Crowd Size, Diversity and Performance

Lionel Robert *
University of Michigan
lprobert@umich.edu

Daniel M. Romero *
University of Michigan
drom@umich.edu

ABSTRACT
Crowds are increasingly being adopted to solve complex problems. Size and diversity are two key characteristics of crowds; however their relationship to performance is often paradoxical. To better understand the effects of crowd size and diversity on crowd performance we conducted a study on the quality of 4,317 articles in the WikiProject Film community. The results of our study suggest that crowd size leads to better performance when crowds are more diverse. However, there is a break-even point – smaller, less diverse crowds can outperform more diverse crowds of similar size. Our results offer new insights into the effects of size and diversity on the performance of crowds.

Author Keywords
Wikipedia; performance; team size; diversity

ACM Classification Keywords
H.5.2 Information Systems: Information Interfaces and Presentation (e.g., HCI)—*User Interfaces: Theory and methods*; K.4.3 Computing Milieux: Computer and Society—*Organizational Impacts: Computer-supported collaborative work*

INTRODUCTION
The "wisdom of crowds'" argues that the "many are smarter than the few" [2, 8]. This concept, along with new forms of organizing, has ushered in the era of "The Crowds." Crowds are being adopted to perform work formally done by individuals. For example, crowds are being used to produce software, generate new ideas for organizations and aid in solving complex problems [2, 8]. Yet, there is still much to learn about what makes crowds more or less effective [2].

Size and diversity are two key characteristics of crowds. Yet, in the literature, their relationship to performance is often paradoxical [4]. They have been linked to increases and decreases in group performance. To better understand their effects we conducted a study on 4,317 articles in the WikiProject Film community on crowds. Our results indicate that the impact of size and diversity on crowd performance are co-dependent. That is, crowd size led to better performance when the crowds were more diverse. However, in some cases smaller, less diverse crowds outperformed similar size more diverse crowds. Results of our study offer new insights into the factors that lead to better crowd performance.

*Equal contribution

CHI 2015, April 18 – 23 2015, Seoul, Republic of Korea.

http://dx.doi.org/10.1145/2702123.2702469

Crowd Size and Crowd Diversity
Theories on the impact of group size offer conflicting views regarding the relationship between size and performance. One view states that increases in size can increase performance [4]. As size increases, more individuals are available to perform the group's task. The other view argues that increases in size can decrease performance [1]. One reason is the difficulties coordinating larger groups [8]. The group diversity literature is just as perplexing. Diversity is often labeled as the double-edge sword [20]. On one hand, diversity should lead to better performance because more diverse members offer a broader range of knowledge and skills [6, 15, 16]. On the other hand, some hypothesize that as diversity increases performance decreases because diverse members often find it difficult to work together [10, 14, 17]. This difficulty in working with diverse others often results in conflict and other types of process losses [9]. Taken together, both the literature on group diversity and group size offer conflicting views on what should lead to better crowd performance.

However, Oliver and Marwell put forth a mathematical model that posited the relationship between size and diversity in the context of collective action regarding the critical mass needed for public goods. Based on their calculations diverse groups can achieve the same level of collective action with fewer individuals than homogeneous groups [13]. Building on a similar assertion, we argue that to fully understand the effects of size and diversity both must be taken into account jointly. In other words, the effects of crowd size should depend on crowd diversity and vice versa. Specifically, we argue that size will be associated with better performance in more diverse crowds but not so in less diverse crowds. We present our arguments related to "why" in the following section.

The Effects of Crowd Size and Diversity on Performance
Increases in size should lead to better performance when crowds are more diverse for two reasons. First, as stated earlier, when diversity increases, crowds have access to a broader range of knowledge and skills [2, 14, 20]. Second, as size increases, the negative effects of diversity are likely to be diluted. The negative effects of diversity materialize when different individuals are forced to work together [9, 14, 20]. However, in larger crowds, individuals have more opportunity to find similar others to work with in the crowd. Taken

together, both claims explain why crowds are likely to benefit rather than suffer from their diversity as size increases. Therefore we propose the following:

Hypothesis: *Crowd diversity should moderate the relationship between crowd size and performance. Increases in size should be positively related to performance when crowds are diverse.*

DATA

We collected data on 4,378 articles from Wikipedia's WikiProject Film community. Article quality was our performance measure and was assessed by the Wikipedia community. In descending order of quality the article classes were *FA, GA, B, C, Start,* and *Stub*. Articles in the class *Start* and *Stub* had little information and very few editors. They were dropped from the analysis because we wanted to study active crowds. For each article, we obtained a complete list of the editors and their edits. Nearly 350,000 editors contributed at least one edit to an article.

METHOD

We considered three types of diversity based on the interest and the amount of work that editors contributed to Wikipedia. Many other factors could be related to article quality such as an editors interest and ability. We did not include these factors because our goal was to examine the relationship between crowd size and diversity and performance.

Article Quality

Each articles class was used as the performance outcome [1]. This measure has been used as a proxy for performance in other Wikipedia studies [7, 8, 11, 18, 19] and there is evidence that article class is related to the quality assessments of outside reviewers [7]. We assigned each article a digit from 1 to 4 corresponding to the C, B, GA, and FA classes, respectively. There were 123, 629, 476, and 3,089 articles in the FA, GA, B, and C classes, respectively.

Topical Diversity

We measured the similarity in topical interests of two editors on Wikipedia by the similarity of the articles they edited across Wikipedia. Given editor u, we let A_u be the set of articles u has edited on Wikipedia. For every pair of editors (u_1, u_2), we measured their Jaccard similarity as $J_{u_1,u_2} = \frac{|A_{u_1} \cap A_{u_1}|}{|A_{u_1} \cup A_{u_1}|}$. J_{u_1,u_2}. This measure indicates the overlap between the articles edited by u_1 and u_2, while controlling for the total number of articles that the pair edited. For each article a, we let P_a be the set of all pairs of editors of article a . We defineed the *topical diversity* of an article a, TD_a, as one minus the average Jaccard similarity of all pairs of editors. That is, $TD_a = 1 - \frac{1}{|P_a|} \sum_{(u_1,u_2) \in P_a} J_{u_1,u_2}$. The average topical diversity is 0.994 with standard deviation of 0.009. When TD_a is high, the editors of a tend to have low overlap in the set of articles they edit, making them a more topically diverse crowd.

[1]More details about the assessment process and quality classes can be found here: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Film/Assessment

Inner Workload Diversity

The workload among editors was not always split the same way. Some articles were characterized by having a small set of editors who made most of the edits, while other articles had editors who split the work more evenly. To measure the extent to which the editors of an article have a diverse or uniform workload, we used the Gini coefficient of their edits. We let E_a be the set of editors of article a and $W_a(e)$ be the number of times editor e contributed to article a. We defined the *inner workload diversity* of article a as the Gini coefficient of the set $\{\cup_{e \in E_a} W_a(e)\}$. An article where all editors contributed a similar number of edits had a low inner workload diversity, while an article where a few editors produced significantly more edits than the rest had a high inner workload diversity. The average inner workload diversity was 0.528 with standard deviation of 0.144.

This measure was proposed as a measure of implicit coordination [8]. We argue that it also serves as a proxy for a type of diversity. An article with high inner workload diversity has different types of editors, some who contribute very little and some who contribute a lot. An article with low inner workload diversity has only one type of editor because they all produce roughly the same amount of work.

Outer Workload

We measured the level of engagement of editors across Wikipedia. For each article a, we defined the *outer workload* of its editors as the mean number of edits they contributed to Wikipedia articles, other than a. The average outer workload was 2809 edits with standard deviation of 995.

Outer Workload Diversity

We also measured the diversity of outer workload of the editors of an article similarly to how we measured inner workload diversity. Outer Workload Diversity measures whether the article has a combination of editors who are heavy contributors to Wikipedia in general as well as editors who focus mostly on a single or very few articles, or whether most editors have about equal outer engagement. We defined the *outer workload diversity* of an article as the Gini coefficient of the number of contributions to other Wikipedia articles by each editor. The average outer workload diversity was 0.649 with standard deviation of 0.084.

Crowd Size

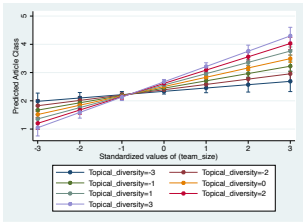
We measured crowd size by the log of the number of editors for each article. The average log of number of editors was 4.66 with a standard deviation of 1.22. The actual average number of editors per article was 198, with as many as 2,618.

RESULTS

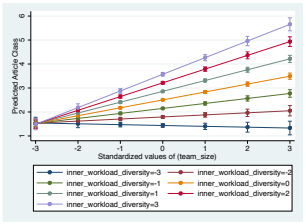
The analysis was conducted using linear regression. We standardized all the independent variables. Topical diversity, inner workload diversity, crowd size, and outer workload were positively related to quality, and outer workload diversity was negatively related to quality. The R^2 was 0.22. Then we included the interactions between all the independent variables and crowd size. All interactions were significant. Table 1

Variable	Main Effects		Crowd Size Interactions	
	Coeff.	**SE**	**Coeff.**	**SE**
Topical diversity	0.041^{**}	0.013	0.056^{***}	0.014
Outer workload	0.180^{***}	0.016	0.265^{***}	0.024
Outer workload diversity	-0.098^{***}	0.011	-0.060^{*}	0.028
Inner workload diversity	0.320^{***}	0.012	0.356^{***}	0.012
Crowd size	0.322^{***}	0.019	0.330^{***}	0.021
Topical diversity X Crowd size	—	—	0.071^{***}	0.016
Outer workload X Crowd size	—	—	0.060^{***}	0.015
Outer workload diversity X Crowd size	—	—	0.046^{***}	0.014
Inner workload diversity X Crowd size	—	—	0.122^{***}	0.012

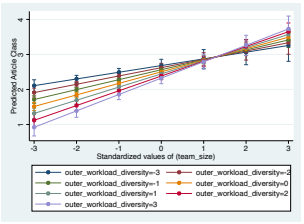
Table 1. Results of regression analysis to predict quality from diversity measures, outer workload, and crowd size. Model contains a total of $4,317$ observations. Significance key: *:p-val < 0.05, **:p-val < 0.01, ***:p-val < 0.001



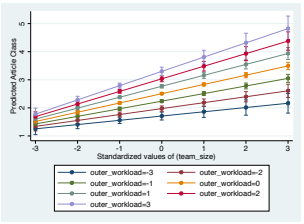
(a) Effect of topical diversity on quality by number of editors



(b) Effect of inner workload diversity on quality by number of editors



(c) Effect of outer workload diversity on quality by number of editors



(d) Effect of outer workload on quality by number of editors

Figure 1. Interaction effects of each measure and crowd size on the quality.

shows our final model, which had an R^2 of 0.25. The increase in variance explained was significant at the 0.001 level.

These results support our hypothesis that crowd size is positively related to crowd performance in diverse crowds. Crowd size had a greater and more positive effect on performance when diversity was high. This held true for all three measures of diversity. We also found that in some cases smaller, less diverse crowds performed better than similar-size, less diverse crowds. This highlights that diversity does not always have a positive effect on performance, as previously found in the literature [2, 6, 10, 15, 17], and that crowd size may be one of the factors that determine when diversity is beneficial.

Finally, the effect of outer workload is also greater for large teams than for small teams, which suggests that larger teams benefit more from having editors who are more engaged in editing Wikipedia articles[2].

DISCUSSION

The objective of this study was to examine the relationship between crowd size and diversity and crowd performance. Results indicate that diversity is vital to understanding when crowd size leads to better performance. Crowd size is positively related to performance in highly diverse crowds but not so in less diverse crowds. The results of this study have several implications for theory and design.

Theoretical Implications

[2]Since more than 70% of the articles in our data set are in quality class C, we repeated the regression analyses only using quality classes B, GA, and FA. We observed that the trends are consistent in both sets.

First, Figures 1(a), 1(b), and 1(c) all demonstrate that the relationship between crowd size and crowd performance is dependent on crowd diversity. Increases in size are positively related to performance when crowds are highly diverse. However, when crowds are not diverse this relationship diminishes. Based on our findings, increases in size need to be accompanied with increases in diversity. In addition, this finding also contributes to the literature on group size and diversity. Our results suggest that to better understand the effects of size and diversity on group performance future studies should take both into account jointly.

Second, our findings contribute to the work by Oliver and Marwell. Aligned with their findings, we discovered that the effects of size are dependent on diversity. However, their work focuses on public goods and explains how smaller groups can achieve the same critical mass as larger groups. Contextually, we examine large online groups, which provides an important layer of external validity.

Third, there appears to be a break-even point between size and diversity. Although the exact points differ, what is consistent across the three figures is that up to the break-even point less diverse crowds perform better than more diverse crowds. These results support our assertion that, as size increases, crowds may not be hurt by diversity because members have less chance of working with diverse others. In smaller crowds there is a greater chance of individuals having to work with diverse others. This, in turn, can lead to the negative effects of diversity that can counter the positive benefits. This explains why smaller, less diverse crowds outperform similar-size, more diverse crowds. In doing so, these results identify

that the benefits of diversity may have a tipping point, which may differ depending on the type of diversity.

Design Implications

Our results have important implications for collaborative systems that support crowd work. The use of crowds is becoming increasingly common across many domains and this trend has inspired the design of recommender systems intended to maximize the benefits of large collaborations by suggesting new partnerships [3, 12]. Currently, recommendation systems use individual attributes like expertise and experience to make suggestions [3]. Our results suggest that recommender systems should also consider the number of editors and the current diversity of the crowd along with how the new member would change both the size and diversity.

Limitations

One, we found a significant interaction between crowd size and diversity on performance. However, our results were only tested on a specific Wikipedia community. More work is needed to determine whether the results hold in other crowds. We note that our measures are general enough that they can be easily applied in other domains where users contribute to multiple projects. Hence, validating our findings on other domains is a feasible extension of the present work. Two, while our results highlight the importance of size relative to diversity, we do not attempt to quantify exactly how much diversity is necessary. In fact, the amount of diversity seems to depend on the type of diversity. Therefore, future research should be directed at determining what dictates the tipping point for a particular type of diversity. Three, the teams that we studied have two important characteristics: they are online and are very large. It is unclear whether the present results have implications for offline and/or smaller teams. Finally, we standardized all the continuous variables, which is normally recommended [16]. We acknowledge that there is debate about the effectiveness associated with such action.

ACKNOWLEDGMENTS

We thank Tawanna Dillahunt, Jon Kleinberg, and Paul Resnick for useful discussions.

REFERENCES

1. Alnuaimi, O.A., Robert, L.P., and Maruping, L.M. (2010) Team size, dispersion, and social loafing in technology-supported teams: A perspective on the theory of moral disengagement. *J. of Management Information Systems*, 27(1):203–230.
2. Chen, J., Ren, Y., and Riedl, J. (2010) The Effects of Diversity on Group Productivity and Member Withdrawal in Online Volunteer Groups. Proc. CHI.
3. Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. (2007) SuggestBot: using intelligent task routing to help people find work in wikipedia. Proc IUI.
4. Cummings, J.N., Kiesler, S., Zadeh, R.B., and Balakrishnan, A.D. (2013) Group Heterogeneity Increases the Risks of Large Group Size: A Longitudinal Study of Productivity in Research Groups. *Psychological Science*, 24(6):880–890.
5. Earley, P.C. and Mosakowski, E. (2000) Creating hybrid team cultures: An empirical test of transnational team functioning. *Academy of Management Journal*, 43:26–49.
6. Jehn, K.A, Northcraft, G.B., Neale, M.A. (1999) Why differences make a difference: A field study of diversity, conflict, and performance in workgroups. *Administrative Science Quarterly* 4:741.
7. Kittur, A., Chi, E., and Suh, B. (2008) Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements. Proc. CHI.
8. Kittur, A. and Kraut, R.E. (2008) Harnessing the wisdom of crowds in Wikipedia: quality through coordination. Proc. CSCW.
9. Van Knippenberg, D. and Schippers, M.C. (2007) Work group diversity. *Annual Review of Psychology*, 58:515–541.
10. Larson, J.R., Christensen, C., Abbott, A.S., Franz, T.M. (1996) Diagnosing groups: charting the flow of information in medical decision-making teams. *J. of Personality and Social Psychology* 2:315–330.
11. Liu, J. and Ram, S. (2011) Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Sys*. 2(2):1–23.
12. McDonald, D.W. (2003) Recommending collaboration with social networks: a comparative evaluation. Proc. CHI.
13. Oliver, P. E., and Marwell, G. (1988). The Paradox of Group Size in Collective Action: A Theory of the Critical Mass. II. *American Sociological Review*. 1–8.
14. O'Reilly, C.A., Caldwell, D.F., and Barnett, W.P. (1989) Work group demography, social integration, and turnover. *Administrative Science Quarterly*. 34:21–37.
15. Page, S.E. (2007). The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. *Princeton University Press*.
16. Robert, L.P. (2013) A multi-level analysis of the impact of shared leadership in diverse virtual teams. Proc CSCW.
17. Stasser, G., Taylor L.A., Hanna C. (1994) Information sampling in structured and unstructured discussions of three- and six-person groups. *J. Personality Social Psych.* 3
18. Stvilia, B., Twidale, M. B., Smith, L.C., and Gasser, L. (2005) Assessing information quality of a community-based encyclopedia. Proc. ICIQ.
19. Wilkinson, D.M., and Huberman, B.A. (2007) Assessing the value of cooperation in Wikipedia. *First Monday* 12(4).
20. Williams, K.Y. and O'Reilly, C.A. (1998) Demography and diversity in organizations: A review of 40 years of research. *Research in Organizational Behavior*, 20:77–140.