The Impact of Group Size on the Discovery of Hidden Profiles in Online Discussion Groups

YLA TAUSCZIK, University of Maryland
XIAOYUN HUANG, University of Maryland

Online discussions help individuals to gather knowledge and make important decisions in diverse areas from health and finance to computing and data science. Online discussion groups exhibit unique group dynamics not found in traditional small groups, such as staggered participation and asynchronous communication, and the effects of these features on knowledge sharing is not well-understood. In this paper we focus on one such aspect: wide variation in group size. Using a controlled experiment with a hidden profile task we evaluate online discussion groups' capacity to share distributed knowledge when group size ranges from 4 to 32 participants. We found that individuals in medium-sized discussions performed the best, and we suggest that this represents a tradeoff in which larger groups tend to share more facts, but have more difficulty than smaller groups at resolving misunderstandings.

CCS Concepts: • **Human-centered computing → Collaborative content creation**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Hidden Profile; Online forums; Collective intelligence; Knowledge sharing; Collective information processing

ACM Reference Format:
Yla Tausczik and Xiaoyun Huang. 0000. The Impact of Group Size on the Discovery of Hidden Profiles in Online Discussion Groups. *ACM Trans. Soc. Comput.* 0, 0, Article 0 (January 0000), 25 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Individuals today often turn to online discussions to gather knowledge and help make important decisions in diverse areas both personal (e.g. health, finance) and professional (e.g. software engineering, data science). For example, the online health discussion board TuDiabetes [30] allows those dealing with the disease to seek experience and feedback from others in similar situations. On question and answer (Q&A) sites like Stack Overflow, professional software developers ask for help debugging code and using new libraries [49]. On Kaggle, data scientists work to identify which machine learning techniques are best for a particular application. These discussions are important because they influence treatment choices, shape code and inform business decisions, among many other contributions.

Discussion on these sites and others takes place in *online discussion groups*, which we define as a set of individuals who assemble and respond to an open call for discussion (e.g., post, question) on an online platform. Online discussion groups form (and dissolve) on an ad hoc basis and are characterized by group dynamics different from those in traditional small groups. Discussion is online, using computer-mediated communication. It is also open to a large pool of potential

Authors' addresses: Yla Tausczik, University of Maryland; Xiaoyun Huang, University of Maryland.

2469-7818/0000/1-ART0 $15.00
https://doi.org/0000001.0000001

participants, in which the choice to participate is voluntary and self-directed. Individuals enter discussion at different times, may choose to come and go, and leave at different times. Thus, participation is staggered, with high turnover and (partially or fully) asynchronous communication.

Although the potential audience for a discussion is very large, often the actual number of contributors may vary widely. For example, on Stack Overflow 61% of questions received 5 or fewer responses, 39% received between 6 and 50 responses, and less than 0.1% of questions received more than 50 responses (extracted using Stack Exchange Data Explorer July 8, 2018). Even these numbers are aggregated across time, so that measures of overlapping participation (reflecting perceived group size) may be much lower. Majchrzak and colleagues [28] have characterized knowledge sharing in this type of medium as decentralized, asynchronous, ongoing, emergent and unplanned.

With the growing importance of online discussion it is important to understand how online discussion groups collectively process information and the conditions under which they do so optimally or sub-optimally. In decision-making, groups exchange information, selectively attend to some information over others, use that information to reason about alternatives, and negotiate individual opinions to form collective judgements [18, 26]. Successful online discussion groups can produce more reliable recommendations than the same group of individuals working independently. For example, a three-year experiment showed that online discussion groups could predict geopolitical events with higher accuracy than either individuals or prediction markets [6]. However, when online discussions groups fail they can produce poor judgements that mislead individuals and organizations and result in false beliefs and poor decisions. For example, parents may skip or delay important vaccinations for their children based on (mis)information and reasoning found on online parenting forums.

Hidden Profile Tasks (HPTs) are the most widely studied method for testing a group's capacity to collectively process information when knowledge is widely distributed across group members [42, 52]. Hidden Profile Tasks are a specific class of problem designed to assess the ability of a small discussion group to share relevant information in a laboratory setting. In HPTs relevant information is distributed across group members such that no person has all the relevant information, thus discovering the optimal solution requires group members to share information. HPTs are designed to simulate a class of real-world problems that require knowledge sharing and collaboration between individuals with different expertise. For example, a data science problem which requires domain knowledge about public health record keeping and knowledge of machine learning. A large body of work in social psychology has shown that traditional small groups are more likely to discuss information known to most people in the group (common) than information only known to a few people in the group (rare) [27]. As a result of this bias traditional small groups often fail to share distributed, rare knowledge and perform sub-optimally on HPTs [27]. The literature also suggests that group size affects knowledge sharing in traditional small groups. On balance larger groups are more biased in how they share knowledge and as a result perform worse on HPTs than smaller groups [27].

HPT can also be used to evaluate how well online discussion groups share and process distributed knowledge. One study partially (although not fully) replicated the information-exchange bias associated with HPTs in online discussion groups of size 30 [45]. This study found that groups were more likely to share common facts than rare facts, when rare facts were known to only one person, but equally likely to share rare facts if they were known to a third of the group. They also found that the unique dynamics of online discussion affected information processing. Due to staggered participation groups sometimes shifted from correct solutions to incorrect solutions because individuals who arrived late discounted opinions shared before they had arrived.

The current study replicates and expands on this prior study [45] by experimentally evaluating the effect of group size on the capacity of online discussion groups to share distributed knowledge using a HPT. We focus on group size in particular because group size can vary widely in online discussions, group size has been shown to affect this particular bias in knowledge sharing, yet group size may be perceived differently in online discussions than in traditional small groups due to the unique dynamics, such as staggered participation. This paper contributes to a nascent research area investigating collective information processing in online discussion groups [30, 45, 46] and the long standing investigation of distributed knowledge sharing in groups using HPTs [52].

2 LITERATURE REVIEW

2.1 Online Discussion Groups

Online discussion groups have been studied most extensively using a community building paradigm [21]. Researchers have shown that online communities, like those offline, face a set of universal social problems that must be overcome to build community, including socializing newcomers, encouraging contribution, and regulating community member behavior [21]. However, the unique nature of online social interaction, including anonymity, high turnover and computer-mediated communication, may make building community more difficult online [21]. Online communities rely on social norms, like norms of civility, and technological affordances, like collaborative filtering, to overcome some of these difficulties and to foster constructive conversation in online discussion forums [20, 22, 23].

In this paper we focus on online discussion using an information processing rather than community building paradigm; a perspective that has received less attention. A few studies have catalogued patterns of discourse in online discussion groups. Researchers find that individuals engage in discourse acts that are suggestive of collective information processing (also referred to as collective sensemaking), such as providing information, reframing problems, constructing arguments, elaborating on others comments, synthesizing comments, on a variety of online discussion platforms [30, 46, 54]. Other work has demonstrated the importance of knowledge sharing in online discussion groups for specialized fields, such as software engineering [2, 44].

A few studies suggest that online discussion groups may not be processing information optimally. Online discussion groups are vulnerable to biases and inefficiencies, such as underprovisioning [13] and social influence [32], which can lead users to attend to less relevant comments at the expense of other more relevant comments [8].

2.1.1 Online Groups & Group Size. While very large online groups are hypothesized to be able to solve problems well because they can draw on ideas from experts with diverse and specialized knowledge [19], observational studies have found a complex, and sometimes inconsistent relationship between group size and performance in online groups. For example, prior work on groups in Wikipedia have found positive, negative, or no relationship between group size and performance depending on the study [3–5]. Which has led researchers to argue that the relationship between group size and performance may depend on other variables, such as whether large groups are diverse [36]. In Q&A communities, one study found that solution quality increased as group size increased, however the benefit of group size had diminishing returns [47]. In taking an experimental approach, this study can better control for group composition and thus better understand how group size relates to performance in online discussion groups.

2.2 Hidden Profile Tasks

Groups have the advantage of being able to draw on different people's knowledge and experience in making a decision. Despite this advantage, foundational work in small group research has shown that small groups are not always able to capitalize on their potential to share distributed knowledge [27, 42]. The most widely used technique to evaluate a groups' capacity to share and process information has been what are called Hidden Profile Tasks (HPTs). In HPTs a group is asked to make a decision, such as choosing the best job candidate [51]. To make the decision the group is given a set of facts; some facts are given to everyone in the group (**common facts**) and some facts are given to only one person in the group (**rare facts**). By design, the set of facts given to each individual (their **information profile**), considered in isolation, supports a sub-optimal solution; to discover the optimal solution groups must share the distributed, rare facts in the scenario (the **hidden profile**).

Stasser and colleagues discovered over a series of experiments [41–43] that groups performed very poorly on this task. Most tended to discuss common facts more than rare facts (**information-exchange bias**), even though rare facts were more important for identifying the best solution. A meta-analysis of 101 independent effects from 65 studies found consistent evidence of an information-exchange bias and poor performance on HPTs; on average these effects are very large [27].

Prior research has supported three main explanations for why groups often fail to share distributed facts in HPTs [52]. First, common facts are known to more people than rare facts, so chance alone indicates that common facts will be shared more frequently, this is referred to as information sampling [42]. Second, people trust information that aligns with their opinions. By design in HPTs common facts align with pre-discussion opinions whereas rare facts conflict with them [15]. Third, due to social comparison processes people trust information that others believe and trust, so that the importance and relevance of common facts seems amplified when they are shared by others [52].

While the majority of prior work on HPTs has used face-to-face (FtoF) communication, several studies have compared FtoF with computer-mediated communication (CMC). Researchers argued CMC would improve information sharing and discussion quality because it provided new affordances, such as a persistent record of information exchanged and the ability to share information without having to wait for others to speak [12]. However a meta-analysis shows no significant difference between the two mediums [27]. There may be advantages for particular types of CMC. For example, Murthy and Kerr [33] found that groups shared more information when using bulletin boards than when using chat. Goyal and colleagues [14] found groups performed better using CMC tools when they were given an interface to externalize and share their sensemaking process. In prior studies, CMC groups almost always had the same temporal dynamics as FtoF groups: all members started at the same time, were present for a fixed period of time, and ended at the same time. When HPTs were tested with online discussion groups, which has distinct temporal dynamics (e.g. staggered arrival, high turnover) one study found that in addition to failing to share facts online discussion groups had difficulty reaching consensus, later members discounted early members opinions, and group decisions shifted over time, often getting worse [45]. However, this study only partially replicated prior work, since they found an information-exchange bias for rare facts given to only one person out of 30, but not rare facts given to only a third of the group.

2.2.1 *Hidden Profile Tasks & Group Size.*

Researchers have proposed two factors that could cause different degrees of information-exchange bias in small and large groups: information sampling and social loafing [27]. According to the information sampling explanation an information-exchange bias emerges because common facts are more likely to be sampled because the proportion of the group that knows a common fact is higher than the proportion of the group that knows a rare fact [41]. If rare facts are known to only one person regardless of the group size (as is the case in the vast majority of prior work), then the information-exchange bias should get worse for larger groups because common facts become even more frequent relative to rare facts. However, if rare facts are known to the same proportion of the group, regardless of size, then the information-exchange bias should be about the same in different sized groups.

Social loafing suggests that individuals in larger groups will be less motivated to contribute because they assume someone else will [24]. If social loafing affects behavior then individuals in larger groups will be less likely to share rare facts because they will be less likely to contribute at all, again making the information-exchange bias worse for larger groups [9].

Prior work evaluating the effect of group size on HPTs has led to mixed findings. Four studies have experimentally tested the effect of group size on the information-exchange bias, comparing groups of size 3-4 to groups of size 6-8 [9, 31, 39, 41]. In two of these studies the information-exchange bias was worse in larger groups, though for different reasons. The first study found that larger groups discussed common facts more than smaller groups, consistent with the information sampling explanation [41], while the second study found that larger groups discussed rare facts less than smaller groups, consistent with the social loafing explanation [9]. The other two studies found no difference in the information-exchange bias across the different sized groups [31, 39].

Two meta-analyses have considered the effect of group size on the information-exchange bias by synthesizing across multiple studies with different sized groups. Both meta-analyses found that information-exchange bias was worse for larger groups; the second also tested the effect of group size on performance and found that larger groups performed worse [27, 35]. However, almost all the studies in the meta-analyses used groups of size 3-8, with the exception of one study which included groups of size 10 [11]. Only one study has examined the HPTs for the much larger groups, however they only examined groups of 30 and were only partially able to replicate the information-exchange bias [45].

3 CURRENT STUDY

A lauded advantage of online discussion forums is the ability to gather knowledge and experiences from many people. We investigated the capacity of online discussion groups' to share distributed knowledge by conducting an experiment with a Hidden Profile Task (HPT) on Amazon's Mechanical Turk (AMT). We formed groups of size 4 to 32 with most of the *ad hoc* features of online discussion groups. Individuals arrived at different times, they could come and go as they pleased, and finish when they wanted. Individuals could get credit without actively participating in the discussion (although they were encouraged to contribute). They communicated using a online discussion forum with typical features of popular platforms (e.g. Reddit, Hacker News). This aligns with many of the characteristics of online discussions: staggered participation; partially asynchronous communication; high turnover; selective participation and computer-mediated communication. Based on this design, we investigated how well online discussion groups of different sizes pooled distributed facts (Research Question 1) and selected the optimal solution (Research Question 2) as well as the factors that influenced their performance (Research Question 3).

Research Question 1: How well do online discussion groups of different sizes pool distributed facts?

Despite their differences, we expected that online discussion groups would also suffer from the information-exchange bias known to affect traditional small groups. The mechanisms that drive this bias–information-sampling, preference-informed evaluation of facts, social comparison processes–should affect these groups as well. Earlier work [45] found only partial support for information-exchange bias in this type of group, but this was based on a very weak version of a HPT. Thus, we predicted:

Prediction 1A: Online discussion groups will share more common facts than rare facts.

As group size varies, the concepts of rare and common facts becomes more complicated. In a group of four, one person knows each rare fact; in a group of sixteen, should we give a rare fact to just one individual or to one quarter of the group? Previous research involving medium-sized groups has used the first interpretation [39, 41]. Which is why they hypothesized that information sampling would lead to more biased information sharing and performance as group size increased (i.e., the ratio of common to rare facts gets worse as group size increases). Here, instead, we chose to maintain the relative ratio between rare and common facts (25%:75%) assuming instead that a constant fraction of the population knows rare facts. Thus, larger groups have an advantage over smaller groups. If more people know a fact, there are more chances for someone to decide to share the fact with the group. We also predicted:

Prediction 1B: Larger online discussion groups will share more facts.

The more complex issue was how group size would influence the degree to which online discussion groups shared common facts relative to rare facts. Although mixed, prior work suggested that larger groups should have a greater information-exchange bias than smaller groups [9, 27, 35, 41]. While information sampling would not be expected to worsen bias in larger online discussion groups due to the way we operationalized rare facts, social loafing would still be expected to worsen bias for larger groups. However, we had the most uncertainty about this prediction, since online discussion groups are fundamentally different than traditional small group discussions, in ways that might affect perception of group size.

Prediction 1C: Larger online discussion groups will tend to share even more common facts relative to rare facts than smaller online discussion groups.

Research Question 2: How well do online discussion groups of different sizes perform on a hidden profile task?

Groups that fail to share rare facts or focus on common facts more than rare facts perform poorly on HPTs [27]. We expected that online discussion groups would share more common facts than rare facts and as a result we expected that they would perform poorly on the HPT. In the prior study of a HPT in online discussion groups which used a weak version of HPT, only 62% of individuals reported the correct solution [45].

Prediction 2A: Online discussion groups will perform poorly on the hidden profile task.

Similarly, we predicted based on prior work that larger online discussion groups would exhibit a greater information-exchange bias, so we also predicted that they would perform worse on the task. As explained in Prediction 1C we had some uncertainty about this prediction due to differences between traditional small group discussions and online discussions.

Prediction 2B: Larger online discussion groups will tend to perform worse than smaller online discussion groups on the hidden profile task.

Research Question 3: Which factors explain why online discussion groups of different sizes perform differently on a hidden profile task?

On balance, prior work suggests larger groups should exhibit a larger information-exchange bias than smaller groups [27], but the empirical evidence is mixed and there are multiple factors that might lead groups of different sizes to pool facts and make decisions differently. As our third research question we tested specific predictions associated with a few of these factors to see if they could explain the observed patterns of behavior.

Another body of work has extensively documented the relationship between group size and social loafing [24]. Social loafing has been hypothesized to affect groups solving HPTs, as well [9]. Individuals in larger groups should be less motivated to contribute due to social loafing, thus we would expect that individuals in larger groups to share fewer common and rare facts. However, we also recognized that unique qualities of online discussion groups, such as different perceptions of group size when there is high turnover and different social norms about when to participate in discussion, might interfere with social loafing.

> *Prediction 3A: Social loafing should result in individuals in larger online discussion groups sharing fewer common and fewer rare facts than individuals in smaller online discussion groups.*

According to the information sampling explanation if the ratio of common and rare facts is held constant across groups of different sizes this suggests that both common and rare facts will be shared more often by larger groups. If each individual shares two facts at random from the facts they know, in larger groups more facts of both types will be shared. As a result bias in information sharing should not get worse as group size increases. (The information-exchange bias should decrease very slightly in larger groups as fact sharing approaches 100% under the information sampling model, but due to the gradual nature of the change and our modest sample size we did not test this prediction explicitly.)

> *Prediction 3B: Information sampling should result in larger online discussion groups sharing more common and rare facts than smaller online discussion groups and no worse bias in information sharing.*

Information sampling suggests individuals in larger groups are equally likely to share a rare fact as individuals in smaller groups, but at a group level larger groups will share more rare facts. Social comparison processes may combine with information sampling to increase the likelihood of an individual sharing a rare fact. As larger groups discuss rare facts, this will increase the perceived importance of rare facts, leading individuals to be more likely to share a rare fact (including ones not already discussed).

> *Prediction 3C: Social comparison processes should result in individuals from larger online discussion groups sharing more rare facts than individuals in smaller online discussion groups.*

Although the experiment kept the distribution of facts constant across groups of different sizes, we expected that larger online discussion groups would have more disagreement because they contained more individuals and individuals hold different experiences and perspectives that lead them to interpret facts differently. We also expected that online discussion groups would have more difficulty reaching a consensus when there was disagreement due to staggered participation and more relaxed consensus norms. Staggered participation and turnover in online discussion groups means individuals are not always present at the same time, which makes it difficult to have a back-and-forth conversation that resolves disagreements. If forming a consensus in online discussion groups is perceived to be difficult, there may not be an expectation to form a consensus. Researchers have observed that online discussion groups often value multiple perspectives and opinions [30]. Therefore we predicted that larger online discussion groups would have more disagreement than smaller groups and that disagreement would be less likely to be resolved through consensus processes.

> *Prediction 3D: Larger online discussion groups should exhibit less agreement than smaller online discussion groups.*

Some or all of these factors might affect behavior in online discussion groups. To complicate matters further, the different factors have different implications for research questions 1 and 2. Social loafing suggests larger online discussion groups will exhibit a larger information-exchange bias, whereas social comparison processes and information sampling suggest larger groups will exhibit a smaller (or about equal) information-exchange bias. Differences in consensus suggest fewer group members will identify the correct solution in larger online discussion groups compared to smaller online discussion groups even when the groups pool all facts. In summary, to explain why group size might differentially affect knowledge sharing we evaluated the two factors described in prior work that investigated the effect of group size in HPTs–information-sampling and social loafing–we also evaluated two new factors–social comparison processes and agreement. These factors were drawn from well-studied social psychological processes (e.g. social loafing) and two of the three mechanisms believed to explain the information-exchange bias (information-sampling, social comparison processes) we did not evaluate the effect of the third mechanism (preference-consistent evaluation of information) because it was not expected to differ in its effect across groups of different sizes.

4 METHOD

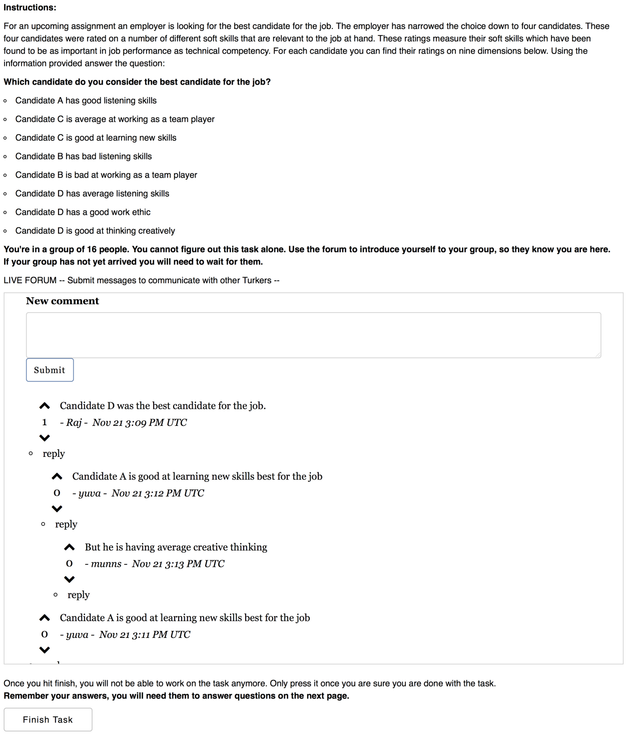
We experimentally evaluated a hidden profile task (HPT) with groups of four different sizes ranging from 4 to 32. The differences between our methods and prior work included the following: 1) the temporal dynamics of members' arrival and departure, 2) the distribution of rare facts in larger groups, 3) the way that we measured performance and 4) the specific interface of our online discussion forum. These differences were created to better represent the way that knowledge is shared on modern discussion platforms like Reddit and Hacker News and on Q&As like Stack Overflow; each difference is explained in more detail below. In all other respects our methods are similar to a majority of prior work in which a HPT has been given to a group working online using computer-mediated communication (CMC).

4.1 Experimental Design & Discussion Platform

The groups had four different sizes: 4, 8, 16, 32, allowing comparison across multiple-sized groups. We chose 4 as the smallest group size because it is a typical group size used in prior studies of traditional small groups [27] and 32 was the largest group size practical for an experiment. This group size range is also typical of online discussions; for example, 90% of groups on www.reddit.com/r/programming and 100% of groups on stackoverflow.com were 32 or less for posts made on an arbitrary day (i.e. June 15, 2018).

In online discussions individuals arrive at different times, leave at different times, and selectively attend to and participate in the conversation by choosing which messages to read and when to comment. These actions are supported by the affordances of a new generation of CMC tools often referred to as social media [28]. Thus, groups sharing knowledge in modern discussion forums have unique temporal dynamics that are distinct from almost all prior work using HPT with the exception of [45].

In this experiment we chose to examine knowledge sharing under these distinct conditions by allowing users to stream into the discussion at different times, come and go as they pleased, and leave as they felt satisfied with the solution. These actions were encouraged through instructions and supported by the affordances of the CMC tool, which mimicked modern discussion forums on other platforms (e.g. Reddit, Hacker News). Using the discussion forum users could read, post, reply, and up or down vote comments (see Fig. 1). Comments were visible and persistent. Comments were displayed in order by time, with child comments nested below parent comments. Each comment included the text of the message, timestamp, and the screen name selected by the sender. No other information was provided about individual participants. In addition, to these common features for asynchronous communication, comments were updated in real-time as is typical for a live forum which are sometimes used when a large number of people are expected to be active at the same time on these platforms. However, no notification system was used to attract individual's attention or alert a user to new content once they had left the platform or shifted to other work (e.g. email alerts, highlighted text, mailbox alerts).



Instructions:

For an upcoming assignment an employer is looking for the best candidate for the job. The employer has narrowed the choice down to four candidates. These four candidates were rated on a number of different soft skills that are relevant to the job at hand. These ratings measure their soft skills which have been found to be as important in job performance as technical competency. For each candidate you can find their ratings on nine dimensions below. Using the information provided answer the question:

Which candidate do you consider the best candidate for the job?

- Candidate A has good listening skills
- Candidate C is average at working as a team player
- Candidate C is good at learning new skills
- Candidate B has bad listening skills
- Candidate B is bad at working as a team player
- Candidate D has average listening skills
- Candidate D has a good work ethic
- Candidate D is good at thinking creatively

You're in a group of 16 people. You cannot figure out this task alone. Use the forum to introduce yourself to your group, so they know you are here. If your group has not yet arrived you will need to wait for them.

LIVE FORUM -- Submit messages to communicate with other Turkers --

New comment

Submit

Candidate D was the best candidate for the job.
1 *- Raj - Nov 21 3:09 PM UTC*

- reply

  Candidate A is good at learning new skills best for the job
  0 *- yuva - Nov 21 3:12 PM UTC*

  - reply

    But he is having average creative thinking
    0 *- munns - Nov 21 3:13 PM UTC*

    - reply

Candidate A is good at learning new skills best for the job
0 *- yuva - Nov 21 3:11 PM UTC*

Once you hit finish, you will not be able to work on the task anymore. Only press it once you are sure you are done with the task.
Remember your answers, you will need them to answer questions on the next page.

Finish Task

Fig. 1. A screenshot of the discussion page, which includes task instructions, an individual information profile, a comment box, and threaded comments.

4.2 Task Design

We designed and created a HPT that asked groups to select the best candidate for a job from four candidates ("you will be working together to identify the best person for a job"); this is a commonly used decision task [42]. Groups were collectively given a set of 16 facts that provided information about how each candidate rated on a job skill ("Candidate A is good at thinking creatively"). However, each group member was only given 8 of these facts. By design the facts given to each individual made a non-ideal candidate appear to be the strongest candidate; only by pooling facts would the group discover the ideal candidate (Table 3). Furthermore, the individual profiles supported two different non-ideal candidates in equal numbers, in order to generate dissent and discussion [38].

HPTs study the differential sharing between two classes of information: rare and common knowledge. Common knowledge is presented to most of the participants while rare knowledge is only given to a few. In a typical study with 3-4 group members common and rare knowledge is operationalized by giving some facts to everyone (common knowledge/facts) and giving other facts to only one person (rare knowledge/facts). Things are more complicated in larger groups.

	Candidates			
Soft Skills	A	B	C	D
learn new skills	+	-	0	0
team player	+			+
listening skills	-	-	-	+
creative thinking	0	0	+	+
work ethic			+	+

Table 1. Example distribution of soft skills for four job candidates.

	group size				
Fact Type	**4**	**8**	**16**	**32**	**%**
Common	3	6	12	24	75%
Rare	1	2	4	8	25%

Table 2. The number of group members given each type of fact for the four group size conditions.

Profile	**Common Facts**	**Rare Facts**	**Preference**
1	A_1+, B_1-, B_2-, B0, C_1+, C_2+	C0, D_1+	C
2	A_1+, A_2+, B_1-, B_2-, C_2+, D0	A0, D_2+	A
3	A_2+, B_1-, B0, C_1+,C_2+, D0	A-, D_3+	C
4	A_1+, A_2+, B_2-, B0, C_1+, D0	C-, D_4+	A
	Ideal Candidate: D		

Table 3. The distribution of common and rare facts among group members in groups of size 4. Facts are represented symbolically (letter = candidate, +/0/- = positive/neutral/negative, subscripts = different facts). Larger groups had an analogous distribution with a unique information profile for each member.

The usual way to operationalize rare facts in larger groups is to provide them to only one person regardless of the size of the group [41]. There are two problems with this method: 1) if rare knowledge is known to a certain fraction of the population, then larger groups ought to contain more individuals with that knowledge and 2) it creates a confound in which rare facts are given to a smaller percentage of the group for larger groups (e.g. 25% for group of 4 vs. ~3% for a group of 32). To correct for these problems we operationalized common knowledge as facts known to 75% of the group and rare knowledge as facts known to 25% of the group (See Table 2). We expect common facts to be discussed more than rare facts because they are known to many more people (See Limitations for discussion of exact percentages).

Each participant was given an information profile containing 8 facts (6 common and 2 rare). Facts were given as written statements in a bulleted list. Information profiles were designed so that, within each group, all group members had different information profiles. Participants were told they needed to work with others to solve the problem ("you cannot figure this task out alone"), but were not explicitly told that they had different information profiles or rare facts. Facts were positive (e.g. "is good at thinking creatively"), negative (e.g. "is a bad team player"), or neutral (e.g. "is average at learning new skills"). To ensure participants would consider these job skills important and of similar value we pretested 24 skills on AMT. Of these we selected two sets of 5 skills that were distinct and rated as equally important to job performance.

4.3 Participants & Procedure

Participants were recruited with a Human Intelligence Task placed on AMT; each person was allowed to participate only once. All 836 participants who were randomly assigned to these 55 groups were included in the analysis (Group Size 4: 15 groups; 8: 11 groups; 16: 15 groups, 32: 14 groups). 53.1% of the participants were female, and they ranged in age from 18 to 78 ($M = 34.7$, $SD = 10.6$). 91.4% of participants had completed some college or higher education. Most of the participants came from the United States (86.1%) and India (11.2%).

Participants accepted the HIT, consented to participate in the experiment, and then moved on to the task. Next, participants were assigned to a group; an existing group if one was available (less than 1 hour old and below the assigned capacity) or a new group was created. Groups were randomly assigned to be one of four sizes when created.

Individuals in the same group arrived at different times. They were all directed to the same task page which contained general instructions, a private information profile, and a shared discussion forum. The forum, whose technical characteristics are described above, was the only method provided to communicate and share information. When each individual felt he or she was satisfied with their group's solution, he or she could elect to leave the discussion permanently and fill out the post-task questionnaire. Individuals decided to leave the discussion at different times.

The questionnaire asked participants to report their group's solution to the task, task satisfaction, and general demographics (e.g. age, sex, education). Participants were paid $1.50 for the HIT and a bonus up to $1.00 if they submitted the correct solution. The amount of the bonus was proportional to the percent of the group that submitted the correct solution, incentivizing collaboration.

4.4 Qualitative Coding

To measure fact sharing raters examined the content of each message sent in the 55 groups. Two groups per condition were selected and the two raters independently coded whether each message shared facts ("Does the person share facts about one or more candidates"). There was substantial agreement between the raters (Cohen's Kappa = 0.90). All disagreements on those messages were then resolved through discussion and each half of the remaining messages were coded by one of the raters. After coding was completed, one of the raters took the statements in which facts were shared and identified which facts were shared by matching the statements to the information profiles given to participants. Unique mentions of facts were tallied at both the group and individual level.

4.5 Statistical Analysis & Variables

Groups of four different sizes (**Group Size**) were given two types of facts–common and rare–(**Fact Type**) corresponding with a 4×2 between-within subjects experimental design. In addition to these independent variables, we measured three dependent variables:

Percent of Facts Shared At a group level we calculated the number of different facts mentioned in discussion divided by the number of facts given to the group binned by fact type (e.g. 3 different common facts shared out of 8 common facts given is 37.5%). This provides a measure of information coverage, since the full content of a fact was always shared if mentioned.

Agreement We measured the group agreement by calculating the largest percentage of the group that reported the same group solution, that is the size of the plurality.

Solution Quality We calculated for each group member whether they had submitted the correct solution in the post-task questionnaire.

Our statistical approach was informed by the data. First, we evaluated whether or not to treat group size as a numeric or categorical variable. We built models for RQ1 and RQ2 using both approaches and used Akaike Information Criteria (AIC) to determine which model was more parsimonious. In both cases the model that treated group size as a categorical variable scored better. Thus, for consistency all reported analyses treat group size as categorical.

We addressed each research question at the appropriate level of analysis. Information exchange, information-sampling, and consensus are group-level processes and outcomes, so these were analyzed at the group level. Social loafing and social comparison are individual-level processes, so these were analyzed at the individual level. As mentioned above, performance was evaluated at an individual level because there was substantial disagreement among group members in some groups. Due to different levels of analysis, different statistical techniques had to be used for different research questions. ANOVA was used to analyze the effect of group size and fact type on group-level outcomes, with the exception of information-sampling which required bootstrapping. Generalized mixed effects regression models were used to analyze the effect of group size for individual-level outcomes. In these models group ID was included as a random effect to control for dependencies in the data. Missing values were omitted.

5 RESULTS

As expected in online discussions participants began the task at staggered times. On average half of group members had arrived within 5 minutes (SD = 6.3), 75% within 8 minutes (SD = 8.5) and 100% within 12 minutes (SD = 12.5). We only asked participants to at a minimum submit a solution to the task. Nonetheless most participants showed at least minimal engagement with the group discussion. 89.6% of participants sent at least one message; total messages sent per person ranged from 0 to 58 ($M = 4.78$, $Med. = 3$). We told participants that they were free to enter and exit the task as many times as they wanted and we encouraged them to check back as more people arrived. 10.8% of participants entered and exited more than once. In total, on average participants spent 10.2 minutes on the task ($Med. = 7.5$ min), often leaving for the last time at staggered intervals. As a result discussions ranged from 8 minutes to 187 minutes ($M = 64$, $Med. = 40$).

In total there were 3,962 comments exchanged in all 55 groups. Larger groups tended to exchange more comments than smaller groups (4: $M = 18.7$ comments, SD = 12.4; 8: $M = 43.0$, SD = 30.8; 16: $M = 74.5$, SD = 32.0; 32: $M = 151.9$, SD = 38.2). Groups made use of threading; the depth of comments ranged from 0 to 13. While 60.2% of comments were parent comments (depth = 0), 39.8% of comments were child comments (depth >= 1). A variety of types of comments were posted: introductions (e.g. *"Hello. How is everyone doing?"*), opinions (e.g. *"I think A is the best choice."*), facts (e.g. *"Candidate C is good at thinking creatively."*), synthesis (e.g. *"Here's the info I gather from thread putting all our info together..."*), and meta-discussions (e.g. *"We need to work out what everyone has, and pool the resources, before we come to a conclusion."*). We often observed redundant statements of fact in the comments. Individuals sometimes repeated facts others had already posted, perhaps because they had not read earlier statements or were repeating facts in agreement with earlier statements. Individuals also sometimes made redundant statements of facts when they were trying to organize and synthesize information. For example, a participant said *"Let's list everything we know about A"* and other participants replied to this comment with previously shared comments about Candidate A.

5.1 Research Question 1: How well do online discussion groups of different sizes pool distributed facts?

We performed repeated measures ANOVA examining the effect Group Size (4, 8, 16, 32) and Fact Type (Common, Rare) on the Percent of Facts Shared; results are presented in Tables 4 and 5 and Figure 2.

	df	F	p	Partial η^2
Fact Type (Common, Rare)	1	40.9	< 0.0001	0.45
Group Size (4, 8, 16, 32)	3	13.5	< 0.0001	0.44
Group Size X Fact Type	3	2.8	0.05	0.14
Residuals	51			

Table 4. Repeated measures ANOVA results evaluating the effect of group size and fact type on the percent of facts shared during discussion. 55 groups were analyzed.

Group Size	Common	Rare	t	df	p	Cohen's d
4	48% (40%)	28% (32%)	5.3	14	0.0001	1.36
8	58% (37%)	52% (35%)	0.9	10	0.41	0.26
16	89% (21%)	75% (19%)	5.3	14	0.0001	1.36
32	98% (5%)	91% (9%)	3.3	13	0.006	0.88

Table 5. Means and standard deviations for the percent of common and rare facts shared during group discussion. Paired t-tests comparing the percent of common and rare facts for each level of group size. 55 groups were analyzed separated by condition.

There was a significant main effect of fact type on the percent of facts shared. On average, 74% of common facts were shared compared to 62% of rare ones. This is consistent with the information-exchange bias known from small group work on HPTs. Although rare facts are critical to solving the problem, common facts are more likely to be discussed.

There was also a significant main effect of group size on the percent of facts shared. Linear and quadratic polynomial contrasts revealed a significant linear effect of group size on sharing facts (Estimate = 1.83, SE = 0.31, z = 5.9, $p < 0.001$) and no significant quadratic effect ($p = 0.92$). Larger groups shared more facts on average than smaller groups. Groups of size 4 shared only 38% of facts on average, while groups of size 32 shared 95% of facts on average. This is consistent with the information sampling explanation (more details below). In larger groups, there are more opportunities for facts to be shared because each fact is given to a larger number of people.

However, there was a marginally significant interaction (p = 0.05) between group size and fact type on the percent of facts shared. Paired t-tests comparing the percent of common versus rare facts shared for each group size using a Bonferroni corrected significance level of 0.0125 showed significantly more common facts were shared than rare facts for groups of size 4, 16, and 32 (Table 5), but not groups of size 8. In other words we observed a large information-exchange bias for groups of size 4, 16, and 32. Surprisingly while we observed that groups of size 8 shared more common facts than rare facts, this difference did not reach statistical significance. These results are consistent with our prediction that degree of bias would depend on group size, but inconsistent with our specific prediction that larger groups would display more bias.

In conclusion, larger online discussion groups were better at pooling information than smaller online discussion groups. However, even large online discussion groups suffered from the information exchange bias observed in traditional small groups.

5.2 Research Question 2: How well do online discussion groups of different sizes perform on a hidden profile task?

Results from prior work suggested that online discussion groups often generate multiple answers to a question rather than a single answer [30, 45]. For this reason, we asked individuals to report their

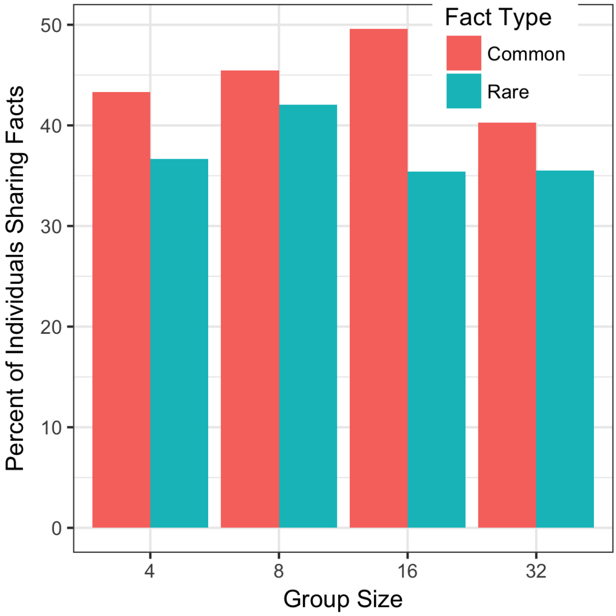


Fig. 2. Mean (SE) percent of common and rare facts for groups of different sizes.

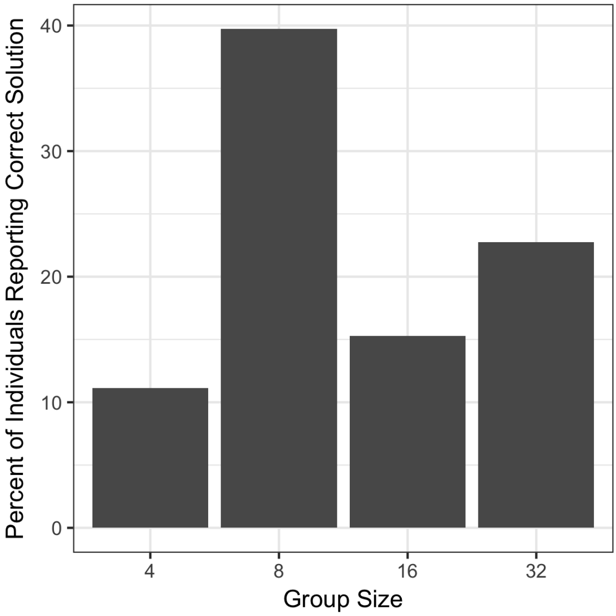


Fig. 3. Percent of individuals reporting the correct solution as their groups' solution in each group size condition.

Group Size	Coef.	SE	z	p
Intercept	-4.54	1.18	-3.83	0.0001
4 vs 8	3.61	1.50	2.41	0.02
4 vs 16	0.55	1.36	0.41	0.68
4 vs 32	2.46	1.38	1.78	0.07

Table 6. Mixed effects logistic regression model predicting the likelihood of an individual reporting the correct solution as their groups' solution depending on the size of their group. 712 individual responses from 55 groups.

group's solution and we report performance based on these individual-level reports. On average, we found that only 21% of participants reported the correct solution.

To compare performance across groups of different sizes we built a mixed effects logistic regression model to predict the likelihood of an individual reporting the correct solution to the problem given their group's size; group ID was included as a random effect to control for dependencies in the model (see Table 6 and Figure 3). We found a significant difference in performance based on group size (Table 6). Linear and quadratic polynomial contrasts revealed a significant quadratic effect of group size on performance when the inflection point was centered on groups of size 8 (Estimate = 7.83, SE = 3.4, z = 2.25, p = 0.02), but no linear effect (p = 0.31). These results suggest a quadratic relationship between group size and performance, in which individuals in mid-sized groups of 8 tended to perform the best.

The findings from research question 2 are consistent with those of research question 1. In addition to observing a trend in which online discussion groups of size 8 had a smaller information-exchange bias, individuals from groups of size 8 were also more likely to reach the correct solution in the HPT. However, this still leaves open the question of why mid-sized online discussion groups would perform better than either smaller or larger online discussion groups.

There are two other interesting results with respect to this research question. First, in general, individuals did very poorly on this task. Second, although online discussion groups of size 16 and 32 tended to share the most rare facts, individuals in these groups did not perform better than those in online discussion groups of size 8.

5.3 Research Question 3: Which factors explain why online discussion groups of different sizes perform differently on a hidden profile task?

We proposed several explanations for why sharing facts and performance might depend on group size: social loafing, information sampling, social comparison processes and consensus issues.

5.3.1 Social loafing. If social loafing were affecting online discussion groups we would expect to see a decrease in the likelihood of an individual sharing at least one fact (regardless of type) as group size increased. As group size increases individuals should be expected to be less likely to contribute, because they assume someone else will. We used mixed effects logistic regression to investigate the impact of group size on the likelihood of an individual sharing at least one fact; we included group ID as a random effect to control for dependencies in the model. The distribution of the percent of facts shared at an individual level was zero-inflated and right skewed to correct for this violation of assumptions we dichotomized the dependent variable. The results were the same regardless of the operationalization. We found no statistically significant relationship between group size and the likelihood of sharing a fact (Table 7). Regardless of group size around half of individuals shared at least one fact.

5.3.2 Information sampling. As groups grow in size, all else being equal, we expected that more facts would be shared overall. In larger groups there are more opportunities for facts to be shared because there are more individuals who can share facts. We used bootstrapping to build a model based on what we would expect to see at a group level if individuals shared facts at random. We took the number of facts each individual shared (ignoring group size and fact type) as our distribution for the number of facts shared per individual. We simulated groups of size 4, 8, 16, and 32. For each group we created the appropriate number of users and for each user we randomly sampled (with replacement) from our distribution of the number of facts shared per individual to obtain the number of facts shared in the group. Then we drew that number of facts from that individual's information profile at random (without replacement), pooled the facts shared at a group level, and tallied the total number of common and rare facts shared. We performed this simulation for equal sized samples (i.e. samples of size 55) 1000 times to achieve sampling means and 95% confidence intervals. Figure 4 displays the bootstrapped means and 95% confidence intervals as well as our observed means.

From our bootstrapped estimates we see that: 1) Common facts are shared more than rare facts because they appear more frequently across information profiles. 2) As groups increase in size more facts are shared, such that close to 100% of both common and rare facts are shared in groups of size 32. Our observed means somewhat follow the means and 95% confidence intervals (CI) generated by our bootstrapped estimates. We found no significant differences between simulated data, which models information sampling, and the observed data for the percentage of rare facts shared (4: p = 0.67, 8: p = 0.36, 16: p = 0.37, 32: p = 0.89). We also found no significant difference between the simulated data and the observed data for the percentage of common facts shared in groups of size 4 (p = 0.22); for larger groups the observed groups tended to share significantly fewer common facts than expected given the simulated data (8: p = 0.001, 16: p = 0.001, 32: p < 0.001); these differences remained significant even after using a Bonferroni corrected significance level of 0.00625. These results suggest that the information sampling hypothesis partially explains the pattern of results, with the exception that larger groups tend to share slightly less common facts than would be expected by the information sampling model.

5.3.3 Social comparison processes. If social comparison processes affect online discussion groups we would expect individuals from larger online discussion groups to be more likely to share at least one rare fact due to a positive-feedback loop between sharing more rare facts at a group level and individuals increased trust in rare facts. We constructed a mixed effect logistic regression model to investigate the impact of group size on the likelihood of an individual sharing at least one rare fact; we included group ID as a random effect to control for dependencies in the model. The distribution of the percent of rare facts shared at an individual level was zero-inflated and right skewed to correct for this violation of assumptions we dichotomized the dependent variable. The results were the same regardless of the operationalization. We found no significant relationship (Table 7). Regardless of group size, around 40% of individuals shared at least one rare fact.

5.3.4 Consensus. In larger online discussion groups we expected that the group would have more trouble coming to a consensus solution. For each group we calculated the largest percentage of the group that gave the same response (i.e. the size of the plurality). We then performed ANOVA at a group level to test the effect of group size on the size of the plurality. We found a significant relationship between group size and the size of the plurality (F(1,53) = 24.8, $p < 0.0001$, $\eta^2 = 0.32$). Linear and quadratic polynomial contrasts revealed a significant linear effect (Estimate = -96.9, SE = 20.0, t = -4.8, $p < 0.001$) and no significant quadratic effect ($p = 0.41$). Figure 5 shows that for larger groups, smaller proportions agreed with one another. Groups of size 4 showed an average of

	Model 1				Model 2			
Group Size	Coef.	SE	z	p	Coef.	SE	z	p
Intercept	-0.21	0.28	-0.74	0.46	-0.58	0.30	-1.92	0.06
4 vs 8	0.30	0.38	0.81	0.42	0.24	0.40	0.59	0.56
4 vs 16	0.31	0.33	0.95	0.34	-0.05	0.35	-0.13	0.89
4 vs 32	0.16	0.32	0.52	0.61	-0.04	0.34	-0.13	0.90

Table 7. Mixed effects logistic regression models predicting the likelihood of an individual sharing a fact (Model 1) and the likelihood of an individual sharing a rare fact (Model 2). 836 individual responses from 55 groups.

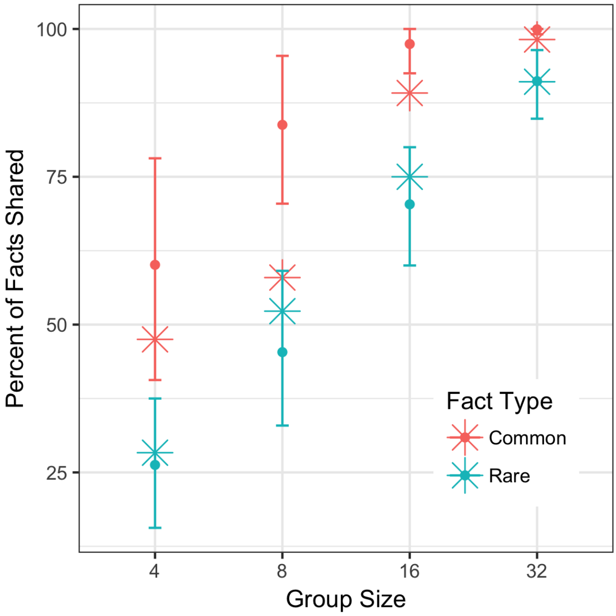


Fig. 4. Estimated percent of common and rare facts shared (means and 95% confidence intervals) at a group level for simulated data that assumes individuals share facts at random. Stars display actual, observed means.

88% of the group reported the plurality response, whereas for groups of size 32 only 57% of the group reported the plurality response.

We expected that larger groups would have more disagreement and have more difficulty forming a consensus. Larger groups contain more individuals who may vary in their experiences and perspectives. In addition, there are a few reasons why larger online discussion groups may not reach consensus. First, larger online discussion groups experience higher turnover, which means more group members may leave early before all facts are shared and discussion ends. Second, as factions grow in size within a group group polarization may lead individuals to be less likely to change their mind; furthermore in a larger group more minds need to be changed to reach consensus. Third, as reaching an agreement become more difficult there may be less of an expectation and pressure to form a consensus. Some other studies show less agreement in larger groups (e.g. [17]), however findings are mixed (e.g. [10]).

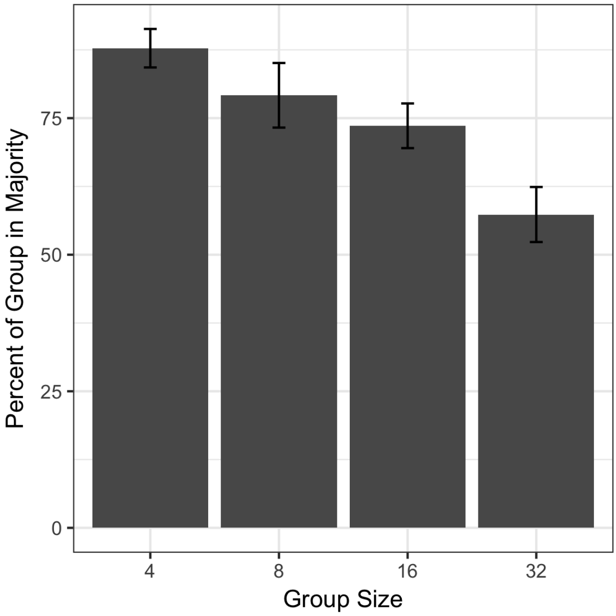


Fig. 5. The largest percent of group members reporting the same solution per group by condition (means, standard errors).

We examined whether turnover alone could explain lower agreement in larger groups for the ten groups that shared all facts (Size 4: 0, Size 8: 2, Size 16: 2, Size 32: 6). We found that a greater fraction of the group left before all facts were shared in larger groups of 16 and 32 (M_{16} = 45% and M_{32} = 35%) than in smaller groups of 8 (M_8 = 0%). However, we also found that there was less agreement in larger groups of 16 and 32 (M_{16} = 80% and M_{32} = 54%) than smaller groups of 8 (M_8 = 100%) even among those people who stayed until all facts were shared. These results suggest that multiple factors, including turnover, but not exclusively turnover, may play a role in the failure of the online discussion groups to reach consensus.

5.3.5 Summary. We find that information sampling and consensus issues better explain the results than social loafing or social comparison processes. In addition, information sampling and consensus issues, when combined, help to explain why medium-sized online discussion groups (8 members) performed better. Likely due to information sampling large online discussion groups pooled distributed facts much better than small online discussion groups. However, small online discussion groups were able to reach consensus much better than large online discussion groups. These two competing consequences of group size resulted in the best performance for medium sized online discussion groups of 8. This tradeoff is illustrated when we examine a cross section of the data (Table 8). Compared to groups of size 4, more groups of size 8 pooled facts well. 18% of groups of size 8 versus 7% of groups of size 4 shared more than three quarters of facts. Of the groups of size 8 that pooled facts well, they achieved higher levels of performance compared to groups of size 16 and 32 because they were able to better establish a group consensus. 100% of individuals in these groups of size 8 reported the correct solution, compared to 36% and 25% for individuals in these groups of size 16 and 32 respectively.

Groups sharing >75% rare facts		
Group Size	% Groups	Solve Rate
4	7%	75%
8	18%	100%
16	40%	36%
32	86%	25%

Table 8. The percentage of groups that shared over 75% of rare facts by condition. The average percentage of individuals reporting the correct solution in these groups.

RQ1: How well do online discussion groups of different sizes pool distributed facts?	
Online discussion groups will share more common facts than rare facts	Supported
Larger online discussion groups will share more facts	Supported
Larger online discussion groups will tend to share even more common facts relative to rare facts	Unexpected Result
RQ2: How well do online discussion groups of different sizes perform on a hidden profile task?	
Online discussion groups will perform poorly on the hidden profile task	Supported
Larger online discussion groups will tend to perform worse on the hidden profile task	Unexpected Result
RQ3: Which factors explain why online discussion groups of different sizes perform differently on a hidden profile task?	
Individuals in larger online discussion groups will share fewer facts (Social Loafing)	Not Supported
Larger online discussion groups will share more common and rare facts (Information Sampling)	Partially Supported
Individuals in larger online discussion groups will share more rare facts (Social Comparison)	Not Supported
Large online discussion groups should exhibit less agreement (Consensus)	Supported

Table 9. Summary of research questions, predictions, and findings.

6 DISCUSSION

As individuals and organizations increasingly rely on online discussion groups to gather information and make important decisions, it is critical to examine how groups pool facts and form decisions through discussion on these platforms. As a starting place we examined the way that online discussion groups exchanged distributed information and its effect on decision quality. We found a novel empirical result: medium-sized groups performed best and there was a non-linear relationship between group size and performance. We discuss implications of this finding, as well as how we might use these results to extend theories of information exchange from traditional small groups to this new setting.

Our results suggest that theories about information exchange developed for traditional small group discussions are applicable to online discussion groups despite differences in temporal dynamics, communication tools, social norms and sometimes size. Our findings in this study are generally consistent with the main theoretical arguments made by small group researchers about HPTs [42, 52]. Consistent with prior research we find that groups in this setting often fail to exchange all facts, can be biased in which facts they share, and perform sub-optimally on HPTs [27, 52]. One implication of these results is that individuals may be inadvertently withholding relevant information from online discussions.

Although decision quality was poor for individuals in online discussion groups of all sizes, it was best for groups of size 8. Medium-sized groups were large enough, but not too large. Larger groups are better at sharing facts because there are more people who can contribute rare knowledge and expertise [19, 29]. However smaller groups, so long as more facts are shared, have an easier time clearing up misunderstandings, focusing on the most relevant information, and convincing everyone of the right answer. This latter point is consistent with prior work which shows that groups perform the best on problems and situations when they can explain the reasons why one answer is better than the other, convincing other group members who hold erroneous views to change their minds [25]. Large groups had trouble forming a consensus, in part because it was difficult to convincing everyone of the right answer in a large group, but also because of the temporal dynamics of online discussion groups in which individuals choose to leave at different times. Another implication of these results is that medium-sized groups may have an advantage over small or large groups, particularly for online discussion groups.

6.1 Extending Theory

Information exchange theories from small group research must be extended to fully model behavior in online discussion groups. Findings in this study contradict specific predictions and findings about the impact of group size on information exchange and performance developed studying traditional small group discussions [27, 35].

First, most previous studies have assumed that rare facts would be known to a single person even in larger groups [41], we argued in this paper that in larger groups we should assume that rare facts may be known to more than one person. If rare facts are known to 25% of people in the general population then, on average, we should expect one person in a group of size 4 to know the fact and eight people in a group of size 32. This explains why this study and Stasser and colleagues' study [41] both find results that can be partially explained by information sampling, but opposite results. If a consistent fraction of the population knows rare facts, than information sampling should lead to more fact sharing in larger groups and the same or slightly less bias in larger groups. This seemingly minor discrepancy in how to operationalize the distribution of rare facts has a major impact on how groups perform on this task. Our reasoning explains why our findings of better performance for groups of size 8 over groups of size 4 may be more relevant to real world groups than prior contradictory findings [35, 41]. This discrepancy also represents a larger issue neglected by current theories–how are groups formed from individuals in a larger population–that becomes critical when applying these theories to online discussion groups.

Second, consistent with the prior study testing HPTs in a similar environment we found that members of large online discussion groups performed sub-optimally in part because large groups failed to reach consensus [45]. Current theories of information exchange in traditional small groups always assume a group reaches a consensus, in contrast multiple decisions often emerge in online discussions. Due to high turnover group decisions can shift as group composition changes [45]. Due to large size and asynchronicity groups may not expect to reach consensus and may not put as much effort into consensus building. Further, surfacing multiple competing opinions is sometimes a valued characteristic of these discussions [30]. Extended theories will need to contend with the impact multiple decisions have on different stakeholders, including participants in the discussion and passive readers of the discussion.

Finally, differences in temporal dynamics and communication tools can create other differences in expectations and norms. For example, contrary to theory we did not observe social loafing in large online discussions groups [9], perhaps because perceptions of group size are different in online discussion groups. Due to staggered arrival and high turnover the group's efforts are distributed across time with a smaller number of people present at a given time.

6.2 Design Implications

We argue that medium-sized groups performed the best in this experiment because they were a compromise between the advantages of being large (more knowledge is shared) and small (easier to clear up misunderstandings). If this explanation is correct than the most direct implication for design from this study is that discussion platforms focusing on knowledge sharing should encourage group sizes large enough that individuals share rare knowledge, but no larger than necessary to extract this knowledge. Based on this one study medium-sized groups (e.g. 8) seem preferable, however the optimal group size will depend on the discussion platform and how knowledge is distributed among the community. Online discussion platforms, like StackOverflow and Reddit, can control the size of the group directly by only recommending questions to be answered by specific people or indirectly by changing the placement of a question on the front page. For example, a question could be promoted until it had attracted a medium number of contributors (e.g. 8) and then hidden from new users until that existing group had time to engage in back-and-forth discussion. Future work, should also explore the effect of using more complex algorithms to recommend posts to users. For example, there may be the most benefit from keeping online groups medium-sized contemporaneous (e.g. same day) and deliberately rotating membership over time. Other work suggests that specific patterns of group rotation encourage innovation [37]. This approach has the advantage of drawing knowledge from a much larger group. However, collaborative filtering algorithms which rank comments would need to be improved to better account for shifts in opinion over time.

Secondly, discussion platforms should modify their design in ways that help to combat the identified information-exchange bias and sub-optimal performance. Successful interventions from the small group literature may be applicable to online discussion forums as well. Studies have found that individuals are more likely to share rare knowledge if they are told that they are the only person in the conversation with expertise in that area [40]. Discussion forums sometimes make users aware of relative expertise of contributing authors by using historical performance on the site, such as by including reputation points next to someone's name (e.g. Stack Overflow). Rather than providing a global measure of expertise there may be more value in providing information on specific areas of expertise (e.g. Baysian statistics, machine learning etc.). Individuals may be more likely to share their unique knowledge if they can see that no one else in the discussion has their same expertise.

Performance might also be improved by separating knowledge accumulation from deliberation. Voigtlaender and colleagues [50] found that groups pooled more knowledge if they kept a structured list of facts separate from discussion. For example, a discussion forum could be designed to label individual pieces of information embedded within a comment (this could be done manually by the original commenter, by others in the community or done automatically using natural language processing tools). The forum could treat these pieces of information as individual objects separate from the original comments and could display them in a list next to the threaded comments and/or enable users to manipulate and organize these pieces of information (e.g. knowledge map). Several tools have been built on top of discussion forums to tackle the problem of reusing dense, unstructured online discussion [16]. These tools enable users to label, extract, summarize and distill content from online discussions [1, 34, 53]; while they have typically been used to summarize discussions that have already taken place, the current study suggests they may be useful to individuals while the discussion is ongoing. Future work is needed to investigate whether these tools could help clear up misunderstandings which become more pernicious in larger online discussion groups. Similarly, other successful interventions could be incorporated into discussion platforms, such as interfaces to externalize the sensemaking process [14] or tools to help plan how a decision will be made before discussion begins [48].

Large groups struggled to consistently identify the correct solution and form a consensus even when all relevant information was shared in the online forum. Relevant facts and arguments can get buried in discussion threads [30]. Collaborative filtering through voting can help surface the best content from lengthy discussions [7]. However, more sophisticated collaborative filtering methods are needed for hidden profile-like problems which involve cumulative knowledge (facts must be pooled across many different comments) and shifting opinions (by design early preferences are incorrect and correct decision must be discovered). With better collaborative filtering large groups may be able to consistently identify the correct answer.

6.3 Limitations & Future Work

There are several limitations of this work. Due to the small number of groups per group size condition (11-15) there is a risk of Type 2 errors. In addition, this study has a higher risk of Type 1 errors because we conducted multiple hypothesis tests and contrasts to investigate several competing predictions. Due to the limited scope of this experiment we do not know how these results would generalize to the large variety of knowledge sharing discussions online. In particular, the largest groups we tested were of size 32 which is much smaller than the largest groups on these platforms, which can reach hundreds of participants. We expect that information exchange will operate similarly in groups of size 100 as they did in groups of size 32 due to similar conditions however we cannot be certain without further experimentation.

In addition, in many online discussions individuals do not get paid (e.g. Q&A, online forums) or receive payment only for submitting a winning solution (e.g. innovation contests). AMT is an atypical context because individuals were paid in part for task participation. It is unclear how intrinsic versus extrinsic motivations may differentially affect information exchange.

HPTs are contrived problems with a very specific structure, in part chosen arbitrarily by the experimenters. For example, we arbitrarily elected to make common facts known to 75% of the group and rare facts known to 25% of the group. In real-world problems knowledge has a variety of distributions, including problems in which rare knowledge is known to more and less than 25% of people and problems in which the ratio of common to rare knowledge is different. Future work, should examine whether groups pool information when facts are distributed in variety of different ways. It is unclear the extent to which problems analogous to hidden profile problems occur in the wild [52].

Future work should also consider other important aspects of online discussions. In our experiment each discussion was a one-shot event, but in online forums some users engage in multiple discussions and repeatedly encounter each other. This could create interesting group dynamics that affect decision making. Furthermore users in real-world platforms select which discussions to join based on the current state of discussion, which might impact their decision to join and participate in decision making.

7 CONCLUSION

In this study we tested the hidden profile paradigm in online discussion groups of four different sizes. We found that many groups were biased in how they shared facts and that many groups produced sub-optimal decisions. Surprisingly, we found that groups of size 8 performed better than either larger or smaller groups. We argue that this is a result of a tradeoff between a discussion being too small, in which not enough facts are shared, and being too big, when it is harder to achieve consensus. From a practical perspective, these findings suggest that designers of online discussions should encourage medium-sized groups and test out potential interventions known to reduce the information-exchange bias in traditional small groups. From a theoretical perspective, these findings suggest specific ways theories need to be expanded to model behavior in online discussion forums, including considering how group formation affects the distribution of rare facts in groups and how the emergence of multiple, competing opinions in online discussions impacts beliefs and actions.

ACKNOWLEDGMENTS

We thank our participants, Mark Boons, and the National Science Foundation (IIS #1657308) for support.

REFERENCES

[1] Mark S Ackerman, Anne Swenson, Stephen Cotterill, and Kurtis DeMaagd. 2003. I-DIAG: From Community Discussion to Knowledge Distillation. In *Communities and Technologies*. Springer, 307–325.

[2] Mauricio Aniche, Christoph Treude, Igor Steinmacher, Igor Wiese, Gustavo Pinto, Margaret-Anne Storey, and Marco Gerosa. 2018. How modern news aggregators help development communities shape and share knowledge. In *Proc. International Conference on Software Engineering*. ACM, IEEE, 1–12.

[3] Ofer Arazy, Wayne Morgan, and Raymond Patterson. 2006. Wisdom of the crowds: Decentralized knowledge construction in Wikipedia. In *Workshop on Information Technologies & Systems*. AIS, Phoenix, AZ, 1–6.

[4] Ofer Arazy and Oded Nov. 2010. Determinants of Wikipedia quality: The roles of global and local contribution inequality. In *Proc. Conference on Computer Supported Cooperative Work*. ACM, 6–9.

[5] Ofer Arazy, Melanie Lisa Yeo, and Oded Nov. 2013. Stay on the Wikipedia task: When task-related disagreements slip into personal and procedural conflicts. *Journal of the American Society for Information Science and Technology* 64 (2013), 1634–1648.

[6] Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. 2017. Distilling the wisdom of crowds: Prediction markets versus prediction polls. *Management Science* 63 (2017), 691–706.

[7] Jurgen Buder, Christina Schwind, Anja Rudat, and Daniel Bodemer. 2015. Selective reading of large online forum discussions: The impact of rating visualizations on navigation and learning. *Computers in Human Behavior* 44 (2015), 191–201.

[8] Keith Burghardt, Emanuel F. Alsina, Michelle Girvan, William Rand, and Kristina Lerman. 2016. The Myopia of Crowds: A Study of Collective Evaluation on Stack Exchange. *PLoS ONE* 12 (2016), 1–19.

[9] Michael G. Cruz, Franklin J. Boster, and Jose I. Rodriguez. 1997. The impact of group size and proportion of shared information on the exchange and integration of information groups. *Communication Research* 24 (1997), 291–313.

[10] Larry L Cummings, George P Huber, and Eugene Arendt. 1974. Effects of Size and Spatial Arrangements on Group Decision Making. *Academy of Management Journal* 17, 3 (1974), 460–475.

[11] Alan R. Dennis. 1996. Information exchange and use in group decision making: You can lead a group to information but you can't make it think. *MIS Quarterly* 20 (1996), 433–457.

[12] Alan R. Dennis, Kelly M. Hilmer, and Nolan J. Taylor. 1997. Information exchange and use in GSS and verbal group decision making: Effects of minority influence. *Journal of Management Information Systems* 14 (1997), 61–88.

[13] Eric Gilbert. 2013. Widespread underprovision on Reddit. In *Proc. Conference on Computer Supported Cooperative Work*. ACM, 803–808.

[14] Nitesh Goyal and Susan R. Fussell. 2016. Effects of sensemaking translucence on distributed collaborative analysis. In *Proc. Conference on Computer-Supported Cooperative Work*. ACM, 287–301.

[15] Tobias Greitemeyer and Stefan Schulz-Hardt. 2003. Preference-consistent evaluation of information in the hidden profile paradigm: Beyond group-level explanations for the dominance of shared information in group decisions. *Journal of Personality and Social Psychology* 84 (2003), 322–339.

[16] Derek L Hansen. 2009. Overhearing the Crowd: An Empirical Examination of Conversation Reuse in a Technical Support Community. In *Proceedings of the Fourth International Conference on Communities and Technologies*. ACM, 155–164.

[17] A Paul Hare. 1952. A Study of Interaction and Consensus in Different Sized Groups. *American Sociological Review* 17, 3 (1952), 261–267.

[18] Verlin B. Hinsz, R. Scott Tindale, and David A. Vollrath. 1997. The emerging conceptualization of groups as information processors. *Psychological Bulletin* 121 (1997), 43–64.

[19] Lars B. Jeppesen and Karim R. Lakhani. 2010. Marginality and problem solving effectiveness in broadcast search. *Organization Science* 21 (2010), 1016–1033.

[20] Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. *Designing for civil conversations: Lessons learned from ChangeMyView.* Technical Report. GVU Technical Report. 1–11 pages.

[21] Robert E. Kraut and Paul Resnick. 2012. *Building Successful Online Communities: Evidence-Based Social Design.* MIT Press.

[22] Cliff Lampe and Paul Resnick. 2004. Slash(dot) and burn. In *Proc. Human Factors in Computing Systems*. ACM, 543–550.

[23] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31

(2014), 317–326.

[24] Bibb Latané, Kipling Williams, and Stephen Harkins. 1979. Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology* 37 (1979), 822–832.

[25] Patrick R. Laughlin. 1999. Collective induction: Twelve postulates. *Organizational Behavior and Human Decision Processes* 80 (1999), 50–69.

[26] Patrick R. Laughlin, Bryan L. Bonner, and Andrew G. Miner. 2002. Groups perform better than the best individuals on Letters-to-Numbers problems. *Organizational Behavior and Human Decision Processes* 88 (2002), 605–620.

[27] Li Lu, Y. Connie Yuan, and Poppy L. McLeod. 2012. Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review* 16 (2012), 54–75.

[28] Ann Majchrzak, Samer Faraj, Gerald C. Kane, and Bijan Azad. 2013. The contradictory influence of social media affordances on online communal knowledge sharing. *Journal of Computer-Mediated Communication* 19 (2013), 38–55.

[29] Thomas W. Malone, Robert Laubacher, and Chrysanthos N. Dellarocas. 2009. Harnessing crowds: Mapping the genome of collective intelligence. *MIT Sloan Research Paper* 4732-09 (2009), 1–21.

[30] Lena Mamykina, Drashko Nakikj, and Noemie Elhadad. 2015. Collective sensemaking in online health forums. In *Proc. Conference on Human Factors in Computing Systems*. ACM, ACM, 3217–3226.

[31] Brian E. Mennecke. 1997. Using group support systems to discover hidden profiles: An examination of the influence of group size and meeting structures on information sharing and decision quality. *International Journal of Human-Computer Studies* 47 (1997), 387–405.

[32] Lev Muchnik, Sinan Aral, and Sean J. Taylor. 2013. Social influence bias: A randomized experiment. *Science* 647 (2013), 647–651.

[33] Uday S. Murthy and David S. Kerr. 2004. Comparing Audit Team Effectiveness via Alternative Modes of Computer-Mediated Communication. *Auditing: A Journal of Practice and Theory* 23 (2004), 141–152.

[34] Kevin K Nam and Mark S Ackerman. 2007. Arkose: Reusing Informal Information from Online Discussions. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*. ACM, 137–146.

[35] Torsten Reimer, Andrea Reimer, and Uwe Czienskowski. 2010. Decision-making groups attenuate the discussion bias in favor of shared information: A meta-analysis. *Communication Monographs* 77 (2010), 121–142.

[36] Lionel P. Robert and Daniel M. Romero. 2017. The influence of diversity and experience on the effects of crowd size. *Journal of the Association for Information Science and Technology* 68 (2017), 321–332.

[37] Niloufar Salehi and Michael S. Bernstein. 2018. Hive: Collective Design Through Network Rotation. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 151 (Nov. 2018), 26 pages.

[38] Stefan Schulz-Hardt, Felix C. Brodbeck, Andreas Mojzisch, Rudolf Kerschreiter, and Dieter Frey. 2006. Group decision making in hidden profile situations: Dissent as a facilitator for decision quality. *Journal of Personality and Social Psychology* 91 (2006), 1080–1093.

[39] Garold Stasser and Dennis Stewart. 1992. Discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment. *Journal of Personality and Social Psychology* 63 (1992), 426–434.

[40] Garold Stasser, Dennis Stewart, and Gwen M. Wittenbaum. 1995. Expert roles and information exchange during discussion: The importance of knowing who knows what. *Journal of Experimental Social Psychology* 31 (1995), 244–265.

[41] Garold Stasser, Laurie A. Taylor, and Coleen Hanna. 1989. Information sampling in structured and unstructured discussions of three- and six-person groups. *Journal of Personality and Social Psychology* 57 (1989), 67–78.

[42] Garold Stasser and William Titus. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology* 48 (1985), 1467–1478.

[43] Garold Stasser and William Titus. 1987. Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology* 53 (1987), 81–93.

[44] Margaret-Anne Storey, Alexey Zagalsky, Fernando Figueira Filho, Leif Singer, and Daniel M. German. 2017. How social and communication channels shape and challenge a participatory culture in software development. *Transactions on Software Engineering* 43 (2017), 185–204.

[45] Yla Tausczik and Mark Boons. 2018. Distributed knowledge in crowds: Crowd performance on hidden profile tasks, In Proc. International Conference on Web and Social Media. *Proceedings of International Conference on Web and Social Media*, 405–414.

[46] Yla Tausczik, Aniket Kittur, and Robert E. Kraut. 2014. Collaborative problem solving: A study of MathOverflow. In *Proc. Conference on Computer Supported Cooperative Work*. ACM, 355–367.

[47] Yla Tausczik, Ping Wang, and Joohee Choi. 2017. Which size matters? Effects of crowd size on solution quality in big data Q&A communities. In *Proc. International Conference on Web and Social Media*. AAAI, 260–269.

[48] J Lukas Thürmer, Frank Wieber, and Peter M Gollwitzer. 2015. A self-regulation perspective on hidden-profile problems: if–then planning to review information improves group decisions. *Journal of Behavioral Decision Making* 28, 2 (2015), 101–113.

[49] Christoph Treude, Ohad Barzilay, and Margaret-Anne Storey. 2011. How do programmers ask and answer questions on the web?. In *Proc. International Conference on Software Engineering*. ACM, 804–807.

[50] Denise Voigtlaender, Felix Pfeiffer, and Stefan Schulz-Hardt. 2009. Listing and structuring of discussion content: as a means of improving individual decision quality in hidden profiles. *Social Psychology* 40, 2 (2009), 79–87.

[51] Gwen M. Wittenbaum. 1998. Information sampling in decision-making groups: the impact of members' task-relevant status. *Small Group Research* 29, 1 (1998), 57–84.

[52] Gwen M Wittenbaum, Andrea B. Hollingshead, and Isabel C. Botero. 2004. From cooperative to motivated information sharing in groups: Moving beyond the hidden profile paradigm. *Communication Monographs* 71 (2004), 286–310.

[53] Amy X. Zhang and Justin Cranshaw. 2018. Making Sense of Group Chat Through Collaborative Tagging and Summarization. *Proc. ACM Human-Computer Interaction* 2, CSCW, Article 196 (Nov. 2018), 27 pages.

[54] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging discussion forums and wikis using recursive summarization. In *Proc. Conference on Computer Supported Cooperative Work*. ACM, 2082–2096.

Received XXX; revised XXX; accepted XXX