Involvement in Online Crowdsourcing Communities

Abstract

The recent rise of online communities significantly transforms the ways of social interactions and information exchange. Individuals’ voluntary community involvement contributes to growth and success of these online communities by creating economic and social values to individuals and to the community. Motivated by this phenomenon, this research investigates how community involvement impacts a member’s peer-recognition and task performance within an online crowdsourcing community. We collected secondary data from the discussion forums of a crowdsourcing platform that focuses on data analytics projects. Our results reveal several important findings. First, community involvement improves both peer-recognition and performance ranking of a member in the community. Second, as a symbol of social status, peer-recognition is found to be negatively associated with performance ranking in the competitive setting. Our findings offer strategic implications to solvers, seekers, and designers of online crowdsourcing communities, as well as other professional online communities.

Keywords
Crowdsourcing, online communities, community involvement, peer-recognition.

Introduction

Social media technologies have facilitated the rapid growth of online communities, and such communities provide a platform for knowledge and information sharing among their members. Without formal governance, the growth and success of these communities depend on their members’ voluntary community involvement. Such involvement is mainly manifested through sharing information and knowledge, and asking/answering questions via posting messages in discussion forums. For example, an emerging form of online communities, Internet-based communities of practice, has become popular and attracted a large group of professionals with a shared practice but geographically distributed to exchange knowledge (Wasko and Faraj 2005). Members of these communities are likely to learn from their peers, develop professional skills, build social connections, and/or gain peer-recognition (Wenger and Snyder 2000, Huang and Zhang 2013). Many organizations also encourage their employees to participate in communities of practice to improve their skill levels or intellectual capital. For example, IT-hosting company Rackspace has recently changed its policy to encourage its employees to contribute to any public open source project, including those that compete with Rackspace (Henderson 2014).

Prior research on online communities has primarily explored what motivates members’ voluntary involvement within a community (Ardichvili et al. 2003, Wasko and Faraj 2005, Sun et al. 2012). However, few studies have empirically quantified the benefits that may be brought for a member by his/her community involvement activities (Roberts et al. 2006, Huang and Zhang 2013). This could be mainly due to the difficulty in developing proper metrics and collecting empirical data regarding the actual impact of community involvement. Based on secondary data from a crowdsourcing platform that
provide objective measures for actual impact of community involvement, this study makes it feasible to empirically quantify the benefits of community involvement.

Using a large dataset collected from a major data analytic crowdsourcing platform, Kaggle.com, we empirically quantify the benefits of community involvement in online crowdsourcing communities. In crowdsourcing competitions, solvers compete against each other to come up with the best solution which is entitled to a reward. Additionally, they may also voluntarily involve in community forums by sharing information and knowledge with their rivals. Sharing knowledge among members in these communities, though likely to benefit the entire community including the contributors, may also result in knowledge spillover and possibly weaken contributor's competitiveness. We investigate how solvers' voluntary involvement with the community, by sharing information in forums, influences their performance ranking and peer-recognition. Our data provide objective performance evaluations, while most of the prior community studies used subjective measures of benefits from surveys or interviews.

This study makes three unique contributions. First, it is among the few that empirically quantifies the benefits of community involvement in terms of its impact on one's performance ranking and peer-recognition. Second, the community investigated in this paper is unique and differs from the other online communities explored in prior literature that are mostly collaborative in nature. In our study, members of the community are direct competitors for monetary rewards for different crowdsourcing projects. We suggest that a contributor could lose his or her competitive advantage if valuable knowledge were shared. This is reflected by the negative relationship between peer-recognition and performance in our results.

The remainder of this paper is organized as follows. First, we review the literature related to our study and develop hypotheses. Second, we describe the data and variables. Third, we outline the research model and data analyses. Fourth, we discuss the key results and implications. Finally, we conclude and suggest directions for future research.

**Literature Review and Hypothesis Development**

Voluntary community involvement can be considered as a type of pro-social behavior, which is often defined as “voluntary intentional behavior that results in benefits for another” (Sproull et al. 2005). The research on pro-social behavior suggests that contributing to, sharing knowledge with, and helping others may indeed influence a contributor's gain in both intellectual and social benefits. For example, scholars have shown that people who help others in online software support communities by answering questions enjoy learning benefits, reputational benefits, and other benefits related to advancing the group (Sproull et al. 2002).

Grant and Sonnentage (2010) found that the experience of helping others may increase employees' psychological well-being, protect them from emotional exhaustion, and help them to maintain performance when working on demanding jobs. Grant and Berry (2011) found that when driven by pro-social motives to help others, people are able to come up with ideas that are both novel and useful, resulting in higher levels of creative performance. In addition, that encourages them to invest more time and energy in task, and enhance work effectiveness and productivity (Grant 2008). Weinstein and Ryan (2010) found that the voluntary engagement in pro-social behaviors may energize people by satisfying their fundamental psychological needs of competence, relatedness, and autonomy. Grant (2008) showed that the pro-social motivation to help others, when it is based on people's free choice, increases employees' persistence, productivity, and performance on their jobs.

Pro-social behaviors may also enhance a person's reputation in the community that an individual is contributing to (Ariely et al. 2009, Bolino 1999, Griskevicius et al. 2010). Social recognition is identified as one of the main reasons for pro-social behaviors on the net. Thankful response to a helpful post is visible to everyone and thus creates social recognition (Sproull et al. 2005). Ozer (2011) demonstrated that voluntary behaviors to help other employees contribute to the establishment of trusting and high-quality relationships between the focal individual and other employees. In his recent New York Times bestseller, Grant (2013) presented ample evidence that people who give to others may become more successful by building reputation and social capital.

How would community involvement influence individual performance in the specific context of online crowdsourcing? In keeping with the pro-social literature, studies of online communities highlight many
individual benefits of community involvement including gaining reputation, improving professional status, increasing professional contacts, enhancing self-image, gaining access to expert advice, and becoming more confident in their own knowledge (Bateman et al. 2011). In our context, we identify two such benefits: performance improvement and peer-recognition.

Kaggle.com mainly deals with predictive modeling tasks. The tasks are very intellectual and the success of these projects depends on the skills and creativity of solvers. Pro-social motivation may enhance an individual’s cognitive processing and ability to generate creative ideas (Grant and Berry 2011, Forgeard and Mecklenburg 2013), and encourage them to spend more time and energy on task that eventually enhances their performance through increasing effectiveness and productivity (Grant 2008). Even though community involvement in crowdsourcing is not a required aspect of task performance, such behavior may foster an other-focused mentality that helps an individual to perform more creatively on his/her own problem-solving task. Moreover, intellectual tasks are cognitively demanding. Engaging in community involvement may help solvers to reduce stress, feel more confident, and therefore persist and perform better on their tasks (Grant and Sonnentag 2010, Mogilner et al. 2012, Xu et al. 2009). Taken together, we argue that community involvement should enhance a solver’s performance in intellectual crowdsourcing tasks. Thus, we hypothesize:

Hypothesis 1 (H1): Solvers’ community involvement is positively related to their task performance.

Peer-recognition in our scenario refers to how well a solver is acknowledged by other members for his/her contributions. Studies that have shown giving to others in the same community may enhance the giver’s reputation and facilitate the development of social network. For example, Flynn (2003) found that employees who are generous in giving information, goods, and services to other employees are perceived to have higher social status. Furthermore, pro-social literature suggests that the reputational benefit of pro-social behaviors is particularly salient when the helping behaviors are visible to the public (Bolino 1999, Hardy and Van Vugt 2006). In our scenario, solvers contribute to the community through sharing information and knowledge, and asking and answering questions by posting messages in discussion forums that are open to the public. Following the studies in pro-social literature, the high visibility of these social exchanges indicates that community involvement in this context is likely to result in an increase in a solver’s reputation and status. Thus, we hypothesize:

Hypothesis 2 (H2): Solvers’ community involvement is positively related to their peer-recognition.

Peer-recognition of community involvement may help a solver to establish a personal reputation in both technical competence and generosity. With this reputation, solvers have an opportunity to increase their resources through acquiring facets of social capital such as trust, reciprocity, and ties/social network. Peer-recognition leads to informational and knowledge benefits due to reciprocity. When a solver shares knowledge with peers, that will enhance the likelihood of his/her peers share knowledge and information with the solver. In addition, peer-recognition helps to improve trust, that in-turn leads to high quality intellectual exchanges (Chiu et al. 2006). Further, participation in crowdsourcing projects is recurring. Generally, solvers participate in crowdsourcing competitions more than once. A solver with high peer-recognition is more likely to be invited by the others to collaborate on future projects. This increases a solver’s social network, opportunity to participate with capable others, and contribute to a solver’s overall performance.

Hypothesis 3 (H3): Solvers’ peer-recognition is positively related to their performance.

Data

We collected data from a specialized crowdsourcing platform that focused on data analytics projects (Kaggle.com). Kaggle has a pool of more than 100,000 data scientists from over 100 countries and 200 universities. They are experts in various quantitative fields such as computer science, statistics, economics, mathematics, and physics. Since its launch in 2010, Kaggle has served many companies including GE, Allstate, Merck, Ford, and Facebook to run analytics competition for the best predictive models in order to improve sales forecasting, increase customer retention, reduce operating costs, accelerate product development, and gather information from social media.

Companies, government, and researchers provide datasets to Kaggle along with their problems and the amount of reward they are willing to pay to the winners. Based on these inputs, Kaggle sets up contests.
Each participant or participating team can submit multiple solutions within the contest duration. Kaggle evaluates all submissions in real time using a test dataset and provides participants instant feedback through a live score card, where participants receive information on the predictive accuracy of their models and their relative positions (ranks) in the contest.

Kaggle’s website features each solver with an online profile, which shows a solver’s personal information and overall performance score based on ranking positions in previous contests. When Kaggle calculates the performance score, they take into account the competition intensity of each contest by considering factors such as the number of teams competing in the contest and their relative position.

In addition, each contest has a forum where solvers can initiate or participate in multiple discussion threads in various topics. This is the prime method of sharing knowledge and information among the solvers. Participation in these forums is completely voluntarily. The contest forum also shows how many times each discussion thread has been viewed and how many replies have been received for each post. If another solver finds a post is helpful, he/she can recognize the contribution by sending thanks to the contributor. All the thanks a post received are displayed below the post. Summary statistics on the number of posts by a solver, the number of thanks received, and the number of thanks given are also shown in a solver’s profile page.

We collected the information of 10,312 solvers. The information includes contests that solvers have participated in from the beginning of Kaggle.com through July 2012.

**Dependent Variables**

To study the impacts of community involvement on both individual performance and peer-recognition, we assess the dependent variables in the following ways.

**Performance:** In our main research model, we used a solver’s profile score to measure his/her performance. Kaggle uses a formula to calculate each solver’s profile score based on his/her performances in prior competitions. The maximum achievable points in a competition are based on the total number of solvers who participate in that competition, and adjusted by the difficulty level of the competition. According to Kaggle, “the current formula for each competition splits the points among the team members, decays the points for lower finishes, and adjusts for the number of teams that entered the competition...”¹ A solver’s profile score is dynamically updated after each competition. It ranges from 0 to 563,500, with a mean of 6,055.

**Peer-recognition:** In on-line context, thankful response to a helpful post is visible to all and generates social recognition (Sproull et al. 2005). In addition, prior studies have identified online “likes” as an indicator of peer encouragement and recognition (Gaved et al. 2013). In our context “thanks” is similar to “likes” in those contexts. Hence, we used the total number of thanks a solver received as a proxy for peer-recognition. The number of thanks received in our data set ranges from 0 to 175 with a mean of 0.32.

**Independent Variables**

**Community Involvement:** We used the total number of posts made by a solver on forums as a proxy for community involvement. This measure is consistent with the measurements used in prior studies. Sproull et al. (2005)’s study in pro-social behaviors on the net has identified message as the basic unit of contribution and “active participation” is measured by number of posts. Bateman et al. (2011) used the number of replies posted as a measurement of community participation; Tsai and Bagozzi (2014) used the number of messages posted to measure quantity contribution. The number of posts in our data set ranges from 0 to 506 with a mean of 1.2.

**Control Variable**

We control for the number of contests participated and the number of thanks given to the others when estimating the impact of community involvement on peer-recognition. We control for the number of

---

¹ Kaggle user ranking and tier system is found at https://www.kaggle.com/wiki/UserRankingAndTierSystem
contests participated, the number of software tools mastered, and education background when estimating the impact of community involvement on performance. We use the number of views or replies to posts as instrument variables when estimating our models.

Number of contests: We control for the total number of contests a solver has participated in. It ranges from 1 to 33 with a mean of 1.5.

Number of views/replies: The number of views/replies is the total number of times each topic thread is viewed/replied, respectively. The number of views ranges from 0 to 22,864 with a mean of 192.2, while the number of replies ranges from 0 to 52 with a mean of 0.3.

Thanks given: The numbers of thanks sent by a solver to other contributors to appreciate their valuable contributions. It ranges from 0 to 105 with a mean of 0.3.

Number of software tools: This is the total number of software tools/statistical packages listed in a solver’s profile. This variable has a range of 1 to 8 with a mean of 1.

Education: This is an ordinal variable derived based on the level of education, which ranges from 1 to 4.

Results

Simultaneous Equation Model

We used a simultaneous equation model to estimate coefficients, because the Ordinary Least Square (OLS) regression analysis may not be appropriate in our context. The OLS regression is based on the assumption that the independent variable (community involvement) affects the dependent variables (performance and peer-recognition) without any feedback loop between the dependent and independent variables. However, this may not be true in our context because our explanatory variables could be endogenous due to simultaneity. For example, Sproull et al. (2005) noted that receiving a positive response to a post will increase the likelihood for a contributor to engage in the community. Hence, in our context, community involvement can lead to peer-recognition, which may in turn lead to further involvement in the community. This feedback issue could make our OLS estimators biased and inconsistent. Hence, we address this concern by using a simultaneous equation system; that is, we used the Three Stage Least Square (3SLS) technique to estimate the coefficients of the model. The 3SLS technique has several advantages. 3SLS extends 2SLS by including all the equations simultaneously and takes into account correlations between error terms. Hence, it is more efficient than 2 SLS. Further, 3SLS estimators are robust to non-normality.

Our model consists of three separate structural equations for performance, peer-recognition, and community involvement. First, the performance equation includes endogenous community involvement, peer-recognition as well as exogenously determined variables such as education, software tools, and contests that are performance-related. Second, the peer-recognition equation includes endogenous community involvement as well as exogenously determined variables such as thanks-given, and contests that are peer-recognition related. We expect thanks given to influence peer-recognition due to reciprocity (i.e., if you received more thanks from your peers for your contribution, you may give them more thanks for their contribution). Third, the community involvement equation includes endogenous peer-recognition, performance, and exogenously determined community involvement related variables such as views and contests. In all the three equations, the number of excluded exogenous variables from the equation is at least as large as the number of right-hand side endogenous variables. Thus, the order condition for identification is satisfied and all three equations can be estimated.

\[ Performance_i = \alpha_{11} \text{Community Involvement}_i + \alpha_{12} \text{Peer Recognition}_i + \beta_{11} \text{Education}_i + \beta_{12} \text{Software Tools}_i + \beta_{13} \text{Contests}_i + \epsilon_1 \]  

\[ Peer \text{ Recognition}_i = \alpha_{21} \text{Community Involvement}_i + \alpha_{22} \text{Thanks Given}_i + \beta_{21} \text{Contests}_i + \epsilon_2 \]

\[ \text{Community Involvement}_i = \alpha_{31} \text{Peer Recognition}_i + \alpha_{32} \text{Performance}_i + \beta_{31} \text{Views}_i + \epsilon_3 \]
$\beta_3 \text{Contests}_{1} + \varepsilon_3$

<table>
<thead>
<tr>
<th></th>
<th>Community Inv.</th>
<th>Peer-recognition</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community Inv.</td>
<td></td>
<td>0.2004 ***</td>
<td>1.5045 ***</td>
</tr>
<tr>
<td>Peer-recognition</td>
<td>1.2718 ***</td>
<td></td>
<td>-1.8674 ***</td>
</tr>
<tr>
<td>Performance</td>
<td>0.2872 **</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Views</td>
<td>0.0667 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td>0.2868</td>
<td></td>
</tr>
<tr>
<td>Software tools</td>
<td></td>
<td>0.3945</td>
<td></td>
</tr>
<tr>
<td>Thanks given</td>
<td></td>
<td>0.3649 ***</td>
<td></td>
</tr>
<tr>
<td>Contests</td>
<td>-1.2087 **</td>
<td>0.0730 ***</td>
<td>4.2877 ***</td>
</tr>
<tr>
<td>Chi2</td>
<td>2614.75</td>
<td>10440.99</td>
<td>1694.42</td>
</tr>
<tr>
<td>Sample size</td>
<td>10312</td>
<td>10312</td>
<td>10312</td>
</tr>
</tbody>
</table>

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 1: 3SLS Regression Results

Table 1 summarizes the results of 3SLS regression using the cumulative number of posts as a measure of community involvement. Community involvement has a positive and significant impact on performance ($\alpha_{i} = 1.50$, $p < 0.01$) and peer-recognition ($\alpha_{i} = 0.20$, $p < 0.01$) supporting Hypotheses 1 and 2, respectively. Contrary to our expectations, peer-recognition shows a negative and significant impact on performance ($\alpha_{i} = -1.87$, $p < 0.01$). Hence, Hypothesis 3 is not supported. Community involvement can happen in many ways such as asking and answering questions and interacting with peers. But, a member’s contribution gets recognized by others only when others feel they are being helped and/or information shared is truly useful. Hence, in this context, peer-recognition (i.e., thanks) is also an indicator of the value of information given to other contestants. The crowdsourcing context is competitive in nature. Sharing useful knowledge with others who are contestants may give others an advantage in a competition by increasing others’ chances of winning the competition. Thus the more peer-recognition (thanks) an individual receives from community involvement, the more advantage this solver has given other contestants and the more performance sacrifices he or she has made in the competition. From this perspective, peer-recognition indicates the competitive loss a contributor suffers and this explains the negative impact on the contributor’s performance.

Discussion

This study contributes to a better understanding regarding the way in which community involvement influences a contributor’s performance and peer-recognition in specific online community in crowdsourcing competitions. Our results suggest that the impact of community involvement on the contributor is two-fold: it positively influences both performance and peer-recognition; that is, both intellectual capital and social capital. Peer-recognition was found to have a significantly negative relationship with performance. As with any empirical study, this study is subject to limitations. This study does not consider how an individual’s reputation outside the community impacts his/her peer-recognition within the community. Moreover, our cross-sectional data limit our capability to investigate progressive “learning” process of community members.

Implications for Theory

Our findings have a number of theoretical implications. First, our findings extend the growing research on online communities by quantifying the benefits of community involvement in a new type of online communities that is crowdsourcing communities. Crowdsourcing communities are different from other online communities, because participants are rivals. Contributors of online communities investigated in
prior studies are not direct competitors. Prior research has mostly addressed the possible motivations for people to participate in online communities. Little research has theorized and tested the actual benefits of community involvement to the contributor. Our findings provide evidence that the involvement in online communities does have important impact on individual performance and reputation.

Second, conflicting with prior literature, our results suggest that the cost effect of peer-recognition on performance dominates the positive effect on social capital in the crowdsourcing setting. This could be due to the fact that gaining peer-recognition requires time and effort which is scarce in crowdsourcing contests, and the valuable knowledge contributed when gaining peer-recognition is likely to enhance the rivals’ performance and impair the relative position of the contributors. Nonetheless, the positive direct effect of community involvement on performance is much stronger than the negative indirect effect through peer-recognition. Thus the total net effect of community involvement on performance is still positive. That is, those who contribute actively in crowdsourcing forums may sacrifice their performance ranking when they strive for peer-recognition, though the loss is compensated by the gains in intellectual capital.

Finally, our study contributes to the literature on pro-social behaviors by applying theories regarding these behaviors to the crowdsourcing competition setting. Moreover, our rich data set from a predictive modeling crowdsourcing website provided an objective and unbiased measure of individual performance in the community, while prior studies commonly used subjective measures. This is unique compared with the online communities in other settings and allows us to estimate our models based on real cases.

Implications for Practice

Our findings have a number of practical implications to seekers, solvers, and platform providers of crowdsourcing communities as well as other online communities. Crowdsourcing platform providers can use the insights from our research when designing crowdsourcing contests. Platform providers can develop and foster intellectual exchanges through forums in online crowdsourcing communities to improve the participants’ performance and strengthen the connections among them. Platform providers can also implement rules and mechanisms to encourage and facilitate knowledge contribution among members.

Participants of crowdsourcing competitions can use the findings in the study to improve their performance and gain social capital. Our results suggest that individuals can enhance their social and intellectual capital, and increase their performance by actively engaging in information and knowledge exchanges in online crowdsourcing communities. Though improving personal reputation comes with a cost, the accumulation in social capital may gain more support and skills, which help advance performance in the long run. Understanding the benefits of pro-social behaviors on the net provide insights to other online communities. In organizational settings, project managers may also apply these insights to enhance the performance of individuals in their work teams.

Conclusion

This study takes initial steps towards examining the impact of participants’ community involvement on their performance and peer-recognition in online crowdsourcing communities. Our findings offer guidelines to both participants of crowdsourcing contests and contest organizers on how to improve the performance of participants. Given the rapid development of technologies and practice, more data and cases about online communities in other settings will become available for further research on cost and benefits of pro-social behaviors on the net.

REFERENCES


