Network Structures of Collective Intelligence: The Contingent Benefits of Group Discussion

Joshua Becker\textsuperscript{a,1}, Abdullah Almaatouq\textsuperscript{b}, and Emőke-Ágnes Horvát\textsuperscript{c}

\textsuperscript{a}UCL School of Management, University College London
\textsuperscript{b}Sloan School of Management, Massachusetts Institute of Technology
\textsuperscript{c}School of Communication, Northwestern University

\textsuperscript{1}Correspondence to: joshua.becker@ucl.ac.uk

WORKING PAPER: last updated 15 September 2020

Abstract. Research on belief formation has produced contradictory findings on whether and when communication between group members will improve the accuracy of numeric estimates such as economic forecasts, medical diagnoses, and job candidate assessments. While some evidence suggests that carefully mediated processes such as the “Delphi method” produce more accurate beliefs than unstructured discussion, others argue that unstructured discussion outperforms mediated processes. Still others argue that independent individuals produce the most accurate beliefs. This paper shows how network theories of belief formation can resolve these inconsistencies, even when groups lack apparent structure as in informal conversation. Emergent network structures of influence interact with the pre-discussion belief distribution to moderate the effect of communication on belief formation. As a result, communication sometimes increases and sometimes decreases the accuracy of the average belief in a group. The effects differ for mediated processes and unstructured communication, such that the relative benefit of each communication format depends on both group dynamics as well as the statistical properties of pre-interaction beliefs. These results resolve contradictions in previous research and offer practical recommendations for teams and organizations.

Keywords: belief accuracy, communication, group dynamics, Delphi method, wisdom of crowds

The fundamental paradox in group decision-making is that interaction between group members is often an integral part of the decision-making process, and yet social influence dynamics pose the risk of undermining decision quality through processes such as herding (Da & Huang, 2019; Lorenz et al., 2011) and groupthink (Janis, 1982; Turner & Pratkanis, 1998). As a result, a great deal of research has sought to understand why group decision-making sometimes fails (Janis, 1982) and what practices, if any, can allow interacting groups to produce accurate beliefs (Green et al., 2007). However, despite decades of interest, researchers have offered only contradictory answers on whether and how social influence impacts belief accuracy.
One critical question is how to mitigate the risks associated with unstructured, free-flowing conversation such as conformity pressure and overly domineering discussants (Dalkey, 1969). These concerns motivated the development of the “Delphi method” (Dalkey, 1969) and other strategies for mediating communication through controlled processes (Gustafson et al., 1973; Ven & Delbecq, 1974). While techniques varied across implementation, they all generally allowed people to communicate information indirectly, e.g. by writing numbers down on a piece of paper. As an experimental paradigm, this process offers a very simple design: first collect independent pre-communication estimates; then after some mediated process such as the exchange of numeric estimates, collect post-communication estimates. Researchers then pose a simple question: does the group become more accurate? And, additionally, which process produces the greatest increase/decrease in accuracy? However, only a relatively small number of experiments employed questions that could be objectively verified for accuracy. Moreover, these few experiments yielded conflicting results on whether such controlled, numeric communication offers any benefit over unstructured discussion or whether either process improved accuracy at all (Hastie, 1986).

We resolve these contradictory results by showing how network theories of belief formation developed in research on “collective intelligence”\(^1\) can explain why unstructured discussion will sometimes outperform numeric communication and why the outcome is sometimes reversed. In particular, we draw on two key observations. First, Becker et al (2017) demonstrated that the potential benefit or risk of social influence depends on the network structure of communication. In particular, they reported that decentralized networks where everyone is equally connected reliably improve belief accuracy, while centralized networks risk undermining the wisdom of crowds. However, Almaatouq et al. (2020) challenged this claim, showing that the effect and potential risk of influence network centralization depends on the statistical properties of

\(^1\) There is no single definition of collective intelligence, though a common one is the broad notion of “collectives acting intelligently” (Malone & Bernstein, 2015). For our context, the most relevant idea is that the whole is greater than the sum of its parts, and that groups can be collectively rational even when composed of irrational individuals (Krafft et al., 2016).
the pre-interaction belief distribution. As we note in both our analysis and our discussion, it is important when assessing the effects of social influence to distinguish between communication networks—i.e., binary networks that define who can communicate with whom—and influence networks, in which the tie between two people is represented as a weighted value between zero and one.

Our key insight is the observation that unstructured discussion, which can be influenced disproportionately by domineering individuals (Dalkey 1969), can be modeled as a centralized influence network. Because numeric communication eliminates the possibility of any variation in persuasive influence, the Delphi method and other mediated communication techniques can be modeled as a relatively decentralized network. We show that in this context, the results developed by Almaatouq et al. (2020) can be approximated with a simple heuristic: when a majority of individuals are more accurate than the group, unstructured discussion is most likely to improve accuracy; when a majority are less accurate, numeric communication is superior. As we discuss, however, this behavior is distinct from other majority-driven dynamics such as the risky shift (Stoner, 1961) and group polarization (Myers & Bishop, 1971; Myers & Lamm, 1976). Specifically, those models assume that any given group is drawn towards the majority opinion, whereas we find this result only on average across many groups. We demonstrate this hypothesis with a re-analysis of previous experimental data as well as a novel, pre-registered experimental dataset that replicates these findings with a new population, new design, and systematically sampled question set.

The primary contribution of this paper is to resolve contradictions in previous research about a question that is fundamental to understanding team behavior and organizational performance: how does communication impact belief accuracy? Our findings also show how a network theoretical framework can help us to understand belief formation in otherwise unstructured (in a network sense) communication, including the Delphi method and also everyday interactions such as committee meetings. These dynamics are driven by emergent networks of influence which impact the behavior of both numeric exchange and informal discussion beyond the initial determination of who communicates with whom. The primary implication is that unstructured discussion, like centralized networks, may be considered unreliable under wide range
of conditions. In the conclusion of this paper, we describe how this network theoretical framework reveals possible solutions that allow groups to engage in free-form discussion while mitigating the risks associated with centralized influence.

DECESSIONS AND BELIEF ACCURACY

Broadly speaking, decision-making is a complex behavior that involves many distinct processes such as idea generation, idea evaluation, and idea selection (Davis, 1973). Each of these individual processes has been the subject of extensive research. For example, idea generation has been studied through research on brainstorming (Stroebe et al., 2010) and is subject to unique concerns such as the exploration-exploitation tradeoff (Lazer & Friedman, 2007; March, 1991). The present paper focuses specifically on the formation of accurate beliefs, which is important to idea evaluation—e.g., assessing the expected payoff of an investment. As a basic process, belief formation plays a fundamental role in many kinds of decision-making.

Although decision-making can be a complex process that involves the synthesis of many different sources and types of information, accurate numeric estimates are one common form of belief critical to many kinds of decisions. One of the most widely studied estimation tasks is forecasting (Atanasov et al. 2017, Dalkey and Helmer 1963, Jansen et al. 2016), which is a critical component of strategic decision-making and can include forecasting business metrics like revenue (Da and Huang 2019) and sales (Cowgill & Zitzewitz, 2015), predicting the success of an advertising campaign (Hartnett, Kennedy, Sharp, & Greenacre, 2016), or estimating future macro-economic indicators (Jansen, Jin, & de Winter, 2016). One particularly widely studied form of forecasting is software development effort, i.e. how much it will cost to develop a new project (Jørgensen, 2004). In one case study, Cyert and March (1963) describe a construction firm for whom expectations of future business volume played a central role in the decision to move operations to a new location.

Although formal models may often be available to support forecasting and estimation, subjective or “gut” estimates remain a central source of predictions in both strategic arenas such as economic forecasting (Jansen et al., 2016) as well as operational decisions such as software
development (Jørgensen, 2004). One reason subjective estimates remain important is because statistical models are limited in their ability to handle new contexts, and thus subjective estimates tend to outperform statistical models in unfamiliar contexts such as crises—precisely those times when it is most important to reduce uncertainty (Jørgensen, 2004). While people don’t always quantitatively important decisions in practice, the goal of the present research is to develop prescriptive theories about how we should make decisions. This endeavor is thus similar to the prescription for individuals to develop quantified scoresheets for important decisions [Bazerman and Moore 1994] or to engage in debiasing interventions.

Numeric estimates can play a role in even in decisions where there is not an obviously “correct” choice, such as hiring. Hiring ideally involves predicting in advance how candidates will later be measured on such factors as performance, cultural fit, and likelihood of remaining employed. While many important employer characteristics are subjective, employees at lower levels of organizations are frequently evaluated with explicitly quantified metrics. Industries such as customer service call centers have a long history of quantifying employee productivity with metrics such as quality assurance scores and upsell rates (Holman et al. 2002). More recent advances in monitoring technology are allowing performance quantification for employees as varied as restaurant waitstaff (Pierce et al. 2015) and warehouse workers (Moore 2019). In scenarios where employees are evaluated on a quantitative metric, as is common in customer service call centers, hiring a person is tantamount to forecasting that they will have a high performance score.

**To group or not to group?**

While these decisions certainly can (and often are) made by individuals, groups play an increasingly important role in decisions that were once dominated by individuals, such as medical decision-making (Christensen et al., 2003) and even hedge fund management (Massa et al., 2006). Importantly, the rise of teams offers enormous potential benefit—indeed, the core motivation of “collective intelligence” theory (Da & Huang, 2019; Lazer & Friedman, 2007; Weick & Roberts, 1993; Woolley et al., 2010) is that groups can produce better (e.g., more accurate) decisions than individuals (Krafft et al., 2016; Sunstein, 2006; Weick & Roberts, 1993). For example, studies on
belief accuracy have found that the aggregated beliefs of multiple physicians can outperform even the most skilled individuals (Kurvers et al., 2016), and that aggregate beliefs of amateur investors can outperform expert analysts (Chen et al., 2014; Drogen & Jha, 2013).

One key argument for the accuracy benefits of groups over individuals is the expectation that no single individual is likely to provide an exactly correct numeric estimate. Both statistical principles (Hogarth, 1978; Page, 2007) and empirical data (Atanasov et al., 2017; Galton, 1907; Ven & Delbecq, 1974) suggest that when group members pool their beliefs, errors cancel out in aggregate, allowing groups to generate more accurate beliefs (and thus produce better decisions) than individuals acting in isolation. In fact, the “crowd beats average” principle (Page, 2007)—commonly known as the “variance bias decomposition” in statistical estimation theory—provides a mathematical guarantee that the average belief in a group will be more accurate (in terms of squared error) than a randomly selected individual. The intuitive interpretation of these principles is that when people have diverse information and perspectives, errors “cancel out” and produce accurate group beliefs.

However, groups in practice often fail to take advantage of all the diversity of information held by their individual members. Some examples of how group decision-making can fail emerge from case studies describing “Groupthink” dynamics (Janis, 1982) in organizational decisions. In these studies, norms favoring group cohesion produced conformity pressures which prevented individuals from sharing information that contradicted existing group beliefs (Janis, 1982; Surowiecki, 2004). Importantly, social influence can harm belief accuracy even in the absence of normative pressure. Herding models have shown that strictly rational individuals who would produce accurate decisions independently can produce inaccurate decisions when their beliefs are formed sequentially (Banerjee, 1992). This process has been shown empirically to impact financial crowdsourcing. For example, early versions of Estimize allowed contributors to see the estimates of other forecasters before providing their own forecast, reducing overall accuracy; when they revised their design to require independent contributions, accuracy increased (Da & Huang, 2019).

One especially popular strategy for taking advantage of a group’s collected intelligence while minimizing the risks associated with social influence is the “wisdom of crowds” approach (Atanasov et al., 2017; Da & Huang, 2019; Golub & Jackson, 2010; Lamberson & Page, 2012;
Minson et al., 2018; Mollick & Nanda, 2016). This approach is simple and yet powerful: simply collect the independent beliefs of a large number of individuals, thus harnessing the statistical benefits of group beliefs without exposing the decision to the risks of herding or groupthink. This approach is reflected in the Estimize platform (Drogen & Jha, 2013) as well as other crowdsourcing efforts by firms such as Google (Cowgill & Zitzewitz, 2015), Ford (Cowgill & Zitzewitz, 2015), Dell (Bayus, 2013), and Best Buy (Dvorak, 2008). More advanced statistical methods can improve upon simple aggregation by employing weighting methods that, for example, try to identify people who are consistently more accurate (Budescu & Chen, 2014).

However, while the wisdom of crowds strategy can be extremely effective, it is limited in two respects. First, it may be infeasible or undesirable to prevent contributors from interacting, as when a decision must be made by people who naturally interact in the course of their work and information sharing is unavoidable. In such a context, a process which requires independence may violate the norms or expectations of the organization in question. Second, even if social interaction is potentially avoidable, strategies that harness only the “collected” intelligence of independent individuals fail to take advantage of the potential benefits of group interaction, i.e. “collective” intelligence. In particular, prior research suggests that properly structured communication processes may allow groups to produce even more accurate beliefs than could be obtained by independent individuals (Becker et al., 2017; Dalkey & Helmer, 1963; Gustafson et al., 1973; Ven & Delbecq, 1974). The key element of these processes is that people do not communicate directly via informal conversation, but instead their communication is mediated through some intervening process such as a survey.

Mediated Communication
Motivated by the potential benefits of group interaction, Dalkey developed the “Delphi” method as a process that can theoretically harness the collective intelligence\(^2\) of interacting groups while mitigating the risks associated with social influence dynamics (Dalkey & Helmer, 1963). Although

\(^2\) Though the phrase “collective intelligence” has been introduced relatively recently, the spirit is the same.
the exact method varies widely across implementations (Green et al., 2007; Humphrey-Murto & de Wit, 2019), the Delphi method generally involves allowing decision-makers to share limited information indirectly—mediated by a person or process. Some versions allow participants to write down motivating arguments, while others are limited to the exchange of numeric estimates. E.g., the Delphi method might proceed as follows: a survey collects participants’ independent estimates; a facilitator calculates the mean, median, and inner-quartile and distributes that information to the participants; participants then submit final estimates via survey. In laboratory studies utilizing such methods, researchers have studied topics ranging from the trivial, such as guessing average height/weight ratios (Gustafson et al., 1973) to the more serious, such as optimizing military campaigns (Dalkey & Helmer, 1963). However, most experiments utilizing the Delphi do enable accuracy measurement (e.g. by using questions with no known true answer) and their focus is on other factors such as consensus formation (Dalkey & Helmer, 1963) and the perceived ability to contribute diverse perspectives without repercussion (Ven & Delbecq, 1974).

Critically, however, the few experiments with mediated communication that did actually measure accuracy yielded inconsistent and at times contradictory results (for a review see Hastie, 1986; Rowe & Wright, 1999). In the experiments, people are typically asked to complete numeric estimates before and after some form of group information exchange, allowing the researchers to assess the effect of communication on belief accuracy. Gustafson (1973) found that open discussion produced the most accurate estimates while the Delphi method produced the least accurate estimates, with independent individuals falling in the middle. Findings by Gough (1975; as cited by Hastie, 1986) also found that open discussion outperformed the Delphi method, but found that both methods outperformed independent judgements. Still others (Larreche & Moinpour, 1983) found that the Delphi method outperformed both independent individuals and unstructured discussion.

There are several possible explanations for these inconsistencies. One possible explanation that must be considered is statistical error. In addition to the studies cited above, several researchers (Fisher, 1981; Boje & Murnighan, 1982; Snizek 1990) found no statistically significant difference between different modes of communication. Coupled with the fact that the analytic strategies varied considerably across those studies which did find results, increasing the possibility for Type
I error, this research as a whole may plausible be interpreted as a null result—i.e., that there is no effect of communication on belief accuracy under such controlled processes. However, we show instead that social influence does have intrinsic effects on belief accuracy, but that attempts to measure a “main effect” of social influence for any given form of communication produced inconsistent results due to unobserved heterogeneity across experimental tasks.

Our attempt to revisit this question is motivated in part by more recent research (Almaatouq et al., 2020; Becker et al., 2017; Kao et al., 2018b; Lorenz et al., 2011) that has begun to implement these same methods again (estimate, mediated exchange, estimate) to study social influence, though with the increased control afforded by virtual laboratories. This research is also typically conducted with a different theoretical perspective (and often without reference) to earlier social psychological research on judgement and decision-making. Critically, these experiments have begun to identify ways in which mediated social influence can have systematic effects on belief accuracy by identifying hidden variables impacting communication and belief accuracy such as network structure. While these experiments don’t address the basic question of whether informal conversation is productive, they do provide a framework that can provide the answer. The present paper builds on this framework to answer a simple question: when a group sits down at a table to have a conversation, would they be better off communicating via numbers on a slip of paper?

**The Network Dynamics of Belief Formation**

Prior research on collective intelligence (Almaatouq et al., 2020; Becker et al., 2017) has identified two key principles relating social influence to belief formation, both of which build on a formal model of opinion formation developed by DeGroot (1974). We note that the term “belief” is used in a variety of colloquial and scientific ways (Converse, 1962; Gawronski, 2012; Tversky & Kahneman, 1975). From the perspective of strictly rational choice, as in e.g. game theory, a belief represents a probability distribution over a set of possible events that can be used to calculate which decision generates the maximum estimated payoff. Here, however, we are interested in beliefs more broadly as “what someone thinks about the world.” We don’t need to assume rational decision-makers to take an interest in quantitative beliefs, since people and organizations regularly make decisions based on beliefs about numbers, explicit or not, rational or not: expected return on
investment, how many units will sell, the revenue potential for a sales job candidate. While our theoretical interest is broadly in how people form and share beliefs, our analytic and empirical approach operationalizes beliefs as numeric estimates. We use the term “estimates” to refer to those specific numbers relevant to decisions, both in terms of what people think about the world and what they share with each other. This term is equivalent to ‘judgement’ (Brenner et al., 1996; Mannes, 2009) or ‘opinion’ (DeGroot, 1974; Hogarth, 1978) in other literature. Thus for our theoretical model and experimental results, we assume a person’s “belief” (about, e.g., the number of jellybeans in a jar) to be equivalent to their estimate and thus to be represented entirely by the number they provide.

The DeGroot (1974) model assumes that each individual in a population starts with some initial opinion, i.e. a response to some numeric estimate. Each individual can then observe the beliefs of some or all other members of the population. Each individual then updates their belief as a weighted average, combining their own initial belief with the opinions of their peers. This weighting is completely flexible, in that it allows someone to ignore (or be disconnected) from some peers by assigning that peer a weight of zero, or to ignore social information all together, or to simply adopt the average. After updating, subjects then observe the revised belief of their peers and update again.

The weight that each individual places on each of their peers can be represented as a social network, where the weight is a directed weighted edge between two people. This social network is a weighted influence network, indicating who influences whom and how much. DeGroot shows that if this revision process is repeated indefinitely, the group will asymptotically converge on a weighted average of initial independent beliefs, with each individual weighted proportionately to their centrality score in the corresponding influence network. In other words, centralized networks will be dominated by central individuals, but decentralized networks—where everyone has an equal centrality score—will converge on the initial group mean, thus preserving the wisdom of crowds.

As Becker et al. (2017) argue, this property of belief formation means that centralized networks are inherently unreliable—the effect of social influence will depend only on the accuracy of the central node, generating the ‘wisdom of the few’ rather than the wisdom of the crowd. In
contrast, they observed a reliable increase in the accuracy of decentralized networks; instead of converging on the initial mean, they were drawn toward the belief of accurate individuals. This dynamic can be explained by noting that there are two ways for a person to gain centrality in the influence network. First, they can simply be more influential on their peers, e.g. through status or argumentation. Second, people can increase their centrality through stubbornness, which is operationalized by DeGroot as placing weight on one’s own self—thus giving oneself greater incoming network weight. And in Becker et al.’s experiment, subjects who were more accurate also appeared to be more stubborn. As a result, their networks which were decentralized by design ended up as weakly centralized network in which central individuals were reliably accurate. I.e., a decentralized communication network produced a centralized influence network.

However, where Becker et al. (2017) characterized centralized communication networks as simply unpredictable and thus unreliable, Almaatouq et al. (2020) demonstrated that the effect centralized centralization depends on the statistical properties of the initial belief distribution. To explain this result intuitively, consider the conditions in Becker et al.: subjects are randomly assigned to a location in a ‘star’ network with one central node observed by all their peers. Assume, consistent with empirical data, that the overall movement of the group mean is relatively small compared to the overall distribution. Then after social influence, the group mean will have moved a relatively small amount in the direction of the central node. Thus, the probability that the group improves depends only on the probability that the central node is on the same side of the mean as the true answer.

This analysis is an approximation: it is possible for the mean to “overshoot” the true answer, if a central node is in the right direction but also wildly inaccurate. However, Becker et al.’s analysis show that this approximation effectively characterizes empirical data—group outcomes can be predicted by whether or not the central node is in the relative direction of truth. Almaatouq et al. (2020) provide methods for calculating the exact probability of improvement as a function of the belief distribution and the relative influence of the central individual. Here we show that the intuitive approximation given above can be described formally as a limiting case of Almaatouq et al.’s analysis. We then use this formal demonstration to generate testable hypotheses relating unstructured discussion to numeric exchange.
THEORY AND HYPOTHESES

Assume for simplicity (following Almaatouq et al., 2020) a network with one high influence individual (with belief H) and N-1 equally low influence individuals (with mean belief $\bar{L}$). Let $\theta$ be the true answer. Assume also (again following Almaatouq et al.) some belief updating process (e.g. discussion) such that the mean belief after communication can be calculated as a weighted mean, $\mu_{post} = CH + (1 - C)\bar{L}$ where $C$ represents the centrality\(^3\) of the high influence individual, which in DeGroot (1974) would be a weight proportional to network centrality. Let $\mu_{pre}$ be the pre-communication belief, and note that $\mu_{pre} = \frac{1}{N}H + \frac{N-1}{N}\bar{L}$. Thus when $C = \frac{1}{N}$, everybody is equally influential, and when $C = 1$, the group simply adopts the high influence individual’s belief. Assume that $C \geq \frac{1}{N}$.

Define “group error” as the distance between the mean belief and the true answer, i.e. $|\mu - \theta|$ where $\theta$ is the true answer. Note that, as observed by Becker et al., (2017), there are two primary cases of interest—when H and $\theta$ are on the same side of $\mu_{pre}$, meaning that H is in the direction of truth relative to $\mu_{pre}$; and when H and $\theta$ are on opposite sides $\mu_{pre}$. When H is on the opposite side from truth, the group error will increase. When H is on the same side as truth, then group error will decrease, i.e. the group will become more accurate, as long as $\mu_{post} - \mu_{pre} < 2(\theta - \mu_{pre})$. Note that the right side of the equation is independent of C, and the left side equals 0 when C=0 and increases in magnitude monotonically as C increases. In the case where $H - \mu_{pre} < 2(\theta - \mu_{pre})$, i.e. where the central individual is relatively accurate, then communication will increase group accuracy regardless of the value of C. In contrast, when

\(^3\) Almaatouq et al. (2020) present a slightly rearranged version of the weighted sum in their equation S.5: $\mu_{post} = \omega H + (1 - \omega)(\frac{H}{N} + \bar{L})$. \(\omega\) represents a form of network centralization, a measurement of the overall inequality in influence which we measure below as the Gini coefficient. Our measurement of network centrality, the influence of a single individual, is equivalent in this case as $C = \frac{1}{N} + \frac{N-1}{N}\omega$. We make this rearrangement of terms in order to represent the post-influence estimate as a weighted sum of H and L.
$H - \mu_{pre} > 2(\theta - \mu_{pre})$, the central individual is in the direction of the truth but also very inaccurate relative to the group as a whole. In such a case, a highly influential $H$ may cause the group to “overshoot” the true value so far that it initially moves towards the true value but then moves past the true value and the group ultimately ends up less accurate than it started. This will occur for example when $C=1$, and the group simply adopts the central individual’s belief. More generally, there is some value $C'$ such that for all $C' \leq C \leq 1$ the group mean will be less accurate. Importantly, this also means that for all $C \leq C'$ the group will, despite being influenced by such a highly inaccurate individual, nonetheless improves by sheer chance.

Assume, then, that this is always the case: either $H$ is on the opposite side of truth, and the group becomes less accurate; or $H$ is on the same side of truth, and $C \leq C'$, and thus the group becomes more accurate. Note that even in experimental settings using where an individual is given highly centralized position in a binary communication network, they may nonetheless hold relatively low C in the weighted influence network (e.g., if people ignore the central node and don’t revise, or make relatively small revisions). This assumption is thus consistent with the behavior of such “star networks” in previous experiments on numeric exchange, i.e. with one highly centralized individual, in which groups did not simply adopt the belief of the central individual but yet moved a relatively small amount in the direction of their belief (Becker et al., 2017).

Finally, let $\phi$ indicate the proportion of individuals in any population whose initial belief is on the same side of the mean as the true answer, as shown in the conceptual diagram in Figure 1. Then, $\phi$ represents the probability that any randomly selected individual will hold a belief on the truth side of the mean. If we further assume that network centrality (influence) is uncorrelated with belief (a point we will empirically assess later) then $\phi$ equals the probability that $H$ is on the truth side of the mean and thus that communication in a centralized network will increase accuracy. This analysis yields a simple empirical heuristic. Centralized networks will improve in expectation when $\phi>0.5$, and become less accurate when $\phi<0.5$. We note that this heuristic, which we formally derive here, is conceptually similar to empirical analyses of centralized networks in lab experiments (Becker et al, 2017). Importantly, this model is not intended as a precise theoretical
Using this framework, we can explain the dynamics of numeric exchange versus unstructured discussion. While numeric exchange does let people can obtain some variation in influence through stubbornness, this level of influence is inherently limited. In contrast, unstructured discussion allows people the opportunity to directly sway each other’s beliefs, allowing any individual potentially unlimited influence over group opinion. Thus while both communication formats can be characterized as demonstrating emergent centralization (even in otherwise decentralized networks), the potential degree of centralization is much greater for unstructured discussion. Unstructured discussion can, in theory, reach maximal centralization (one person dictating group beliefs). As a result, the effect of initial belief distribution—which depends on centralization—will be much stronger for unstructured discussion than for numeric exchange. Moreover, because the centralization (stubbornness) in numeric exchange may be correlated with accuracy, this mild centralization in numeric exchange may be generally beneficial.

Taken together, these observations can explain the apparent contradiction in previous experimental findings. Assume that for some group, numeric exchange has a fixed probability $\omega$ of generating increased accuracy. If an experiment happens to be conducted with estimation tasks that yield initial belief distributions with $\phi<\omega$, then unstructured discussion can be expected to decrease accuracy while numeric exchange increases accuracy; and vice versa when $\phi>\omega$. Because

**Figure 1.** Conceptual diagram showing the calculation for $\phi$, or the proportion of individuals on the truth side of the mean. We find that $\phi$ is a useful heuristic for predicting the effect of group discussion.
previous studies have reported such conflicting results on whether social influence improves accuracy, we assume $\omega \approx 1/2$. Based on this theoretical perspective, we will test the following pre-registered hypotheses:

**H1**: In a logistic regression, $\phi$ will predict the probability of improvement for both unstructured discussion and numeric exchange.

**H2a**: The effect of $\phi$ will be greater for unstructured discussion than for numeric exchange.

**H2b**: Unstructured discussion will outperform numeric exchange when $\phi > 1/2$, and vice versa when $\phi < 1/2$.

**H3**: In unstructured discussion, the effect of $\phi$ will be greater for groups with higher centralization scores.

### EMPIRICAL METHODS

We first tested these hypotheses on publicly available data from previous experiments (Becker et al., 2017, 2019; Gürçay et al., 2015; Lorenz et al., 2011). While 1 experiment tested unstructured discussion, and 3 experiments tested numeric exchange, no data available to us tested both numeric exchange and discussion. As a result, this analysis compares both across experimental procedures as well as across question content, thus limiting the validity of causal inference in comparing the two types of communication. In order to test the robustness of our initial findings, we used this initial data to conduct statistical power tests and generate hypotheses for a pre-registered replication experiment to directly compare numeric exchange with unstructured communication. In this second experiment, we also pre-tested tasks to ensure that our experimental trials included a wide range for $\phi$. Overall, the goal of this analysis is to identify potential heterogeneity in the effect of social influence that may have been overlooked in previous research. Table 1 shows the distribution of trials by $\phi$ for both sets of experiments.
To illustrate the explanatory power of this theory, this paper presents a reanalysis of data from previous experiments that measured belief accuracy in groups before and after interaction. This reanalysis tests for the effects of emergent network centralization. This analysis uses four previously published studies (Becker et al., 2017, 2019; Gürçay et al., 2015; Lorenz et al., 2011). These datasets were all made publicly available through the initial publications. Detailed methods can be found in the initial publications, and each study follows a similar procedure. Subjects were asked to complete estimation tasks (e.g. visual estimation, trivia questions, and political facts) before and after exchanging information via a computer mediated communication process. An example of visual estimation task is an image of a jar of gumballs where subjects are asked to estimate how many gumballs are in the jar. An example of a trivia question is estimating the length of the border of Switzerland. An example of a political fact is asking subjects to estimate the number of undocumented immigrants living in the United States.

In three of the studies (Lorenz et al, 2011; Becker et al, 2017; Becker et al, 2019) subjects only exchanged numeric estimates. These studies therefore represent a method equivalent to a digitally mediated version of the Delphi method. Lorenz et al. (2011) allowed 5 rounds of revision (1 independent estimate and 4 socially influenced estimates) while Becker et al. (2017; 2019) allowed 3 rounds of revision (1 independent estimate and 2 socially influenced estimates). In contrast, Gürçay et al allowed subjects to engage in continuous, unstructured discussion via a computer chat interface, so that subjects provided only two answers, a pre-discussion and a post-discussion estimate (1 independent estimate and 1 socially influenced estimate). The data from Gürçay et al.

| Reanalysis | Replication |
|------------|-------------|
| Toward     | 63%         | 49.5%       |
| Away       | 29%         | 46.5%       |
| Split      | 8%          | 4%          |

Table 1. The distribution of trials by the majority opinion as measured by $\phi$. While the data in previous experimental trials was heavily skewed toward a majority-correct task set, our replication experiment successfully produced a mixed balance of outcomes for $\phi$. 

Reanalysis

To illustrate the explanatory power of this theory, this paper presents a reanalysis of data from previous experiments that measured belief accuracy in groups before and after interaction. This reanalysis tests for the effects of emergent network centralization. This analysis uses four previously published studies (Becker et al., 2017, 2019; Gürçay et al., 2015; Lorenz et al., 2011). These datasets were all made publicly available through the initial publications. Detailed methods can be found in the initial publications, and each study follows a similar procedure. Subjects were asked to complete estimation tasks (e.g. visual estimation, trivia questions, and political facts) before and after exchanging information via a computer mediated communication process. An example of visual estimation task is an image of a jar of gumballs where subjects are asked to estimate how many gumballs are in the jar. An example of a trivia question is estimating the length of the border of Switzerland. An example of a political fact is asking subjects to estimate the number of undocumented immigrants living in the United States. In three of the studies (Lorenz et al, 2011; Becker et al, 2017; Becker et al, 2019) subjects only exchanged numeric estimates. These studies therefore represent a method equivalent to a digitally mediated version of the Delphi method. Lorenz et al. (2011) allowed 5 rounds of revision (1 independent estimate and 4 socially influenced estimates) while Becker et al. (2017; 2019) allowed 3 rounds of revision (1 independent estimate and 2 socially influenced estimates). In contrast, Gürçay et al allowed subjects to engage in continuous, unstructured discussion via a computer chat interface, so that subjects provided only two answers, a pre-discussion and a post-discussion estimate (1 independent estimate and 1 socially influenced estimate). The data from Gürçay et al.
is missing chat transcripts from 5 groups, and therefore those trials are omitted from analyses where the chat transcripts are necessary (i.e., measuring emergent centralization).

**Replication Study**

Our replication study follows the same general research paradigm as previous experiments, with two key design features. First, we pre-tested questions to ensure that our tasks covered a wide range of $\varphi$, in order to identify heterogeneity that may have been overlooked in previous research comparing unstructured discussion and numeric exchange. Second, our design allows for direct comparison between unstructured discussion and numeric (Delphi) exchange, where previous studies in our dataset only allowed either one or the other. Following previous research, subjects first provided an initial independent estimate for a numeric estimation task, then engaged in social information exchange, and then provided a final estimate. Subjects were paid based on their accuracy. A complete list of questions is provided in the Appendix.

Subjects were recruited from a panel of 21,000 Amazon Mechanical Turk members maintained for a virtual behavioral laboratory at the host institution. Prior to the beginning of the experiment, subjects were randomly assigned to be recruited either to unstructured discussion groups or Delphi groups. This method allowed us to run the two experiments in parallel while maintaining random assignment to conditions. Subjects were recruited in batches by sending an email with a link to access the experimental website. All subjects who arrived at the website at for a given experimental session were randomly assigned to a group of 20 individuals. If the number of subjects who arrived at the page was not divisible by 20 (e.g., if 59 subjects arrived) individuals would be randomly assigned either to participate or not. Subjects who were not assigned to a group were returned to the pool and invited at a later time.

For unstructured discussion, subjects were given 60 seconds to read the question and provide their initial estimate. They were then placed into a chatroom with the other subjects in their group and given 3 minutes to discuss the question. After this period, subjects were given 30 seconds to provide their final answer. At each stage, a countdown timer indicated the time remaining for that stage.
For numeric (Delphi) exchange, subjects were given 60 seconds to read the question and provide their initial estimate. Subjects were then shown a list of the answers provided by other subjects in their group as well as the average of answers provided by other subjects. Answers were randomized for each subject, so that each subject saw answers in a different order. Subjects were given 60 seconds to review the responses of other subjects and provide an updated (second) estimate. This process was then repeated, allowing subjects to respond to the revised answers of their peers and provide a third estimate. Finally, subjects were shown the revised answers of their peers and given 30 seconds to provide a final answer. At each stage, a countdown timer indicated the time remaining for that stage.

By this method, subjects in each condition had 60 seconds to provide an initial estimate, 3 minutes to respond to the estimates of their peers, and 30 seconds to provide a final answer. In total, we collected 10 trials (i.e., 10 groups of 20 individuals) each for 10 unique questions for each condition, producing a total of 100 trials of unstructured discussion and 100 trials of Delphi exchange. In total, we collected data from 4,000 unique subjects. This pre-registered sample size was based on power tests using the re-analysis of previous data to achieve 80% power on our main hypothesis.

**Analysis**

For the purposes of our analysis, a single experimental consists of a single group of individuals completing a single estimation task. For each experimental trial, we ask a simple question: after social influence, was the mean estimate more accurate, i.e. closer to the true answer? Our primary outcome of interest, for any given experimental condition, is the proportion of trials in which the average answer became more accurate. We assess these outcomes using one- and two-sample proportion tests as well as logistic regression. For our pre-registered reanalysis of previously published data, we used cluster-robust logistic regression with fixed effects intercepts due to the structure of the data, since each recruited group answered multiple questions (i.e., completed multiple trials). For previously published data on numeric exchange, we combined all the data and analyzed it as if collected in a single experiment, clustering on dataset as well as groups within each dataset. In order to test the hypothesis that the effect of \( \varphi \) is moderated by centralization, we
measure centrality for each individual in the discussion groups as the number of chat messages they sent. We measure network centralization as the Gini coefficient on centrality scores (Badham, 2013).

The analysis presented here differs slightly from our pre-registered analysis in order to simplify the presentation of results, improve robustness, and report additional tests of interest. We report the deviations in the Appendix along with full details for the pre-registered analysis. We find comparable results for both our reanalysis of previous data as well as our replication analysis, and therefore present the results of both analyses simultaneously. Our pre-registered analysis included hypothesis tests but did not specify specific “statistical significance” cut-offs and we instead report all relevant tests and the accompanying statistical certainty of those tests.

RESULTS
The primary question facing many prior researchers, for any given communication format, is simply “does this process increase belief accuracy?” We therefore begin our analysis testing whether there is a main effect of social influence for each condition in each dataset. However, we do not find any consistent pattern. In our reanalysis, we found that social influence increased accuracy for a majority of trials in both numeric exchange (61% improvement, \( P<0.001, \chi^2=14.6 \)) and unstructured discussion (55% improvement, \( P=0.11, \chi^2=2.54 \)). In our replication, numeric exchange instead decreased accuracy in a majority of trials (45% improvement, \( P=0.37, \chi^2=0.81 \))
while discussion again increased accuracy (54% improvement, \(P=0.48, \chi^2=0.49\)). These results are consistent with the contradictory results of previous literature, and consistent with our expectation that social influence does not have a single main effect on belief accuracy.

We next test the hypothesis that the effect of social influence is determined by the initial belief distribution. To test this hypothesis, we estimate the effect of \(\phi\) on the probability that social influence improves belief accuracy. Our pre-registered analysis consists of a logistic regression predicting the binary outcome “did error decrease” as a function of \(\phi\). In both the re-analysis and the replication analysis, we found that the probability of improvement increases with \(\phi\) for both discussion and numeric exchange (\(P<0.01\), all four tests, table S1). This analysis shows that the effect of social influence is determined by the initial belief distribution but does not yet explain the inconsistent comparisons of numeric exchange and unstructured discussion.

Consistent with our hypothesis that the effect of initial belief distribution is magnified by network centralization, we find that the effect of \(\phi\) is greater for unstructured discussion than for numeric exchange. Following our pre-registered analysis plan, Figure 2 shows the probability of improvement as a function of the majority opinion by dividing outcomes based on whether \(\phi > \frac{1}{2}\) (majority correct) or \(\phi < \frac{1}{2}\) (majority incorrect). For unstructured discussion, the initial belief distribution determined whether information exchange helped or harmed accuracy. In discussion trials, the majority opinion significantly predicted outcomes (\(P<0.01\) both datasets, proportion test): when \(\phi > \frac{1}{2}\), discussion increased accuracy in 66% of reanalysis trials and 69% of replication trials, but when \(\phi < \frac{1}{2}\), discussion increased accuracy in only 36% and 40% of trials, respectively. In contrast, the majority opinion has a negligible effect on improvement for numeric exchange (\(P>0.44\), both datasets, proportion test). While the two datasets disagreed regarding the main effect of numeric exchange (suggesting no main effect at all), the location of the majority opinion did not tip the scales for numeric exchange in either sets of trials. Most importantly, Figure 2 shows how previous experiments could at times yield conflicting results: when the majority is away from truth, numeric exchange improves accuracy more than discussion (\(P<0.01\), reanalysis; \(P=0.82\), replication; \(\chi^2=7.8, 0.05\)); when the majority is toward truth, discussion is superior (\(P=0.59\), reanalysis; \(P=0.06\), replication; \(\chi^2=0.3, 3.4\)). We note that while these two-sample tests are not
statistically clear in every case, the overall effect is consistent across replication: when $\varphi > \frac{1}{2}$, the probability of improvement in discussion is greater than numeric exchange for all cases, and vice versa when $\varphi < \frac{1}{2}$.

We note that these empirical analyses so far, regardless of the underlying mechanisms, are sufficient in principle to explain prior results. However, we further hypothesize that this variation is driven by emergent centralization in discussion networks, which amplifies the effects of initial distribution. To measure emergent centralization in unstructured discussion, we calculate the Gini coefficient on individual contribution as measured by the number of chat messages sent. We then test for an interaction between centralization and $\varphi$ in a logistic regression, finding results that are statistically uncertain but in the correct direction for both our reanalysis and our replication trials ($P=0.31, 0.32$ respectively; see Table A3).

As an additional exploratory test for the effect of centralization, we also estimate the extent to which the most talkative (central) person in each group predicts group outcomes, as shown in Figure 3. To measure this, we determine for each discussion trial who the most talkative person is, and whether their initial estimate is in the direction of truth. We then measure whether groups are more likely to improve when the most talkative (central) individual is initially in the correct direction compared to the group mean. We find for both reanalysis and replication that discussion
is significantly more likely to improve group accuracy when the most talkative person is in the
correct direction (P<0.01, both analyses, proportion test). Finally, we also test whether individuals
who are more initially accurate are likely to be more talkative. To compare across tasks with
different error ranges, we convert each person’s error and talkativeness into quartiles\(^4\) grouped by
task (for error) and dataset (for talkativeness). We then measure the correlation between these two
values (using cluster-robust regression as above for the re-analysis, clustering data by trial) finding
a near-zero correlation between talkativeness and accuracy for both datasets (-0.01, \(P>0.64\) for the
reanalysis, 0.05, \(P>0.06\) for our replication).

DISCUSSION
Our results offer an explanation for why prior research sometimes showed that mediated processes
improved group belief accuracy over unstructured discussion, and sometimes showed the opposite.
We show through both re-analysis and a novel pre-registered experiment that unstructured
discussion will sometimes increase group accuracy and sometimes decrease accuracy, depending
on \(\phi\), i.e. the proportion of individuals on the true side of the mean. In contrast, numeric exchange
is only minimally (if at all) impacted by \(\phi\). As a result unstructured discussion sometimes
outperforms numeric exchange and sometimes numeric exchange is optimal. We note that one
simplifying interpretation of this dynamic would be to call it a “majority effect,” due to the
tendency for the statistical majority to predict outcomes. However, our results display a critical
paradox: while collectively groups are drawn—on average—to the majority as represented by the
median belief, each individual group is drawn not to the majority for that group but to the most
central individual. In this way, the dynamic we describe differs fundamentally from majority rules
models as in e.g. polarization (Sunstein, 2006).

Because these results depend on variation in the statistical properties of the pre-discussion
belief distribution, another way of interpreting these results is that the relative optimality of
discussion or numeric exchange depends on the estimation task in which the group is engaging.

\(^4\) Due to the relatively small group size, bins smaller than quartiles were not possible for all trials.
We note, for example, research on such simple estimation tasks as counting (e.g., jellybeans in a jar) that belief distributions for similar classes of estimation task yield similar and reliably predictable shapes (Kao et al., 2018a). Thus for a group that may regularly need to complete the same class of estimation tasks routinely (e.g., estimating the success of an advertising campaign, a problem notoriously resilient to statistical aggregation [Hartnett et al., 2016]) it may even be possible to calibrate group behavior, and choose an appropriate communication strategy, based on long term feedback. For cases where this is not possible, or where highly mediated numeric exchange is infeasible or undesirable, our theoretical perspective offers additional possibilities.

In addition to resolving contradictions in previous findings, our results also show how network theoretical results can help to explain belief formation in otherwise unstructured communication such as committee discussion or board meetings. While prior empirical analyses on belief accuracy (Almaatouq et al., 2020; Becker et al., 2017) emphasized a binary network model focused on the pattern of who could communicate with whom—as would be important to studying, e.g., organizational structure—we find that emergent networks of influence can shape dynamics in the same way as explicit communication networks. While the networks in our analyses were spontaneous, emerging from the independent characteristics of anonymous strangers, communication in practice is likely shaped by many pre-existing patterns such as status relations, claims to expertise, and prior social ties. As a result, we expect that future research will find that traditional methods in organizational network research may have additional use in explaining the relative optimality of decision-making by different teams and groups.

Limitations
One clear limitation of this analysis is that our results are relatively uncertain from a statistical perspective. Results are relatively clear for our main hypothesis regarding the differential effect of $\phi$ on discussion versus numeric exchange. The dynamics of unstructured discussion show a strong effect of $\phi$ consistent across both our reanalysis and our replication, and numeric exchange varies in the main effect of social influence but nonetheless shows a consistent and relatively weak moderating effect of $\phi$. However, our attempt to directly measure the hypothesized mechanism, emergent centrality, generated statistically uncertain results. Specifically, we did not see a strong
interaction effect between $\phi$ and centralization as measured by the Gini coefficient on talkativeness. However, we do note that the effects were in the expected direction, and were again consistent across both our reanalysis and our replication. Thus, results are at the very least consistent with our theory and hypothesis. Moreover, we note that our additional analysis showed that the belief of central (talkative) individuals are highly predictive of group outcomes. Most importantly, our observation that outcomes depend on $\phi$, as shown in Figure 2, is minimally sufficient to explain contradictions in previous research. Additional research may be warranted to further identify the dynamics of centralization in group communication.

The biggest challenge facing our research is the limited ability to draw generalizable conclusions outside of the laboratory. In particular, our reliance on networks of anonymous strangers makes it difficult to generalize to organizational settings where people interact within the context of previously established networks. One limitation in this respect is that we measured only one type of centrality, talkativeness, and while influence networks in practice will be determined by mechanisms such as group norms and status relations. On the one hand, this limitation suggests that our results may have underestimated the true effects of emergent networks, as group dynamics such as status and norms may operate more strongly and deeply than factors affecting the interaction of anonymous strangers. As a result, we may reach a fairly reliable conclusion that unstructured discussion in organizational settings acts like centralized networks, and that the effect on accuracy is likely to vary widely by task. However, this limitation also means additional research is needed to understand the extent to which these dynamics are predictable and thus what may be optimal strategy in any given situation.

Conclusion
In the simple context of our experimental design, a network theoretical perspective may seem like methodological overkill—as described in our theoretical analysis, the dynamic of centralized networks can be approximated by a simple weighted combination of high and low influence contributors. Thus, although the theoretical model and predictions presented here may have emerged from the literature of network scholarship, the details of the models are not necessary for solving the specific problem posed presented here. However, we do wish to close by observing
that the flexibility of this model means that future research can develop more complex approaches while maintaining theoretical compatibility with the paradigm presented here and thus earlier research on the Delphi method and related work. This flexibility will be especially important when accounting for sources of influence, such as status, that are not reflected in the laboratory conditions we study. We hope that this theoretical approach will help to draw a bridge between the laboratory results on belief accuracy and organizational communication in practice.

This theoretical approach also points the way to a possible communication strategy that can mitigate the variability we observe in unstructured discussion. Suppose that numeric exchange or other carefully mediated communication processes are, for some normative or operational reason, infeasible or undesirable. This limitation may occur if, for example, people need to share complex information that cannot be transmitted by numbers alone. This limitation may also occur when already have exchanged information informally, e.g. if a process of sensemaking (Weick et al., 2005) where a problem is defined occurs prior to the formal decision. In such a case, network structure may provide a solution. By embedding individuals in small interaction groups with just a handful of members, a group may maintain overall connectivity via overlapping group membership while nonetheless limiting the ability for any one individual to become overly central. Both theoretical hypotheses and analysis of such a strategy would require the more flexible model presented here.

REFERENCES
Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. Proceedings of the National Academy of Sciences, 117(21), 11379–11386.
Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. Management Science, 63(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374
Badham, J. M. (2013). Commentary: Measuring the shape of degree distributions. Network Science, 1(2), 213–225.
Banerjee, A. V. (1992). A simple model of herd behavior. The Quarterly Journal of Economics, 107(3), 797–817.
Bayus, B. L. (2013). Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community. Management Science, 59(1), 226–244.
Becker, J., Brackbill, D., & 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., & Centola, D. (2019). The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences, 201817195.*

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in Probability and Frequency Judgments: A Critical Examination. *Organizational Behavior and Human Decision Processes, 65*(3), 212–219. https://doi.org/10.1006/obhd.1996.0021

Budescu, D. V., & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science, 61*(2), 267–280.

Chen, H., De, P., Hu, Y. J., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies, 27*(5), 1367–1403.

Christensen, C., Abbott, A. S., Gretchen, B., & Chapman, F. A. (2003). Team Medical Decision Making. In *Decision making in health care: Theory, psychology, and applications* (pp. 267–285).

Converse, P. E. (1962). *The nature of belief systems in mass publics.* Survey Research Center, University of Michigan.

Cowgill, B., & Zitzewitz, E. (2015). Corporate prediction markets: Evidence from google, ford, and firm x. *The Review of Economic Studies, rdv014.*

Da, Z., & Huang, X. (2019). Harnessing the Wisdom of Crowds. *Management Science, Forthcoming.*

Dalkey, N. (1969). *The Delphi method: An experimental study of group opinion.* RAND CORP SANTA MONICA CALIF.

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science, 9*(3), 458–467.

Davis, J. H. (1973). Group decision and social interaction: A theory of social decision schemes. *Psychological Review, 80*(2), 97–125.

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

Drogen, L. A., & Jha, V. (2013). *Generating Abnormal Returns Using Crowdsourced Earnings Forecasts from Estimize.*

Dvorak, P. (2008, September 6). Best Buy Taps ’Prediction Market’—WSJ. *Wall Street Journal.* https://www.wsj.com/articles/SB122152452811139909

Galton, F. (1907). Vox populi (The wisdom of crowds). *Nature, 75,* 450–51.

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition, 30*(6), 652–668.

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

Gough, R. (1975). The effect of group format on aggregate subjective probability distributions. *Utility, Probability, and Human Decision Making.*

Green, K. C., Armstrong, J. S., & Graefe, A. (2007). Methods to elicit forecasts from groups: Delphi and prediction markets compared. *Foresight: The International Journal of Applied Forecasting, 8,* 17–20.
Gürçay, B., Mellers, B. A., & Baron, J. (2015). The power of social influence on estimation accuracy. *Journal of Behavioral Decision Making, 28*(3), 250–261.

Gustafson, D. H., Shukla, R. K., Delbecq, A., & Walster, G. W. (1973). A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. *Organizational Behavior and Human Performance, 9*(2), 280–291.

Hartnett, N., Kennedy, R., Sharp, B., & Greenacre, L. (2016). Marketers’ Intuitions about the Sales Effectiveness of Advertisements. *Journal of Marketing Behavior, 2*(2–3), 177–194.

Hastie, R. (1986). Experimental evidence on group accuracy. *Decision Research, 2*, 129–157.

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

Humphrey-Murto, S., & de Wit, M. (2019). The Delphi method—More research please. *Journal of Clinical Epidemiology, 106*, 136–139.

Janis, I. L. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes* (Vol. 349). Houghton Mifflin Boston.

Jansen, W. J., Jin, X., & de Winter, J. M. (2016). Forecasting and nowcasting real GDP: Comparing statistical models and subjective forecasts. *International Journal of Forecasting, 32*(2), 411–436.

Jørgensen, M. (2004). A review of studies on expert estimation of software development effort. *Journal of Systems and Software, 70*(1–2), 37–60. https://doi.org/10.1016/S0164-1212(02)00156-5

Kao, A. B., Berdahl, A. M., Hartnett, A. T., Lutz, M. J., Bak-Coleman, J. B., Ioannou, C. C., Giam, X., & Couzin, I. D. (2018a). Counteracting estimation bias and social influence to improve the wisdom of crowds. *BioRxiv*, 288191.

Kao, A. B., Berdahl, A. M., Hartnett, A. T., Lutz, M. J., Bak-Coleman, J. B., Ioannou, C. C., Giam, X., & Couzin, I. D. (2018b). Counteracting estimation bias and social influence to improve the wisdom of crowds. *Journal of the Royal Society, Interface, 15*(141). https://doi.org/10.1098/rsif.2018.0130

Krafft, P. M., Zheng, J., Pan, W., Della Penna, N., Altshuler, Y., Shmueli, E., Tenenbaum, J. B., & Pentland, A. (2016). Human collective intelligence as distributed Bayesian inference. *ArXiv Preprint ArXiv:1608.01987.*

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Argenziano, G., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences, 201601827.*

Lamberson, P. J., & Page, S. E. (2012). Optimal forecasting groups. *Management Science, 58*(4), 805–810.

Larreche, J.-C., & Moinpour, R. (1983). Managerial judgment in marketing: The concept of expertise. *Journal of Marketing Research, 20*(2), 110–121.

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

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, 108*(22), 9020–9025.
Malone, T. W., & Bernstein, M. S. (2015). Handbook of collective intelligence. MIT Press.
Mannes, A. E. (2009). Are we wise about the wisdom of crowds? The use of group judgments in belief revision. Management Science, 55(8), 1267–1279.
March, J. G. (1991). Exploration and exploitation in organizational learning. Organization Science, 2(1), 71–87.
Massa, M., Reuter, J., Zitzewitz, E., & Stanford, G. S. B. (2006). The rise of teams in fund management. Working paper, INSEAD.
Minson, J. A., Mueller, J. S., & Larrick, R. P. (2018). The Contingent Wisdom of Dyads: When Discussion Enhances vs. Undermines the Accuracy of Collaborative Judgments. Management Science, 64(9), 4177–4192. https://doi.org/10.1287/mnsc.2017.2823
Mollick, E., & Nanda, R. (2016). Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts. Management Science, 62(6), 1533–1553. https://doi.org/10.1287/mnsc.2015.2207
Myers, D. G., & Bishop, G. D. (1971). Enhancement of dominant attitudes in group discussion. Journal of Personality and Social Psychology, 20(3), 386.
Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. Psychological Bulletin, 83(4), 602.
Page, S. E. (2007). The difference: How the power of diversity creates better groups, firms, schools, and societies. Princeton University Press.
Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. International Journal of Forecasting, 15(4), 353–375.
Stoner, J. A. F. (1961). A comparison of individual and group decisions involving risk [PhD Thesis]. Massachusetts Institute of Technology.
Stroebe, W., Nijstad, B. A., & Rietzschel, E. F. (2010). Chapter four—beyond productivity loss in brainstorming groups: The evolution of a question. Advances in Experimental Social Psychology, 43, 157–203.
Sunstein, C. R. (2006). Infotopia: How many minds produce knowledge. Oxford University Press.
Surowiecki, J. (2004). The wisdom of crowds. Anchor.
Turner, M. E., & Pratkanis, A. R. (1998). Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory. Organizational Behavior and Human Decision Processes, 73(2–3), 105–115.
Tversky, A., & Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. In Utility, probability, and human decision making (pp. 141–162). Springer. http://link.springer.com/chapter/10.1007/978-94-010-1834-0_8
Ven, A. H. V. D., & Delbecq, A. L. (1974). The effectiveness of nominal, Delphi, and interacting group decision making processes. Academy of Management Journal, 17(4), 605–621.
Weick, K. E., & Roberts, K. H. (1993). Collective mind in organizations: Heedful interrelating on flight decks. Administrative Science Quarterly, 357–381.
Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. Organization Science, 16(4), 409–421.
Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. Science, 330(6004), 686–688.
Replication code and data for all analyses presented in this paper are available at:

- [https://github.com/joshua-a-becker/emergent-network-structure](https://github.com/joshua-a-becker/emergent-network-structure)

Pre-registration available at:

- [https://osf.io/9xq2j](https://osf.io/9xq2j)

**APPENDIX**

Replication code and data for all analyses presented in this paper are available at:

- [https://github.com/joshua-a-becker/emergent-network-structure](https://github.com/joshua-a-becker/emergent-network-structure)

Pre-registration available at:

- [https://osf.io/9xq2j](https://osf.io/9xq2j)

**Departure from Pre-registered Analysis**

The pre-registered tests for the finding shown in Figure 2 included a series of descriptive one-sample and two-sample proportion tests for the probability of improvement for each point in the figure. However, to simplify the readability of our manuscript, we focus the main text on more central tests and include only the two-sample proportion tests for each line in Figure 2. Here we present the full pre-registered descriptive analysis. Table A1 presents a series of one-sample proportion tests, asking at each point in the parameter space whether social influence significantly improves (or decreases) group error.

The main text also reports a slightly modified version of our test for an interaction effect between $\varphi$ and talkativeness centralization. Our pre-registered analysis only calculated the Gini coefficient for participants present in the discussion, which incorrectly omitted zeros from the calculation—i.e., people who did not contribute to discussion (and thus had minimal influence)

| Majority Relative to Truth | Format | Replication | % Improved | P.val |
|---------------------------|--------|-------------|------------|-------|
| Away Delphi               | reanalysis | 0.584 | 0.171 |
| Away Delphi               | replication | 0.44 | 0.480 |
| Away Discussion           | reanalysis | 0.363 | 0.008 |
| Away Discussion           | replication | 0.395 | 0.222 |
| Toward Delphi             | reanalysis | 0.629 | 0.000 |
| Toward Delphi             | replication | 0.489 | 1.000 |
| Toward Discussion         | reanalysis | 0.661 | 0.000 |
| Toward Discussion         | replication | 0.692 | 0.008 |

**Table A1.** Pre-registered proportion tests. Each row represents a one-sample proportion test for a different part of the parameter space.
but nonetheless contributed to the collective belief by including an opinion. We note that while this test was statistically significant in its original version, it is not significant in the revised version. We provide code for both tests in our replication materials.

Supplemental Analysis

Table A2 shows the regression results for the effect of Phi. Table A3 shows the regression results for an interaction between phi and centralization as measured by Gini coefficient.

Replication Question Text

Crowdfunding 1. Consider this crowdfunding campaign: The goal of this app is to promote new music discovery in a fun and different way. This app would allow musicians to “drop” songs at specific physical locations. Anyone using the app would then be able to listen to the song by visiting that location. The app sought £30,000 (British pounds) and offered funders equity in the company, with a total equity of 35% for the whole campaign. How much money do you think the campaign raised? Answer: 30,000. Source: https://www.seedrs.com/gigdropper

Crowdfunding 2. Consider this crowdfunding campaign: The product is headphones designed for dance music. The goal of the product is to replicate the sound style of being in a club or party. The campaign followed a successful prior round of funding, and the company has already sold
thousands of units. This campaign sought an additional £100,000 (British pounds) in exchange for equity in the company, and ended up exceeding their goals. How much money do you think the campaign raised? Answer: 142,770. Source: https://www.seedrs.com/pump-audio

**Socio-Economic 1:** In 2009, approximately 690 million passengers boarded a plane. (So a round-trip flight counts for 2 passengers boarding.) How many of these passengers boarded out of an airport in the New York City area (JFK, La Guardia, and Newark)? (Give your answer in millions—e.g., enter 1 for 1 million.) Answer: 41. Source: http://infochimps.org/datasets/d35-million-us-domestic-flights-from-1990-to-2009

**Socio-Economic 2:** Across all colleges where the US Department of Education collected data, the average tuition revenue per full time (or equivalent) student was $10,438 per year. In terms of dollars, how much money do you think was spent on instruction, per student? Answer: 7912. Source: https://collegescorecard.ed.gov/data/

**Art 1:** [display image of Planteuse des Betteraves] This drawing by Vincent Van Gogh sold at auction in May, 2018. It is 18 inches tall by 20 inches wide, charcoal on paper. How much did it sell for? (Answer in millions of dollars, e.g. enter 1 for $1 million or 0.5 for $500,000) Answer:
Art 2: [Image of La Lampe] This painting by Pablo Picasso sold at auction in November, 2018. It is 64 inches tall by 51 inches wide, oil on canvas. How much did it sell for? (Answer in millions of dollars, e.g. enter 1 for $1 million or 0.5 for $500,000). Answer: $29.6 million. Source: https://www.christies.com/lotfinder/Lot/pablo-picasso-1881-1973-la-lampe-6169488-details.aspx

Geopolitics 1: The Armed Conflict Location & Event Data Project (ACLED) is a non-governmental organization that tracks violent conflict in Asia, the Middle East, and Africa. One type of event they track is those where civilians were intentionally targeted. In 2018, they recorded 841 such events in Somalia. How many of this type of event do you think they recorded in Yemen for 2018? Answer: 609. Source: https://www.acleddata.com/data/

Geopolitics 2: The Armed Conflict Location & Event Data Project (ACLED) is a non-governmental organization that tracks violent conflict in Asia, the Middle East, and Africa. One type of event they track is those where civilians were intentionally targeted. In 2018, they recorded 841 such events in Somalia. How many of this type of event do you think they recorded in Syria for 2018? Answer: 1501. Source: https://www.acleddata.com/data/

Gun Violence 1: Gun Violence Archive (GVA) is a not for profit corporation that tracks gun-related violence in the United States. In 2018, GVA recorded 1,113 events in Baltimore. (A single event might involve more than 1 person.) How many events do you think they recorded in Chicago in 2018? Answer: 2812. Source: https://github.com/awesomedata/awesome-public-datasets#socialsciences

Gun Violence 2: Gun Violence Archive (GVA) is a not for profit corporation that tracks gun-related violence in the United States. In 2018, GVA recorded 1,113 events in Baltimore. (A single
event might involve more than 1 person.) How many do you think they recorded in Philadelphia in 2018? Answer: 570. Source: https://github.com/awesomedata/awesome-public-datasets #socialsciences