% female, 2.70% other/preferred not to say) aged 18–69 ($M = 34.28$, SD = 9.87) from Amazon's Mechanical Turk crowdsourcing platform via the CloudResearch service[4]. No restrictions on participant location were applied, but participants were required to be fluent English speakers. Participants were assigned into 16-person networks in one of the four network conditions (static, mean-extreme, polarize, or scheduled) and tasked with a "Collaborative Prediction Game" that consisted of 10 rounds with five stages each. Each round of the game involved predicting the probability of one near future event occurring in reality (see Table 1 for the list of events and outcomes). First, participants provided a probabilistic prediction and a short rationale for their prediction, and then proceeded through four stages of social exchange (communication) where each participant would view the responses of their network neighbor(s) and revise their own prediction and rationale (see Figure S3 for screenshots of the user interface). Each stage was limited to 60 seconds to prevent idle individuals from stalling the group, and the entire study took approximately 50–60 min. Participants were given a base payment of $7.25 and monetary incentives for collective accuracy: 2x pay for the top three most accurate networks, 1.67x pay for the fourth- through sixth-most accurate networks, and 1.33x pay for the seventh- through ninth-most accurate networks. A total of 44 networks completed the study, 11 per treatment. All data collection took place between 11 and 19 January 2021, and treatments were run in parallel so that any advantage gained by making the predictions closer to the resolution date of the events was distributed equally across treatments.

The four network treatments in the experiment were identical to those simulated with our agent-based model, described in the previous section. Participants in *static* networks (the control condition) were placed in a randomly generated small-world network structure for each round, and this network structure remained unchanged over each stage of communication. Participants in the *mean-extreme*, *polarize*, and *scheduled* treatments followed an identical procedure, but their network neighbors were subject to change between stages of communication, as determined by the given rewiring algorithm.

Experimental results

Our analyses of the empirical data focus on the accuracy of the collective mean responses of each network pre- and post-communication as assessed by the Brier scoring metric. In particular, we asked the following two questions: (1) How did the networks' average collective error squared (*CES*) differ between treatments post-communication? (2) How did communication affect *CES* within each network, between treatments?[5]

To address the first question, we followed the procedure we preregistered as the main analysis, which involved a linear mixed effects model with each groups' average collective error squared (*CES*) across all events predicted as the dependent variable, the network treatment as a fixed effect, and random intercepts by group (Figure 3, panel A). This analysis suggests that there is no significant effect of the rewiring algorithms on collective accuracy ($F(3, 436) = 0.78$, $p = 0.503$), suggesting that, on average, networks to which a rewiring algorithm was applied did not achieve lower *CES* post-communication. The model's intercept, corresponding to the static network condition (our control condition), is at 0.27 (95% CI [0.23, 0.30], $t(434) = 16.02$, $p < 0.001$), and the main effects of the rewiring algorithms
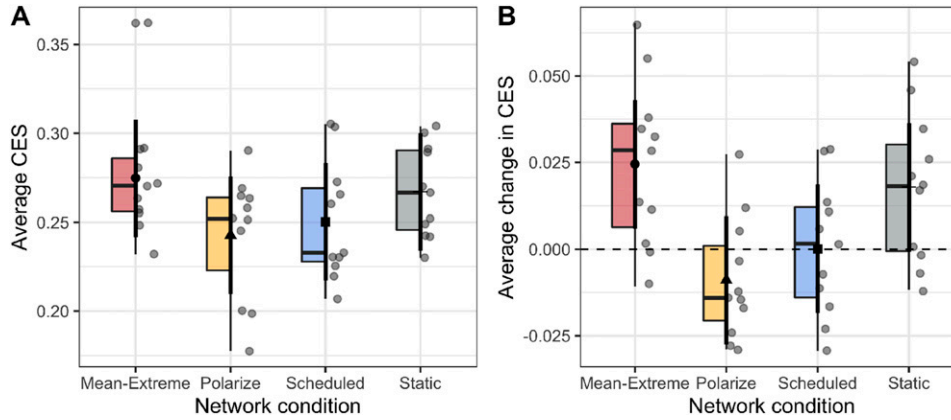
Figure 3. Results of linear mixed effects models. Boxplots and semi-transparent points in the background display the spread of the raw data, and shaped, solid points indicate the model prediction with 95% confidence intervals represented by thick vertical bars. (a) Model with each group's average collective error squared (*CES*) as the dependent variable, network treatment as a fixed effect, and random intercepts by group. (b) Model with each groups' average change in *CES* as the dependent variable (i.e., the difference between post-communication *CES* and pre-communication *CES*), network treatment as a fixed effect, and random intercepts by group.

are statistically non-significant (mean-extreme: $\beta = 0.008$, 95% CI [-0.04, 0.05], $t(434) = 0.33$, $p = 0.742$; polarize: $\beta = -0.024$, 95% CI [-0.07, 0.02], $t(434) = -1.03$, $p = 0.303$; scheduled: $\beta = -0.017$, 95% CI [-0.06, 0.03], $t(434) = -0.71$, $p = 0.479$).[6] However, this analysis does not account for certain key confounding variables—namely, the initial network structure and initial predictions in each network. While we could explicitly control for these in our modeling and simulation work by starting each iteration with perfectly identical networks, it was not possible to match these variables across treatments in the empirical study because each participant only completed the study one time, in one particular network, and in one particular treatment.

In addressing the second question, we conducted an unregistered analysis to side-step the potential confounding effects of initial network structure and initial predictions by evaluating the effect of communication *within* each network. That is, instead of directly comparing the accuracy of networks' collective predictions post-communication between treatments, we compare the *change in accuracy* between each network's prediction pre- and post-communication. Upon refitting our linear mixed effects model with the networks' change in *CES* as the dependent variable, we find a significant treatment effect ($F(3, 436) = 2.72$, $p = 0.044$) that suggests networks mediated by our polarize algorithm were most likely to benefit from communication than (Figure 3, panel B). This model's intercept, corresponding to the static network treatment, is at 0.02 (95% CI [0.00, 0.04], $t(434) = 1.90$, $p = 0.058$), the main effect of the polarize algorithm is statistically significant and negative (i.e., communication decreased error more) ($\beta = -0.03$, 95% CI [-0.05, 0.00], $t(434) = -2.02$,

$p = 0.044$), and the main effects of the mean-extreme and scheduled algorithms are statistically non-significant (mean-extreme: $\beta = 0.01$, 95% CI [-0.02, 0.03], $t(434) = 0.50$, $p = 0.615$; scheduled: $\beta = -0.02$, 95% CI [-0.04, 0.01], $t(434) = -1.34$, $p = 0.181$). This result is encouraging because it suggests not only that the polarize algorithm prevented communication from leading groups astray through deleterious social influence but also that, in many cases, the algorithmic mediation actually led groups towards more accurate predictions than those that would have been produced by aggregating the individuals' pre-communication predictions.

To further investigate this statistically significant but unregistered analysis, we refit the mixed model with alternative specifications. First, we added *events* (i.e., stimuli) as a random factor since statistical literature has argued that the failure to do so can inflate Type I error rates on fixed effect estimates (Judd et al., 2012; Yarkoni, 2022). This analysis also returns a significant treatment effect and does not change the interpretation of our results ($F(3, 427) = 4.29$, $p = 0.005$). Next, we checked whether our result is robust to other loss functions measuring collective accuracy, and found that it is not: we observed statistically insignificant results when applying the mixed effects model with groups as a random factor and either change in collective *square root error* ($F(3, 436) = 1.02$, $p = 0.385$) or collective *absolute error* ($F(3, 436) = 1.54$, $p = 0.204$) as the dependent variable. Finally, we noted that, although we planned and preregistered the use of mixed effects models, the models displayed singular fits—even when only random intercepts for groups are included—which indicates that our model specifications are unnecessarily complex for the data. We thus also tested for a treatment effect on the change in

CES with a one-way ANOVA, which returned the same results as those reported in the previous paragraph.

Despite some fragility in the result, the finding of a significant treatment effect on how communication influenced *CES* within groups suggests that mediating communication in online social networks with different rewiring algorithms can—under specific conditions and operationalizations of error—steer the accuracy of collective beliefs. As such, these findings can be taken as a basic proof of concept. But on the other hand, our main preregistered hypothesis that there would be a statistically significant main effect between network treatments on postcommunication *CES* was not supported, and our initial simulation results do not directly map onto the experimental

results. In order to reconcile these findings, there are three key considerations for future work: (1) more closely controlling for the confounding effects of initial network structure and individuals' differences, (2) applying the rewiring algorithms to networks of more knowledgeable individuals, and (3) better accounting for potential context-dependent effects of each algorithm.

In the experimental design we originally conceived, we sought to control for the confounding effects of initial network structure and individuals' differences by randomly reassigning each participant into one of four identically structured but differently treated networks between each round. Unfortunately, because this procedure involves running 64 participants simultaneously on a single server, and because our experiment necessarily involves algorithmic computation between each stage of each round, we were unable to run this design with the software used because participants experienced significant lags and crashes. This unexpected obstacle forced us to adjust our design such that participants were randomly assigned to a network condition upon signing up for the experiment, and then sent to a separate server depending on the condition (i.e., one network per server at a time). Though this adjustment was necessary to ensure participants could provide quality responses, it means our analysis of a main effect between network conditions may be confounded. To remedy this in future work, one could opt for different software or replicate the experiment with increased statistical power.

A second limitation of our experiment is that the participants may not have possessed much relevant knowledge on the events being predicted. This can be noted in the poor performance across all collective predictions, whereby the probability of producing an accurate binary prediction

Table 2. Tally of groups in each treatment that made the correct binary prediction (0.5 cutoff) on each event post-communication. A correct prediction means that the group's collective prediction what greater than 0.5 if the true outcome was one and vice versa. maximum of 11 per cell.

Event ID	Static	Mean − Extreme	Polarize	Scheduled
uk_covid	3	0	4	3
youtube_subs	10	11	11	10
biden_approval	0	0	1	1
us_uk_vax	6	9	6	7
Bitcoin	8	9	7	10
super_bowl	1	1	1	0
us_climate	11	9	10	11
sp500	11	11	11	11
Epl	4	4	4	6
americas_covid	2	1	2	1
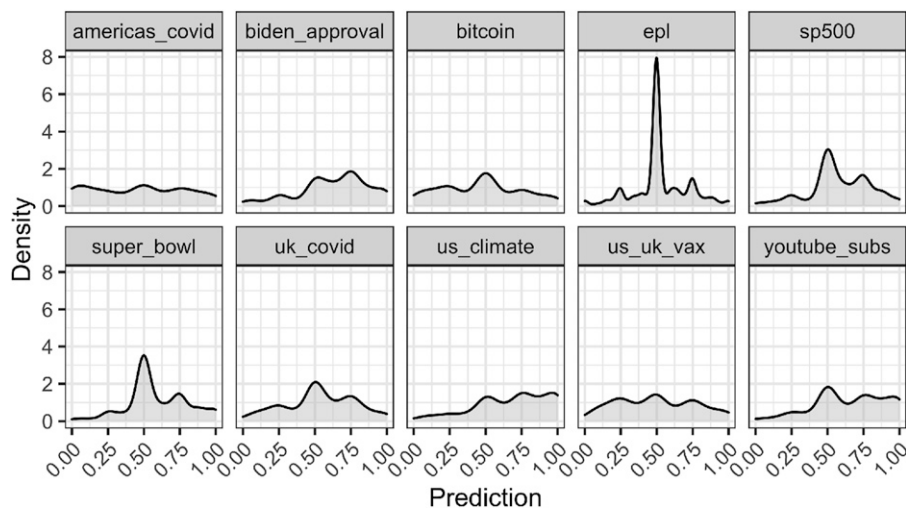
Figure 4. Aggregate distributions of participants' initial predictions for each event in the empirical study. See Table 1 to match event IDs to the actual event prompt. See Table S3 for the mean and standard deviation of participants' initial predictions split out by treatment and event.

(i.e., a collective prediction greater than 0.5 if the event occurred in reality and vice versa) was barely above chance (Table 2, 228/440 correct predictions, 51.81%; also see Table S2 for the average post-communication *CES* for each event in each condition). In principle, this general poor performance of the participants is inconsequential, because random assignment balances incompetence across treatments and we then focus on between treatment effects. However, the underlying logic of rewiring algorithms assumes that there exists some relevant, varied information to be communicated amongst individuals in the group. While an inspection of the rationales entered by participants suggests that a vast majority of individuals engaged in good faith participation, it seems that our participants did not possess many unique pieces of evidence that could be amplified or discounted by a rewiring algorithm. Future work could thus benefit from evaluating the effects of rewiring algorithms on networks of more knowledgeable individuals.

Finally, it is important to note that our experiment focused on one particular prediction context: probabilistic estimates on events where individuals' initial estimates display little to no skew towards one alternative or another (Figure 4). It may be worthwhile to experimentally explore how the rewiring algorithms presented here would fare in other prediction contexts (e.g., continuous, numeric predictions rather than probabilistic predictions); however, several past studies already provide data from crowd-wisdom experiments in which communication took place over static network structures that are ripe for re-analysis (e.g., Becker et al., 2017, 2019; Gürçay et al., 2015; Lorenz et al., 2011). In fact, a re-analysis of the data from those studies demonstrates that the optimal network structure for eliciting the wisdom of the crowd depends on the estimation context—the specific population of individuals faced with a specific estimation task (Almaatouq et al., 2022). Specifically, that work shows that when a group's initial estimates are highly skewed or heavy-tailed, a centralized network structure can promote collective accuracy, whereas decentralized network structures might hinder collective accuracy in such contexts (and vice versa). Given that our rewiring algorithms affect network centralization in different ways—namely, the mean-extreme algorithm increases it while the polarize algorithm decreases it—this insight could explain our experimental results and why they differ from our simulations. This is because in our simulations with optimal Bayesian agents, the networks' initial estimates always display a skew towards the truth, but in our experiment, initial estimates displayed no such skew (Figure 4). Thus, the polarize algorithm may simply have been better suited to the particular prediction tasks considered in our experiment, and the mean-extreme and scheduling algorithms may be better suited to other contexts, such as those simulated with our initial modeling. To explore this point on potential context-dependent effects, we next conducted follow-up modeling using empirical data from past crowd-wisdom experiments to initialize simulations of our algorithms in numeric prediction contexts, which characteristically display highly skewed distributions of initial predictions (also see Figure S4).

## Follow-up simulations of numeric estimation contexts

We set out additional simulations to explore how the rewiring algorithms might perform in numerical estimation contexts—where the 16-agent networks estimate (or predict) some unknown positive number—rather than probabilistic estimation contexts. Such tasks map onto classical crowd wisdom scenarios such as estimating the weight of an ox, as well as high-stakes, real-world scenarios like forecasting the number of ICU admissions per week during a pandemic.

We follow the procedure described in the previous section on "modelling and simulations" and initialize our model by randomly generating an undirected small-world network (Watts and Strogatz, 1998), have our agents follow the same updating rule borrowed from Becker et al. (2017), and consider the same four network conditions (static, mean-extreme, polarize, and scheduled). However, instead of having each agent integrate evidence (represented as samples from a Bernoulli distribution) via Bayes' theorem to establish their initial estimate, we assign each agent an initial estimate by sampling from a compilation of empirical data from four previously published experiments (Becker et al., 2017, 2019; Gürçay et al., 2015; Lorenz et al., 2011). This compiled dataset spans a total of 54 estimation tasks on which 2,885 individuals provided independent estimates (Almaatouq et al., 2022). Each task—or "estimation context"—in this dataset is represented by a distribution of independent estimates and a true value. For example, one task contains 278 participants' estimates of the London population in July 2010, with the true value of 7,825,200 (Gürçay et al., 2015). Note, however, that we normalize the estimates for each task to be between 0 and 1 in order to suit our belief updating rule and mean-extreme rewiring algorithm while maintaining the distributions' shape.

Following 500 iterations of each of the 54 estimation tasks in which four matched networks are simulated (i.e., one of each network condition starting from an identical initial network), we assess collective accuracy by calculating the squared error of the mean estimate post-communication (*CES*; i.e., the Brier score). While other loss functions such as absolute error and square root error may be applicable in some task domains, our pattern of results is consistent across these loss functions; thus, and also because of the theoretical link of *CES* to the Diversity Prediction Theorem, we focus on *CES* for the sake of this paper.
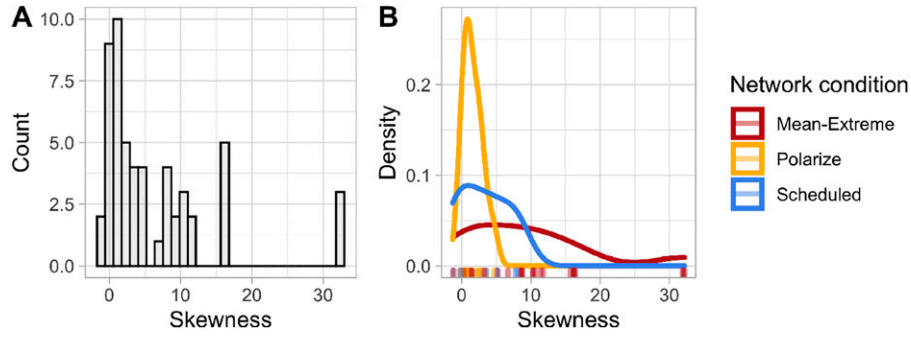
Figure 5. The skewness parameter space. (a) The distribution of skewness in the 54 estimation tasks considered. (b) The distribution of skewness where each network condition produced the lowest collective error as compared to the other conditions.
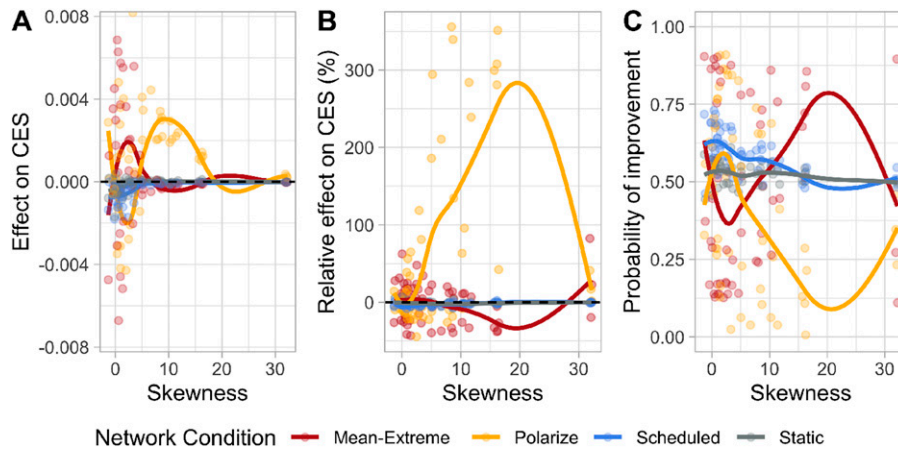
Figure 6. Network performance over skewness as compared to matched static networks. (a) The average effect on *CES* across skewness (i.e., the average change in *CES* compared to matched static networks). (b) The average relative effect on *CES* across skewness (i.e., the average change in error divided by the average error of matched static networks). (c) The probability of improvement across skewness (i.e., the proportion of the 500 iterations of each task where a given network condition was more accurate than matched static networks). Lines represent local regressions (LOESS) fitted to the data with a polynomial degree of 2.

Across all of the estimation tasks considered, the four network conditions' *CES* was nearly equal on average (static networks, $M = 0.016$, $SD = 0.031$; mean-extreme networks, $M = 0.017$, $SD = 0.033$; polarize networks, $M = 0.016$, $SD = 0.029$, scheduled networks, $M = 0.016$, $SD = 0.031$). However, these averages overlook potential context-dependent effects. An analysis of *CES* task-by-task, rather than in aggregate, reveals that mean-extreme networks achieved the highest accuracy on 31 tasks, polarize networks achieved the highest accuracy on 15 tasks, and scheduled networks achieved the highest accuracy on eight tasks. Static networks did not achieve the highest accuracy on any tasks.

To further understand these context-dependent effects of the rewiring algorithms, we characterized each task by the skewness of the distribution of individuals' initial estimates and then observed how each network condition's average *CES* varied across the skewness parameter space. As shown

in Figure 5, the rewiring algorithms display a clear favoritism for certain regions of the skewness parameter space: mean-extreme networks were the most accurate for tasks with highly skewed estimate distributions ($n = 31$, $M = 9.47$, $SD = 9.33$), polarize networks were the most accurate for tasks with estimate distributions that display low skewness ($n = 15$, $M = 1.56$, $SD = 1.38$), and scheduled networks were the most accurate on tasks with mid-range skewness ($n = 8$, $M = 3.21$, $SD = 3.55$).

In Figure 6, we further investigate how the effects on collective accuracy produced by the rewiring algorithms track over skewness. Using the *CES* of static networks as a baseline condition, we calculated three measures for each of the 54 estimation tasks for mean-extreme, polarize, and scheduled networks: the average effect on *CES* (i.e., the average change in error), the average relative effect on *CES* (i.e., the average change in error divided by the average error of matched static

networks), and the probability of improvement (i.e., the proportion of the 500 iterations of each task where a given network condition was more accurate than a matched static network). This analysis suggests not only that the different rewiring algorithms prefer different estimation contexts, but that there is an important interaction: the mean-extreme algorithm actively increases collective error on tasks with low skewness and the polarize algorithm actively increases collective error on tasks with high skewness.

The results of these follow-up simulations add value to our investigation of rewiring algorithms as a tool for collective intelligence design in three ways. First, we once again find evidence suggestive of a basic proof of concept: rewiring algorithms can boost the accuracy of collective estimates/predictions in social networks under certain conditions. Second, we gain clarity around the context-dependence of rewiring algorithms' effects and around why our experimental results seem to depart from the results of our initial simulations: whether a rewiring algorithm helps or hinders collective accuracy (or has no effect) depends on the distribution of initial pre-communication estimates. In our original modelling where agents first integrated evidence independently via Bayes' theorem, the distribution of initial estimates was always skewed, which suits the mean-extreme algorithm (Figure 1, Figure 5). Whereas in our experiment, the distribution of initial estimates displayed no such skew and were more uniformly or normally distributed (Figure 4), which suits the polarize algorithm (Figure 5). This reconciliation of our results in turn brings us to the third insight gained from the follow-up simulations: it may be possible to identify distributional characteristics of estimates (e.g., skewness) that allow one to select a rewiring algorithm capable of increasing the accuracy of social networks' collective estimations before communication takes place. Crucially, this means that algorithms could be efficiently selected and applied in contexts where there is no track record of individuals' predictive success and the truth or falsity of individual estimates is not (yet) known. Where sufficient ground truth data on accuracy exists, such as in expert judgments of medical scans, that data can unquestionably be used to fine-tune networks of judges (Kurvers et al., 2019). However, that leaves many of the most pressing real-world judgment tasks unaccounted for. In particular, we may want collective judgments to derive high-quality predictions for consequential unique events for which, by definition, ground truth data will be unavailable. Rewiring algorithms, as a method that enhances collective accuracy in such contexts, may thus provide a valuable prediction tool for many domains.

Conclusion

Can rewiring algorithms enhance collective decision-making in online social networks? In the context of modern-day social media, where algorithms are already deployed to mediate social interactions for commercial interests, it seems worthwhile to investigate whether algorithms can be (re-)designed for epistemic interests. Naturally, this first requires exploratory research to test whether such effects can be made to occur (as opposed to whether such effects will necessarily occur) (Brauer and Kennedy 2023; Mook 1983), as we have sought to deliver here. While our results are not to be interpreted as a suggestion that the specific algorithms presented should be deployed in practice, the present findings provide a proof of concept and encourage continued research.

Author contributions

J. W. B.: Funding acquisition, conceptualization, software, methodology, analysis, writing - original draft. **A. A.**: Software, resources, writing - review and editing. **M. A. R.**: Software, resources, writing - review and editing. **U. H.**: Funding acquisition, conceptualization, methodology, writing - review and editing, supervision.

Ethical statement

The experimental work presented in this paper received ethical approval from the Departmental Ethics Committee of the Department of Psychological Sciences, Birkbeck, University of London. Reference number: 181950. All participants in the experiment provided informed consent.

Open science statement

All data and code used in this paper is publicly available on GitHub: https://github.com/jwburton/bbk-nesta-ci. The preregistration for the online multiplayer experiment can be accessed here: https://aspredicted.org/9ny8i.pdf.

ORCID iDs

Jason W Burton https://orcid.org/0000-0002-6797-2299
Abdullah Almaatouq https://orcid.org/0000-0002-8467-9123
M Amin Rahimian https://orcid.org/0000-0001-9384-1041

Data Availability Statement

All data and code used in this paper is publicly available on GitHub: https://github.com/jwburton/bbk-nesta-ci. The preregistration for the online multiplayer experiment can be accessed here: https://aspredicted.org/9ny8i.pdf.

Supplemental Material

Supplemental material for this article is available online

Notes

1. To generate small world networks, we follow the algorithm presented by Watts and Strogatz (1998), which entails starting with a ring lattice and randomly rewiring each edge with probability p. For each iteration of our model, we randomly sample a rewiring probability p from a uniform distribution between 0 and 1, meaning that our simulations include iterations with starting network structures ranging from an orderly ring (when $p \approx 0$) to a disorderly random network (i.e., an Erdős–Rényi graph, when $p \approx 1$). The average degree of the resulting networks is 2.
2. We use the DeGroot model as a tractable, descriptive framework of behavior in communicating groups to provide a starting point for this exploratory, proof-of-concept work. We do not use the DeGroot model as a normative framework.
3. An alternative baseline condition is *random rewiring*, where the network structure is entirely regenerated at each time step (see Figure S2). However, this random rewiring condition is not a realistic network condition. There is no real-world scenario in which a communicating group's underlying network structure is repeatedly, entirely, and randomly regenerated between social exchanges. We thus argue that our static network condition is a more relevant, ecologically valid baseline condition.
4. Our sample size was determined by how much research funding was available for this study and ended up being smaller than the preregistered sample size of 1280 participants because we had underestimated the amount of funds that would be needed to pilot test our experiment.
5. We also preregistered that we would conduct an exploratory analysis looking at possible effects on confidence calibration. However, this analysis was not conclusive and is the focus of ongoing work. It has thus been excluded from the present paper.
6. Using the same mixed model specification, no significant treatment effect is observed on average individual accuracy when measured as squared error, square root error, or absolute error. However, a statistically significant treatment effect on variance is observed ($F(3, 426) = 5.31, p = 0.001$). See Table S1 for full results of these models.

References

Almaatouq A, Noriega-Campero A, Alotaibi A, et al. (2020) Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences* 117(21): 11379–11386.

Almaatouq A, Becker J, Houghton JP, et al. (2021) Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods* 53: 2158–2171.

Almaatouq A, Rahimian MA, Burton JW and Alhajri A (2022) The distribution of initial estimates moderates the effect of social influence on the wisdom of the crowd. *Scientific Reports* 12: 16546.

Arrow KJ, Forsythe R, Gorham M, et al. (2008) The promise of prediction markets. *Science* 320(5878): 877–878.

Becker J, Brackbill D and Centola D (2017) Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences* 114(26): E5070–E5076.

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

Brauer M and Kennedy KR (2023) On effects that do occur versus effects that can be made to occur. *Frontiers in Social Psychology* 1: 1193349.

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78: 1–3.

Condorcet MJANC (1785) *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Bulletin Des Sciences Mathématiques et Astronomiques.

Fischhoff B, Slovic P and Lichtenstein S (1977) Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance* 3(4): 552–564.

Galton F (1907) Vox populi. *Nature* 75: 450–451.

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

Grofman B, Owen G and Feld SL (1983) Thirteen theorems in search of the truth. *Theory and Decision* 15: 261–278.

Gürçay B, Mellers BA and Baron J (2015) The power of social influence on estimation accuracy. *Journal of Behavioral Decision Making* 28(3): 250–261.

Hahn U (2022) Collectives and epistemic rationality. *Topics in Cognitive Science* 14(3): 602–620.

Hahn U, Hansen JU and Olsson EJ (2018a) Truth tracking performance of social networks: how connectivity and clustering can make groups less competent. *Synthese* 197: 1511–1541.

Hahn U, Merdes C and von Sydow M (2018b) How good is your evidence and how would you know? *Topics in Cognitive Science* 10(4): 660–678.

Hahn U, von Sydow M and Merdes C (2019) How communication can make voters choose less well. *Topics in Cognitive Science* 11(1): 194–206.

Hogarth RM (1978) A note on aggregating opinions. *Organizational Behavior and Human Performance* 21(1): 40–46.

Hong L and Page SE (2012) Some microfoundations of collective wisdom. In: Landemore H and Elster J (eds). *Collective*

Wisdom: Principles and Mechanisms. Cambridge: Cambridge University Press, 56–71.

Howe J (2006) The rise of crowdsourcing. *Wired Magazine* 14(6): 176–183.

Jönsson ML, Hahn U and Olsson EJ (2015) The kind of group you want to belong to: effects of group structure on group accuracy. *Cognition* 142: 191–204.

Judd CM, Westfall J and Kenny DA (2012) Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology* 103(1): 54–69.

Kurvers RHJM, Herzog SM, Hertwig R, et al. (2019) How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances* 5(11): eaaw9011.

Ladha K (1992) The Condorcet Jury Theorem, free speech, and correlated votes. *American Journal of Political Science* 36(3): 617–634.

Ladha K (1993) Condorcet's Jury Theorem in light of de Finetti's theorem: Majority-rule voting with correlated votes. *Social Choice and Welfare* 10: 69–85.

Lazer D (2015) The rise of the social algorithm. *Science* 348(6239): 1090–1091.

Lichtenstein S, Fischhoff B and Phillips LD (1977) Calibration of probabilities: the state of the art. *Decision Making and Change in Human Affairs*. Dordrecht: Springer, 16, 275–324.

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

Mook DG (1983) In defense of external invalidity. *American Psychologist* 38(4): 379–387.

Morgan MG (2014) Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences* 111(20): 7176–7184.

Muchnik L, Aral S and Taylor SJ (2013) Social influence bias: A randomized experiment. *Science* 341(6146): 647–651.

Mulgan G (2018) Artificial intelligence and collective intelligence: The emergence of a new field. *AI & Society* 33: 631–632.

Murphy AH (1973) A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology* 12(4): 595–600.

Page SE (2008) *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. New Edition. Princeton: Princeton University Press.

Rufibach K (2010) Use of Brier score to assess binary predictions. *Journal of Clinical Epidemiology* 63(8): 938–939.

Simon J, Bass T, Boelman V and Mulgan G (2017) *Digital Democracy: The Tools Transforming Political Engagement*. London: Nesta, 1–87.

Surowiecki J (2005) *The Wisdom of Crowds*. New York: Anchor.

Watts DJ and Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393: 440–442.

Wei H, Pan Z, Hu G, et al. (2018) Identifying influential nodes based on network representation learning in complex networks. *PLoS One* 13(7): e0200091.

Wolfers J and Zitzewitz E (2004) Prediction markets. *The Journal of Economic Perspectives* 18(2): 107–126.

Yarkoni T (2022) The generalizability crisis. *Behavioral and Brain Sciences* 45(e1): 1–78.

Zollman KJS (2013) Network epistemology: Communication in epistemic communities. *Philosophy Compass* 8(1): 15–27.