A Social Capital Perspective of Participant Contribution in Open Source Communities: The Case of Linux

Ray M. Chang
McGill University
myong.chang@mcgill.ca

Sung-Byung Yang
Hansung University
sbyang@hansung.ac.kr

Jae Yun Moon
Korea University
jymoon@korea.ac.kr

Wonseok Oh
School of Business
Yonsei University
wonseok.oh@yonsei.ac.kr

Alain Pinsonneault
McGill University
alain.pinsonneault@mcgill.ca

Abstract

Drawing upon a perspective of social capital, we investigate the extent to which several key dimensions of social capital accumulated during the growing stage of thread evolution influence both the quantitative and qualitative aspects of participant contribution in open source software development communities. To validate our hypotheses, we collected data from 223 Linux kernel threads, in which intensive intellectual and social interactions occur among the participants. The results suggest that the structural dimension of network capital (network centralization) is significantly associated with contribution quality, but not with contribution quantity. In contrast, the relational (network strength) and the dynamic (network growing speed) dimensions of network capital are significantly associated with contribution quantity, but not with contribution quality. The governance dimension (administrator participation) of network capital was found to be negatively significant on both the quality and quantity of contribution. Finally, no significant relationship was found between contribution quantity and contribution quality.

1. Introduction

Over the past decade, open-source software development (OSSD) has become a popular and powerful mechanism for the development of IT applications. Since many public and private organizations have been in support of the movement, OSSD has gained substantial market credibility and legitimacy [12]. Researchers in various disciplines have studied a wide range of OSSD-related topics, including the motivation for developers’ participation [e.g., 8] and factors leading to either OSS failures or successes [e.g., 22].

Although many of the individual-level factors regarding developers’ motivation to participate in an OSS community have been identified, there is a paucity of research available regarding the participants’ levels of contribution once they have joined the community. While some members actively participate, their level of participation varies widely in terms of both the quantity and quality of their contributions. In this study, using a social capital perspective, diverse forms of “network capital” are examined, reflecting how members’ day-to-day interactions influence participant contribution. Special attention is paid to the network capital co-produced by the participants during the growing stages of an OSS reply-to message network within a thread. Although OSS members can freely enter into or leave the communities at any time, we argue that the shape of the overall network structure is determined at an early stage of thread evolution. Further, it may not change substantially over time, even with the addition of new nodes (e.g., a new member’s entry) and links (e.g., a new message exchange between existing members). In this respect, the initial communication network structure may serve as a pivotal communicative conduit for participant contribution.

We drew upon social network studies to identify several key dimensions of initial network formation, such as degree of centralization, strength of network ties, number of primary nodes, and speed of network growth. We then examined how each dimension is associated with the degree of participant contribution at the thread level. To empirically validate our hypotheses, we collected data from 223 Linux kernel threads, in which intensive interactions occur among
the participants to bolster and improve the operating systems.

The rest of this paper is structured as follows. Section 2 presents the theoretical background of the study and Section 3 discusses the theoretical model and the hypotheses. Section 4 details the research method, and the results are provided in Section 5. Section 6 offers the implications of the results and Section 7 suggests avenues for future research.

2. Theoretical background

Recently, a social capital perspective has been used as a theoretical framework to understand the performance of individuals and groups who interact in online communities. For example, several researchers have focused on the structural properties of social capital, such as position or structure in the network, to investigate an individual’s knowledge sharing behavior [9, 21], while others have examined the role of relational properties in learning and knowledge transfer [13, 34]. Wasko and Faraj [38] demonstrated that knowledge sharing in electronic networks of practice is facilitated by three properties of social capital: structural, cognitive, and relational. Kuk [17] found that various types of structural social capital are significantly associated with the degree of knowledge sharing behaviors in OSS communities.

At least, three aspects of the existing literature in this field of research are noteworthy. First, although valid and useful, the extant literature seems to offer only a fragmented view of social capital’s impact on an individual’s performance. More specifically, most studies have focused solely on one of the three categories (structural, relational, or cognitive) with respect to the mode of social capital investigated. A more comprehensive framework that not only integrates the various dimensions of social capital, but that also offers new categories is therefore necessary in order to understand the multi-faceted nature of the relationship between network resources and performance. In recognition of this need, we developed a holistic model that combines diverse aspects of social capital: structural (network centralization), relational (network strength), governance (administrator participation), and dynamic (network growing speed) dimensions. These four characteristics represent the key components that constitute a social network, including nodes, links, structures, and evolution. Given that social capital is embedded and cultivated within a complex web of social networks, it is essential to understand all aspects of the networks. Figure 1 is a schematic representation of the four types of social capital, each of which was derived from the four dimensions of a network.

Second, most network studies have investigated a variety of network elements (e.g., actor, dyadic link, or ego-centric) in isolation. The actor and dyadic level studies link the attributes of either each actor or tie, separately to individual performance [5]. Similarly, ego-centric-level studies focus on the activities of the actors, while the other two dimensions are left unexplored [5].

Although studies conducted at the local level are conducive to exploring the impact of either actors or ties on individual performance, they may provide a limited view of how these network components work in tandem at a holistic level. Since these studies do not fully take the diversity of network elements into account, they may overestimate the role of individuals, as well as the significance of their relationships, while underestimating the importance of the actors’ collective behavior [26]. In fact, the literature contains only a conceptual discussion of a network-level approach. With the exception of Kuk [17] and Kudaravalli and Faraj [16], no empirical assessments have been conducted at the network (e.g., thread) level to date [26]. This may be due in part to the fact that collecting data at the network level is far more difficult and costly than gathering data at either the dyadic or ego-centric levels. To fill this gap, the present study was conducted at the network level. A series of hypotheses were developed to examine how the various dimensions of network capital at the thread level were related to the amount of participant contribution.

Finally, the majority of the previous studies in this area have focused on static aspects of the social
network by assessing only one particular point in time when investigating the amount of social capital. However, social capital tends to evolve naturally over time as the number of nodes and links constantly change. In this study, we have made an initial attempt to address this issue by using the moving averages of the amount of network capital produced, and by tracing the evolutionary patterns of social capital over the entire lifecycle of a given thread. Moreover, we included the speed at which networks expand as one of the elements inherent to a social network to represent a dynamic dimension of social capital [10].

3. Research model and hypotheses

The intellectual contribution of participants of an OSS community, whether in the form of suggesting design concepts, writing program code, or brainstorming ideas occurs through participation in a discussion thread. Although intellectual contribution might take a variety of forms, this study focuses on two dimensions of intellectual contribution: contribution quantity and contribution quality. Each dimension assesses a different but complementary aspect of participants’ intellectual contribution. The research model is shown in Figure 2. Based on the four dimensions of network capital articulated earlier, we formulate hypotheses regarding the relationships between each of the constructs and intellectual contribution as well as the relationship between the two key properties of contribution.

![Figure 2. Research Model](image)

3.1. Structural capital: network centralization

Social capital theory proposes that the compositional characteristics of a network is an important predictor of a participant’s intellectual contribution [19]. Network centralization refers to the degree to which the distribution of the connections is concentrated on a few central nodes [28]. Social networks become centralized when one or a few focal individuals dominate the discussion. This structure typically exhibits a star-like or hub-shaped power-law distribution, reflecting the participation inequality among the participants regardless of their role (e.g., administrators or developers). In contrast, a decentralized network shows a random structure that lacks dominant leaders [20].

Two contrasting views emerge when assessing the impact of network centralization on thread-level intellectual contribution. One view cited in the social network literature suggests that network centralization has a negative impact on a network’s creative performance (e.g., knowledge creation, creativity, or innovation) because centralization discourages the expression of diverse views; views from which innovative ideas are often generated [40]. In such environments, participants might remain passive and follow one highly influential individual. Consequently, participants are not motivated to commit and do not develop a habit of cooperation [32]. Such centralized networks reduce the autonomy of non-central (or peripheral) individuals.

A second perspective holds that network centralization is an effective structure that encourages knowledge exchange and learning and positively affects intellectual contribution as well as the rapid diffusion of innovative knowledge [35]. A hub-like, centralized network is efficient and participants can readily seek expertise with a lower search cost [20]; the technical authorities in such a network are easily identified by a participant who needs assistance. According to this view, easy access to an expert highly motivates the member to participate. In a decentralized network, however, communication spreads evenly among the participants (i.e., the network has a ring structure), the coordination and search costs may not be trivial, and communication between participants becomes indirect because no agent mediates their interactions. Moreover, in a decentralized network, integration of diverse views and opinions is challenging.

This study takes the approach that in a centralized network, the benefits accrued by the participants outweigh the costs incurred from the lack of idea diversity. Also, a hub-structure structurally facilitates member contribution. In the early stage of a discussion thread, when diverse views emerge, a centralized network results in a greater number of more insightful contributions by participants [17]. Continuous review and feedback processes are maintained in such a network, and the multiplicity of views and ideas are more easily integrated. Thus, we hypothesize:
**H1a:** Levels of initial network centralization will be positively related to the thread’s contribution quantity.

**H1b:** Levels of initial network centralization will be positively related to the thread’s contribution quality.

### 3.2. Relational capital: network strength

The nature of the relationships between nodes is another crucial factor that might explain a participant’s intellectual contribution [19]. The strength of the tie between two participants is one of the most commonly adopted variables to represent the quality of the dyadic relationship, and is typically illustrated in the literature by the thickness of the link. A strong dyadic relationship increases with increasingly frequent communication [33], and many social network studies have shown that strong ties are useful conduits for knowledge exchange between individuals [e.g., 13, 34]. An increasingly strong tie is in essence an increasingly strong contribution.

Relational capital at the dyadic level can represent network-level relational capital. A network that contains many strong ties is a stronger network than one with fewer (or weaker) ties. In this study, network strength is represented as the network-wide aggregated value of the strength of all of the ties; this was adopted as the key construct of the relational dimension of network capital. Despite the absence of formal management structures, the participants can develop an effective normative environment that fosters collaboration and coordination by communicating with each other frequently [31]. Once such an environment is formed during the early stages of the lifecycle of a thread, a sense of reciprocity forms among the participants, which in turn ensures continuing supportive exchanges. Therefore, networks with many strong tie relationships advance further and generate more insightful discussion. Thus, threads with higher levels of initial network strength should be positively related to network-level intellectual contribution in both quantitative and qualitative aspects. Based on these arguments, we posit:

**H2a:** Levels of initial network strength will be positively related to the thread’s contribution quantity.

**H2b:** Levels of initial network strength will be positively related to the thread’s contribution quality.

### 3.3. Governance capital: administrator participation

Although no formal governance mechanisms exist in OSS communities, administrators are typically responsible for both the substantive and social management of the community. Substantive duties include initiating projects, setting visions and goals, and coordinating activities while social duties include motivating members to contribute, resolving conflicts, and sanctioning behaviors [e.g., 6]. In traditional organizational settings, a leader’s involvement is often considered effective for eliciting from the employees both affective (i.e., employee willingness to work for the organization stemming from emotional attachment) and continuance contributions (i.e., employee willingness to continue to work stemming from perceived costs of membership in the organization) [2, p. 728]. Moreover, strong governing behaviors might encourage self-concept-based motivation by increasing the intrinsic benefits of voluntary efforts [30]. However, in OSS communities, which are characterized by voluntary contributions and lack of obligation, active involvement of leaders might not be welcome. Administrators maintaining tight control over project directions and goals in the initial phases of discussion may evoke images of a cathedral-type decision making structure, and thus undermine participants’ autonomy and sense of ownership, both of which are crucial motivators for participation [37]. A lesser degree of administrator participation in the growing stage allows for more developers to participate actively and bring a variety of perspectives. This motivates especially peripheral participants to increase their contribution and promotion [14]. Therefore, to promote participants’ intellectual contribution in terms of both quantity and quality, administrators should keep their involvement to a minimum, in particular during the initial stages of the project [23]. Previous studies have shown that members do not participate when most of the work is conducted by an administrator [e.g., 11, 29]. Moreover, maintaining multiple administrators in the project should also be discouraged as doing so will make participants more susceptible to the influence of the administrators. Based on these arguments, we propose the following:

**H3a:** Levels of initial administrator participation will be negatively related to the thread’s contribution quantity.

**H3b:** Levels of initial administrator participation will be negatively related to the thread’s contribution quality.

### 3.4. Dynamic capital: network growing speed

Research on interpersonal communications [e.g., 7, 39] suggests that the quantity and quality of the knowledge exchanged are highly associated with the rate at which knowledge is delivered. Although faster responses might allow the receivers to act upon urgent and pressing concerns in a timely manner, in the case
of knowledge-intensive communities, the level of detail or extensiveness of the responses is more important than response time [39]. Therefore, rapid answers might decrease the perceived value of the knowledge exchanged when this has a negative impact on level of accuracy or detail.

OSS networks grow as replies are made to the discussion threads. The literature indicates that rapid responses tend to be more emotional than are slow responses, which are more rational and cognitive [e.g., 36]. In terms of contribution quantity, threads with rapid responses are likely to receive relatively shorter messages than do threads that receive slower responses.

Developers place more weight on accuracy, which is usually higher in the later responses to a topic. It might be argued that an individual is able to respond quickly regardless of the degree of accuracy or detail, but if these early behaviors are considered at the network level, slower responses are perceived as more valuable. In particular, slower responses might reduce uncertainty and equivocality by increasing the likelihood of providing more complete knowledge [39]. Therefore, when the aggregated response speed, or network growth rate in the initial stage, is high, it is more likely to restrict network-level contribution in both quantitative and qualitative terms. Thus, we postulate:

- **H4a:** A high initial network growth rate will be negatively related to the thread’s contribution quantity.
- **H4b:** A high initial network growth rate will be negatively related to the thread’s contribution quality.

### 3.5. Impact of contribution quantity on quality

Previous studies [e.g., 4, 25] have found a relationship between quantity and quality, demonstrating that as the number of proposed ideas increases the probability of developing, a high-quality idea also increases. The same logic was applied in this study. In the later stages of a thread’s lifecycle, those participants who are intellectually committed contribute more insightful and more useful suggestions [39]. Based on these arguments, we suggest:

- **H5:** Higher levels of contribution quantity will be positively related to higher levels of contribution quality.

### 4. Research methodology

#### 4.1. Data collection

The data were collected from the open source Linux operating system kernel development community, in which globally distributed developers perform intellectual exchanges to develop and refine Linux operating systems. The Linux community was chosen because the development of operating systems requires sophisticated knowledge and intellectual collaboration at a level not required for many other software applications [18]. Developers use a designated mailing list for their communications, and for our purposes, a thread was defined as a series of email messages that contain the same subject. Our sample covered the ninety-month period between January 2000 and June 2008. During this period, numerous threads were initiated and maintained by Linux developers. Our analysis focused on Request for Comments (RFC) threads, where OSS developers propose significant upgrades to the current kernel system, since the paper focuses on participants’ intellectual contributions that involve a degree of sophistication beyond that involved with patches and bug reports. From the pool of threads in the archive, a total of 6,862 RFC threads were identified, among which nontrivial threads with sufficient volume that contained at least 25 messages with a minimum of four participants were selected. After this screening process, a total of 223 threads with approximately 11,000 messages were included in the final analysis. As our goal is to investigate how the initial formation of the network influences participant contribution, our methodological task involved two major components: (1) to identify initial network structures, from which we derived a set of independent variables, and (2) to operationalize the level of participant contribution, our key dependent variables, based on content analysis. In the following sections, we provide detailed descriptions of the procedures employed for these two aspects.

The network evolution stages were divided on the basis of the number of messages which reflects the activity level over a thread’s lifetime. The lifecycle of a thread was divided into three stages: growing, stable, and mature. The first one-third of the entire message thread was identified as the growing stage and the remaining two-thirds as the stable and the mature stage, respectively. With regard to the causalities among the variables, it is assumed that the various forms of network capital embedded in the initial or growing network structure (e.g., during the initial 1/3 of the thread) affects the participant contribution during the remaining state of the network (e.g., during the remaining 2/3 of the thread).

#### 4.2. Growing stage reply-to message networks

An RFC thread is created when one of the members initiates a new subject of discussion by sharing his or
her ideas with the community. The collaboration network grows over time as other members reply to the initial and subsequent threaded messages. We captured the reply-to message networks based on those reply relationships within the thread. Although all messages are broadcast to all subscribers of the mailing list, only a few highly interested members actually respond to it. In the context of software development, passive readers who do not participate in the discussion cannot contribute as much as those who participate. Therefore, consistent with other studies [e.g., 22], we included only the members who directly participated in the threaded discussion. Multiple message exchanges between two specific participants were counted as a multiplicity factor that reflects the strength of the interaction between the participants. Table 1 displays the specific definitions used to compute the network measures for the growing stage network of the thread.

4.3. Content analysis: contribution quantity and quality

We analyzed the actual content of the messages by using an automated content analysis software, Linguistic Inquiry and Word Count (LIWC) (http://www.liwc.net/) [24]. This application has been used in a broad range of fields, including social, psychological, medical, and educational studies as well as in information and management sciences [15]. The LIWC program searches the text for target words or word stems based on pre-defined linguistic categories. The word usage of each category is summarized as the percentage of total word count.

Prior to using the tool, the sample messages were refined by removing non-communicational elements such as html tags, program code, and signatures. We also removed any quoted text in order to retain only original content written by the sender in each message. On average, 12.5 (S.D. = 7.7) participants produced 47.2 (S.D. = 31.4) messages per thread.

We focused on two aspects of participant contribution - quantity and quality - and used these as the dependent variables. To avoid any potential problems that might arise from causality [40], only those messages exchanged during the stable and the mature stages were included in the content analysis. Contribution quantity was measured with the average word count per message for each thread. Following Wasko and Faraj [38], we assumed that the more a developer expresses his or her thoughts, the more contribution the person makes. As the RFC threads typically deal with new ideas and thoughts such as sophisticated technical proposals, clarifications, and questions/answerers, more detailed messages are found to be more helpful for the members to better understand the new ideas proposed.

To gauge the quality aspect of participant contribution, we used the LIWC category of “Insight”, which consists of 195 words and word stems, including such words as realize, understand, know, and reason. Pennebaker et al. [24] define the category as the words representing “learning” and “understanding.” The category has been used to substantiate the degree of experiential learning [1]. To check the validity of the results from the automated text analysis, two of the authors coded 10% of the threads for the extent to which insightful and knowledgeable discussions were shared. The correlation coefficients between each coder’s score and the output from LIWC were 0.72 and 0.74 respectively with 0.84 of the inter-coder reliability.

4.4. Control variables

Several control variables were included in the model to attenuate their potential impacts on the results. The thread was coded as 1 if its initial message contains source code which propose more concrete ideas to call for more contribution [27], and was coded as 0 otherwise. In addition, a variable representing the culture of a thread was included. Specifically, in line with Bagozzi and Dholakia [3], a variable that reflects members’ inhibiting behavior was included to control for the impact of inhibition on the quantity and quality of contribution. Intuitively, such negative atmosphere would structurally prevent members from making contributions. For this, we used the LIWC category “Inhibit”. Table 1 shows the means and standard deviations for the variables included in the analysis.

[Insert Table 1. here]

5. Results

We chose the partial least squares (PLS) method as the main analytical approach. Similar to the alternative structural equation modeling (SEM) techniques (e.g., LISREL, EQS, AMOS), PLS aims to answer a set of interrelated research questions by simultaneously modeling the relationships between multiple independent and dependent constructs. PLS is particularly suitable for the intent of our study, as it requires fewer assumptions regarding the measurement scales and the distribution of data. Table 2 illustrates the results of the exploratory factor analysis. Each indicator has a higher loading on its construct than on any other construct.
The goodness of the structural model was evaluated using path coefficients and their statistical significance. Figure 3 depicts the path coefficients ($\beta$), $t$-values, and the amount of variance explained ($R^2$) by the endogenous variables comprising the structural model.

**Figure 3. Results**

As shown in Figure 3, five of the nine causal paths were supported, although at different levels of statistical significance. More specifically, of the four hypotheses that pertain to contribution quantity, three of the constructs (network strength, administrator participation, and network growing speed) were significantly related. Thus, H2a, H3a, and H4a were supported by the data. However, no such significance was detected between network centralization and contribution quantity; therefore, H1a was not supported. Overall, approximately 12% ($R^2 = 0.120$) of the variance in contribution quantity was explained by the combination of the four independent variables and the two control variables.

Regarding the hypotheses focusing on contribution quality, two of the four paths were statistically significant. A significant relationship was found between network centralization and contribution quality (H1b), albeit at the 0.1 significance level. In addition, a significant negative relationship was observed between administrator participation (H3b) and contribution quality. Neither network strength (H2b) nor network growing speed (H4b) was significantly associated with contribution quality. Interestingly, contrary to our expectations, no significant association was found between contribution quantity and quality, rejecting H5. Regarding the two control variables, only the existence of codes in the initial message was significantly associated with contribution quantity. Four independent variables, contribution quantity, and two control variables in tandem explained approximately 9% ($R^2 = 0.085$) of the variance in contribution quality.

### 6. Discussion and implications

The results suggest that the structural dimension of network capital (degree of network centralization) is significantly associated with contribution quality, but not with contribution quantity. The lack of empirical support for H1a may be due to the presence of the two opposing forces (diversity versus learning), each of which influences participant contribution in a different direction as articulated in the hypothesis development. The relational (network strength) and the dynamic (network growing speed) dimensions of network capital were significantly associated with contribution quantity, but not with contribution quality. One speculation for the insignificant relationship with contribution quality is that strong ties may produce and preserve many redundancies, which structurally prevent the generation of innovative ideas. Moreover, the lack of support for H4b suggests that contribution quality does not depend on the physical time that elapses during the growing stage. This finding indicates that contribution quality does not automatically occur with the progression of time. As expected, the governance dimension (administrator participation) of network capital had a significant negative effect on both the quantity and quality of contribution. Interestingly, contribution quantity had no significant impact on contribution quality. This indicates that OSS developers do not find voluminous discussions essential for producing qualitative contributions.

Our findings provide important managerial implications regarding the role of administrators. Although centralized communication patterns lead to superior discussion performance, any deliberate interventions or manipulations by the administrator may result in adverse outcomes. More specifically, centralized communication structures can serve as effective learning conduits for eliciting quality discussions among members only when they are "naturally" established and evolved by peer members. Conversely, when administrators "artificially" craft such structures through their frequent interventions and controls, adverse outcomes may result, as they discourage member participation. Furthermore, the findings suggest that administrators should flexibly manage their level of involvement contingent on the network's developmental phase. In particular, it may be wise to keep their intervention to a minimum during the growing stage of a thread (H3). Yet, since a centralized pattern is desirable during the growing stage (H1a), several peer participants with relevant expertise (including the thread initiator) should lead the discussion, rather than the administrator. These initiators should proactively play an anchoring role to
establish a centralized communication structure. Our results suggest that too intense and frequent interactions between peer members who did not initiate the discussion may weaken the quality of discussion. Similarly, in order to elicit high quality contribution, all participants are advised to fully utilize their pre-established communication relationships instead of creating new relationships with other participants during the thread’s growing stage. Finally, since instant and quick replies during the growing stage of the network may be negatively related to contribution quantity in the maturity stage, participants should contemplate their responses and provide accurate and detailed information.

This study contributes to the growing body of literature devoted to the performance of OSS communities. Although several studies [e.g., 3, 27] have identified the factors that motivate developers to join OSS communities, our understanding of what influences developers’ levels of contribution after joining remains limited. A developer’s initial motivation may wane as he or she interacts with other developers, many of whom may have disparate views and goals. OSS communities are, after all, virtual forums where people socially and intellectually establish and terminate relationships. In other words, their level of post-entry contribution may be affected by the social fabric within which their relationships are embedded. This is one of the first studies to investigate the role of initial network structures in determining the quantity and quality of an OSS participant’s subsequent contribution. We argue that the overall interaction pattern among the participants is likely to be determined at the early, growing stage of the thread’s evolution.

This study further focuses on the evolutionary nature of such networks, tracing changes in network capital throughout the entire lifecycle of an OSS thread. Although attention is primarily given to the initial structure of the network, insights into how the network structure is transformed over time, with the supplementation of new nodes and links, are provided as well. Finally, this study analyzes the actual content of the messages exchanged by participants, which has not yet been attempted in the literature. Previous studies of network exchanges are often criticized because they merely analyze the patterns of message exchange while largely ignoring the content of the messages. An automated tool was used to examine the actual content of the intellectual exchanges between participants, and the outputs were used as proxies for participant contribution in our study.

7. Limitations and future research

Our study has several caveats that should be addressed in future studies. First, each stage in the thread lifecycle was measured by using the number of messages exchanged rather than with real-time data, such as the time required for each thread to reach the mature stage. While our measures allowed us to control for thread length and potential differences in task complexity, using the real time to maturity could provide interesting insights into thread evolution. However, the real-time approach has its own limitations; it is too subjective to determine, and quite difficult to justify the cut-off points (e.g., one week). Nevertheless, future studies should employ this approach to validate our findings.

Another limitation involves the use of an automated text-analysis tool to analyze the large volume of messages. The strength of this tool lies in its ability to provide objective analytical assessments, which can be difficult to achieve when human evaluators perform the task manually. Although LIWC has been widely used in other research disciplines, the tool is relatively new to IS research and its accuracy is continuously improving as new contextual words are added to the existing categories. It is possible that the list of contextual words is still incomplete. To address this issue, we performed manual assessments on a fraction of the sampled messages (i.e., 10% of total messages) and found high consistency between the two approaches.

Finally, the variance explained by our model is relatively limited. This suggests that many other factors affect a member’s contribution once he or she has joined the community. A formal survey should be carried out to identify these specific factors. To advance our understanding of the evolutionary aspects of reply-to message networks, future research should explore changes in the structural formations resulting from the entry of new participants. For example, with whom do these new entrants tend to interact? Among these new entrants, who contributes the most in terms of network capital and participant contribution and why? Another interesting question may be to determine the extent to which one can predict thread length based on the initial interaction patterns. While this study has answered several questions, it has also raised a number of additional questions that must be addressed in future research. We hope that our study will lay the foundation for future research on this important topic.

8. References


Table 1. Operational definitions and descriptive characteristics

<table>
<thead>
<tr>
<th>Stage</th>
<th>Construct</th>
<th>Item</th>
<th>Definition</th>
<th>Avg.</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growing (Initial 1/3) Stage</td>
<td>Network Centralization</td>
<td>Degree centralization</td>
<td>$\chi$: degree (betweenness) centrality of node / $\chi^<em>$: maximum degree (betweenness) centrality centrality Centralization $= \sum (\chi - \chi^</em>) / \max \sum (\chi - \chi^*)$</td>
<td>0.538</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Betweenness centralization</td>
<td></td>
<td>0.379</td>
<td>0.146</td>
</tr>
<tr>
<td></td>
<td>Network Strength</td>
<td>Multi-link ratio</td>
<td>proportion of strong ties; # strong links / # total links</td>
<td>0.397</td>
<td>0.243</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multi-message ratio</td>
<td>message overloading per link; # total messages / # total links</td>
<td>2.152</td>
<td>0.741</td>
</tr>
<tr>
<td></td>
<td>Administrator Participation $^a$</td>
<td>adm-node ratio</td>
<td>proportion of administrators; # admins / # total participants</td>
<td>0.308</td>
<td>0.191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>adm-message ratio</td>
<td>proportion of administrators’ messages; # messages of admins / # total messages</td>
<td>0.395</td>
<td>0.270</td>
</tr>
<tr>
<td></td>
<td>Network Growing Speed</td>
<td>Message growing speed $^b$</td>
<td>log(message occurring rate per unit time); log(# total messages / total elapsed hours)</td>
<td>-0.165</td>
<td>0.565</td>
</tr>
<tr>
<td>Stable and Mature (Later 2/3) Stage</td>
<td>Contribution Quantity</td>
<td>Word count per message $^b$</td>
<td>log(word count per message) $^c$</td>
<td>1.982</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>Contribution Quality</td>
<td>Insight words percentage</td>
<td>insight words percentage $^c$</td>
<td>2.108</td>
<td>0.556</td>
</tr>
<tr>
<td>1$^{st}$ message</td>
<td>Code in Initial Message</td>
<td>Ini-code</td>
<td>1 = initial message contains technical codes 0 = otherwise</td>
<td>0.637</td>
<td>0.482</td>
</tr>
<tr>
<td>Growing Stage</td>
<td>Inhibiting Climate</td>
<td>Inhibit</td>
<td>inhibit words percentage $^c$</td>
<td>0.944</td>
<td>0.681</td>
</tr>
</tbody>
</table>

$^a$To identify the administrators, we used the file named “MAINTAINERS” provided in Linux kernel sources.
$^b$Due to its highly skewed distribution, we used log-transformed measure.
$^c$LIWC outputs were used to capture the values.

Table 2. Results of factor analysis

<table>
<thead>
<tr>
<th>Construct</th>
<th>Items</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Centralization (1)</td>
<td>Degree centralization</td>
<td>.975</td>
<td>.036</td>
<td>.004</td>
<td>-.010</td>
<td>.013</td>
<td>.039</td>
<td>-.004</td>
<td>.015</td>
</tr>
<tr>
<td></td>
<td>Betweenness centralization</td>
<td>.972</td>
<td>-.098</td>
<td>-.018</td>
<td>-.001</td>
<td>.014</td>
<td>.054</td>
<td>-.020</td>
<td>.021</td>
</tr>
<tr>
<td>Network Strength (2)</td>
<td>Multi-link ratio</td>
<td>.062</td>
<td>.941</td>
<td>.064</td>
<td>.000</td>
<td>.111</td>
<td>.055</td>
<td>.080</td>
<td>.029</td>
</tr>
<tr>
<td></td>
<td>Multi-message ratio</td>
<td>-.123</td>
<td>.951</td>
<td>-.030</td>
<td>-.043</td>
<td>-.017</td>
<td>-.025</td>
<td>-.016</td>
<td>-.005</td>
</tr>
<tr>
<td>Administrator Participation (3)</td>
<td>Adm-node ratio</td>
<td>.049</td>
<td>.004</td>
<td>.922</td>
<td>.042</td>
<td>-.107</td>
<td>-.080</td>
<td>-.027</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>Adm-message ratio</td>
<td>-.062</td>
<td>.027</td>
<td>.931</td>
<td>.020</td>
<td>-.010</td>
<td>-.062</td>
<td>.039</td>
<td>-.030</td>
</tr>
<tr>
<td>Network Growing Speed (4)</td>
<td>Message growing speed</td>
<td>-.009</td>
<td>-.036</td>
<td>.054</td>
<td>.989</td>
<td>-.092</td>
<td>-.027</td>
<td>-.063</td>
<td>-.070</td>
</tr>
<tr>
<td>Contribution Quantity (5)</td>
<td>Word count per message</td>
<td>.023</td>
<td>.081</td>
<td>-.086</td>
<td>-.094</td>
<td>.980</td>
<td>.052</td>
<td>-.102</td>
<td>.011</td>
</tr>
<tr>
<td>Code in Initial Message (7)</td>
<td>Insight</td>
<td>.081</td>
<td>.025</td>
<td>-.129</td>
<td>-.027</td>
<td>.052</td>
<td>.984</td>
<td>-.020</td>
<td>.066</td>
</tr>
<tr>
<td>Inhibiting Climate (8)</td>
<td>Inhibit</td>
<td>-.021</td>
<td>.054</td>
<td>.010</td>
<td>-.063</td>
<td>-.099</td>
<td>-.020</td>
<td>.990</td>
<td>.017</td>
</tr>
</tbody>
</table>