Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Which Size Matters? Effects of Crowd Size on Solution Quality in Big Data Q&A Communities

Yla Tausczik, Ping Wang, Joohee Choi
iSchool, University of Maryland, College Park
{ylatau, pwang, jchoi27}@umd.edu

Abstract

Question Answering websites have evolved into one of the most important platforms for knowledge sharing and problem solving online. Despite widespread adoption of Q&As by technical communities as well as an abundance of domain experts, many questions fail to attract a sufficient audience to obtain a good solution or any solution at all. We investigate the effects of crowd size on solution quality in Stack Exchange Q&A communities on topics related to big data. We find that three distinct levels of group size in the crowd (topic audience size, question audience size, and number of contributors) affect solution quality. Therefore, we argue that group size in the crowd is not unitary, but rather a multi-level construct. This work advances a theoretical model of group size in the crowd and the relation between crowd size and performance. The work also has practical implications for system designers trying to route crowds to problems efficiently.

Question answering (Q&A) sites have evolved into one of the most important platforms for knowledge sharing and problem solving online. On a question answering site individuals pose questions directly to the site's community, which is often organized around a common, domain-specific interest. Popular Q&A sites are known for providing good answers quickly (Mamykina et al. 2011), and their efficacy and efficiency has led to widespread adoption. Open source software user support, which used to be conducted primarily through mailing lists, has largely migrated to Q&A communities (Vasilescu et al. 2014). Mature Q&A sites have fostered devoted communities of experts with deep domain knowledge. In their responses, these experts create rich content of lasting value (Anderson, Huttenlocher, and Kleinberg 2012). In aggregate, these Q&A sites use the crowd to generate documentation that rivals and exceeds official descriptions, especially because Q&A responses are often more up-to-date and complete (Parnin et al. 2012). Q&As are particularly valuable for new and emerging technical practices, in which software, techniques, and tools change rapidly.

Despite widespread adoption of Q&As by technical communities as well as an abundance of domain experts, many questions submitted to these sites go unanswered, are not answered in a timely fashion, or are answered only partially (Asaduzzaman et al. 2013). One reason questions are not answered may be that they never reach a large enough audience of experts. While many questions on these platforms attract thousands of views, others only attract a handful. Directing the audience's attention is a practical consideration for both individual question askers and system designers. Question askers must attract an audience which is large and targeted enough to generate a good solution. System designers want to allocate community resources efficiently by ensuring that community attention is distributed efficiently to the questions that will benefit most from greater visibility.

How large of an audience is needed to attract a good solution? The relationship between group size and problem-solving performance is complex. Theoretically, small group researchers have argued that performance should increase monotonically with group size, because larger groups possess more knowledge from which to generate solutions (Steiner 1966). However, the positive effects of group size on performance are often reduced or eliminated by group inefficiencies (Hill 1982). On average, groups only perform as well as their best member and additional group members add relatively little. At worst, groups perform worse than the individuals would have working on their own. Crowds are assumed to benefit from their large size and sidestep some group inefficiencies, but studies on the relationship between crowd size and performance are mixed. Larger crowds have been associated with better performance, but only under some conditions (Kittur and Kraut 2008). Mixed empirical findings in small groups showed the relationship between group size and performance is complicated and context dependent, requiring more complex theoretical models, this is even truer for the crowd. In particular, we argue that the nested, multi-layered nature of groups in the crowd require a more sophisticated theoretical approach.

Previous studies on group size in the crowd tend to treat group size as a unitary construct (Robert and Romero 2015). In contrast, we argue that group size in the crowd manifests at multiple levels, a concept which we refer to as ***stratified group size***. The Q&A process is a type of broadcast search, and one of its central features is problem self-selection (Jeppesen and Lakhani 2010). A problem is broadcast to a large crowd, a smaller group of individuals click through to actually read it, and a still smaller subset chooses to actively work on the problem to submit a comment or so-

lution. This process of self-selection leads to the stratified nature of crowds: when we talk about group size, are we talking about the community who might see the problem, the audience who reads it, or the contributors who (try to) solve it?

All of these might plausibly effect one another and overall performance. A larger community has access to more expertise, but this only matters if the right people see a question. A larger audience ensures broader exposure for a question, but only matters if some of these readers provide useful responses. A larger number of contributors should presumably increase the completeness and quality of the solution, but may be susceptible to group inefficiencies.

The contribution of this paper is two-fold. First, we advance a theoretical model of group size in the crowd. Second, we empirically test this model by investigating the relationship between the various levels of group size and solution quality.

Related Work

Our research on the relationship between group size and solution quality draws on previous research conducted with small groups, crowds in other settings (e.g. wikis, Amazon's Mechanical Turk), and crowds in Q&A communities.

Small Group Problem Solving

Small group researchers have studied the effect of group size on problem-solving performance in small groups for over half a century. While group size was initially believed to increase group problem-solving performance, empirical evidence showed an inconsistent relationship between the two (Kerr and Tindale 2004). Some studies have found that larger groups performed better than small groups (Taylor and Faust 1952), while others have found no relationship between group size and performance (Lorge and Solomon 1959). Where an effect of group size on performance was found, the relationship was often negatively accelerating, meaning each additional group member added less value than the one before (Gibb 1951).

Small group researchers proposed increasingly complex theoretical models to explain mixed findings (Hill 1982). The basic argument can best be explained by the following equation: *group performance = sum of individual performance + collaboration benefits - process losses*.

Group size was believed to have a strong positive effect on performance when the gains from collaborating were high and/or process losses were low and to have little to no effect when process losses were high. A large body of research has investigated different types of collaboration benefits, such as assembly bonus effects (e.g. collective induction), and process losses, such as motivational losses (e.g. social loafing) and coordination losses (e.g. failure to pool information).

Building from this basic framework other researchers built more complicated theoretical models to explain when and how group size might increase performance. Steiner (1966) argued that potential collaboration benefits should depend on the type of task and that some types of tasks could potentially benefit more from larger groups. He proposed five different models relating group size to potential performance for five types of tasks–additive, disjunctive, conjunctive, compensatory and complementary. For example, the disjunctive model explained how performance would increase with group size for problems that could be solved by a single member of the group knowing the right answer. The complementary model explained how performance would increase with group size for problems that required multiple group members to work together to combine their knowledge. Steiner's argument that the relationship between group size and performance should depend on task type was persuasive. However, not many studies have empirically tested Steiner's theoretical models, in part because they are hard to apply to real-world problems.

Crowds and Collective Intelligence

Crowds are believed to succeed because they draw on work from a very large group. The principle of "wisdom of the crowd" is that aggregating independent contributions of many individuals results in better end products (Surowiecki 2005). In theory, aggregating over many diverse judgements concentrates useful, correct information while canceling out erroneous information. Researchers have found that larger crowds outperform smaller crowds because larger crowds are more diverse, improving aggregate judgements (Krause et al. 2011).

A different but complementary explanation for why size helps crowds to succeed is that it allows the best solvers and the best solutions to float to the top through problem self-selection. Jeppesen and Lakhani (2010) argue that self-selection in particular enables crowds to attract high-quality solutions. On InnoCentive, R&D problems are broadcast to a large crowd; the openness of the problem and the ability of crowd members to self-select problems removes barriers to entry. As a result they find that people who otherwise might not be assigned to work on a problem, such as those farther from the domain of the problem or those who are socially marginalized have a higher likelihood of submitting winning solutions. The combination of broadcasting to a large crowd and allowing self-selection means the final contributors have more relevant expertise than we would expect from assigned participants in a small group.

Kerr and Tindale (2011) argued that the factors that make the relationship between group size and performance complex for small groups should apply to large crowds on online platforms as well. The effect of group size on performance has been investigated in wikis and crowdsourcing platforms such as Mechanical Turk. Consistent with the empirical evidence for small groups, the relationship between group size to crowd performance has been mixed. Many studies have found that performance does increase with group size (Kittur and Kraut 2008; Robert and Romero 2015), but a few have reported finding no effect of group size (Zhu et al. 2014). Furthermore, group size was often moderated by other aspects of collaboration, including group diversity (Robert and Romero 2015), whether coordination was implicit or explicit (Kittur and Kraut 2008), and whether work was done sequentially or simultaneously (André, Kraut, and Kittur 2014). As was found with small groups, group size may only

benefit crowds under specific circumstances when the benefits of collaboration are high and process losses low.

Q&A Communities

Question answering websites enable information seekers to ask questions and community members to provide answers and share knowledge. They are becoming the dominant platform for providing software support (Vasilescu et al. 2014) and software documentation (Parnin et al. 2012). Through asking and answering questions, communities accumulate repositories of useful knowledge with lasting value (Anderson, Huttenlocher, and Kleinberg 2012). Community members provide answers through a mix of independent work and collaboration (Tausczik, Kittur, and Kraut 2014; Zagalsky et al. 2016). Collaboration happens when individuals iteratively contribute to solving a problem by adding complementary information, correcting mistakes, and/or improving on answers.

To date, only a few studies have investigated crowd size and performance in the context of Q&A communities. Typically, most questions on popular Q&A sites such as Stack Overflow (SO) are answered quickly (Mamykina et al. 2011). However there are also questions that are never answered, or receive only incomplete or belated answers (Asaduzzaman et al. 2013). Two studies have investigated the way that tag audience size affects response time for questions on SO and other Stack Exchange Q&A sites. Bhat and colleagues (2014) found that questions labeled with tags that had a larger audience, as defined by a larger historical pool of solvers, received answers more quickly. Ortega and colleagues (2014) partially confirmed Bhat and colleagues' findings using survival analysis, a more sensitive and appropriate statistical technique. They found that questions labeled with tags with larger audiences received an accepted answer more quickly in 3 out of 8 communities that they investigated, including SO, Math, and Server Fault.

Summary

Summarizing, the evidence from both small group and crowd research suggests larger groups can perform better, but only when collaboration is beneficial and efficient. The conditions needed for group size to increase performance are only sometimes met. Problem type is hypothesized to be one of these conditions, but has rarely been studied.

It is important to note that studies analyzing the effect of crowd size on performance have used incompatible definitions of crowd size. Studies with crowds on Q&A platforms have studied the crowd as the *potential* set of contributors, or audience, while studies focusing on wikis have measured actual contributors. No study has considered crowd size at multiple levels simultaneously. The current paper addresses these two gaps in the literature by simultaneously investigating the effect of multiple levels of group size on performance in the crowd, in the context of several different types of problems.

Research Questions

In previous work on group size in crowds, researchers have operationalized group size in different ways, with some looking at audience size while others considered the number of contributors. This paper aims to understand the relationship between group size and performance in Q&A communities by developing and testing a model of group size that includes both. This more complex model better captures the relationship between group size and performance in the crowd for a few reasons. First, research that only focuses on one type of group size (e.g. contributor size but not audience size) may underestimate the effect of group size on performance. Second, when relationships are context dependent, more complex models may be required to elucidate this dependence. Third, our model allows us to partially explain *why* group size affects performance.

We propose that, under the right conditions, more contributors and a larger audience can both improve performance in Q&A communities. A larger number of contributors can improve performance directly by providing new answers or enhancing old ones (Tausczik, Kittur, and Kraut 2014). A larger audience can improve performance through two distinct pathways. One, a larger audience increases the likelihood that more users will contribute to a question (Quantity Pathway). As stated above, more contributors can improve performance. Two, a larger audience increases the likelihood of reaching more qualified users who are better suited to contribute a solution (Quality Pathway). When audience size is large, there is a higher likelihood of a qualified expert seeing the question. When audience size is small, users without ideal expertise will try to answer questions as time elapses without a better response. In other words, as audience size increases there is a better fit between a question and its respondents, improving performance.

We investigate how both audience size and the number of contributors affects performance in Q&A communities. In Q&A communities we identify three levels of group size: there is a pool of potential experts on a topic that could attend to a question (topic audience), there is the a subset of these individuals who attend to a question (question audience) and there is a smaller set of individuals who actively help to answer the question (contributors). We measure the effect of three levels of group size on performance:

Research Question 1: How do the three levels of group size in the crowd–topic audience size, question audience size, and number of contributors–affect solution quality?

The effect of group size on performance is often context dependent. Steiner (1966) hypothesized that problem type would affect the relationship between group size and performance in small groups. The specifics of Steiner's models are hard to apply in practice. We evaluated the substance of the argument by investigating the question:

Research Question 2: Does the relationship between group size and solution quality depend on problem type?

Method

Analysis focused on three Stack Exchange (SE) Q&A communities in which big data was an active topic: Stack Overflow (SO), Cross Validated (CV), and Data Science (DS). Different communities attract different audiences, have different norms, and different participation patterns (e.g. number of questions, number of answers, total traffic). SO is the

oldest and most popular Q&A, focusing on programming. CV focuses on statistics and machine learning. DS is newest of the three, has the least traffic, and focuses on interdisciplinary methods related to data science. We investigated our research questions by drawing data from multiple communities, demonstrating that our results generalize across the particulars of specific communities. We chose to focus on big data because, as an interdisciplinary topic, it was an active topic in multiple communities. In addition, we selected big data because it is an area of rapid innovation in which software, methods, and techniques are rapidly evolving and being disseminated to a large community of practitioners, a context in which Q&A discussions are particularly valuable (Parnin et al. 2012).

Data Collection

We made use of user generated and curated tags to identify questions related to big data. We began with a set of seed tags clearly related to big data (e.g. 'big data', 'large data'). Using Chi Squared tests we gathered a larger set of tags that reliably co-occurred on the same questions as our seed tags. The final set of tags related to big data covered four general areas: cloud computing (e.g. 'amazon-spark', 'bigtable'), machine learning and text analysis (e.g. 'decision-tree', 'sentiment-analysis'), and two database management systems: Cassandra and Apache Kafka. Questions that had one or more of these tags were considered to be related to big data.

We collected questions with big data tags that were posted to SO, CV, or DS during approximately 2 weeks in early June 2016. Using the SE API we collected the number of views each question received hourly for the first 24 hours. In total we collected view data for 6,432 questions (SO: 4,936, 91%; CV: 385, 7%; DS: 111, 2%). We then omitted questions that were closed as inappropriate (2%), that were deleted by the question-asker (12%), or were missing hourly counts (14%). That left 3,918 questions (72%). Data about the questions, comments, answers, users, and votes were collected using the SE Data Explorer. Exactly 1 month after each question was posted we collected performance data including the (optional) best answer selected by the question asker and the number of votes which each answer received.

We selected a 24 hour activity window for this study because it balanced the capture of most solution-directed activity while avoiding most of the activity using the question/solution as an online resource. One of the primary measures of audience size was page views, which worsen as a measure of audience size as people visit the page for reasons other than answering a question. Most solving activity happens within the first 24 hours[1].

Types of Problems

We used a typology of SO questions (Treude, Barzilay, and Storey 2011), which consisted of ten different types of questions. A random sample of questions were hand coded by two raters to determine the type of problem. In total, 465 questions were coded, randomly sampled with a slightly higher percentage coming from CV and DS questions to achieve a more balanced data set across communities (SO: 58%, CV: 32%, DS: 10%). Interrater reliability indicated moderate agreement (62% Agreement, Cohen's Kappa 0.48). Consensus was reached through discussion or by bringing in a third coder when consensus could not be reached.

For the purposes of this paper we combined question types that were similar in their solutions and we did not consider question types that were infrequent in our dataset. Specifically, we identified three classes of questions that we expect would be solved differently: how-to (153, 33%), conceptual/decision-help (118, 26%), and error/review questions (103, 22%)[2]. How-to problems asked specific questions about how to perform a specific task. Typical questions asked about programming commands or statistical tests (e.g. *"If I input an image, how can [I] get a bounding box over the regions where a specific neuron is activated using keras or theano?"*). Conceptual and decision-help questions were grouped together because they both asked open-ended questions in which there was not necessarily a correct answer. In these questions, the question asker wanted an explanation of the underlying concepts in order to use tools and statistical approaches correctly (e.g. *"Can Random Forest regression handle non-stationary input variables?"*). Error and review questions asked for help solving a particular bug and/or reviewing code that was not behaving as desired. These questions asked the community to troubleshoot a problem (e.g. *"Currently i am trying to learn Secondary Sort in map reduce but getting an error of null pointer exception while running my mapper"*).

Statistical Models

We constructed regression models to relate group size in the crowd to solution quality. We considered three different levels of crowd size. We took multiple steps to ensure that our results are robust. We operationalized solution quality in two different ways, and showed that the pattern of results is consistent for both. Because one measurement of solution quality was binary while the other was a count, we used logistic regression and negative binomial regression models as appropriate to measure solution quality. Many different factors influence audience and group size on SE sites, and several of these are plausible confounds to the relationship between group size and solution quality. We constructed models with and without these control variables and showed that the results are consistent even controlling for potential confounds. We explain each variable in our models below:

Group Size On SE sites there are at least three levels of group size. Users reported finding questions to answer by searching for questions by tags. We defined **Topic Audience Size** as the set of solvers who might potentially answer a question with a set of specific tags. We operationalized

[1] A random sample of 6 month old questions showed that 75% of questions that received a good solution did so within 24 hours.

[2] For simplicity the grouped categories conceptual/decision-help and error/review are referred to as conceptual and error respectively in the results and discussion sections.

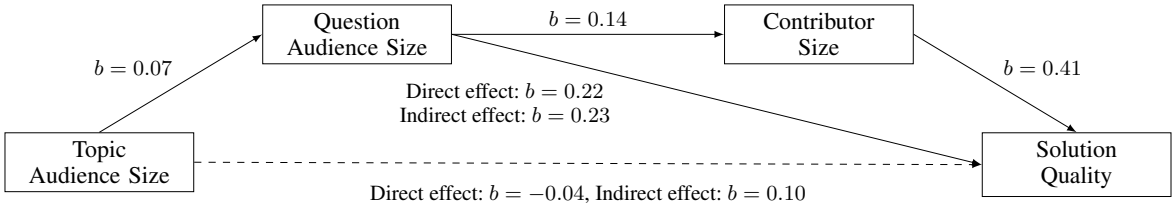

Figure 1: Results of mediation analysis showing the relationship between the three levels of group size–topic audience size, question audience size, and contributor size– and whether a question received a good solution (Solution Quality). Indirect effects are computed as $c - c'$ which are more robust for non-traditional mediation approaches. Dotted lines reflect negative direct effects.

		Step 1 $X \rightarrow Y$	***Step 2*** $X \rightarrow M$	***Step 3 & 4*** $X + M \rightarrow Y$
Good Solution (Y)	Topic Audience (X)	0.06** (0.02)	0.07*** (0.005)	-0.04* (0.02)
	Question Audience (M)			0.46*** (0.08)
	Question Audience (X)	0.45*** (0.08)	0.14*** (0.01)	0.22*** (0.09)
	Contributors (M)			0.41*** (0.11)
Solution Score (Y)	Topic Audience (X)	0.01† (0.004)	0.07*** (0.005)	-0.02*** (0.004)
	Question Audience (M)			0.31*** (0.01)
	Question Audience (X)	0.29*** (0.01)	0.46*** (0.01)	0.17*** (0.01)
	Contributors (M)			0.27*** (0.02)

Table 1: Mediation analysis results testing whether the effect of topic audience size on solution quality is mediated by question audience size and whether the effect of question audience size on solution quality is mediated by contributor size. † $p < 0.10$ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

topic audience size for a question as the total number of people who had provided a good answer (see definition below) to a question with one of the target question's tags in the last three months. We recorded the number of page views a question received in the first 24 hours as a measure of **Question Audience Size**. Unfortunately SE does not make unique page views available, but total page views approximate unique views and represent audience attention given to a question. Finally, we recorded the number of users (other than the question asker) who contributed a comment or answer in the first 24 hours as a measure of **Contributor Size**. All measures of group size were log normalized to control for skew. All three measures varied in magnitude based on the question (Topic Audience Size: 0-29,398, Question Audience Size: 1-1,341, Contributor Size: 0-7).

Solution Quality We measured solution quality in two ways. A question was considered to have received a **Good Solution** if either the question asker had accepted an answer as the best or the question had received at least one answer scoring at least one. Previous studies have tended to use accepted answer as a measure of whether a question had a good solution. We found a substantial proportion of questions that received high scoring answers that were not accepted by the question asker. Question askers sometimes received good answers but did not accept them probably because they did not know how to, they never returned to their question, or did not want to choose between multiple good answers. Our measure accounts for these biases by considering a solution to be good if it is rated as good by either the question asker or the community. Questions sometimes receive multiple solutions, each of which provides additional value. The second measure of solution quality we used was **Solution Score**: the sum total of scores for all answers to a question[3]. Anderson and colleagues (2012) found that this measure of solution quality best predicted the lasting value of a set of solutions.

One concern is that our two measures of solution quality might be confounded with audience size. We addressed this potential bias in three ways to ensure it was minimal. First, we checked for robustness by using a measure of solution quality unbiased by audience size–whether the question asker accepted an answer, and found the same patterns of results[4]. Second, although solutions and group size were only recorded when they were within the study's activity window of 24 hours, votes were tallied for 1 month to allow up and down votes to accrue, more accurately measuring a solution's true value to the community and the question asker, unbiased by early audience size. Third, we report analyses for two measures of solution quality, the first of which uses a very low threshold for the number of votes, substantially

[3]Answers with a negative score were treated as having a score of zero. The distribution of solution scores was skewed; thus negative binomial regression was used when possible and solution score was log normalized when linear regression was required for traditional mediation analysis.

[4]We don't report this measure because it undercounts solved questions since some question askers never accept an answer regardless of whether they have received a good answer

reducing dependence on audience size. We discuss this confound in more detail in the limitations section.

Control Variables We measured several potential confounds and included them as control variables in the models, including variables associated with site traffic (community site, creation hour, creation day of week), question readability (title and body length, title and body automated readability index), question quality (question score, question asker's reputation score), and topic specification (number of tags, whether tags changed).

Results

Research Question 1: How does group size in the crowd affect solution quality?

We examined the relationship between the three levels of group size and solution quality. As expected, all three levels of group size were correlated with each other (ρ: 0.12-0.55). As single predictors in separate logistic regression models all three levels of group size were significant positive predictors of whether a question received a good solution.

We predicted a complex relationship between levels of group size and solution quality, in which some levels of group size might be partially or fully mediated by other levels of group size. For example, we tested whether the effect of question audience size on solution quality was mediated by contributor size. Mediation helps us to understand the interrelationships between levels of group size. If contributor size fully mediates the relationship between question audience size and solution quality, this would suggest that audience size only improves performance by increasing the number of contributors. In contrast, if audience size is not mediated or only partially mediated by contributor size it suggests that there are other ways in which audience size improves performance (we argue through self-selection of better contributors). With three levels of audience size we tested mediation between levels. Specifically, we tested the effect of topic audience size on performance mediated by question audience size, followed by the effect of question audience size on performance mediated by contributor size (Figure 1). We used traditional mediation analysis (Baron and Kenny 1986) to test for mediation when solution quality was operationalized as a continuous outcome variable using solution ratings (Solution Score) and an extension of mediation analysis appropriate for mixture of continuous and dichotomous variables when solution quality was measured as a dichotomous outcome variable (Good Solution). The latter approach necessities using a mixture of logistic and linear regression models and standardizing coefficients so they are comparable across these two different types of models (MacKinnon and Dwyer 1993).

The effect of topic audience size on solution quality was almost entirely mediated by question audience size as demonstrated by results of four mediation analysis steps (Baron and Kenny 1986) (Table 1). Questions that were on a topic with a larger audience had significantly higher quality solutions (Step 1; marginally for solution score). Questions that were on a topic with a larger audience attracted significantly more views (question audience size) (Step 2).

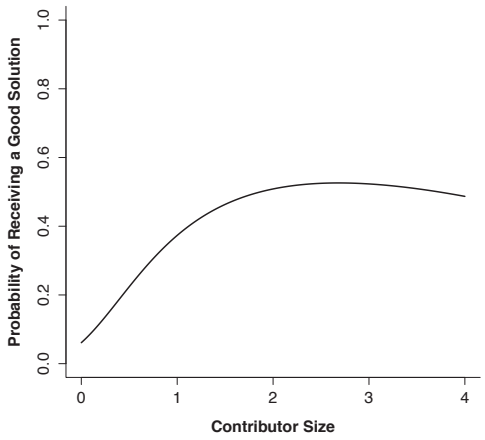

Figure 2: Logistic regression model showing the relationship between the number of contributors to a question (Contributor Size) and the probability of receiving a good solution.

When topic audience size and question audience size were entered into a model, question audience size was a significant positive predictor of solution quality, while topic audience size was no longer a positive predictor (Step 3 & 4). These results show that question audience size mediated the relationship between topic audience size and solution quality as measured by good solution or solution score albeit though inconsistent mediation (topic audience size switched signs). For solution quality as measured by good solution, topic audience size had predominately an indirect and not a direct effect on solution quality (Figure 1). In other words, the results show topic audience size increased performance by increasing the number of users who viewed the question (Quantity Pathway) and not by increasing the fit of the users who viewed the question (Quality Pathway).

The effect of question audience size on solution quality was partially mediated by contributor size (Table 1). Questions with more views (question audience size) had significantly higher quality solutions (Step 1). Questions with more views had more contributors (Step 2). When both question audience size and contributor size were entered into a model together, both remained significant positive predictors of solution quality, however the effect of question audience size on solution quality was reduced (Step 3 & 4). The effect of question audience size on solution quality was reduced by half by adding contributor size into a model, which suggests contributor size partially but not fully mediates the effect of question audience size on solution quality. These results suggest that question audience size increases performance by increasing the number of contributors to a question (Quantity Pathway) as well as by other means, which we suggest is due to better fit between contributors and questions (Quality Pathway).

Final models were developed using all three levels of group size and considering potential quadratic terms when

these models outperformed others.[5] Models 1 and 3 present the final models predicting whether a questions receives a good solution and the solution score, respectively (Tables 2 & 3). We found that contributor size had the strongest relationship with solution quality, followed by question audience size and topic audience size, which had the weakest effect. The best models for both measures of solution quality had a significant quadratic term for contributor size[6]. Figure 2 shows that the likelihood of getting a good solution increases as the number of contributors increases, however the benefit of adding contributors plateaus at two contributors. We observe the same pattern of results for solution score, as the number of contributors increases the solution score increases, however the benefit of adding contributors plateaus at five contributors.

Research Question 2: Does the relationship between group size and solution quality depend on the type of problem?

We predicted that the effect of group size would be different depending on the type of question. We tested models with interaction terms between each level of group size and question type. We found a significant interaction between question type and question audience size for both measures of solution quality[7] (Tables 2 & 3 Models 2, 4) and no significant interaction for the other two measures of group size[8]. Three observations can be made about the significant interaction between question type and question audience size. First, error problems are not as sensitive to audience size as how-to problems (Observation 1, Figure 3). Second, error problems are more likely to be solved than how-to problems when they receive only a few views and are much less likely to be solved than how-to problems when they receive many views (Observation 2). Third, as question audience size increases, solution score increases at a much faster rate for how-to problems than conceptual or error problems (Observation 3). We examined some individual questions in more detail to try to understand these observations.

Error problems ask the community to help debug code, so good answers to these problems are all-or-nothing, either the community finds a solution that fixes the bug or they do not. On the one other hand, this means that solvers can often get lucky and fix the bug with little effort. For example, *user37760* was getting negative numbers when he shouldn't have; *miindlek* was able to quickly identify that the negative numbers resulted from scaling binary features and provide instructions to *user37760* on how to scale some but not all features (User37760 2016). This was a very easy bug that could be identified quickly and had an easy solution. On the other hand the all-or-nothing nature of error problems means that difficult bugs may absorb substantial work without finding a solution. For example, even after many suggested explanations from the community to address the question *"Why is only one of spark jobs running using only one executor"* the community could not pinpoint the correct explanation for that user's situation (KikiRiki 2016). Error problems in particular may be more difficult than other types of questions to solve. The all or nothing nature of error problems may explain Observations 1 and 2. Because some error problems are easy, roughly 25% of error problems can be solved regardless of how big an audience they attract, but because most error problems are very difficult and require a correct solution to a specific situation, even when they attract a large audience many may not be solved. How-to problems ask the community how to perform a specific task, and so require finding a contributor with familiarity and expertise

Predictor	**Model 1**	**Model 2**
Intercept	-1.33	-1.67
Contributors Size	2.15***	1.46**
Contributors Size^2	-1.90***	-3.91***
Ques. Audience Size	1.76***	0.52
Topic Audience Size	0.07**	0.18*
Ques. Type: concept (vs. error)		-0.17
Ques. Type: how-to (vs. error)		-0.61***
Ques. Audience Size X Type: concept		1.02
Ques. Audience Size X Type: how-to		2.47**
McFadden R^2	0.20	0.24

Table 2: Logistic regression models showing the relationship between group size and whether a question received a good solution (Model 1: full data set, Model 2: interaction with question type). * p $<$ 0.05, ** p $<$ 0.01, *** p $<$ 0.001

Predictor	**Model 3**	**Model 4**
Intercept	-1.44	-2.03
Contributors Size	1.46***	1.45***
Contributors Size^2	-0.63***	-1.07**
Ques. Audience Size	1.66***	2.92***
Topic Audience Size	0.16*	0.13
Ques. Type: concept (vs. how-to)		0.64
Ques. Type: error (vs. how-to)		0.62
Ques. Audience Size X Type: concept		-1.61**
Ques. Audience Size X Type: error		-1.84**
McFadden R^2	0.15	0.18

Table 3: Negative binomial regression models showing the relationship between group size and solution score (Model 3: full data set, Model 4: interaction with question type). * p $<$ 0.05, ** p $<$ 0.01, *** p $<$ 0.001

[5]Results are presented using a hierarchical regression approach. Measures of group size were entered into the model in order from most upstream to most downstream–topic audience size followed by question audience size followed by contributor size. Thus indirect effects of group size are reported without being eclipsed by mediators. See mediation results for a full model showing the relationship between measures of group size.

[6]We looked for quadratic effects for all three levels of group size; the other quadratic terms were non-significant and thus omitted from the final models.

[7]We looked for interactions with question type for all three levels of group size; the other interaction terms were non-significant and thus omitted from the final model.

[8]The interaction between question type and question audience size remained consistent even considering confounding variables such as how well written the question was.

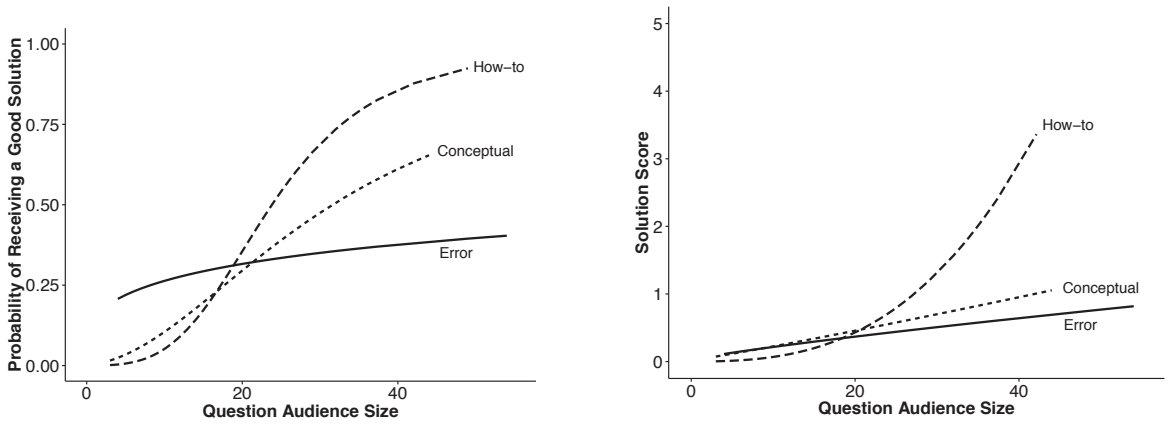

Figure 3: Models showing the interaction between question type and audience size on the probability of receiving a good solution (left) and mean solution score (right).

related to the task. Once such an expert is found, the problem can almost always be answered. Even in the rare case that a task cannot be performed, explaining what can and what cannot be done is a good answer in itself. For example, *Ninja* asked *"How to get the last value of a result efficiently"* using a hive query without iterating through the results. In contrast to the error problem mentioned above, where contributions were unhelpful because they did not solve the bug, here the how-to contributions provided a next best alternative for the question asker's goal. Large audiences are more likely to contain users who have the right expertise needed to solve how-to problems. This means that question self-selection can aid how-to problems more than error problems. It is easy for a user to quickly judge if they have the relevant expertise needed to solve a how-to problem, but tricky for them to know if they can track down the cause of a bug.

Large audiences are also more likely to contain a diversity of experts. How-to problems, in particular, benefit from these diverse viewpoints. Diverse answers to how-to problems sometimes provide alternative ways to perform a task, giving the question asker several options. They also sometimes provide multiple perspectives on a single approach, making it easier for the question asker to apply it. For example, *fmalussena* asked *"How do I compare the performance of random forests for regression?"*. *hxd1011* provided an abbreviated answer that addressed the question, and which was later supplemented by *EngrStudent*, who provided another perspective on the same answer including additional examples, code, and guidance (Fmalaussena 2016). This may explain Observation 3: as question audience size increases solution score rapidly increases for how-to problems. Large audiences provide a more diverse set of answers which is particularly valuable for how-to problems.

Conceptual problems fall somewhere between how-to and error problems. Conceptual problems are rarely solved with small audiences, but are often solved by large audiences. Conceptual problems are trickier than how-to problems because they are open ended.

Discussion

Question answering sites have become an important platform for drawing up-to-date knowledge from a large crowd of experts. Experts can solve user problems, troubleshoot, provide advice, and augment official documentation. Getting good answers to a question depends on attracting a large enough set of experts. We investigated how many users in various capacities were needed to provide good solutions.

We developed a more complex model of group size in the crowd that helped to align and address gaps in prior literature. Prior work on the effect of group size on performance in the crowd had focused either on the number of contributors as a measure of group size (e.g. Kittur and Kraut 2008) or on audience size as a measure of group size (e.g. Ortega et al. 2014). Each type of research had found that group size increased performance, but only some of the time. We proposed and tested a model of group size in the crowd that considered the effect of group size at multiple levels. This model provides a framework that aligns previous literature by including both types of measures of group size simultaneously and extends this work by better estimating the true effect of group size on performance. By only focusing on a single level of group size, previous work potentially underestimated the effect of group size on performance. Stratified group size is more likely to capture the effect of group size on performance because it considers multiple pathways for group size to increase performance.

The model we developed also considered context dependent effects of group size on performance. Previous work had found that attracting a larger audience only sometimes increased performance in Q&A communities (Ortega et al. 2014). By drawing inspiration from the small group literature and Steiner (1966) we showed that the effect of audience on solution quality depends on the type of problem. Some types of questions like how-to benefit from larger audiences and more diverse viewpoints, whereas other types of questions like error problems are more likely to be either easy to solve or intractable and do not benefit much from

larger audiences. Context dependent models are needed to explain the inconsistent relationship between group size and performance in the crowd.

A multi-level model of group size also helps to identify *why* increasing group size increases performance. The explanation for the effect of contributor size on performance is straightforward: without contributors there is no solution and additional contributors provide more information, fix mistakes, and improve answers (Tausczik, Kittur, and Kraut 2014). A larger number of contributors is associated with increased solution quality, but only with rapidly diminishing returns. Solution quality quickly plateaus as contributor size reached between 2 and 5 contributors. Process losses are not likely to explain the rapidly diminishing returns because Q&A platforms encourage brief and efficient communication and coordination. Instead, the diminishing returns may be explained by the idea of low-hanging fruit. Adding a few more contributors may improve the final solution, but most of the important points are already addressed by the first few contributors.

The explanation for the effect of audience size on performance is multifaceted. We argue that there are two pathways through which improves audience size increases performance. Questions that attract a larger audience are likely to attract more contributors, and as we argue above, increasing the number of contributors is likely to increase solution quality (Quantity Pathway). Questions that attract a larger audience are also more likely to attract contributors with more appropriate expertise because they can draw users from a larger crowd (Jeppesen and Lakhani 2010), thus questions with a large audience are likely to have better qualified set of contributors (Quality Pathway). We found equal evidence for audience size affecting performance through both pathways on Stack Exchange Q&As.

We expect both audience and contributor size to be potentially important for any crowd work in which task self-selection is important, which includes Q&A communities as well as many other forms of crowd work, including design contests, crowdsourcing competitions, prediction markets, and wikis. However, the degree to which audience size, contributor size, or both affect performance will depend on the dynamics of the platform, including communication tools, coordination, tasks, and user expertise. A stratified group size model can be useful to identify the pathways by which group size increases performance and to quantify how much each of these pathways contributes to performance. This, in turn, helps to explain the mechanism(s) by which group size improves performance.

Design Implications

This study suggests some immediate changes that could be made to Q&A systems like Stack Exchange (SE) sites to efficiently allocate the attention of the large number of users to maximize performance as well as some deeper design implications for structuring group work in crowds.

Audience attention is a limited resource. From a system design perspective, our objective is to optimally route audience attention to the problems that need it the most. Our results suggest question type is an important consideration for such allocation. How-to problems with large audiences were almost always solved, whereas how-to problems with small audiences were rarely solved. Error problems were solved at about the same rate regardless of audience size. One implication is that system designers may want to route audience attention toward some types of questions. In particular, they may want to route more attention toward (unsolved) how-to problems and less toward error problems. If question type were built into the platform, SE-like sites could be designed to treat questions differently based on this information. Questions could be automatically categorized (e.g. how-to vs. conceptual) by training machine learning models and/or directly requesting this categorization from during question submission. In this study, we focused on large categories of common types of questions, future work could investigate how group size affects a broader and more nuanced set of question types.

Our findings also have practical implications for the design of group work platforms in general. In traditional groups a small number of individuals are assigned to a problem. Our results argue that one advantage of crowd work is that contributors self-select questions, ensuring that they are well suited to the problem. Because some types of problems benefit from this self-selection process, and many organizations make use of traditional groups, there may be some benefit in allowing these traditional groups to assemble on the fly on the basis of self-selection.

Limitations

The reliance on SE data and the observational design of this study created several sources of potential bias and confounds; these are limitations of our approach and design. First, we used SE data to measure audience size and solution quality, which had some disadvantages. We operationalized question audience size as the number of page views received in the first 24 hours, but would have preferred to use unique page views. Page views approximate unique page views, but introduce error because some individuals may visit the page more than once (e.g. question asker checking for updates). We used two measures of solution quality that depended on user votes, which can be biased by audience size because questions that attract a larger audience have the opportunity to receive more votes. To minimize bias we took steps to reduce potential sources of bias and checked robustness across a variety of measures with varying types of error. For example, to reduce the impact of repeated page views, we limited the activity window to the first 24 hours, in which fewer repeat visits are likely to happen. To increase the validity of votes as a measure of quality and reduce its dependence on audience size, we measured votes over a much longer time period (1 month). We also measured solution quality in multiple ways, including one metric that did not rely on user votes (acceptance by question asker), and found the same pattern of results regardless of which measure was used. As a result we believe these sources of bias are minimal.

Second, we presume a directionality of effects based on mechanisms proposed here and in other research. However, because the study design is observational we cannot demonstrate causality. In spite of these limitations, tracking group

size and performance in over 3000 naturalistic groups would not have been practical without the use of observational data.

Conclusion

Despite widespread adoption of Q&As by technical communities and an abundance of domain experts, many questions do not attract a large audience. We investigated the effect of group size on solution quality in Q&As. We found that multiple levels of group size in the crowd–topic audience size, question audience size, and contributor size–affected solution quality. We argue that group size in the crowd is a multi-level construct and not unitary. In crowd work with task self-selection audience size can affect performance by increasing the number and quality of contributors. This work advances our theoretical model of group size in the crowd and its relation to performance.

Acknowledgments

This work is supported by NSF (CISE IIS-1546404).

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*, 850–858. ACM Press.

André, P.; Kraut, R. E.; and Kittur, A. 2014. Effects of Simultaneous and Sequential Work Structures on Distributed Collaborative Interdependent Tasks. In *Proceedings of the Conference on Human Factors in Computing Systems*, 139–148. ACM Press.

Asaduzzaman, M.; Mashiyat, A. S.; Roy, C. K.; and Schneider, K. 2013. Answering Questions about Unanswered Questions of Stack Overflow. In *Proceedings of the Conference on Mining Software Repositories*, 97–100. IEEE Press.

Baron, R., and Kenny, D. 1986. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51:1173–1182.

Fmalaussena. 2016. How do I compare the performance of random forests for regression? http://stats.stackexchange.com/questions/218299. Accessed: 2016-09-13.

Gibb, J. R. 1951. Effects of group size and threat reduction on creativity in a problem-solving situation. *American Psychologist* 6:324.

Hill, G. W. 1982. Group Versus Individual Performance: Are N + 1 Heads Better Than One? *Psychological Bulletin* 91:517–539.

Jeppesen, L. B., and Lakhani, K. R. 2010. Marginality and problem solving effectiveness in broadcast search. *Organization Science* 21:1016–1033.

Kerr, N. L., and Tindale, R. S. 2004. Group performance and decision making. *Annual Review of Psychology* 55:623–655.

KikiRiki. 2016. Why is only one of spark jobs running using only one executor? http://stackoverflow.com/questions/37834560. Accessed: 2016-09-13.

Kittur, A., and Kraut, R. E. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the Conference on Computer Supported Cooperative Work*, 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.

Lorge, I., and Solomon, H. 1959. Individual performance and group performance in problem solving related to group size and previous exposure to the problem. *Journal of Psychology* 48:107–114.

MacKinnon, D. P., and Dwyer, J. H. 1993. Estimating mediated effects in prevention studies. *Evaluation Review* 17:144–158.

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*, 2857–2866. ACM Press.

Ortega, F.; Convertino, G.; Zancanaro, M.; and Piccardi, T. 2014. Assessing the Performance of Question-and-Answer Communities Using Survival Analysis. arXiv preprint.

Parnin, C.; Treude, C.; Grammel, L.; and Storey, M.-A. 2012. Crowd documentation: Exploring the coverage and the dynamics of API discussions on Stack Overflow. Georgia Tech Technical Report.

Robert, L., and Romero, D. M. 2015. Crowd Size, Diversity and Performance. In *Proceedings of the Conference on Human Factors in Computing Systems*, 1379–1382. ACM Press.

Steiner, I. D. 1966. Models for inferring relationships between group size and potential group productivity. *Behavioral Science* 11:273–283.

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 Computer-Supported Cooperative Work*, 355–367. ACM Press.

Taylor, D. W., and Faust, W. L. 1952. Twenty questions: Efficiency in problem solving as a function of size of group. *Journal of Experimental Psychology* 44:360–368.

Treude, C.; Barzilay, O.; and Storey, M.-A. 2011. How Do Programmers Ask and Answer Questions on the Web? In *Proceedings of the International Conference on Software Engineering*, 804–807. ACM Press.

User37760. 2016. Avoid scaling binary columns in sci-kit learn StandsardScaler. http://stackoverflow.com/questions/37685412. Accessed: 2016-09-13.

Vasilescu, B.; Serebrenik, A.; Devanbu, P.; and Filkov, V. 2014. How Social Q&A Sites are Changing Knowledge Sharing in Open Source Software Communities. In *Proceedings of the Conference on Computer Supported Cooperative Work*, 342–354. ACM Press.

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*, 441–451. ACM Press.

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*, 1445–1455. ACM Press.