Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

Distributed Knowledge in Crowds: Crowd Performance on Hidden Profile Tasks.

Yla Tausczik
College of Information Studies
University of Maryland, College Park
ylatau@umd.edu

Mark Boons
Rotterdam School of Management
Erasmus University Rotterdam
mboons@rsm.nl

Abstract

Individuals today discuss information and form judgements as crowds in online communities and platforms. "Wisdom of the crowd" arguments suggest that, in theory, crowds have the capacity to bring together diverse expertise, pooling distributed knowledge and thereby solving challenging and complex problems. This paper concerns one way that crowds might fall short of this ideal. A large body of research in the social psychology of small groups concerns the shared information bias, a tendency for group members to focus on common knowledge at the expense of rarer information which only one or a few individuals might possess. We investigated whether this well-known bias for small groups also impacts larger crowds of 30 participants working on Amazon's Mechanical Turk. We found that crowds failed to adequately pool distributed facts; that they were partially biased in how they shared facts; and that individual perception of group decisions was unstable. Nonetheless, we found that aggregating individual reports from the crowd resulted in moderate performance in solving the assigned task.

Individuals and organizations increasingly use online communities to broadcast difficult problems to crowds. These online communities specialize in a diversity of problems from R&D (InnoCentive) to software development (Stack Overflow) to data science (Kaggle) to mathematics (Polymath Projects). On many of these platforms the crowd constructs a solution through open discussion. Individuals share relevant knowledge, suggest approaches, provide partial and complete solutions, and critique and discuss solutions.

Crowd problem-solving discussions have led to remarkable successes, in which novel solutions were provided for difficult problems. On Stack Overflow, a software development Q&A, questions typically receive more than one answer and receive a satisfactory answer within 21 minutes (Mamykina et al. 2011). Polymath projects are open collaborations among many mathematicians that take place on blogs and wikis, and have generated new proofs for several open research questions in mathematics (Cranshaw and Kittur 2011). 90% of questions on MathOverflow, a Q&A community devoted to solving small, novel problems in mathematics, receive a solution (Tausczik, Kittur, and Kraut 2014); content from this site has been referenced in over 90 academic mathematics papers (Tausczik 2016).

Despite the clear successes of many crowd platforms at facilitating problem solving, there are also examples of monumental failures. When police released pictures of the Boston Marathon bombers to see if any citizens could identify the suspects, a crowd on Reddit identified the wrong culprit, leading to confusion and misinformation (Lee 2013). Researchers have confirmed experimentally that online forums which support discussions among very large crowds, like Reddit, are heavily biased by early opinions, even when those early opinions are chosen arbitrarily (Muchnik, Aral, and Taylor 2013). Failures in crowd problem solving have the potential to waste resources, generate additional misinformation, and provide individuals and organizations with faulty, misguided solutions. Unfortunately, there is very little research that focuses on how crowds go about solving problems.

A substantial body of research in social psychology considers the way that smaller, more traditional groups work together to solve problems. This research has shown that groups often have biases and inefficiencies that prevent them from performing to their potential. Even when group members collectively possess all the information needed to solve a problem, they often fail to share all relevant knowledge (Stasser and Titus 2003), focus on the wrong information (Gigone and Hastie 1993), and discount minority points of view even when they are correct (Laughlin 1999). In addition, group communication can derail individual work (Diehl and Stroebe 1987); individuals often do not speak out when they disagree with the majority (Wood et al. 1994) and are less motivated to work as hard in a group (Latané, Williams, and Harkins 1979). Because of these biases and inefficiencies, groups often perform worse than would be expected based on aggregate individual potential.

Crowd problem solving shares many of the characteristics of the small groups studied by social psychologists: it involves a collection of individuals studying a common problem, working in the open, and actively collaborating to identify a solution. At the same time, other aspects of crowd problem solving are fundamentally different from the small group context studied in traditional social psychology. This research typically assumes a fixed set of actors, all of whom contribute to a discussion, and through this discussion arrive

at a consensus solution.

Online crowds pose a radically different way of engaging problem solvers. Problem-solving groups on open platforms are not fixed; participants decide when and where to focus their attention among a wide array of options, and proactively choose which of their observations (if any) they will share with the group. Actors come and go, making crowd problem-solving an asynchronous process with high turnover. While the total number of actors who see a problem is usually much higher than in traditional groups (sometimes thousands), the actual number of contributors to a solution tends to vary from one to dozens.

In this paper we focus on how crowds share information and form judgements through discussion. Known as the "wisdom of the crowd," one of the most important advantages of a crowd is its capacity to generate accurate judgements by pooling distributed knowledge from a large set of diverse individuals (Surowiecki 2005). We evaluate the performance of crowds on a hidden profile task, a classic type of problem which tests the ability of a group to share distributed facts (Wittenbaum, Hollingshead, and Botero 2004). This paper contributes to an understanding of how information is shared and how judgements are formed in crowd discussions. We found that crowds failed to adequately pool distributed facts; were partially biased in how they shared facts; and had unstable group decisions. However, in spite of these problems, we found moderate to high performance when we aggregated over individuals' reported judgements.

Related Work & Current Study

One of the biggest advantages of crowds is their capacity to elicit diverse and rare knowledge, combine this knowledge, and co-construct solutions. Crowdsourcing platforms and contests like Innocentive, 99Designs, and Quirky draw divergent ideas from many individuals with specialized knowledge (Yang, Chen, and Pavlou 2009; Jeppesen and Lakhani 2010). Crowdsourcing applications, like prediction markets and prediction polls, leverage the diversity of the crowd by aggregating crowd opinions to produce judgements that are more accurate than those made by individuals or small groups (Krause et al. 2011; Atanasov et al. 2017). Researchers have found evidence of crowds co-constructing solutions through discussion by combining expertise and sharing partial work (Cranshaw and Kittur 2011; Tausczik, Kittur, and Kraut 2014). In summary, crowds have the capacity to produce accurate solutions to challenging problems by eliciting and combining specialized knowledge.

Crowds may not always reach their full potential. Like small groups, crowds may be vulnerable to biases in whether and how they share and discuss information. In a landmark paper, Stasser and Titus (Stasser and Titus 1985) found that small groups performed poorly on hidden profile tasks, which require them to pool distributed facts to solve a problem. In hidden profile tasks each person in a group is given a subset of the facts needed to solve a problem; some of the facts are provided across all or most group members (common) while other facts are given to a single individual (rare). Group discussion tends to focus on common facts while neglecting rare ones. As a result, this process often results in a non-optimal solution to the task. This finding has been tested in over 144 papers and confirmed to be a large effect by a meta-analysis (Lu, Yuan, and McLeod 2012). These experiments illustrate a shocking bias in the way that small groups discuss distributed information that leads directly to poor problem-solving performance.

Despite the differences between crowds and the traditional small groups studied by psychologists, we expect the same bias to operate in the crowd. Researchers have shown that three different effects combine to explain the counterintuitive finding for small groups–preference consistent evaluation, sampling, and social comparison–and all of these are likely to effect crowds as well.

In hidden profile tasks each individual tends to form a preference before discussion, based on the subset of facts they are given. Due to the design of the task, individuals have more common facts than rare facts, more common facts support a non-optimal solution, and as a result individuals tend to favor non-optimal solutions going into discussion (preference consistent evaluations) (Greitemeyer and Schulz-Hardt 2003). As these initial preferences are formed at the individual level based on features embedded in the design of the hidden profile task, we would expect to see a similar effect in crowds.

Because common facts are provided to more individuals than rare ones, these make up a larger proportion of the total statements given to the group. Therefore, we expect these will be shared more often (sampling) (Stasser 1992). For example, consider an experiment with 3 participants, 3 rare facts and 3 common facts. Each participant receives 1 rare fact and 2 common facts; even though there are the same number of common and rare facts (3 each), common facts are twice as prevalent in the group (6 to 3). The sampling effect is present, and may even be exaggerated, in crowds. Generalizing the example above to a crowd, the ratio of common facts to rare ones would be 20:1 for a crowd of 30. Thus we expect the sampling bias in crowds to be as larger or even larger than that in small groups.

Social comparison processes also drive group members to value common facts more highly (Wittenbaum, Hollingshead, and Botero 2004). Individuals use others to evaluate the importance and trustworthiness of information when they are uncertain. Hearing others say common facts makes individuals more confident in the importance and accuracy of those facts for completing the task (Postmes, Spears, and Cihangir 2001). In addition, repeating common facts benefits individuals trying to demonstrate to others that they can do the task and are competent (Wittenbaum and Bowman 2004). It is not clear whether to expect the social comparison effect to be weaker or stronger in the crowd. One the one hand, group size should worsen the effect for two reasons. First, sharing opinions and facts that disagree with others is harder in larger groups (Wood et al. 1994); second, social loafing is worse in larger groups, that is individuals have less motivation to speak up in general (Latané, Williams, and Harkins 1979). On the other hand, some unique aspects of crowd work (e.g. high turnover, asynchronicity) may dampen the effect of social comparison processes. For example, there may be less pressure to conform in the crowd.

With high turnover, group members come and go, leading to an expectation that group composition, and potentially majority opinions, may easily shift. On balance, we would expect to see some bias in crowds due to social comparison.

In summary, despite evidence showing crowds have a capacity to pool distributed facts, theory suggests that crowds will be biased in which facts they share, such that they will share fewer rare facts. In the current study we gave crowds a hidden profile task. We described how crowds discussed facts, quantified the degree to which crowds pooled facts, and evaluated whether crowds were biased in how they shared facts to answer the research question:

RQ1: How do crowds share and discuss distributed facts?

We expect that if crowds share rare facts at low rates than they will perform poorly on a hidden profile task, which by design necessitates sharing rare facts. However, one complexity of crowds, in comparison to traditional small groups, is that crowds may not form a consensus decision. A multiplicity of opinions, which may be at odds with one another, is valued in crowds more than in small groups (Mamykina, Nakikj, and Elhadad 2015). In addition, high turnover in crowds means individuals come and go, so that group decisions may vary as the group composition changes. In the current study we examined whether crowds formed a consensus decision, whether and how group decisions changed over time, and we computed crowd performance to answer a second research question:

RQ2: How well do crowds perform on a hidden profile task?

Method

We took a descriptive approach in which we observed small crowds in an engineered crowdsourcing setting using an artificial task[1]. A few studies have observed how problem solving discussions unfold in crowds on Q&A platforms (e.g. Tausczik, Kittur, and Kraut 2014) and online communities (e.g. Mamykina, Nakikj, and Elhadad 2015). The current study complements and extends these naturalistic observations with more structured experimentation. By using an artificial task we could be certain of who in the crowd knew what facts, and could determine with certainty which facts were shared and which were not shared; this would be impossible when observing crowds in the wild.

We performed our observations on Amazon's Mechanical Turk (AMT). AMT is a specific type of crowdsourcing platform; it is commonly used for crowdsourcing applications, including ones that utilize crowd diversity to pool judgements (e.g. Mitra and Gilbert 2015) and solve problems through discussion (e.g. Zhu et al. 2014). AMT has some similarities and some differences compared to other crowdsourcing settings which will influence the generalizability of our results. We observed crowds of size 30, which are on the smaller size for crowds, but comparable in size to crowds used in many applications (Zhu et al. 2014; Mitra and Gilbert 2015; Atanasov et al. 2017) and the number of people typically contributing to discussions in Q&A sites (Tausczik, Kittur, and Kraut 2014; Tausczik, Wang, and Choi 2017).

[1]Task materials, data, and code are provided at osf.io/3eajf.

Fact Type	Num. Facts	Num. Members	% Members
Common	12	20	66%
Infrequent	3	10	33%
Unique	3	3	3%

Table 1: Distribution of facts across a crowd.

Study Design & Hidden Profile Task

We created crowds of size 30 and had them work on a hidden profile task. We used a group size of 30, roughly matching the smallest crowds found in the wild. We designed our hidden profile task to be suitable for a crowd context, based on tasks used in prior small-group studies. Groups were asked to identify the best candidate for a job from a set of four hypothetical candidates. To make this choice we provided groups with information about candidates' scores on a battery of 18 skill tests (e.g. Verbal Reasoning).

We distributed information about the skill tests among the 30 group members, so that some skill tests were given to a majority of group (common), some to a minority (infrequent), and some to only one person (unique) (see Table 1). Each person in a group was given information about 9 of the 18 skills (called an information profile), 1-2 of these skills were known to a minority of the group (rare or unique), the others were known to a majority of the group (common). See Table 2 for an example of an information profile.

If information from all skill tests were combined, they were designed to favor a single candidate who had the greatest net positive results. Individual information profiles were designed to create a high degree of conflict among group members regarding the preferred candidate. Each information profile included information that suggested one candidate was the best: 1/3 suggested the true best candidate, 1/3 suggested one non-ideal candidate, and 1/3 suggested a different non-ideal candidate.

Because crowds arrive asynchronously the materials were designed to provide information in a specific sequence. Information profiles were assigned to group members sequentially so that each subsequent person to arrive favored a different candidate (e.g. A, B, C, A, B) and the group had access to all the facts by the time the 10th person had arrived (1/3 of the group).

Participants

Participants were recruited with a Human Intelligence Task (HIT) placed on Amazon's Mechanical Turk (AMT). On crowdsourcing platforms like AMT, a large number of individuals often drop out of a task and/or fail to complete even the most basic requirements of the task. To focus our attention on groups that had minimal participation we excluded any crowd that did not have at least two thirds of members either post a message or fill out the questionnaire, and at least half of members report a group solution. This left 13 crowds for analysis.

		Candidates				
Id	**Skills**	A	B	C	D	**Fact Type**
2	Verbal Comprehension	-	+	+	+	Infrequent
3	Verbal Reasoning	+	-	+	+	Common
5	Grammar	0	0	+	-	Common
9	Algebra	+	-	-	+	Common
10	Trigonometry	+	+	0	-	Common
11	Data Interpretation	-	-	+	-	Common
15	Arithmetic	0	-	+	+	Common
16	Analogies	+	+	0	-	Common
17	Antonyms	-	0	+	-	Common
	Profile Score	1	-1	**5**	-1	
	Overall Score	**5**	1	2	0	

Profile Preference: **C**, Correct Solution: **A**

Table 2: Example information profile.

We describe the demographics for individuals in crowds who met the inclusion criteria. In total 293 individuals participated in our study. 53% were male, ranging in age from 18 to 70 (M = 34, Median = 31). A majority of participants had completed an undergraduate degree or higher (55%). Most participants were from the United States (86%), a smaller percentage were from India (9%) and a variety of other countries (5%).

Procedure

After accepting our HIT and receiving informed consent, participants proceeded to the main task. Participants were assigned to an active group and each person was given a unique information profile. In the main task, participants had access to general instructions, a table that presented their personal information profile, and the group discussion forum. Participants could use the forum to discuss the task by reading and posting messages to the group; it allowed threading and voting; and displayed a persistent history of messages. The discussion forum was typical of communication tools used by crowds (e.g. Reddit). Participants were allowed to join and leave the main task at will. When participants decided that they had completed the task, they were allowed to proceed to the questionnaire.

After completing the main task, participants filled out a questionnaire. The questionnaire asked each participant to report their final group solution, group satisfaction, and general demographics (e.g. age, sex). Participants were paid $1.50 for the main task and up to $1.00 extra for the bonus if they reported the correct solution (in proportion to the number of group members who submitted the correct solution). We allowed participants to complete the task multiple times, because workers in crowd settings often complete the same task more than once (e.g. answering similar questions on a Q&A). Repeat participants could not use prior knowledge to improve their performance, because we used a different set of facts and solutions for each session. We observed no improvement in performance for repeat users.

Content Coding

Individuals' messages were independently scored by two coders who were naïve to the research questions. For each message the coders recorded whether the message included a discussion of the facts and whether it included specific facts. The coders reached high levels of agreement on both measures (discussed facts: kappa = 0.75, shared specific facts: kappa = 0.76). We also asked coders to record whether a message suggested a group decision and, if so, to record the suggested decision (kappa = 0.64). Disagreements were resolved by taking the union of the two coders responses (e.g. a statement was considered to have discussed facts if either coder marked it as such). For messages that were marked as sharing specific facts, we recorded which specific facts were shared and their type (common, infrequent, or unique). A generous approach to coding facts was taken in which specific facts were coded as having been shared even if only part of a fact was shared (e.g. score for one but not all candidates on a skill).

Results

RQ1: How do crowds share and discuss distributed facts?

We observed how crowds discussed distributed facts and answered two sub-questions: did crowds pool distributed facts? and did crowds exhibit bias in which facts they shared? Raters coded the content of discussion statements, which were then aggregated at a crowd level for analysis.

Anatomy of crowd discussions In discussing the task, crowds made several different kinds of statements, such as sharing specific facts (e.g. "C has better verbal comp"), discussing facts broadly (e.g. "I only see 3 pluses for A I see 5 for C"), suggesting a group decision (e.g. "I agree B is best"), greetings (e.g. "Hi everyone!"), meta-discussions of the task (e.g. "Our job is to find the best candidate. Isn't it?"), strategies for making a decision (e.g. "Let's do the final vote"), and asking for information (e.g. "Between your candidates B and C which has better verbals comprehension skill?"). We attended to the first three types of statements.

Sharing and discussing facts was rare. Only 23% of statements discussed facts, compared to 62% which discussed a group decision. An even smaller percentage of statements, 3%, included information about specific facts. Participants discussed facts in a way that was atypical of other hidden profile studies. Individuals tended to aggregate positive and negative facts by candidate (e.g. "I only see 3 pluses for A I see 5 for C"), rather than discussing the details of a candidates strengths (e.g. "C has better verbal comp"). This simplified communication, but obscured details about the facts needed to fully pool information. For example, users may have discussed strengths associated with Candidate A without realizing they were talking about different strengths.

Crowds began discussing a decision before they shared facts. For each crowd we recorded when they first discussed facts, shared a fact, and/or suggested a decision. Paired t-tests showed that on average crowds first discussed a decision 35 seconds before they discussed facts ($t(12) = -3.80$,

Model 1 - Top Voted Statements				
	Coef.	**SE**	**t**	**p**
Intercept	0.06	0.02	3.74	0.0002
Stated Decision	0.14	0.02	6.89	< 0.0001
Discussed Facts	-0.11	0.03	-4.22	< 0.0001
Shared Facts	0.01	0.06	0.21	0.83
R^2	4%			

Table 3: Linear regression model predicting the score for a discussion statement based on the content of the statement[4].

p = 0.003, d = -1.05) and 4 minutes and 32 seconds before they shared facts[2] (t(9) = -2.48, p = 0.03, d = -0.79). Crowds began trying to make a decision from the very beginning, which meant that they typically began discussing a decision before discussing the facts of the problem.

Further, we found that decisions were more visible than discussions of facts. Voting can be used to make important content visible in the discussion. We compared the scores given by the crowd to different types of statements. We found that suggested decisions were given higher scores on average than discussions of facts (See Table 3). We speculate that decisions received higher scores because crowd members used 'up votes' to express agreement with an opinion. However, a consequence was that decisions were promoted to be more visible than discussions of facts.

In summary, we observed two distinctive characteristics of crowd discussions: 1) facts were shared in aggregate, obscuring details, and 2) there was a greater focus on decision-making than discussing or sharing facts.

Information pooling We found that crowds shared very few facts. On average, crowds shared 19% of the facts given (3.46 out of 18; 95% CI 1.84 - 5.00). Some crowds shared no facts at all, while the crowd that shared the most facts only shared half of them (9 out of 18). This is partially a result of the atypical way that crowds discussed facts in this study compared to prior studies. As mentioned above, to reduce the amount of information that needed to be communicated users took the approach of tallying strengths and weakness before sharing information with their group (e.g. "I only see 3 pluses for A"), which meant very few users shared concrete facts (e.g. "A is good at algebra"). While this approach made communication easier, it also inhibited information pooling because the specifics of facts were not shared. As a result there was very little information pooling in any of the crowds.

Information exchange bias We expected that a higher proportion of common facts would be shared. We computed the proportion of common, infrequent, and unique facts shared per crowd. For example, if 2 out of 12 common facts were shared in discussion we recorded that 17% of common facts were shared. On average we found that 21% of common facts, 26% of infrequent facts, and 5% of unique facts were shared. Planned pairwise comparisons showed a statistical difference in the proportion of common and unique facts shared (t(12) = 3.24, p =0.007, Cohen's d = 0.90), but no difference in the proportion of common and infrequent facts shared (t(12) = -0.42, p = 0.68). Crowds shared a greater proportion of common facts than unique facts.

In comparison to prior work, we found all types of facts were shared at much lower rates than is typical for this type of task. In prior work the exact rates of sharing vary depending on the study design, however all comparable design have higher rates than what we observed. For example, (Larson et al. 1996) found small face-to-face groups shared 77% common facts and 67% of unique facts, while (Hightower and Sayeed 1996) found small groups using computer-mediated communication shared 25% of unique facts. The complexity of the current task was similar to tasks used in these other studies. Crowds on AMT engaged in much less information pooling compared to prior studies.

We found that crowds exhibited a partial bias in how they shared facts. Crowds were more likely to share common facts than unique facts, but no more likely to share common facts than infrequent facts. While both infrequent and unique facts were rare (known to less than a third of the group), infrequent facts were known to many more people than unique facts (10 vs. 1). In small groups rare knowledge is necessarily known to only one or two people, in crowds rare knowledge may be known to enough people that it functions more like common knowledge. As a result there may be less bias in the degree to which rare knowledge is shared in crowds.

RQ2: How well do crowds perform on a hidden profile task? Crowds were tasked with using the provided facts to find the best solution as a group. Each crowd member reported the group decision. We examined whether there was consensus among the group, whether group solutions changed over time, and we calculated performance based on the recorded solutions.

Consensus Users reported their group's solution. For many crowds there was a lack of consensus. 29% of users reported a different group solution than was given by the majority. Despite being asked to reach a unanimous group decision, crowds did not always reach consensus, for several different reasons. Sometimes minority opinions persisted despite attempts to reach a consensus. At other times, group members had different perceptions of the discussion and its presumed consensus. Sometimes group groups reached a local consensus, but this agreement changed over time. We discuss the latter in more detail in the next sections.

In a crowd it may be very difficult to reach a true consensus because crowds are large with high turnover. In addition, there may not be as strong a social norm for consensus in the crowd. Previous work shows that a diversity of perspectives are valued in crowd discussions (Mamykina, Nakikj, and Elhadad 2015). Since, we did not find a singular group solution, we examined four different methods for extracting a solution from a crowd, using all individual user reports (Report. User); the majority solution given by users in a crowd (Report. Majority); the decision mentioned the most times in the crowd discussion (Disc. Majority); or the discussion

[2]Crowds that never discussed specific facts were excluded from this analysis.

Solution Metric	RMaj	RUser	DMaj
Report. Majority (RMaj)			
Report. User (RUser)	71%		
Disc. Majority (DMaj)	85%	67%	
Disc. Top (DTop)	62%	56%	69%

Table 4: Percent agreement between different methods of recording a group solution.

Solution Metric	% Correct	95% CI	p-value
Report. Majority	0.77	0.46-0.95	0.002
Report. User	0.62	0.56-0.68	<0.001
Disc. Majority	0.62	0.32-0.86	0.04
Disc. Top	0.46	0.19-0.75	0.38

Table 5: Proportion of crowds which solved the problem correctly. For each proportion a binomial test was conducted to compare the proportion of crowds solving the problem correctly to the proportion expected to solve it correctly based on individual information profiles.

	Num. Crowds (Percent)	
Constant Solution	8 (62%)	
Correct		5
Incorrect		3
Changing Solution	5 (38%)	
Incorrect to Correct		1
Correct to Incorrect		4
Incorrect to Incorrect		0

Table 6: Total number of crowds with a constant solution and with a changing solution as well as sub totals based on whether the solution was correct.

statement that included a decision and was given the highest score by the crowd (Disc. Top). No two methods produced exactly the same results, though there was moderate agreement between the four approaches (Table 4).

Solution quality Table 5 reports group performance using the 4 different methods for recording group solutions. We found that some methods for extracting a solution from a crowd produced more accurate solutions than others. For example, using aggregated user reports produced more accurate solutions than using unaggregated user crowds or using records of the discussion.

We focused our attention on the best performing method, aggregated user reports. Based on this evaluation, we found that 77% of crowds identified the correct solution. We can compare this rate of performance to a few benchmarks. First, we compared it to what we would expect by chance given initial individual preferences. Based on distributed information profiles, one third of crowd members preferred the ideal candidate before discussion. We found that crowds performed significantly better than this benchmark (see Table 5). A one-sample test of proportions showed that crowd performance of 0.77 was significantly greater than 0.33 ($p = 0.002$). Second, performance of 77% is very similar to findings from prior studies that used similar designs; (Hightower and Sayeed 1996) found small groups identified the correct solution 75% of the time.

We found a positive but non-significant correlation between how many facts were shared and group performance (Biserial correlation = 0.52; Logistic regression Coef = 0.46, $z = 1.23$, $df = 11$, $p = 0.22$), which is consistent with prior work in small groups (Lu, Yuan, and McLeod 2012). The relationship between sharing facts and performance is complex and noisy because it depends on how groups share and weight facts in forming a decision (Stasser 1992). Crowds likely did as well as they did because they discussed facts in aggregate and they shared some facts, particularly rare ones. If they had done a better job at pooling facts they could have performed much better.

Stability over time We examined whether the group decisions of our crowds changed over time. Due to the nature of crowds on AMT, arrivals were staggered and exhibited high turnover. Arrivals followed a Poisson distribution with 50% of the crowd arriving by 7 minutes, 75% by 17 minutes, and 100% by 42 minutes on average. Users tended to stay an average of 8 minutes and 19 seconds (Median = 6 minutes, SD = 11 minutes). As a result around half the crowd was present at the beginning and the percent of the crowd present at any given time declined rapidly over the course of the discussion (Table 7). Discussions lasted on average 63 minutes (Median = 54 minutes, SD = 31 minutes).

We expected that group decisions might change over time, especially given that work in open collaborations tends to develop iteratively (Howison and Crowston 2014). We also predicted that when group decisions changed they would tend to get better; edits tend to improve the quality of work on Q&As and wikis (Kittur and Kraut 2008; Li et al. 2015; Tausczik, Kittur, and Kraut 2014). In crowd discussions users who arrive late have a record of all previous discussion, including a record of shared facts and group reasoning. We examined users' reported group decision in the order that users left the discussion. We considered a decision to be a dominant group decision at a particular time if three consecutive users reported it as the group decision. We tallied the number of crowds whose group decision shifted over time (Table 6)

We found that the group decision changed over time for 5 out of 13 crowds. In contrast to our prediction, crowds were more likely to shift from a correct decision to an incorrect decision (80%). In other words, some crowds updated their group decisions as users came and went, but when they did so decisions got worse more often than they got better.

Did early discussion influence users who arrived late? Because of staggered arrival and high turnover, users who arrived late had access to early discussion, but could not interact with the users who left those messages. Instead they interacted with a new set of individuals who may or may not have had the same opinions and perspectives on the problem. For some crowds discussion of opinions changed over time. Figure 1 visualizes discussion of opinions over time for

three types of crowds: crowds with a constant correct decision, a constant incorrect decision, and a changing decision from correct to incorrect[5]. For crowds with a constant correct solution, the correct solution remained a dominant opinion throughout discussion. This was not true for crowds with a constant incorrect decision, the incorrect solution dominated early, but competed with other solutions later in discussion. For crowds that changed group decisions, a different solution became a dominant opinion later in discussion.

For users who arrived after discussion had begun we examined which had a greater influence on their reported group decision: early discussion or late discussion. For each user we gathered the set of statements made before a user entered the discussion (early discussion) and the set of statements made while the user was present (late discussion). For each set of statements and each candidate we calculated the proportion of statements supporting that candidate (e.g. 20% of early statements supported Candidate A). We entered these two variables in a model predicting whether a user reported a particular candidate as the group solution. As a potential control variable we also included whether a users' individualized information profile suggested this candidate was the best (profile preference). We used a generalized mixed effects model to control for dependencies in the data[6].

We found that both early discussion and late discussion influenced users' reports of the group decision (Table 7 Model 2). Early discussion had a greater influence on reported decisions than late discussion (standardized coefficient of 1.55 vs. 0.92). We predicted that the impact of early discussion might decline as time passed. We tested a second model which considered the time a user arrived, and the interaction between when a user arrived and the influence of early and late discussion on the reported decision (Table 7 Model 3). We found significant interactions between arrival time and the influence of early and late discussions on users' report of the group discussion. The later a user arrived the less influence early discussion had on their reported decision and the more influence late discussion had on their reported decision.

In summary, for some crowds a dominant opinion emerged early in discussion and remained dominant over time. In others opinions shifted over time in the discussion. Users who arrived after discussion had begun were influenced by both early and late opinions expressed in the discussion. The more time passed since the start of discussion and when a user arrived the less the user was influenced by early opinions and the more they were influenced by late opinions. Crowd discussions that persist with an influx of users over a long period may be vulnerable to changing decisions.

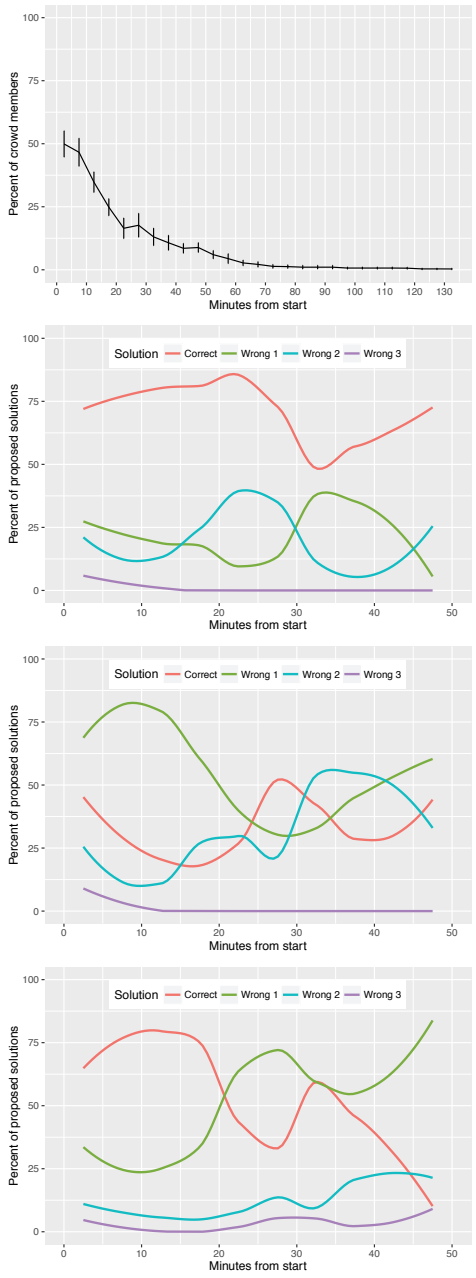

Figure 1: Top: Average proportion of a crowd present over time. Middle Top: Changes in proposed solutions discussed by crowds with a constant correct solution. Middle Bottom: Changes in proposed solutions discussed by crowds with a constant incorrect solution. Bottom: Changes in proposed solutions discussed by crowds with a changing solution.

Discussion

Crowd applications, innovation contests and Q&A platforms are built to capitalize on the "wisdom of crowds" by bringing together diverse and specialized knowledge from many individuals. However, these applications and platforms may fail to realize the full benefits of diversity. Prior work suggests that groups often fail to pool knowledge effectively when that knowledge is distributed. To our knowledge this is the

[5]Since there was only one observation we did not graph the discussion for the crowd whose decision changed from incorrect to correct.

[6]glmer was used from the R package lme4.

	Model 2 - Reported Decision				Model 3 - Reported Decision			
	Std. Coef.	**SE**	**z**	**p**	**Std. Coef.**	**SE**	**z**	**p**
Intercept	-2.10	0.18	-11.75	< 0.0001	-2.45	0.26	-9.67	< 0.0001
Early Discussion	1.55	0.14	10.9	< 0.0001	2.26	0.20	10.6	< 0.0001
Late Discussion	0.92	0.14	6.69	< 0.0001	0.64	0.15	1.31	0.19
Profile Preference	0.41	0.26	1.56	0.12	0.60	0.28	2.13	0.03
Arrival Time					0.14	0.14	1.00	0.32
Time X Early Discussion					-0.78	0.12	-6.52	< 0.0001
Time X Late Discussion					0.54	0.14	3.75	0.0002
Pseudo R^2	63%				69%			

Table 7: Generalized mixed effects models predicting user reported decisions based on early and late discussion (Model 2) as well as moderation by user arrival time (Model 3).

first study to test the shared information bias in a crowd setting. Crowds in this study failed to pool facts effectively, exhibited partial bias in how they shared facts, and often failed to reach a consensus. Nevertheless moderately accurate solutions could be deduced from crowd responses.

Crowds can share information, partial work, and collectively make sense of a problem when they are given tools for discussion such as chat, online forums, Q&As, wikis, collaborative document editors, repositories for sharing products, and annotation and comment systems. However, crowds may not realize the full potential of discussion. In this study we found crowds focused too much on debating decisions and not enough on sharing and discussing facts. As a result at most 50% of known facts were discussed explicitly. While AMT may incentivize minimal communication and expediency, encouraging less sharing, we also see this as a consistent pattern across other crowd platforms such as Q&As and innovation contests. Crowds, in general, may focus on pooling individually generated solutions at the cost of in-depth discussion, sharing of partial information, and collaboration (Zagalsky et al. 2016).

Collective sensemaking is difficult but, when properly supported, can produce better solutions than is possible by aggregating individual responses (Atanasov et al. 2017; Boudreau and Lakhani 2015). However, The right affordances and incentives are needed to make pooling knowledge easier and more rewarding.

A somewhat positive finding of our study is that crowds may be less biased in their discussion of distributed facts than small groups. While previous work has found that small (e.g. N = 3) and medium (e.g. N = 6) sized groups share common facts more than rare facts (Lu, Yuan, and McLeod 2012; Cruz, Boster, and Rodriguez 1997), we found that crowds (N = 30) shared rare (infrequent) facts at about the same rate as common facts, even when those facts were only available to a minority of the crowd. Sharing rare facts is important to problem solving: as they are not widely known and they can radically change majority opinion when they are relevant. In crowds infrequent facts may be available to enough individuals to be shared like common information, even when they are known only to a small fraction participants. However crowds, like small groups, are still biased against sharing unique facts known only to a single person.

In order to interpret these finding in real-world contexts, we need to know more about how information is typically distributed in crowds in the wild. An open question for future research is whether key facts are typically known by only a single person or by a small minority of the crowd?

Consensus was more elusive for the crowds in this study than has been found in prior work on traditional small groups. As users entered and left, crowd composition shifted and different opinions dominated discussion. Later crowd members used both early and late discussion to form a group decision, which meant that for 40% of crowds group decisions shifted over time. Counter to our predictions, updated group decisions got worse more often than they got better. This result has two implications for crowds. First, due to a lack of consensus the quality of crowd solutions will depend on how solutions are inferred from crowd discussions. Second, archived crowd discussions may be difficult for new readers and participants to fully understand.

We found that crowds performed moderately well, about as well as traditional small groups in prior work, when aggregated user reports were used to judge performance. Not all methods for inferring a solution are likely to perform as well. We found a trend in which individual user reports performed worse. Despite the fact that crowds were more likely to identify the correct solution in aggregate, many of the individuals in those crowds still identified the wrong solution. This result suggests that while crowdsourcing requesters may be able to extract good solutions by aggregating a crowd's responses, many individual crowd members and subgroups in the crowd may be misinformed even after the correct information and reasoning has surfaced.

Readers use discussions left behind by crowds to evaluate options, form opinions, and shape behavior. For example, programmers read popular Q&As to help with their work (Anderson, Huttenlocher, and Kleinberg 2012); individuals consult online communities to gather health information and make decisions (Mamykina, Nakikj, and Elhadad 2015); citizens read user comments to understand policies, social issues, and the news (Esau, Friess, and Eilders 2017). These discussions result from many people coming together, sharing information, debating, and trying to make sense of a problem. However, the information, reasoning, and final conclusions may not be easy for individuals who were not active in the conversation to understand. We found that majority opinions and top rated opinions stated in the discus-

sion sometimes disagreed with user reports and were less likely to be correct. In addition, we found that users who joined the discussion late were influenced by early discussion, but sometimes reported different solutions from earlier participants, which were also less likely to be correct. This is consistent with work on online health communities which finds that users have trouble understanding arguments left behind in threaded discussions (Mamykina, Nakikj, and Elhadad 2015). A different structure may be needed to make crowd discussions useful artifacts for future readers.

Limitations and Future Work

As an initial exploratory study of hidden profile task in the crowd we took an observational approach which allowed us to describe behavior; however this approach has a few limitations. First, it does not allow direct comparison to traditional small groups. Given the large number of differences between traditional small groups and crowds, including size, turnover, asynchronicity, communication platform, social norms, and sample population, future work should extend the current study by experimentally testing the impact of each dimension individually on information sharing and performance.

Second, in this study of observed crowd behavior we limited ourselves to only one crowdsourcing context, small to moderate sized crowds (N = 30) addressing a simple task. AMT is one specific crowdsourcing platform, with minimal communication, external incentives in the form of payment and a less educated crowd. Other platforms vary along all these dimensions and more. Studies of crowds on AMT often involve judgements made by crowds of similar sizes, but there is considerable variation in crowd size from only a few active contributors to many thousands. Future studies should consider larger group sizes (100-1000 individuals) and different platforms to see how these design choices might affect findings.

Third, we took a standard approach to studying the way that groups communicated while working on a hidden profile task, counting how many facts were shared in discussion. Most real crowd tasks that involve sharing information also involve more complex reasoning and problem solving. Additionally, in crowds more complex social dynamics may develop (e.g. subgroups). Future work should examine the content and structure of crowd discussions using more sophisticated methods, such as content coding of a broader class of statements and network analysis, especially in crowds working on more complex, realistic problems.

Conclusion

Theoretically one of the most important advantages of crowds problem-solving is the ability to pool diverse knowledge from tens to thousands of people. In this study, we investigated a potential impediment to information sharing in such crowds. Extrapolating from foundational work on information sharing in small groups in social psychology, we tested whether and how crowds on Amazon's Mechanical Turk pooled distributed information. Evidence from this study suggests that crowds only pooled a fraction of the information that they were given and were partially biased in which information they chose to share. Additionally, because group decisions in crowds shifted over time as crowd members came and left, we found that aggregate group decisions were more trustworthy than crowd members' individual understandings of group consensus. Future work should explore how specific dimensions of crowd work, such as high turnover and problem self-selection affect biases in information pooling.

Acknowledgments

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

References

Anderson, A.; Huttenlocher, D.; and Kleinberg, J. 2012. Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD '12)*, 850–858. ACM Press.

Atanasov, P.; Rescober, P.; Stone, E.; Swift, S. A.; Servanschreiber, E.; Tetlock, P.; Ungar, L.; and Mellers, B. 2017. Distilling the Wisdom of Crowds: Prediction Markets versus Prediction Polls. *Management Science* 63(3):691–706.

Boudreau, K. J., and Lakhani, K. R. 2015. "Open" Disclosure of Innovations, Incentives and Follow-on Reuse: Theory on Processes of Cumulative Innovation and a Field Experiment in Computational Biology. *Research Policy* 44:4–19.

Cranshaw, J., and Kittur, A. 2011. The Polymath Project: Lessons from a Successful Online Collaboration in Mathematics. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 1865–1874. ACM.

Cruz, M. G.; Boster, F. J.; and Rodriguez, J. I. 1997. The Impact of Group Size and Proportion of Shared Information on the Exchange and Integration of Information Groups. *Communication Research* 24:291–313.

Diehl, M., and Stroebe, W. 1987. Productivity Loss in Brainstorming Groups: Toward the Solution of a Riddle. *Journal of Personality and Social Psychology* 53:497–509.

Esau, K.; Friess, D.; and Eilders, C. 2017. Design Matters! An Empirical Analysis of Online Deliberation on Different News Platforms. *Policy and Internet* 9(3):321–342.

Gigone, D., and Hastie, R. 1993. The Common Knowledge Effect: Information Sharing and Group Judgment. *Journal of Personality and Social Psychology* 65:959–974.

Greitemeyer, T., and Schulz-Hardt, S. 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* 84:322–339.

Hightower, R., and Sayeed, L. 1996. Effects of communication mode and predicsussion information distribution characteristics on information exchange in groups. *Information Systems Research* 7(4):451–465.

Howison, J., and Crowston, K. 2014. Collaboration through open superposition: A theory of the open source way. *MIS Quarterly*.

Jeppesen, L. B., and Lakhani, K. R. 2010. Marginality and Problem Solving Effectiveness in Broadcast Search. *Organization Science* 21:1016–1033.

Kittur, A., and Kraut, R. E. 2008. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, 37–46. ACM Press.

Krause, S.; James, R.; Faria, J. J.; Ruxton, G. D.; and Krause, J. 2011. Swarm Intelligence in Humans: Diversity can Trump Ability. *Animal Behaviour* 81:941–948.

Larson, J. R.; Christensen, C.; Abbott, A. S.; and Franz, T. M. 1996. Diagnosing groups: Charting the flow of information in medical decision-making teams. *Journal of Personality and Social Psychology* 71(2):315–330.

Latané, B.; Williams, K.; and Harkins, S. 1979. Many Hands make Light the Work: The Causes and Consequences of Social Loafing. *Journal of Personality and Social Psychology* 37:822–832.

Laughlin, P. R. 1999. Collective Induction: Twelve Postulates. *Organizational Behavior and Human Decision Processes* 80:50–69.

Lee, D. 2013. Boston bombing: How internet detectives got it very wrong.

Li, G.; Zhu, H.; Lu, T.; Ding, X.; and Gu, N. 2015. Is it Good to be like Wikipedia?: Exploring the Trade-offs of Introducing Collaborative Editing Model to Q&A Sites. In *Proceedings of the Computer Supported Cooperative Work and Social Computing (CSCW)*, 1080–1091. ACM.

Lu, L.; Yuan, Y. C.; and McLeod, P. L. 2012. Twenty-five Years of Hidden Profiles in Group Decision Making: A Meta-Analysis. *Personality and Social Psychology Review* 16:54–75.

Mamykina, L.; Manoim, B.; Mittal, M.; Hripcsak, G.; and Hartmann, B. 2011. Design Lessons from the Fastest Q&A Site in the West. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2857–2866. ACM.

Mamykina, L.; Nakikj, D.; and Elhadad, N. 2015. Collective Sensemaking in Online Health Forums. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* 3217–3226.

Mitra, T., and Gilbert, E. 2015. CREDBANK: A Large-scale Social Media Corpus With Associated Credibility Annotations. In *Proceedings of the International Conference on Web and Social Media*, 258–267. AAAI.

Muchnik, L.; Aral, S.; and Taylor, S. J. 2013. Social Influence Bias: A Randomized Experiment. *Science* 647:647–651.

Postmes, T.; Spears, R.; and Cihangir, S. 2001. Quality of Decision Making and Group Norms. *Journal of Personality and Social Psychology* 80:918–930.

Stasser, G., and Titus, W. 1985. Pooling of Unshared Information in Group Decision Making: Biased Information Sampling During Discussion. *Journal of Personality and Social Psychology* 48:1467–1478.

Stasser, G., and Titus, W. 2003. Hidden Profiles: A Brief History. *Psychological Inquiry* 14:304–313.

Stasser, G. 1992. Pooling of unshared information during group discussion. *Group Process and Productivity* 48–57.

Surowiecki, J. 2005. *The wisdom of crowds*. New York: Random House.

Tausczik, Y. R.; Kittur, A.; and Kraut, R. E. 2014. Collaborative Problem Solving: A Study of MathOverflow. In *Proceedings of the Computer Supported Cooperative Work and Social Computing (CSCW)*, 355–367. ACM.

Tausczik, Y.; Wang, P.; and Choi, J. 2017. Which Size Matters? Effects of Crowd Size on Solution Quality in Big Data Q&A Communities. In *ICWSM*, 260–269.

Tausczik, Y. R. 2016. Citation and Attribution in Open Science: A Case Study. In *Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, 1524–1534. ACM.

Wittenbaum, G. M., and Bowman, J. M. 2004. A Social Validation Explanation for Mutual Enhancement. *Journal of Experimental Social Psychology* 40:169–184.

Wittenbaum, G. M.; Hollingshead, A. B.; and Botero, I. C. 2004. From Cooperative to Motivated Information Sharing in Groups: Moving Beyond the Hidden Profile Paradigm. *Communication Monographs* 71:286–310.

Wood, W.; Lundgren, S.; Ouellette, J. A.; Busceme, S.; and Blackstone, T. 1994. Minority Influence a Meta-Analytic Review of Social Influence Processes. *Psychological Bulletin* 115:323–45.

Yang, Y.; Chen, P.-y.; and Pavlou, P. 2009. Open Innovation: Strategic Design of Online Contests. In *Proceedings of the International Conference on Information Systems (ICIS)*, 1–42. AIS.

Zagalsky, A.; Gomez Teshima, C.; German, D. M.; Storey, M.-A.; and Poo-Caamaño, G. 2016. How the R Community Creates and Curates Knowledge: A Comparative Study of Stack Overflow and Mailing Lists. In *International Conference on Mining Software Repositories (MSR)*, 441–451. ACM.

Zhu, H.; Dow, S. P.; Kraut, R. E.; and Kittur, A. 2014. Reviewing versus Doing: Learning and Performance in Crowd Assessment. In *Proceedings Conference on Computer Supported Cooperative Work and Social Computing (CSCW '14)*, 1445–1455. ACM Press.