Organizational Research Methods
000(00) 1-31

DOI: 10.1177/1094428110375002
http://orm.sagepub.com

Socialization in Open Source Software Projects: A Growth Mixture Modeling Approach

Israr Qureshi[1] and Yulin Fang[2]

Abstract

The success of open source software (OSS) projects depends heavily on the voluntary participation of a large number of developers. To remain sustainable, it is vital for an OSS project community to maintain a critical mass of core developers. Yet, only a small number of participants (identified here as "joiners") can successfully socialize themselves into the core developer group. Despite the importance of joiners' socialization behavior, quantitative longitudinal research in this area is lacking. This exploratory study examines joiners' temporal socialization trajectories and their impacts on joiners' status progression. Guided by social resource theory and using the growth mixture modeling (GMM) approach to study 133 joiners in 40 OSS projects, the authors found that these joiners differed in both their initial levels and their growth trajectories of socialization and identified four distinct classes of joiner socialization behavior. They also found that these distinct latent classes of joiners varied in their status progression within their communities. The implications for research and practice are correspondingly discussed.

Keywords

latent class analysis, latent class growth models, latent growth models, longitudinal data analysis, quantitative: structural equation modeling

Introduction

The open source software (OSS) development model originated in the 1970s, partially as a defensive reaction to the move by some private software companies to appropriate publicly available software into their proprietary applications (Stallman & Lessig, 2002). Over the last decade, this intriguing software development model has emerged as a viable alternative to commercial software projects (Fitzgerald, 2006) and has attracted increasing academic and corporate attention (Sen, 2007; Stewart, Ammeter, & Maruping, 2006). Some OSS projects have achieved remarkable adoption success. Among the best known OSS projects are the Linux operating system, and the Apache web server, which answers 70% of all the webpage requests through the Internet (Netcraft, 2004). For the

[1] Department of Management and Marketing, Hong Kong Polytechnic University, Hong Kong, China
[2] Department of Information Systems, City University of Hong Kong, Hong Kong, China

Corresponding Author:
Israr Qureshi, Hong Kong Polytechnic University, M801 Li Ka Shing Tower, Hong Kong, China
Email: Israr.qureshi@inet.polyu.edu.hk

commercial market, it is reported that 60% of the largest companies (those with over 5,000 employees) in North America plan to implement OSS applications, half of these perform mission-critical functions such as application servers and web servers (IBM, 2006; Schadler, 2004).

The notable success of OSS projects, however, would not have been accomplished without individual developers' voluntary participation (Roberts, Hann, & Slaughter, 2006). Indeed, research has shown that failure in OSS development is frequently due to the shortage of volunteer participation, whereas successful OSS projects often feature a large, sustainable pool of active participants (Crowston, Annabi, & Howison, 2003; Krishnamurthy, 2002; Markus, Manville, & Agres, 2000).

A key to ensuring that there is a sufficient and sustainable supply of active developers is to motivate, engage, and retain peripheral developers who are involved in OSS projects but who do not yet have authorization to manage a codebase (Fang & Neufeld, 2009; Von Krogh, Spaeth, & Lakhani, 2003). Peripheral developers represent the majority of OSS project membership, yet only a small portion of them, whom we have referred to as "joiners" in this article, will distinguish themselves over time and eventually join the core developer group with authorization to manage a codebase. As candidates, to become core developers, the joiners contribute to maintaining a critical mass of this group and therefore constitute an essential ingredient in the long-term viability of OSS projects (Ducheneaut, 2005; Lee & Cole, 2003). Thus, it is important to understand how joiners socialize themselves and become part of the core of the OSS community.

Although considerable research has focused on understanding how to motivate developers to participate (Franke & von Hippel, 2003; Hertel, Niedner, & Herrmann, 2003; Roberts et al., 2006; Shah, 2006; Von Hippel, 2001; Von Krogh et al., 2003), there is a distinct lack of research on the socialization process, particularly the socialization of joiners, with only a few exceptions (Ducheneaut, 2005; Fang & Neufeld, 2009; Von Krogh et al., 2003). Moreover, these few exceptions, each of them a case study, either take a static approach to prescribe a set of "joining scripts" (Von Krogh et al., 2003) or adopt a longitudinal approach to arrive at a set of sequential steps for joining the core group (Ducheneaut, 2005; Fang & Neufeld, 2009), with the implicit assumption that joiners should follow a homogenous trajectory of socialization.

Although these qualitative findings are useful, their conclusiveness must be viewed with caution. In fact, the results of our systematic investigation, which are supported by prior research (Ye & Kishida, 2003), indicate that joiners vary significantly in terms of their lead time (LT) for core status attainment, implying that unobserved subpopulations of joiners with different socialization trajectories are likely to exist. However, sufficient theoretical and empirical measures have not been undertaken to gain a better understanding of this heterogeneity. Nevertheless, understanding the potential heterogeneity of socialization trajectories and the resulting outcomes is vital, because it can potentially advance our theoretical insights into the latent characteristics that affect the joiners' socialization processes. Given this backdrop, the current study aims to *identify* the various classes (subpopulations) of joiners (those eventually awarded core status) through studying their socialization trajectories and *explore* the effects of joiners' membership in these classes on the time taken for them to attain core status.

To address this research question, we base our empirical exploration on social resource theory (Lin, Ensel, & Vaughn, 1981a, 1981b). By recognizing the criticality of social resources, organizational researchers have highlighted the importance of socializing with those who are higher in the organizational hierarchy to individual career outcomes, such as job placement and mobility (Lin, 1990; Seibert, Kraimer, & Liden, 2001). The naturally evolving patterns of joiners' socialization with core developers—the social resources involved in this interaction—provide a critical point of departure for our exploration. Building on organizational sociology, we suggest that different temporal developments of social resources with higher status developers (i.e., core developers) are likely to influence the joiners' status progression. Thus, our research first identifies the heterogeneity of joiners' socialization patterns and then relates each pattern to the LT for core status attainment.

The nature of our research question entails an analytical technique that supports the modeling of individual trajectories based on intraindividual changes over time and classifies individuals in distinct categories based on interindividual differences in behavioral patterns (i.e., different classes of socialization trajectories).[1] However, the existing methods in OSS research have either focused on the variable-centered approach, such as regression and structural equation modeling (SEM; Roberts et al., 2006), or on the individual-centered approach, such as cluster analysis (Grewal, Lilien, & Mallapragada, 2006) that accounts for interindividual differences only and does not explain intraindividual changes.

These interindividual differences may be either observable or unobservable. For observable differences (such as gender, race, etc.) where the heterogeneity is obvious, the researcher can use multigroup methods (Qureshi & Compeau, 2009; Sörbom, 1974). However, for unobservable differences where the heterogeneity is not observable, there is no apparent a priori basis upon which to form groups. As a result, the data are generally analyzed as a sample drawn from a single population where all the individuals are potentially assumed to have the same set of parameter values (Collins & Lanza, 2010; Muthén, 1989). However, such an assumption of homogeneity could be misleading. For example, the degree of technology adoption could differ between individuals who would otherwise not show any observable heterogeneity. Similarly, individual consumers may perceive the same product differently based on their past purchasing experiences and their perceived expected value of a product. Hence, it is generally difficult to identify such groups a priori based on demographic, personality, or other related data (Jedidi, Jagpal, & Desarbo, 1997; Lubke, 2010; Moore, 1980).

The growth mixture modeling (GMM) technique, however, can assist in identifying this unobserved heterogeneity. This analytical technique summarizes longitudinal data by modeling both intra- and interindividual variability in developmental trajectories through identifying unobserved subpopulations (i.e., a small number of classes) defined by their initial levels and the shape of their growth trajectories (Muthén, 2001; Wang & Bodner, 2007), thus effectively consolidating the two approaches. This technique has found widespread usage in research on development studies and sociology (Kreuter & Muthén, 2008; Pinquart & Schindler, 2007; Wu & Witkiewitz, 2008). However, it has not been commonly used in the management field except for two notable exceptions (Wang, 2007; Wang & Bodner, 2007), which demonstrate the use of GMM with data on a retiree's psychological well-being. In this study, we use archival data to demonstrate the use of GMM in the context of OSS developer socialization. Specifically, we use the SEM framework for GMM because SEM is a powerful method of simultaneously estimating a structural and measurement model (Jöreskog, 1971).

In the next section, we provide the conceptual background and the research hypotheses. Here, we review the OSS literature, introduce social resource theory, and develop the hypotheses. To test these hypotheses, we introduce the data collection and analysis strategy. Particularly, we elaborate on the GMM technique and conduct an empirical investigation by drawing on a longitudinal data set of 133 joiners from 40 projects, whose LT to core status ranges between 7 and over 200 weeks. Finally, the empirical results are discussed in conjunction with the existing OSS literature.

Literature Review: Joiners' Socialization in OSS Communities

An OSS project involves a decentralized community of volunteer developers who collaborate to produce a software product using Internet-based tools such as project websites, mailing lists, and concurrent versioning systems (CVS). Although access to certain tools (e.g., CVS systems) was restricted to core developers who took on key technical activities and demonstrated advanced technical knowledge, access to mailing lists was free and open to everyone, resulting in a large pool of participants (Von Krogh et al., 2003). Although different projects may have several different roles for the participants, a two-tier role structure is most commonly seen in OSS projects: that of the tier

of core developers with authorization to submit code changes and that of peripheral developers with permission to participate in the mailing list, report bugs, and suggest modification in code without authorization to submit code changes (Lee & Cole, 2003). Thus, the most distinct characteristic of peripheral developers is that they have no authority to change codebases unless they first become core developers.

Nevertheless, both types of developers are important for the success of OSS projects. Although core developers have direct administrative responsibilities in maintaining the software codebase, peripheral developers make indirect contributions through their participation in the various mailing lists of the community (Lee & Cole, 2003). Hence, these two tiers of developers are ecologically dependent on each other. The core developers draw intellectual input from peripheral developers by relying on them to generate patches of computer codes and to report bugs. Whereas candidates to become core developers must be drawn from the large pool of peripheral developers based on nominations made by existing core developers based on their evaluation of the peripheral developers' participation activities (Fang & Neufeld, 2009). Following Von Krogh et al. (2003), we use the term "joiners" to represent the group of peripheral developers who eventually join as core developers. Joiners, as future core developers, are essential to sustaining a critical mass of core developers and, hence, are vital to the long-term survival of OSS projects (Ducheneaut, 2005).

However, joining the core developer community is not effortless. Software development, whether close or open source, is a knowledge-intensive activity by nature, which requires high levels of domain-specific knowledge, broad experience, and intensive learning on the part of those wishing to contribute to it (Fichman & Kemerer, 1997; Pliskin, Balaila, & Kenigshtein, 1991). Only those who continuously participate in the development process over an extended period of time can contribute in a meaningful way, and many others find it too difficult to integrate themselves with the core developer team (Kohanski, 1998). Similarly, OSS research has also found that it takes time and effort for peripheral developers to gain the ability to socialize with the core group and only a very small percentage of the peripheral group whose performance is successfully recognized and valued will eventually succeed (Ducheneaut, 2005; Fang & Neufeld, 2009). As such, there is an urgent need to better understand the process by which developers become socially integrated into OSS projects (Von Hippel & Von Krogh, 2003).

Limited research has been directed toward addressing this issue, which was primarily focused on differentiating between the socialization behavior of joiners and that of permanent peripheral software developers. Based on an inductive, qualitative approach (Glaser & Strauss, 1967), researchers found that those who participate according to a "joining script" in terms of their activity type and intensity are more likely to succeed in becoming core developers (Von Krogh et al., 2003). Using an ethnography approach, Ducheneaut (2005) focused on a single joiner (as opposed to multiple peripheral developers) and identified a temporal sequence of socialization activities that contributed to this status progression. Based on situated learning theory (Lave & Wenger, 1990), Fang and Neufeld (2009) characterized socialization in OSS projects as a recursive process of continuous learning, competence demonstration, and role transformation. Joiners iterate between learning and performing, and as a result, their roles progressively evolve within the community. Role advancement, in turn, opens even more opportunities for continued learning and performing.

Despite the advancement in understanding, the value of these a priori studies is limited in several important aspects. First, they focus on the dichotomy of role change (i.e., the role switch between peripheral and core members) as a socialization outcome and neglect the fact that a significant difference exists among joiners in terms of the time they would take to change their roles. For instance, the LT for joiners to achieve the core status in a PhpMyadmin program ranged between 0 and 20 weeks (Fang & Neufeld, 2009). It is important to understand the mechanism behind determining the LT for status attainment because OSS managers with such an understanding could be in a better position to more effectively recruit and develop joiners—steps that are vital to the survival and

prosperity of OSS projects (Ducheneaut, 2005). Second, by studying the distinctions between joiners and nonjoiners (i.e., those peripheral developers who may not become core developers), previous studies assumed that homogenous socialization patterns existed within the group of joiners. However, the noted differences in LT for core status attainment among joiners imply that their socialization processes might be heterogeneous. Finally, previous studies focused on identifying the *types* of socialization activities without focusing on the characteristics of the party with which the joiners sought to socialize. However, since core status change is a decision made by the group of core developers, it is important for peripheral developers to socialize with them.

To develop deeper insights into the socialization process that occurs between joiners and core developers, we will build our exploration on social resources theory (Lin et al., 1981b), which is derived from the conventional organizational setting but is particularly relevant to the context of the current study.

Theoretical Background: Social Resource Theory

Social resource theory focuses on the nature of the resources embedded within the interpersonal relationships of a focal individual and advocates that an individual's relationship can convey advantages when he or she connects with someone who has the type of resource required for that individual to fulfill his or her instrumental objectives (Lin et al., 1981b). A connected individual who possesses or controls the necessary resources for the attainment of the focal individual's goals (e.g., developmental contacts at higher organizational levels) can be considered a social resource (Seibert et al., 2001). For instance, within organizational contexts, managers or peers who provide career development advice and support are considered relevant social resources when focal employees pursue their instrumental career goals (Lin et al., 1981b).

Organizational members access and develop social resources within their organization through organizational socialization, a process by which they acquire the attitudes, behaviors, and knowledge required to perform effectively within the organization by interacting with other members (Van Mannen & Schein, 1979). Although this process is useful to everyone in the organization, it is particularly beneficial to newcomers or individuals with a lower status because it helps them connect with important social resources such as ''insiders'' or more experienced members who are at the higher organizational levels (Louis, 1990; Reichers, 1987). Empirical research has shown that people who connect with others of prestigious status are more likely to reach a prestigious status themselves (Marsden & Hurlbert, 1988).

Experienced organizational members (i.e., insiders), who are acting as important social resources, can provide numerous benefits to newcomers with respect to their career development in several ways. First, these insiders can facilitate the newcomers' learning by sharing knowledge about the general organizational context and the conduct of specific task assignments as well as the role expectations for the newcomer and his or her assigned responsibilities (Morrison, 1993, 2002; Seibert et al., 2001). Second, they can optimize the newcomers' performance by conferring on them an organizational identity and providing social support (Podolny & Baron, 1997). Third, relationships with insiders, particularly with those at higher organizational levels, can help secure career sponsorship by increasing the newcomers' visibility as well as by highlighting his or her credentials and legitimacy within the organization (Lin, 1999; Seibert et al., 2001). Through these mechanisms, the development of social resources can contribute to several aspects of employees' career success, such as achieving prestigious job status (Marsden & Hurlbert, 1988), job satisfaction, and promotion (Seibert et al., 2001).

Research on the development of social resources is, however, still emerging. The limited research that has been conducted in this field has acknowledged that newcomers to an organization may already possess different initial levels of social resources because some of them have prior

connections with employees within that particular organization (Castilla, 2005). Similarly, research has also shown that newcomers with different personal and social backgrounds may be in different positions to connect with high-level organizational contacts, leading to varying stocks of social resources (Lin et al., 1981a; Seibert et al., 2001). Such initial differences may have immediate as well as long-term performance implications for the newcomers (Castilla, 2005).

Hypotheses: Socialization and Status Progression in OSS Communities

Based on social resource theory (Lin et al., 1981a), we conceptualize that the joiners' socialization in OSS communities as a dual process of developing and leveraging various social resources. Consistent with the theory and the focus of our study, we refer to social resources in the OSS context as consisting of access to upper level contacts (i.e., core developers) from whom joiners gain useful information, social support, and sponsorship. Our analysis of social resources focuses on the intensity of the relationship between joiners and core developers, and our goal is to examine the importance of the temporal heterogeneity of such a relationship—the level at which it begins and how it evolves over time in different ways and how it affects the time taken to attain core status.

As discussed earlier, existing research suggests that leveraging social resources is critical to the career development of organizational members because they can benefit from tapping social resources in terms of learning, social support, and career sponsorship (Seibert et al., 2001). Within the OSS context, the joiners' socialization pattern characterizes a process of not only acquiring the necessary skills for task accomplishment (Fang & Neufeld, 2009; Shah, 2006) but also becoming cognizant of the community norms, values, and preferences of the core developers (Ducheneaut, 2005). In addition, this socialization process is instrumental to obtaining ''allies'' or acceptance from core developers who can provide sponsorship for proposals with respect to code changes (Ducheneaut, 2005). Furthermore, supporting each other through providing comments, suggestions, and advice through socializing on mailing lists promotes a sense of reciprocity (Roberts et al., 2006). This norm of reciprocity, together with the open source ideology that is typical of developers (Stewart & Gosain, 2006), fosters the culture of mutual support. Through these mechanisms, we contend that possession of vital social resources can contribute to a joiner's status progression.

Nevertheless, the joiners' socialization is also a process of developing even more social resources. Fang and Neufeld (2009) characterize OSS developer socialization as a recursive process of participation, learning, and role transformation. The more joiners learn from and interact with core developers, the more he or she will be given opportunities to socialize even further with the core developers and the more social resources he or she will be able to accumulate. Similarly, the joiners' social resources can help develop allies and the buy-in of core developers more easily (Ducheneaut, 2005). This process creates greater opportunities to socialize with more developers and build even more social resources.

This recursive process of social resource development suggests that social resource building is facilitated by possessing high levels of existing social resources, a law called ''asset mass efficiency'' (Dierickx & Cool, 1989). This fact is reflected in the familiar phrase, ''success breeds success.'' As such, we hypothesize that joiners' socialization with core developers may follow a nonlinear increasing trajectory.

Hypothesis 1: Joiners' socialization with core developers follows a nonlinear increasing trajectory.

However, this socialization trajectory may not reflect a homogenous pattern between different joiners. First of all, such joiners may start with differing initial levels of social resources. As noted earlier, research on organizational newcomers' socialization suggests that newcomers may join an

organization having different levels of prior social resources (Castilla, 2005). Some newcomers, for example, are referred by existing employees and therefore possess a social relationship with the referrers on joining. Similarly, existing OSS research suggests that newcomers to an OSS project are more likely to have had a prior relationship with specific core developers on the project, implying that they are more likely to initially engage at a higher level of socialization within the project upon joining (Hahn, Moon, & Zhang, 2008). In addition, there is considerable variation in other newcomers' characteristics, such as differing motivations to participate (Roberts et al., 2006), and differing professional and educational backgrounds (Hertel et al., 2003). These differences may also affect the initial level of involvement within the community, including the extent of socialization with core developers.

Second, while joiners all sustain a notable level of socialization activities (Fang & Neufeld, 2009), their growth rates may differ. For instance, those who begin with a significant level of social resources, or professional experience, may become socialized into the core group faster than those who do not. Shah (2006) identified the fact that some developers already possessed software development expertise on joining a community. These more advanced individuals can become more effectively engaged in community discussions immediately on their arrival. And, thus, can develop more social resources over time. In contrast, other joiners may come as strangers who lack project-specific knowledge. Such individuals may stay relatively silent on the developer mailing list, at least for a short initial period, to familiarize themselves with the specific project context and gain other project-specific knowledge (Shah, 2006). Having learned more about the technical details of the project, they would tend to contribute more actively to an ongoing technical discussion as a way of increasing their recognition by core developers (Von Krogh et al., 2003). Thus, we hypothesize:

Hypothesis 2: There are significant differences in joiners' (a) initial level of socialization and (b) growth rate of socialization over time with core developers.

To the extent that joiners' socialization trajectories differ in both their initial levels and their subsequent growth rates, we hypothesize that

Hypothesis 3: Significant heterogeneity exists in the socialization trajectory of joiners and thus distinct classes are identifiable based on this unobserved heterogeneity.

As noted earlier, social resource theory suggests that the possession of social resources positively influences the organizational employees' career outcomes, such as reaching a prestigious job status (Seibert et al., 2001). By the same token, we would expect that certain classes of peripheral developers with higher levels of social resources, which are characterized as having higher initial levels of social resources and high growth rates, can attain core developer status sooner than others. Thus, we hypothesize that

Hypothesis 4: Classes of joiners that have high initial social resources with core developers and that continue to socialize at higher levels with core developers will attain core status sooner than those who do not.

Research Method

Sample Selection

We sampled joiners from OSS projects hosted in Source Forge (SF; http://sourceforge.net), the largest web-based hosting service for OSS projects and a major data source for empirical OSS studies (Colazo & Fang, 2009; Koch & Schneider, 2002; Mockus, Fielding, & Herbsleb, 2002; Newby,

Greenberg, & Jones, 2002). Due to the longitudinal nature of our study, which focused on the temporal patterns of peripheral developers, we needed to sample developers from OSS projects that had the following common dimensions: They must be healthy, mature, and collaborative OSS projects with tractable activity data in both the Concurrent Versions System (CVS) repository and the mailing list. To accomplish this, we followed the approach introduced by Colazo and Fang (2009) that focuses on the projects hosted in SF that met three criteria. First, since our focus is on the joiners' socialization process, the sampled projects must be collaboratively developed. Second, the chosen projects must have been used in some computer architecture other than its original development platform (i.e., "ported"), which functioned as an indication of project maturity (Crowston et al., 2003). Third, they must have activity data that are publicly available in CVS and on the mailing list, because we drew the dependent variable of status progression from CVS and the details of the socialization activities from mailing lists. This effort resulted in 62 OSS projects, which were comprised of 870 joiners (those who eventually became core developers), constituted our sample frame. Two hundred and six of them were successfully identified on both the developer mailing list and the CVS repository and were retained for analysis.

The time taken for the 206 joiners to achieve core developer status (hereinafter termed as "LT") ranged from 1 week to 207 weeks. Of the 206 joiners, 29 were promoted within the first 2 weeks, another 12 in the 3rd week, 15 in the 4th week, 9 in the 5th week, and 8 in the 6th week. Thus, a total of 73 joiners were promoted within the first 6 weeks of joining the list. As we used a 7-week period to model the interaction trajectory, these 73 joiners were not included in our analysis (more information on this decision is provided under the subsection "Model identification"). Thus, our effective sample size is 133.[2,3]

To address the issue of similarities (or differences) between those who were included in the analysis and those who were excluded, we performed a significance test for the means of coding activities between the two groups. We captured the weekly CVS commits once these joiners were promoted to core developer status. Of the total of 870 joiners in the sample frame, we were able to identify 867. We compared the weekly CVS commits of the 206 joiners (after they were promoted to core developer status) to the remaining 661 joiners. The mean CVS commit for Week 1 ($M1 = 9.75$, $M2 = 11.17$) was not significantly different ($F = 0.317$; p value $= .573$) for the two groups. It was the same for Week 2 to Week 7 (with F values ranging from 0.002 to 1.38 and p values ranging between .240 and .966). We also performed significance tests for the means of CVS commits of 133 joiners who were included in the final analysis with the remaining 734. These two groups also did not differ with respect to the CVS commits in any of the first 7 weeks we compared, indicating that our final sample was reasonably unbiased.

Measurement

In this study, we measure the level of socialization at a particular week in terms of the number of joiners' interactions with core developers on the mailing list during that week. If a joiner and at least one core developer were involved in a discussion thread, it was counted as one incidence of socialization. This measure is consistent with that which was adopted in prior research (Ducheneaut, 2005; Fang & Neufeld, 2009). We provide additional information about the measurement of the level of socialization under the subsection, *relevant metrics of time.*

We measure the LT for core status attainment by calculating the time period in weeks between a joiner's first message being posted on the mailing list and his or her first CVS submission.

Analytical Technique and Hypotheses Testing

To test these hypotheses, we need to (a) estimate the initial levels of socialization and the socialization trajectories for each individual developer; (b) identify the classes of joiners based on the

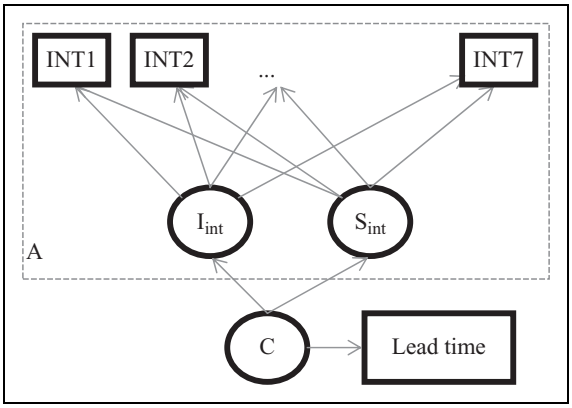

Figure 1. Research model Note: Block "A" represents growth model; INT1 . . . INT7 are cumulative interactions at the end of Week 1 . . . Week 7; I_{int} and S_{int} are intercept and slope latent variables; *C* is latent class variable; Arrow from "*C*" to "lead time" indicate lead time would be different based on class membership.

trajectories (growth patterns) of their interactions with core developers, and (c) examine the differences in the average time required to become a core developer for each class. Thus, Hypotheses 1 and 2 can be tested using the latent curve model (LCM); Hypothesis 3, using latent class analysis of growth trajectories, either GMM or latent class growth analysis (LCGA); and Hypothesis 4, using one-way analysis of variance (ANOVA) or GMM with a distal outcome. We used the Mplus (5.2 version) software for LCM and latent class analysis because Mplus uses generalized SEM frameworks and its implementation is flexible enough to incorporate continuous and categorical variables (Muthén & Muthén, 2007). We used the statistical package for the social sciences (SPSS; version 17) for ANOVA. Figure 1 is a representation of our research model, where block "A" represents the growth model with "measures" INT1 to INT7, which are the cumulative interactions of the joiners with the core developers by the end of Week 1 to Week 7, respectively. I_{int} and S_{int} are the intercept- and slope-latent variables for this growth process. For simplicity, only a single parameter for growth, that is, S_{int}, is shown. A nonlinear growth process may include two (for quadratic growth) or three (for cubic growth) slope-latent variables. *C* is the latent class variable to be estimated using latent class analysis. The arrow from "*C*" to LT indicates that the average LT for each class can be different. This part can be analyzed using ANOVA or GMM with a distal outcome.

The flowchart for the steps involved in the analysis is presented in Figure 2 and is explained in the sections below.

Latent Curve Modeling (LCM). Hypothesis 1 states that the joiners' socialization with core developers follows a nonlinear increasing trajectory. Hypothesis 2 states that there are significant differences in the joiners' initial level of socialization and the growth of their socialization activities over time with core developers. To test these two hypotheses, we use latent curve modeling (LCM). LCM helps the researcher identify the pattern of changes over time by using a set of repeated observed measures to estimate "an unobserved trajectory that gave rise to the repeated measures" (Bollen & Curran, 2006, p. 34). The primary interest is not in the repeated measures themselves but rather in the unobserved path of change, which is referred to as the latent trajectory (Chan, 1998; Collins & Lanza, 2010; MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997). To this extent, LCM resembles the traditional latent variable SEM approach where the indicators of a latent construct are used to gain an understanding of the unobserved construct. LCM models provide an estimate of the random intercepts and random slopes (linear or higher order) for each case (i.e., subject) in the sample so that the trajectories over time for each case can be constructed. As shown in Figure 2, this process

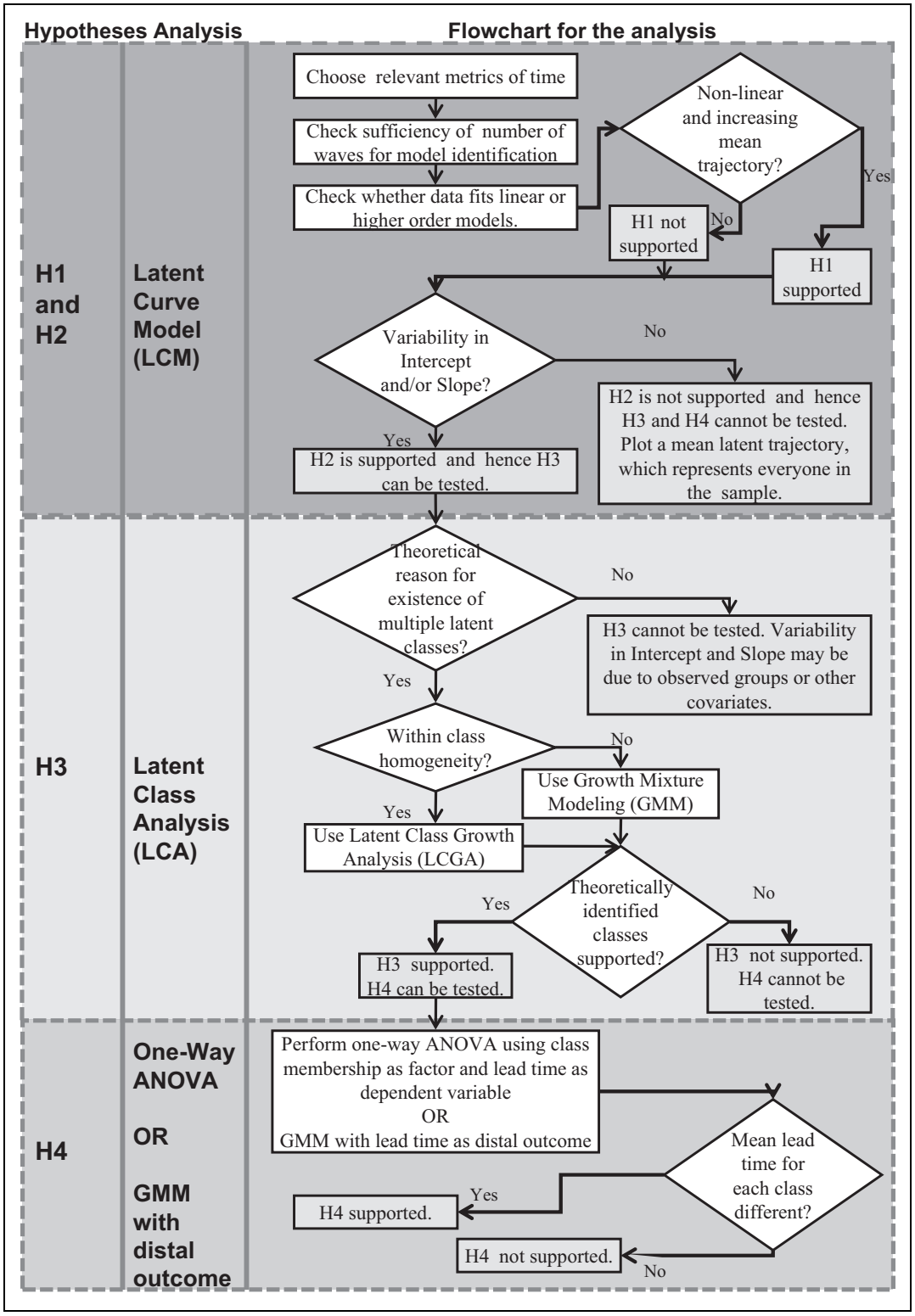

Figure 2. Analysis steps

involves the following major steps: choosing the relevant metrics of time, checking the model identification requirements (i.e., checking the minimum number of "waves" required), testing the fit for linear and higher-order models, and testing the significance of variability in the intercepts and slopes.

If a nonlinear increasing trajectory model shows best fit with the data, then Hypothesis 1 is supported. Whether Hypothesis 1 is supported or not, the next step is to see whether intercepts and slopes have significant variability. If neither intercept nor slope latent variables have significant variability, then Hypothesis 2 is not supported, and all of the cases follow approximately the same trajectory. Thus, there cannot be any unobserved classes based on latent trajectories. Therefore, Hypotheses 3 and 4, which require the existence of a variation in trajectories, cannot be tested. Below, we discuss each step involved in LCM as highlighted in Figure 2.

Relevant metrics of time. There are several issues involved in the selection of the relevant metrics of time. The first issue is the choice of the appropriate unit of time: day, week, month, or year. In some cases, there may be no choice to be made as the unit may be governed by access to data. For example, in the case of annual longitudinal surveys provided by third parties (or government agencies), the unit of time is a year. However, in this study, we had the liberty of choosing the unit of time because we captured mailing list interactions as they actually happened. Although we could have aggregated them on either a daily, weekly, or monthly basis, we used weekly intervals for our study. We chose weeks rather than days as the time unit to avoid the idiosyncrasies associated with a specific day of the week. For example, developers who have full-time jobs may interact more intensely over the weekend than during weekdays. We did not choose the month as an interval because this would have reduced the number of "waves" available for analysis. We will elaborate more on this issue under the section on model identification.

We used cumulative interactions instead of week-to-week interactions for data analysis for two reasons. First, cumulative interaction is aligned with our theorizing with respect to the socialization process. As discussed earlier, we conceptualize that the joiners' socialization is a dual process of developing more social resources on one hand, and tapping into the existing cumulative social resources on the other hand. We argue that it is the dual result of building new and leveraging existing social resources (through cumulative socialization) that is responsible for the joiners' status progression. Second, empirically, the trajectories of cumulative interactions are much easier to model as they follow smooth patterns as compared to those of week-to-week interactions, which might contain spikes.

After the unit of time has been established, the second issue is to decide whether to adopt a chronological (calendar time) order or some other suitable time metric. To explain two possible ways of organizing data for this project, the upper half of Table 1 presents the data structure for the data extracted for this study, which was based on chronological weeks, whereas the lower half of Table 1 presents the same data but is restructured on the basis of the number of weeks after joining OSS mailing lists.

In the upper half of Table 1, W1, W2 ... W208 refer to the chronological weeks beginning at the start of the data collection period (November 1999). This is an arbitrary start date and does not coincide with any important event of interest. A, B, ... G are randomly chosen peripheral developers. The cell values indicate the number of weeks since joining the mailing list and their cumulative interactions with the core developer. For example, the top-left corner cell contains the value 1/2; 1 in this case represents the joining week and 2 indicates the cumulative interactions. The joining weeks, in the upper half of this table, are shown for easy comparison with the lower half; the actual data set, however, need not contain this information. Cells containing "P" (say 21/P) indicate the number of weeks (21 in this case) required for promotion since joining the mailing list. Such data structure may be useful when there is a chronological event of importance. For example, if a researcher was interested in understanding the effect of the dot com bubble burst on OSS developers'

Table 1. Data Structure

Data Structure Based on Chronological Weeks[a]																				
ID	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	...	W52	W53	...	W104	W105	...	W206	W207	W208
A	1/2	2/2	3/3	4/3	5/4	6/5	7/7	8/9	9/11	10/14	11/19	21/P									
B				1/6	2/11	3/16	4/21	5/27	6/33	7/38	8/P										
C						1/0	2/1	3/2	4/2	5/3	6/3	...	47/35	48/P							
D													1/0	2/0	...	53/32	54/P				
E																		...	10/47	11/P	
F		1/0	2/0	3/0	4/0	5/1	6/1	7/1	8/1	9/1	10/2	...	51/4	52/4	...	103/7	104/7	...	205/26	206/30	207/P
G											1/0	...	42/18	43/18	...	94/43	95/P				

Data Structure Based on Weeks After Joining OSS Mailing Lists[b]																					
ID	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	...	w52	w53	...	w104	w105	...	w206	w207	w208
A	2	2	3	3	4	5	7	9	11	14	19	21/P									
B	6	11	16	21	27	33	38	8/P													
C	0	1	2	2	3	3	4	4	5	6	6	48/P									
D	0	0	1	2	2	2	3	3	4	4	4	...	30	32	54/P						
E	3	5	9	13	18	23	28	34	40	47	11/P										
F	0	0	0	0	1	1	1	1	1	2	2	...	4	4	...	7	7	...	30	207/P	
G	0	0	1	1	1	2	2	2	2	2	3	...	6	6	95/P						

Note: [a]W1, W2 ... W208 refers to chronological weeks from the beginning period of data extraction (November 1999); A, B ... G are peripheral developers; Cell values (say 1/2) indicate the number of weeks since joining the mailing list (1 in this case) and cumulative interactions with the core developer (2 in this case); the cell containing "P" (say 21/P) indicates the number of weeks (21 in this case) required for the promotion since joining the mailing list. [b]w1, w2 ... w208 refers to weeks since joining the mailing list. A, B ... G are, respectively, the same peripheral developers as shown in the upper half of the table; Cell values indicate cumulative interactions with the core developer; the cell containing "P" (say 21/P) indicates the number of weeks (21 in this case) required for the promotion since joining the mailing list.

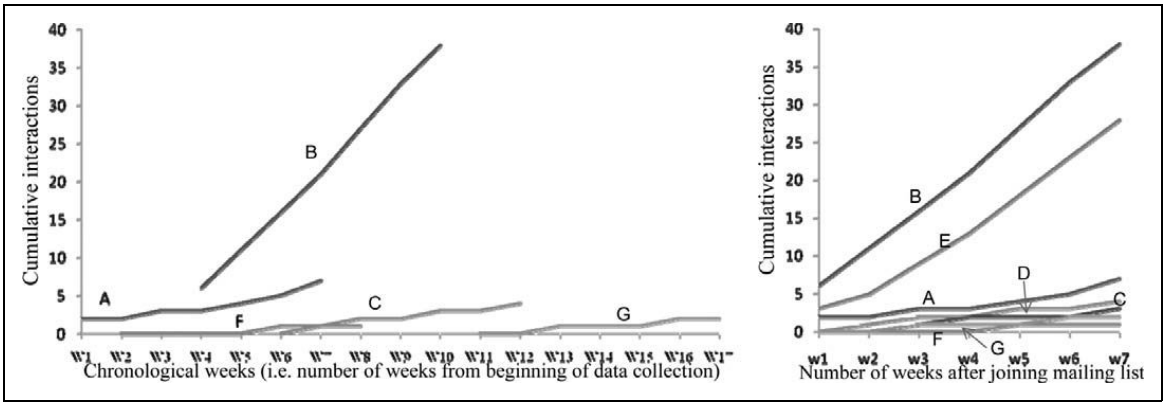

Figure 3A. Cumulative interaction trajectories in chronological weeks. B. Cumulative interaction trajectories in joining weeks.

interaction patterns over a particular period, then such a data structure could be useful. These data are represented graphically in Figure 3a. For convenience, only the first 17 chronological weeks are plotted. The trajectories of developers D and E are not shown as these developers make their first appearance in Weeks 52 and 197, respectively. For easy comparison with Figure 3b, Figure 3a shows the interaction trajectories for the first 7 weeks after joining.

The lower half of Table 1 presents the data in the format that was used for this study. In this table, w1, w2 ... w208 refer to the number of weeks passed for each developer since joining the mailing list. A, B ... G are, respectively, the same peripheral developers as shown in the upper half of Table 1. The cell values indicate the cumulative interactions with core developers (2 in the case of the top left corner cell). The cell containing "P" (say 21/P) indicates the number of weeks (21 in this case) required for promotion since joining the mailing list. Figure 3b shows interaction trajectories for the first 7 weeks, data from the lower half of Table 1. As we were interested in understanding the interaction trajectories after a developer joined the mailing list and its effect on LT, this data structure best suits our requirement.

Model identification. A minimum of three "waves" are required for the identification of a linear LCM (refer to Bollen & Curran, 2006, for an excellent treatment of this topic). Quadratic and cubic LCM exerts a higher demand on the number of "waves." In addition to the model identification requirements, we were also careful about including enough "waves" to capture any latent trajectories. Thus, we decided to use a cutoff of 7 weeks; that is, we included only those joiners whose LT were 7 weeks or more. This step reduced the final sample to 133 peripheral developers spanning over 40 projects. A lower cutoff would have created a model identification problem, while a higher one would have reduced our sample size even further.[4]

The frequency distribution of the LT for these 133 peripheral developers is shown in Table 2. More than 50% of the joiners were promoted within the first 25 weeks of joining the mailing lists.[5] Table 3 provides information about the means and standard deviations for the first 7 weeks of cumulative interactions (INT1 ... INT7) and also for the LT. This table also contains the correlation of the variables used. The correlations among the cumulative interaction variables (INT1 ... INT7) reflect typical time-dependent patterns; that is, the shorter the time lag between the measurements, the higher the correlation (Bliese & Ployhart, 2002; Holcomb, Combs, Sirmon, & Sexton, 2010). As expected, LT has a negative correlation with all the measurements of cumulative interactions; that is, the higher the number of cumulative interactions, the shorter the LT.

As our data were obtained from 40 related OSS projects, we were concerned about clustering issues. We obtained intraclass correlation coefficients (ICC) for all the observed variables used in

Table 2. Frequency Distribution of Lead Time for Status Attainment

Lead Time (weeks)	Number of Developers Promoted
7–25	70
25–50	32
51–75	12
76–100	7
>100	12

Table 3. Mean, Standard Deviation (*SD*), and Correlations

	Mean	*SD*	INT1	INT2	INT3	INT4	INT5	INT6	INT7
INT1	2.98	2.68							
INT2	6.15	5.94	.74***						
INT3	10.01	10.64	.81***	.98***					
INT4	15.79	18.13	.88***	.89***	.96***				
INT5	20.65	22.2	.85***	.65***	.77***	.91***			
INT6	26.79	26.1	.65***	.23**	.39***	.63***	.89***		
INT7	34.27	30.78	.40***	−0.12	.060	.33***	.68***	.94***	
LT	36.87	35.3	−.75***	−.33***	−.43***	−.57***	−.69***	−.68***	−.57***

Note: LT = lead time in weeks. INT1 ... INT7 are cumulative interactions of peripheral developers with core developers, respectively, at Week 1 ... Week 7 after joining the mailing list.
**$p < .01$. ***$p < .001$.

this study and calculated the design effects using the formula suggested by Hox and Mass (2002, p. 5). The ICC and design effects are presented in Table 4. All of the design effects are smaller than "2" indicating that analyzing data at a single level can result in acceptable parameter estimates and inferential tests (Hox & Maas, 2002).

Estimation of LCM. In this step, various LCMs are tested to check which one has the best fit. We tried to fit the linear (LCM1), quadratic (LCM2), and cubic (LCM3) models, as the rate of change may vary over time (Chan, 1998). Table 5 provides the model fit indices (Confirmatory Fit Index [CFI], Tucker-Lewis Index [TLI], and root mean square error of approximation [RMSEA]) for these models. The LCM1 model has a very poor fit (CFI = .562, TLI = .600, and RMSEA=.842). The model fit for LCM2 (CFI = .804, TLI = .784, and RMSEA = .619) is better than LCM1 but still inferior to the standards presented in the SEM literature (e.g., Hu & Bentler, 1999). Based on the model fit indices, it can be concluded that LCM3 is the best fit model (CFI = .992, TLI = .987, and RMSEA = .12) for interaction trajectories. CFI and TLI are both superior than the recommended level (>.95), whereas RMSEA is inferior than the recommended level (<.06).[6]

Table 6 provides information about the means and variances in the LCM3 model. All of the mean trajectory parameters (i.e., intercept, linear, quadratic, and cubic) differ significantly from zero and all of them are positive. Thus, the mean trajectory has a nonlinear shape with increasing growth; hence Hypothesis 1, which stated that on average joiners' socialization with core developers follows a nonlinear increasing trajectory, was supported. Figure 4a shows this mean trajectory graphically.

Table 6 also provides information about the variances in intercepts and slopes. For LCM3, there is a significant variance in the intercepts (i.e., the initial level of the interactions) for the process of

Table 4. Intraclass Correlations (ICC) and Design Effect (Deff)

	ICC	Deff [1+(k − 1) × ICC]
INT1	0.145	1.337
INT2	0.147	1.341
INT3	0.152	1.354
INT4	0.166	1.387
INT5	0.180	1.419
INT6	0.188	1.437
INT7	0.181	1.421
PT	0.294	1.685

Table 5. LCM Models

				Variance			
Models	CFI	TLI	RMSEA	Intercept	Lin	Quad	Cubic
LCM1	.562	.600	.842	26.82***	36.58**	—	—
LCM2	.804	.784	.619	21.57***	30.28***	.65***	—
LCM3	**.992**	**.987**	**.12**	**21.92*** **	**21.82*** **	**.06** **	**.012*** **

Note: LCM1, LCM2, and LCM3 represent linear, quadratic, and cubic Latent Curve Models, respectively; LCM3 is the best fit model. "—" indicates that quadratic and cubic parameters are not required for LCM1, and cubic parameter is not required for LCM2. The numbers in bold provide information about the best-fit model, i.e. LCM3.
**p < .01. ***p < .001.

Table 6. Mean and Variance of Growth Parameters for Accepted LCM3 Model

	Mean		Variance	
	Estimate	*SE*	Estimate	*SE*
Intercept	7.005***	0.402	21.915***	2.670
Linear	1.834***	0.403	21.821***	2.675
Quadratic	0.275***	0.026	0.059**	0.020
Cubic	0.030**	0.010	0.012***	0.002

Note: LCM3 = cubic latent curve models; *SE* = standard errors.
**p < .01. ***p < .001.

interactions with core developers (var(I_{int}) = 21.92, p < .001). All three components of the slope (i.e., linear [var(L_{int}) = 21.82, p < .001], quadratic [var (Q_{int}) = .06, p < .01], and cubic [var(C_{int}) = .012, p < .001]) for the peripheral developers' interactions with core developers have significant variations. Thus, Hypothesis 2 was supported, and we can proceed with testing Hypothesis 3 and then Hypothesis 4.

Latent Class Analysis. The objective of such an analysis is to capture information about interindividual differences in the intraindividual cumulative pattern of interactions (Morin, Morizot,

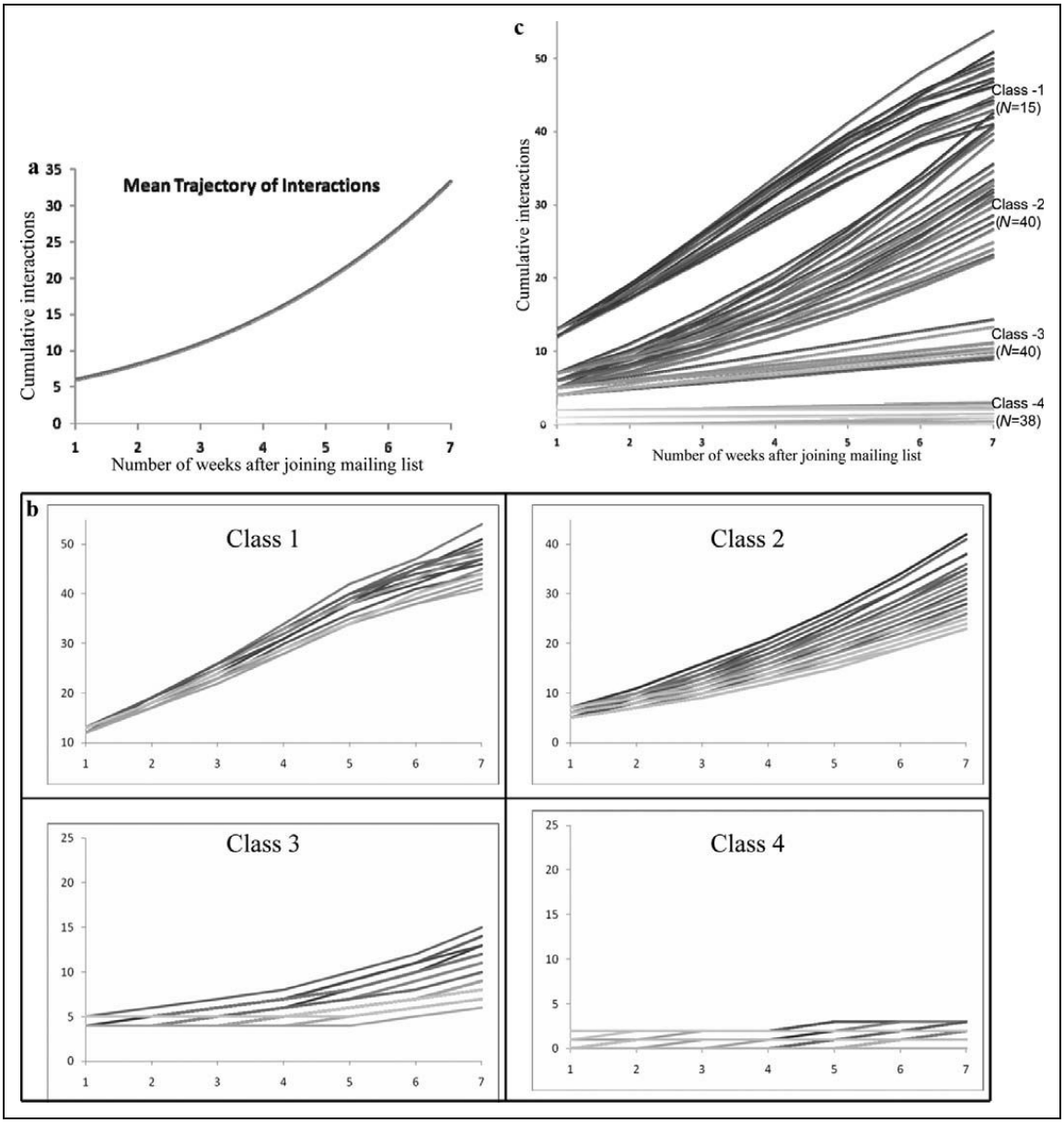

Figure 4A. Mean trajectory of interaction with core developers. B. Individual trajectories within each class. C. Latent trajectory classes for interaction with core developers.

Boudrias, & Madore, 2010; Muthén & Muthén, 2000; Nesselroade, 1991). Such a technique is useful when the observed differences in the patterns are a result of the unobserved heterogeneity of the subject population (Muthén & Muthén, 2000; Nagin, 1999; Wang & Chan, 2010). This heterogeneity in the observed interaction patterns may emerge from the unobserved difference among the developers toward, for example, utility, convenience, ease of use, and other aspects of their interactions within the OSS environment.

Population heterogeneity, such as gender, race, education, and organizational designation, is either observable or available from archival records and thus can be explicitly represented by variables used in a model. When population heterogeneity is unobservable, however, it cannot be accounted for in a model using simple regression or SEM techniques. Nevertheless, a latent

class analysis framework can take this condition into account using latent classes in the model (Muthén & Muthén, 2000; Nagin, 1999; Samuelsen & Dayton, 2010). This is achieved through a use of categorical latent variables that represent latent classes (i.e., unobserved heterogeneity). In latent class analysis that involves growth trajectories, each latent class corresponds to an unobservable subpopulation that has its own growth trajectory, which is defined by a set of parameter values. This analysis can be performed either through LCGA or GMM. Figure 2 shows the various steps involved in performing a LCA of growth trajectories. These are establishing a theoretical basis for the existence of multiple latent classes, choosing to use either LCGA or GMM, and identifying the resulting latent classes. Each of these steps is described below.

Theoretical justification for the classes. Do the latent classes exist, and if they do, how many classes are there?[7] This is not a trivial issue. There is no agreement in the literature with respect to the decision to identify classes a priori based on theory or a posterior based on empirical analysis. Jung and Wickrama (2008) suggest that there should be at least some theoretical justification for the existence of unobserved classes, and they should not be based simply on various fit indices. In the absence of a theoretical justification, however, the existence of multiple classes may simply be due to skewed or otherwise nonnormally distributed data (Bauer & Curran, 2003). However, others believe that latent classes should be extracted empirically rather than be based on theoretical justification (Nagin, 2005). This view is clearly reflected in the work of Luyckx and colleagues.

> Trajectory classes are empirically defined based upon the longitudinal trends—in terms of initial level and rate of change—present in the data. In other words, we did not impose a theoretically derived structure that may or may not fit the data, because such a strategy threatens the statistical validity of the results. (Luyckx, Schwartz, Goossens, Soenens, & Beyers, 2008, p. 599)

However, a consensus of opinion is emerging in the field. Wang and Bodner (2007) suggest that the use of a single theoretical lens might obscure the presence of latent classes and suggest that multiple theoretical lenses are required to appreciate the presence of latent classes and to hypothesize about their antecedents and outcomes. Even after using multiple theories, it may not be possible to hypothesize about the presence of all the latent classes and identify their growth patterns. Thus, the determination of a number of classes may need a combination of such factors as fit indices, "research question, parsimony, theoretical justification, and interpretability" (Jung & Wickrama, 2008, p. 311).

Our position in this article is to start with the existence of latent classes based on theoretical considerations. If the existing theory is not adequate to predict the number of classes, researchers should be open to interpreting the empirical results in light of the existing literature. As noted earlier in this article, we base the identification of the heterogeneity of socialization trajectories on social resource theory combined with the current OSS literature. We use the GMM method to identify the exact number of classes and hypothesize about the relationship between these classes with the dependent variable being based on social resource theory.

GMM or LCGA. GMM builds on LCM in a sense that if there are no variations in either the initial levels or in the slopes of the trajectories, then there is no possibility of classifying them into different classes. GMM represents a latent class analysis in which the latent classes correspond to differences in growth trajectories for a repeatedly measured outcome variable. For example, in a two-class model, one class may have a high intercept and a moderate linear growth, while the other may have a low intercept but a quadratic growth. The objective of the analysis is to estimate the different growth curve patterns, and based on these patterns, estimate the posterior probabilities of the class membership of each individual (Muthén, 2001, 2008).

Table 7. Fit Indices for Latent Class Growth Models

					Class Membership (%)				
#Classes	AIC	BIC	SABIC	Entropy	C1	C2	C3	C4	C5
2-Class	5836.46	5874.33	5833.20	0.98	59	41	–	–	–
3-Class	4892.14	4938.74	4888.13	1.0	30	11	59	–	–
4-Class	**4214.95**	**4276.11**	**4209.68**	**1.0**	11	30	30	29	–
5-Class	4249.99	4305.33	4245.23	0.99	29	30	11	26	4

Note: AIC = Akaike information Criteria; BIC = Bayesian Information Criteria; SABIC = Sample size adjusted BIC. The numbers in bold indicate that the 4-class solution had best fit-indices.

LCGA also uses a similar technique; however, it additionally assumes that there is no variability in the intercepts and slopes among the members within the same latent class. Thus, it assumes that there is within-class homogeneity. If a researcher had a theoretical reason to assume such within-class homogeneity, then LCGA would be a suitable technique; otherwise, GMM should be used. However, it may be a good idea to use LCGA initially and then proceed to test GMM for two reasons. First, under normal conditions, there may be no way of ascertaining the a priori presence or absence of within-class homogeneity. Second, LCA of growth trajectories is data-intensive, and estimating all the parameters (as in case of GMM) may create model identification issues, especially if the number of waves is limited. In LCGA, within-class variances in the intercept and the slope are fixed at zero. This makes the LCGA model relatively simple, and hence the likelihood of model identification is higher than that in GMM.

Estimation of GMM/LCGA. To determine the optimal number of classes, the LCGA models for 2-, 3-, 4-, and 5-classes were analyzed. Several criteria were used to determine the number of classes (Muthén & Muthén, 2000; Nagin, 2005). Table 7 shows the various fit indices. Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and sample size adjusted BIC are interpreted in the same way (Nylund, Asparouhov, & Muthén, 2007). These statistics should be lower for a solution with class k as compared to that for class $k - 1$, indicating that the addition of a class improves the model fit (Luyckx et al., 2008; Nagin, 2005). However, one should not rely only on information criteria, as AIC and BIC are affected by the number of parameters used in the model; in addition, BIC is also affected by the sample size (Wang & Bodner, 2007). Thus, "in selecting growth mixture models, information criteria should be considered [along] with other evidences" (Wang & Bodner, 2007, p. 642). Entropy is another commonly used indicator of classification quality and it ranges from 0.0 to 1.0, where 1.0 represents a better classification (Hix-Small, Duncan, Duncan, & Okut, 2004; Jedidi, Ramaswamy, & Desarbo, 1993; Nagin, 1999) in the sense that there is clear delineation between classes (Celeux & Soromenho, 1996). It is a "standardized summary measure of classification accuracy of placing individuals into trajectory classes based on the posterior probabilities of classification" (Luyckx et al., 2008, p. 606). Entropy provides an assessment of whether individuals are classified into "one and only one category" (Greenbaum, Del Boca, Darkes, Wang, & Goldman, 2005; Muthén, 2004). Based on previous research, Wang and Bodner (2007) concluded that entropy values higher than 0.80 can be viewed as an indication of a good classification.

These fit indices indicate that LCGA yielded four classes (Table 7) as the 4-class solution had a better fit than either the 3-class or the 5-class solution. The class membership, based on posterior probability, for the 4-class solution was reasonably spread (11%, 30%, 30%, and 29%); that is, none of the classes was too small to require its exclusion. The parameter estimates of these classes are provided in Table 8. The mean for the intercept and the growth of the latent variables for all the classes differs significantly from zero. Class 1 has a cubic trajectory, Class 2 has a quadratic trajectory, and both Classes 3 and 4 have linear trajectories.[8]

Table 8. Parameter Estimation of Latent Growth Factors for 4-Class Latent Class Growth Model

	Estimates	*SE*
Class 1 (15)		
Mean of growth factors		
Intercept factor	12.70***	0.20
Linear factor	12.03***	0.30
Quadratic factor	2.57***	<0.01
Cubic factor	3.24***	<0.01
Class 2 (40)		
Mean of growth factors		
Intercept factor	6.06***	0.20
Linear factor	6.96***	0.26
Quadratic factor	2.97***	<0.01
Class 3 (40)		
Mean of growth factors		
Intercept factor	4.90***	0.09
Linear factor	1.17***	0.11
Class 4 (38)		
Mean of growth factors		
Intercept factor	0.86***	0.14
Linear factor	0.10***	0.03

***$p < .001$.

We plotted the trajectories of the members within each class and found that there were visible variations in either their intercepts or their slopes or both (Figure 4b). Thus, the condition for within-class homogeneity was not met, and we decided to implement GMM for the final results. The fit indices for the 2-, 3-, and 4-class GMM analysis are shown in Table 9. The GMM with more than four classes suffered from an identification problem, and in most of the cases, the solutions did not converge even after repeated changes in the starting values. Where the solutions converged, the subjects were placed in four classes.

Table 9 provides information on the AIC, BIC, *N*-adjusted BIC, and Entropy for the 2-, 3- and 4-class GMM. The 4-class GMM had the best fit as compared to the 2-class and 3-class solutions. Thus, the 4-class solution was accepted.[9] The estimates of mean and variance for the intercept and the slope latent variables for each of these classes are shown in Table 10.

These four classes—Class 1, Class 2, Class 3, and Class 4 are shown in Figure 4c. They contain 15, 40, 40, and 38 members, respectively. These four classes are clearly identifiable based on the intercept and slope of their growth trajectories. Members of Class 1 have higher initial levels of interaction with core developers, whereas members of Class 2 and Class 3 have moderate, and Class 4 have lower initial levels of interaction. The growth trajectories also differ for each of these classes. Members of Class 1 have a consistently higher growth rate, members of Class 2 initially have moderate growth and then a higher growth rate; whereas members of Class 3 have moderate growth; and members of Class 4 have a consistently lower growth rate. To the extent that significant heterogeneity exists in the socialization trajectory of joiners, and that distinct classes are identifiable based on this unobserved heterogeneity, Hypothesis 3 was supported.

Relationship of Classes to LT. The next step is to establish whether each of the identified classes differs in terms of LT. We followed the Jung and Wickrama (2008) suggestions that these identified

Table 9. Fit Indices for Growth Mixture Models

					Class Membership (%)				
#Classes	AIC	BIC	SABIC	Entropy	C1	C2	C3	C4	C5
2-Class	3,104.96	3,171.95	3,099.19	0.95	32	68	–	–	–
3-Class	3,048.71	3,124.44	3,042.19	0.99	30	11	59	–	–
4-Class	**2,980.79**	**3,071.08**	**2,973.01**	0.99	11	30	30	29	–
5-Class	Not identified								

Note: AIC = Akaike information Criteria; BIC = Bayesian Information Criteria; SABIC = Sample size-adjusted BIC.

Table 10. Parameter Estimation of Latent Growth Factors for 4-Class Growth Mixture Model

	M		Variance	
	Estimate	*SE*	Estimate	*SE*
Class 1				
Intercept factor	12.93***	0.23	2.41*	2.12
Linear factor	11.96***	0.31	4.89***	1.4
Quadratic factor	2.63***	<0.01	.84**	.29
Cubic factor	3.19***	<0.01	.09+	.048
Class 2				
Intercept factor	6.14***	0.19	2.77**	.98
Linear factor	6.79***	0.24	5.2***	1.19
Quadratic factor	2.89***	<0.01	0.12	.07
Class 3				
Intercept factor	4.68***	0.10	2.51*	1.2
Linear factor	1.38***	0.13	1.44*	.62
Class 4				
Intercept factor	0.87***	0.14	.59+	0.305
Linear factor	0.11***	0.04	.06+	0.031

+p < .1. *p < .05. **p < .01. ***p < .001.

classes can be used as variables "for further analyses, such as conducting a test of mean differences across the classes on the covariates using ANOVA, or using class membership as a predictor for distal outcome" (p. 316). Table 11 presents the mean LT for each of these classes. The mean LT for Class 1 is the lowest (7.5 weeks), and that for Class 4 is the highest (83.2). Using One-Way ANOVA, we found that the mean LT for each class differed significantly from all of the other three classes. Thus, Hypothesis 4 was supported.

It is important to note that if class membership is used as a variable in analysis, then the fact that class membership is based on a probabilistic model rather than on a deterministic model (i.e., class membership is based on posterior probabilities) is not taken into account in further analyses, for example, as in ANOVA[10] (Petras & Masyn, 2009). The implementation of GMM with a distal outcome framework, where LT is a distal outcome, takes into account the probabilistic nature of class membership. This framework takes into account the fact that a distal outcome may have different means in different classes (Muthén, 2008).[11]

Our analysis of GMM with a distal outcome also resulted in four classes with the same class membership (11%, 30%, 30%, and 29%) as that of GMM solution with the distal outcome. We

Table 11. Mean Lead Time for Status Progression in Weeks for Various Classes

Class	Mean lead time
Class 1	7.5***
Class 2	13.4***
Class 3	28.7***
Class 4	83.2***

*** indicates "mean lead time" for this class is significantly ($p < .001$) different from mean lead time of all other three classes.

attribute this similarity to very high entropy (.998) and the average posterior probabilities (ranging from .978 to 1). Estimated means for the classes were 7.5 (C1), 13.4 (C2), 28.7 (C3), and 83.2 (C4). These means are same as those obtained using ANOVA analysis. The Wald test (898.9, $df = 3$) indicated that the means are significantly different from each other (p value < .001). Thus, consistent with our expectation, the results of ANOVA and GMM with a distal outcome do not vary and provide statistical conclusion validity for Hypothesis 4.

Discussion and Conclusion

The current study is designed to study how OSS joiners' socialization patterns relate to their status progression. This issue was understudied in the prior research but is of vital importance to the sustainability and even the survival of OSS projects. To measure status progression, we focused on the LT taken before joiners switched to core developer status. Drawing on social resource theory and the existing OSS literature, we suggest that joiners' socialization patterns with core developers could be nonlinearly increasing and would vary across different peripheral developers, which would in turn affect their status progression within the community.

The empirical results provided general support to the hypotheses. First, using the joiners' cumulative interaction patterns with core developers to measure the level of socialization, we find that joiners' socialization with core developers, as shown in figure 4a, generally follows a nonlinear growth trajectory. Second, we find that individual joiners begin with different initial levels and follow different growth patterns, suggesting the existence of heterogeneity in the socialization trajectories. Third, confirming such heterogeneity, we empirically identify four latent trajectory classes of socialization behavior, that is, the initial level and growth rates: (a) high, high; (b) moderate, high; (c) moderate, moderate, and (d) low, low. Finally, we find strong support for the theory that these latent trajectories classes are associated with the different periods of time taken to attain core developer status. Figure 5 illustrates the four distinct classes and their respective average LT for core developer status attainment.

To discuss further, we must highlight the exploratory nature of this study. As research on joiner socialization in the OSS context is very limited, theoretical insights in this area are just emerging (Von Hippel & Von Krogh, 2003). The prior qualitative research is mainly exploratory, with the aim of building well-grounded theory or providing descriptive insights into the socialization behavior (Ducheneaut, 2005; Von Krogh et al., 2003). We believe that at this point there is a need for more *informed* exploration, with plausible theories being drawn from other related contexts. In this research, we base our scholarly exploration on social resource theory that originated within the organizational contexts and find that significant heterogeneity exists in the joiners' socialization patterns. This empirical finding challenges the implicit assumption made in earlier exploratory studies that the joiners' socialization process is homogenous. We also find that socialization patterns strongly affect the joiners' status progression, a result that should encourage further research in the area.

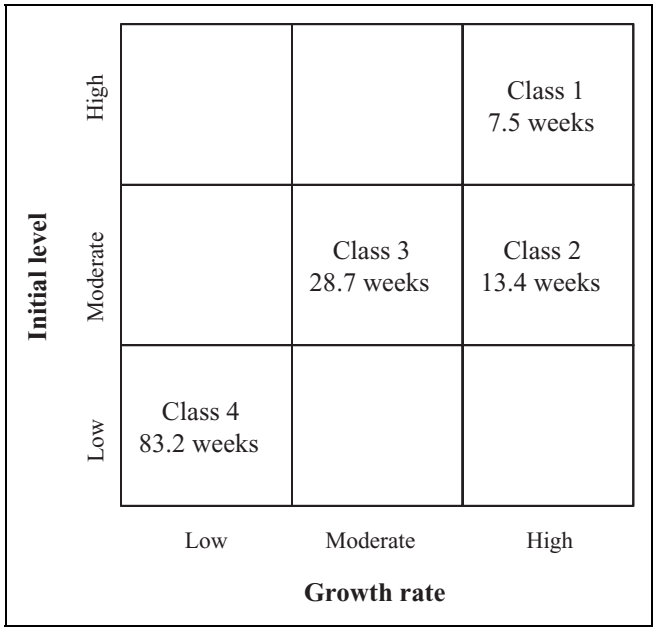

Figure 5. Latent trajectories of classes and average lead time for status attainment

Implications to Theory

From a theoretical point of view, our results suggest several important points for theory development with regards to the role of socialization within OSS contexts. First, it is important to recognize that socialization with core developers has a significant impact on joiners' status progression. The existing research is primarily focused on the intensity and type of socialization in OSS communities as being instrumental to the developers' status progression (Ducheneaut, 2005; Fang & Neufeld, 2009; Von Krogh et al., 2003). In understanding socialization behavior from the perspective of social resources, our study suggests that it is also important to recognize the target toward which such socialization behavior is aimed. Core developers at higher organizational levels have considerable control over the joiners' status progression. By socializing with them, joiners may gain access to and grow to appreciate useful information, receive social support, and solicit sponsorship for initiatives, in the same way that newcomers do in traditional organizations (Seibert et al., 2001). Our empirical finding is largely consistent with existing theoretical development, and future research can explore the mechanisms behind the socialization process more thoroughly.

Second, our study reveals that the joiners' socialization trajectories generally follow nonlinear growth patterns and are heterogeneous. These results are consistent with our theoretical prediction. Hypothetically, there could be nine trajectory classes when we categorize the initial level and the growth rates of socialization as being low, moderate, and high.[12] Yet, we identify only four latent socialization trajectories classes, each with different initial levels and grow rates. As shown in Figure 5, three of the four classes (1, 3, and 4) have growth rates consistent with their corresponding initial levels, implying that a joiner's socialization process may be path dependent, that is, the incremental growth of future socialization may largely depend on the level of current socialization (even after taking into account the cumulative nature of our socialization variable). To the extent that socialization represents a process of developing social resources, this finding implies that the development of social resources in OSS contexts complies with the law of asset mass efficiency: that one could increase the increment added to an existing stock of resources, if one possesses an already existing high level of that stock (Dierickx & Cool, 1989). This strong path-dependent effect offers

Table 12. Descriptive Statistics for Cumulative Interactions that were Achieved in Each Class Prior to Change in Status

Classes	*M*	*N*	*SD*	Min	Max	Median
1	70.20	15	27.92	46	124	51.0
2	111.55	40	37.50	54	180	112.5
3	209.12	40	34.85	132	276	212.0
4	314.47	38	58.30	180	390	325.0

an alternative explanation for the factors influencing developer status progression in OSS communities. Although prior research suggests that it is important for joiners to manage the *process* of socialization (Ducheneaut, 2005; Fang & Neufeld, 2009; Von Krogh et al., 2003), our quantitative results suggests that the *initial condition* for socialization is critical.

However, we do observe an interesting exception. It seems that peripheral developers in Class 2 begin with a moderate level of socialization, followed by maintaining a high level of growth that deviates somewhat from the path-dependency effect. The average LT for status upgrade of this group is 13.4 weeks, which is shortened by half in comparison to that of Class 3 (28.7 weeks). We believe that the distinction of this specific class is likely to be real, rather than random, and research efforts focused on providing theoretical explanations to this distinction would be fruitful. To pursue this further, we conducted a post hoc analysis and did not identify any significant differences in this group of peripheral developers in terms of project membership, suggesting that project-level factors might not play a role. This leads us to conjecture that, at the individual level, it is likely that members of Class 2 might have conducted different types of socialization activities that distinguished them from those in Class 3. These individuals might be particularly successful in implementing "joining scripts" (Von Krogh et al., 2003) or strategically prioritizing the work to which they subscribed within the OSS community (Ducheneaut, 2005). If this is true, the distinction of Class 2 would suggest that, although initial possession of social resources could have a path-dependent effect on peripheral developers' status progression, developers could improve this progression by focusing on certain types of socialization activities.

Third, and in addition to growth pattern of interactions, it would be of interest to examine whether there is a threshold on cumulative interaction that leads to status progression. Through a post hoc analysis, we calculated the total number of interactions (i.e., cumulative interaction) for each joiner up to the time they were promoted to the status of core developer. We found that there was a huge variation in the number of cumulative interactions (min = 46; max = 390; mean = 193; and standard deviation = 98.55). Table 12 provides a classwise mean and standard deviation for the total interactions. A comparison of Table 12 with Table 11 indicates that those who had high initial levels and who follow high growth trajectories (i.e., Class 1) are not only promoted to core developer status sooner (i.e., have a shorter LT) but also they spend fewer social resources (as the mean of their total interactions is significantly lower than the mean of the members of other classes). We found a significant correlation between LT and the total number of interactions ($\gamma = 0.79$, $p < .001$), indicating that, as LT increases, the total amount of social resources (effort) required to become a core developer also increases. This is an important empirical finding. Prior research has suggested that a high level of interactions in an OSS community is essential for a peripheral developer's status change (Von Krogh et al., 2003), without indicating the relationship between the level of interactions and the LT for status change. Our result takes this line of research further by showing that it is perhaps more important to achieve a high level of interaction as quickly as possible.

Finally, it would also be interesting to understand why the other five quadrants do not capture any latent classes. Although it is plausible that individuals with a high initial level of

socialization but with less follow-up socialization might be disappointing to core developers (and therefore can never become joiners), it requires more research into why joiners with low initial levels of socialization are unable to achieve either moderate or high growth rates. One explanation might be that these people come with insufficient domain-specific knowledge and must engage in painstaking learning and purposeful lurking behavior to accumulate sufficient social resources for status progression. If so, the significant knowledge gap might explain the challenge of accelerating socialization, particularly with core developers who demand intellectual and meaningful input from participants. Theoretical efforts to develop such ideas would further enrich the understanding of the factors promoting the socialization process in OSS contexts.

Implications to Method

Methodologically, to our knowledge this is among the few studies in the management discipline that empirically investigates a phenomenon using GMM through a SEM framework. Our literature review indicated that there were only two other studies that use GMM or related techniques within the management discipline. The first study, which was conducted by Wang and Bodner (2007) and Wang (2007), presents an informative account of how to use GMM for identifying and predicting unobserved populations present in longitudinal data. They implement GMM through the SEM framework.

Second, a study by Holcomb et al. (2010) uses random coefficient modeling, which is based on multilevel empirical data. However, they use the SAS framework for multilevel and mixed models (Singer, 1998). Although there is merit in using "an integrated program . . . to perform data reduction, management, and analysis of multilevel longitudinal data within a single statistical package" (Holcomb et al., 2010, p. 4), we concur with the view that the SEM framework provides a more generalized approach to GMM because it allows for the use of latent variables as repeated measures, mediation models, and simultaneous estimation of multiple growth processes (Chan, 1998; Muthén, 2008; Wang & Bodner, 2007).

We provide a nontechnical step-by-step flowchart of how to perform GMM analysis starting with a guideline on how to choose a relevant metric of time and how to ensure that there is a sufficient number of waves for model identification. Our empirical example also provides insight into model identification issues related to GMM. It can be seen from Table 7 (LCGA models) and Table 9 (GMM models), that the Class 5 solution is identifiable for LCGA models but not for GMM models. This is because GMM models require a higher number of parameter estimation as compared to corresponding LCGA models. Thus, our recommendation is that even if there is reasonable evidence of within-class heterogeneity, it is worthwhile to first make an estimate using LCGA models before proceeding to GMM models.

Limitations, Future Research, and Conclusion

Our study has limitations that provide avenues for future research. First, as noted earlier, while our study discovered four distinct latent classes of joiners, future research should focus on investigating potential antecedents to the formation of these latent trajectory classes using qualitative or survey methods. For instance, in-depth interviews or message analyses could be conducted on developers in each class to gain a better understanding of the theoretical reasons behind the observed socialization behavior that differs across the latent classes, such as the motivation to participate, individual backgrounds (e.g., professional and educational backgrounds), and socialization dynamics. Second, future research can investigate the effect of peer support on OSS developer status progression by focusing on socialization with other peripheral developers. Although peripheral developers do not

represent higher hierarchical levels, they may provide other types of benefit. Third, future research may also investigate network positions of peripheral developers and the resulting status outcomes through a social network lens. Fourth, in our study, we measure socialization behavior based on the frequency of interactions. Future research could strengthen this measure by coding the different types of socialization behavior (e.g., buy-in, task advice, or organizational information) and examining their differential effects on status progression. Finally, we believe that future research may replicate this study using a larger sample size.

In conclusion, it is our hope that our initial results will encourage researchers who are studying the open source movement to embrace the social resource perspective and the GMM method and that researchers in the social and managerial disciplines will focus on this domain to provide richer insights into OSS developer socialization behavior and outcomes.

Notes

1. We believe that the growth mixture modeling (GMM) technique (Muthén & Muthén, 2000) combines certain aspects of both latent growth modeling (Bollen & Curren, 2006) and latent class analysis (Muthén, 2001) in the sense that latent classes are based on the similarity/differences in growth patterns. In the latent growth model, that is implemented within the SEM framework, individual variations in growth are captured by the continuous latent variables for intercept and slope, which are random coefficients in the sense that they vary across individuals. These latent variables (growth factors) can be used to estimate the posterior probabilities of membership to various classes. In addition, latent growth model can be implemented in such a way that variation in these latent variables may be predicted by demographic variables. Thus, the GMM framework effectively integrates the technique for identifying interindividual differences (through classes or the effect of demographic variables) in intraindividual variations (Jung & Wickrama, 2008; Nesselroade, 1991).
2. Five of these 133 joiners did not interact with core developers within the first 10 weeks of joining the mailing list. They were included in the analysis.
3. It may be noted that this reduction in sample size does not constitute a direct threat to the generalizability of our conclusions as generalizability is a property of a theory being tested rather than of the specific setting within which the theory is being studied (Chow, 1997). "Because the goal ... is usually to apply the theory beyond the research setting, the degree to which the specific sample represents the population of interest is of less importance" (Highhouse, 2009, p. 556). Building on this premise, we argue that to the extent to which we are able to identify various classes and show that these classes are associated with various level of performance (i.e., LT), our theoretical generalizability holds. We are not indicating that the number of classes we found in this study is exhaustive. Perhaps, using a larger sample size, we may be able to identify more classes; however, that does not invalidate the premise that different classes do indeed exist. Of course, to be able to generalize from one specific study environment to another necessitates an understanding of the new environment (Shapiro, 2002). In other words, to be able to generalize these findings with respect to other OSS repositories, one might require a general understanding of those repositories and of their unique characteristics. The context-specific information helps in identifying boundary conditions to the generalizability of a theory (Hubbard, Vetter, & Little, 1998). Thus, a longer socialization period may be required for more complicated projects to obtain a full understanding of the project and be able to make a meaningful contribution. Thus, even the individual with a higher initial interaction trajectory and a faster growth rate might require a relatively longer lead time (LT) as compared to the joiners in our study.
4. We repeated our analysis for peripheral developers with a promotion time of greater than 6 weeks and the results were comparable. However, the parameters were less stable. For a promotion time of less than 6 weeks, we could only fit a simple latent growth model because one with a quadratic, cubic, and higher order model suffered from an identification problem.

5. If we had used the unit month as the time interval and 7 waves (i.e., 28 weeks), then we would have lost more than 50% of our data points, resulting in a sample size of only 63 developers. This was one of the criteria for not using the month as the unit of time.
6. It should be noted that the goal of this step is to identify the best fit among all the possible models to test the variability in the intercepts and slopes, while at the same time realizing that none of the models might actually pass the stringent model-fit criteria as there might be different models for each potentially unobserved class.
7. At the outset, we would like to acknowledge that no model is true and with the addition of more empirical evidence, each model may be extended, modified, or discarded. However, this does not invalidate the a priori assumption of the presence of classes. Longitudinal mixture models, such as GMM and LGCA, have been commonly used to identify unobserved but distinct groups of individuals (e.g., Geary et al., 2009; Kreuter & Muthén, 2008; Lanza & Collins, 2006;Reinecke, 2006; Shaw, Lacourse, & Nagin, 2005; Wang & Bodner, 2007). We should be careful to avoid pronouncing that we have found a final solution to identifying the actual number of classes present. Because identified classes may be a reflection of the nonnormality of the data set (Bauer & Curran, 2003; McLachlan & Peel, 2000; Nagin, 2005).
8. Table 5 suggested an average cubic trajectory of interaction. We initiated our LCGA with the initial setting that all the classes had a cubic trajectory. However, our initial analysis indicated that the mean cubic factors (for Classes 2–4) and quadratic factors (for Classes 3 and 4) were not significantly different from zero. Hence, we reanalyzed LCGA with Class 1 as cubic, Class 2 as quadratic, and Classes 3 and 4 as linear.
9. Coincidently, the membership for each of these classes was exactly the same as that of the classification obtained using LCGA. This may be on account of very high entropy, (>.99) for both LCGA and GMM results, which is an indicator of a clear delineation between classes (Celeux & Soromenho, 1996; Hix-Small et al., 2004; Jedidi et al., 1993; Nagin, 1999).
10. However, we wish to highlight here that our classification quality was very good (entropy above .99, average posterior probabilities for each class ranged between .99 and 1.0; none of the members had posterior probability lower than .972 with its ''own'' class and higher than .028 with ''other'' classes). As class uncertainty is very low in our study, it can be justified to use the class membership based on most likely (i.e., posterior) probabilities. It is important to note that because there are no fuzzy cases, that is, individuals with nearly the same posterior probability to be assigned to more than one class, we could use the class membership as ''given'' and use it as a variable in ANOVA analysis.
11. An outcome (distal) variable can be incorporated within the GMM framework either as an additional indicator of the latent class variable or as a cause–effect pairing such that the distal outcome is a consequence of latent class membership (Petras & Masyn, 2009). The choice of implementation depends on theoretical consideration. In our study, the reasonable implementation was lead time (distal outcome) as a consequence of latent class membership as we believe that lead time depends on which trajectory of interactions is followed. A detailed discussion on this topic is beyond the scope of this article; interested readers may refer to a few excellent articles on the topic of the incorporation of a distal outcome into a GMM framework (Petras & Masyn, 2009; Wang, Brown, & Bandeen-Roche, 2005).
12. It is worth noting that, consistent with the exploratory nature of this study, we base this categorization on the obtained empirical results to differentiate between the identified latent trajectory classes. The initial levels of low, moderate, or high and the growth rate of a specific class are specified as being relative to those of the other classes. Building on this study, future research can reexamine this categorization using a stronger theoretical grounding and a larger sample size.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: This work was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Grant number CityU 141809).

Acknowledgment

The authors received very helpful and constructive comments from the associate editor, Mo Wang, and three anonymous reviewers that resulted in an enhanced version of the article.

References

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for over extraction of latent trajectory classes. *Psychological Methods*, *8*, 338-363.

Bliese, P. D., & Ployhart, R. E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organizational Research Methods*, *5*, 362-387.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective.* Hoboken, NJ: John Wiley.

Castilla, E. J. (2005). Social networks and employee performance in a call center. *The American Journal of Sociology*, *110*, 1243-1283.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*, 195-212.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal means and covariance structures analysis (lmacs) and multiple indicator latent growth modeling (mlgm). *Organizational Research Methods*, *1*, 421-483.

Chow, S. L. (1997). Science, ecological validity and experimentation. *Journal of the theory of social behavior*, *17*, 181-194.

Colazo, J., & Fang, Y. (2009). The impact of license choice on open source software development activities. *Journal of the American Society for Information Science and Technology*, *60*, 997-1011.

Collins, L., & Lanza, S. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Hoboken, NJ: Wiley.

Crowston, K., Annabi, H., & Howison, J. (2003). *Defining open source software project success*. Paper presented at the International Conference on Information Systems, Seattle, Washington.

Dierickx, I., & Cool, K. (1989). Asset stock accumulation and sustainability of competitive advantage. *Management Science*, *35*, 1504-1513.

Ducheneaut, N. (2005). Socialization in an open source software community: A socio-technical analysis. *Computer Supported Cooperative Work*, *14*, 323-368.

Fang, Y., & Neufeld, D. (2009). Understanding sustained participation in open source software projects. *Journal of Management Information Systems*, *50*, 1-50.

Fichman, R. G., & Kemerer, C. F. (1997). The assimilation of software process innovations: An organizational learning perspective. *Management Science*, *43*, 1345-1363.

Fitzgerald, B. (2006). The transformation of open source software. *MIS Quarterly*, *30*, 587-598.

Franke, N., & von Hippel, E. (2003). Satisfying heterogeneous user needs via innovation toolkits: The case of apache security software. *Research Policy*, *32*, 1199-1215.

Geary, D., Bailey, D., Littlefield, A., Wood, P., Hoard, M., & Nugent, L. (2009). First-grade predictors of mathematical learning disability: A latent class trajectory analysis. *Cognitive development*, *24*, 411-429.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* New York, NY: Aldine.

Greenbaum, P. E., Del Boca, F. K., Darkes, J., Wang, C., & Goldman, M. S. (2005). Variation in the drinking trajectories of freshman college students. *Journal of Consulting and Clinical Psychology*, *73*, 229-238.

Grewal, R., Lilien, G. L., & Mallapragada, G. (2006). Location, location, location: How network embeddedness affects project success in open source systems. *Management Science*, *52*, 1043.

Hahn, J., Moon, J. Y., & Zhang, C. (2008). Emergence of new project teams from open source software developer networks: Impact of prior collaboration ties. *Information Systems Research*, *19*, 369-391.

Hertel, G., Niedner, S., & Herrmann, S. (2003). Motivation of software developers in open source projects: An internet-based survey of contributors to the linux kernel. *Research Policy*, *32*, 1159-1177.

Highhouse, S. (2009). Design experiments that generalize. *Organizational Research Methods*, *12*, 554-566.

Hix-Small, H., Duncan, T. E., Duncan, S. C., & Okut, H. (2004). A multivariate associative finite growth mixture modeling approach examining adolescent alcohol and marijuana use. *Journal of Psychopathology and Behavioral Assessment*, *26*, 255-270.

Holcomb, T. R., Combs, J. G., Sirmon, D. G., & Sexton, J. (2010). Modeling levels and time in entrepreneurship research: An illustration with growth strategies and post-IPO performance. *Organizational Research Methods*, *13*, 348-389.

Hox, J. J., & Maas, C. J. M. (2002). Sample sizes for multilevel modeling. In J. Blasius, J. Hox, E. D. Leeuw, & P. Schmidt (Eds.), *Social science methodology in the new millennium: Proceedings of the fifth international conference on logic and methodology (second expanded edition)*. Opladen, RG: Leske + Budrich Verlag.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Hubbard, R., Vetter, D. E., & Little, E. L. (1998). Replication in strategic management: Scientific testing for validity, generalizability, and usefulness. *Strategic Management Journal*, *19*, 243-254.

IBM. (2006). *IBM open source and linuxline survey: Unisphere Research for IBM*.

Jedidi, K., Jagpal, H. S., & Desarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, *16*, 39.

Jedidi, K., Ramaswamy, V., & Desarbo, W. S. (1993). A maximum likelihood method for latent class regression involving a censored dependent variable. *Psychometrika*, *58*, 375-394.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409-426.

Jung, T., & Wickrama, K. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, *2*, 302-317.

Koch, S., & Schneider, G. (2002). Effort, cooperation and coordination in an open source software project: Gnome. *Information Systems Journal*, *12*, 27-42.

Kohanski, D. (1998). *Moths in the machine.* New York, NY: St. Martin's.

Kreuter, F., & Muthén, B. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology*, *24*, 1.

Krishnamurthy, S. (2002). *Cave or community? An empirical examination of 100 mature open source projects.* Bothell, WA: University of Washington.

Lanza, S., & Collins, L. (2006). A mixture model of discontinuous development in heavy drinking from ages 18 to 30: The role of college enrollment. *Journal of Studies on Alcohol*, *67*, 552.

Lave, J., & Wenger, E. (1990). *Situated learning—Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

Lee, G. K., & Cole, R. E. (2003). From a firm-based to a community-based model of knowledge creation: The case of the linux kernel development. *Organization Science*, *14*, 633-649.

Lin, N. (1990). Social resources and social mobility: A structural theory of status attainment. In R. L. Breiger (Ed.), *Social mobility and social structure* (pp. 247-271). New York, NY: Cambridge University Press.

Lin, N. (1999). Building a new work theory of social capital. *Connections*, *22*, 28-51.

Lin, N., Ensel, W. M., & Vaughn, J. C. (1981a). Social resources and occupational status attainment. *Social Forces*, *59*, 1163-1181.

Lin, N., Ensel, W. M., & Vaughn, J. C. (1981b). Social resources and strength of ties. *American Sociological Review*, *46*, 393-405.

Louis, M. R. (1990). Acculturation in the workplace: Newcomers as lay ethnographers. In B. Schneider (Ed.), *Organizational climate and culture* (pp. 85-129). San Francisco, CA: Jossey-Bass.

Lubke, G. (2010). Latent variable mixture models. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 209). London, England: Routledge.

Luyckx, K., Schwartz, S., Goossens, L., Soenens, B., & Beyers, W. (2008). Developmental typologies of identity formation and adjustment in female emerging adults: A latent class growth analysis approach. *Journal of Research on Adolescence*, *18*, 595.

MacCallum, R., Kim, C., Malarkey, W., & Kiecolt-Glaser, J. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*, 215-253.

Markus, M. L., Manville, B., & Agres, C. E. (2000). What makes a virtual organization work. *California Management Review*, *Fall*, 13.

Marsden, P. V., & Hurlbert, J. S. (1988). Social resources and mobility outcomes: A replication and extension. *Social forces*, *66*, 1039-1059.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York, NY: Wiley-Interscience.

Mockus, A., Fielding, R. T., & Herbsleb, J. (2002). Two case studies of open source software development: Apache and mozilla. *ACM Transactions on Software Engineering and Methodology*, *11*, 309-346.

Moore, W. (1980). Levels of aggregation in conjoint analysis: An empirical comparison. *Journal of Marketing Research*, *17*, 516-523.

Morin, A., Morizot, J., Boudrias, J., & Madore, I. (in press). A multifoci person-centered perspective on workplace affective commitment: A latent profile/factor mixture analysis. *Organizational Research Methods*.

Morrison, E. W. (1993). Longitudinal study of the effects of information seeking on newcomer socialization. *Journal of Applied Psychology*, *78*, 173-183.

Morrison, E. W. (2002). Newcomers' relationships: The role of social network ties during socialization. *Academy of Management Journal*, *45*, 1149-1160.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557-585.

Muthén, B. (2001). Latent variable mixture modeling. *New developments and techniques in structural equation modeling* (pp. 1-33). Mahwah, NJ: Lawrence Erlbaum.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345−368). Thousand Oaks, CA: Sage.

Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. *Advances in latent variable mixture models* (pp. 1–24). Charlotte, NC: Information Age.

Muthén, B., & Muthén, L. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, *24*, 882-891.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric group-based approach. *Psychological Methods*, *4*, 139-157.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In J. L. Horn (Ed.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 92−105). Washington, DC: American Psychological Association.

Netcraft. (2004). *February 2004 web server survey*. Retrieved from http://news.netcraft.com/archives/web_server_survey.html

Newby, G. B., Greenberg, J., & Jones, P. (2002). Open source software development and lotka's law: Bibliometric patterns in programming. *Journal of the American Society for Information Science and Technology*, *54*, 169-178.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535.

Petras, H., & Masyn, K. (2009). General growth mixture analysis with antecedents and consequences of change. In A. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology.* New York, NY: Springer.

Pinquart, M., & Schindler, I. (2007). Changes of life satisfaction in the transition to retirement: A latent-class approach. *Psychology and Aging*, *22*, 442.

Pliskin, N., Balaila, I., & Kenigshtein, I. (1991). The knowledge contribution of engineers to software development: A case study. *IEEE Transactions on Engineering Management*, *38*, 344-348.

Podolny, J. M., & Baron, J. N. (1997). Resources and relationships: Social networks and mobility in the workplace. *American Sociological Review*, *62*, 673-693.

Qureshi, I., & Compeau, D. (2009). Assessing between-group differences in information systems research: A comparison of covariance- and component-based SEM. *MIS Quarterly*, *31*, 197-214.

Reichers, A. E. (1987). An interactionist perspective on newcomer socialization rates. *Academy of Management Review*, *12*, 278-287.

Reinecke, J. (2006). Longitudinal analysis of adolescents' deviant and delinquent behavior: Applications of latent class growth curves and growth mixture models. *Methodology*, *2*, 100-112.

Roberts, J. A., Hann, I.-H., & Slaughter, S. A. (2006). Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the apache projects. *Management Science*, *52*, 984-999.

Samuelsen, K. M., & Dayton, C. M. (2010). Latent class analysis. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 173). New York, NY: Routledge.

Schadler, T. (2004). *Open source moves into the mainstream.* (Trends report). Cambridge, MA: Forrester Research, Inc.

Seibert, S. E., Kraimer, M. L., & Liden, R. C. (2001). A social capital theory of career success. *Academy of Management Journal*, *44*, 219-237.

Sen, R. (2007). A strategic analysis of competition between open source and proprietary software. *Journal of Management Information Systems*, *24*, 233-257.

Shah, S. K. (2006). Motivation, governance, and the viability of hybrid forms in open source software development. *Management Science*, *52*, 1000-1014.

Shapiro, M. A. (2002). Generalizability in communication research. *Human Communication Research*, *28*, 491-500.

Shaw, D. S., Lacourse, E., & Nagin, D. S. (2005). Developmental trajectories of conduct problems and hyperactivity from ages 2 to 10. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *46*, 931-942.

Singer, J. D. (1998). Using SAS proc mixed to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *23*, 323-355.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure across groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229-239.

Stallman, R. M., & Lessig, L. (2002). *Free software, free society: Selected essays of Richard M. Stallman.* Boston, MA: GNU Press.

Stewart, K. J., Ammeter, A. P., & Maruping, L. M. (2006). Impacts of license choice and organizational sponsorship on user interest and development activity in open source software projects. *Information Systems Research*, *17*, 126-144.

Stewart, K. J., & Gosain, S. (2006). The impact of ideology on effectiveness in open source software development teams. *MIS Quarterly*, *30*, 291-314.

Van Mannen, J., & Schein, E. H. (1979). Toward a theory of organizational socialization. In B. Staw (Ed.), *Research in organizational behavior* (Vol. 1, pp. 209-264). Greenwich, CT: JAI Press.

Von Hippel, E. (2001). Innovation by user communities: Learning from open source software. *MIT Sloan Management Review*, *Summer*, 82-86.

Von Hippel, E., & Von Krogh, G. (2003). Open source software and the "private-collective" innovation model: Issues for organization science. *Organization Science*, *14*, 209-223.

Von Krogh, G., Spaeth, S., & Lakhani, K. (2003). Community, joining and specialization in open source software innovation. *Research Policy*, *32*, 1217-1241.

Wang, M. (2007). Profiling retirees in the retirement transition and adjustment process: Examining the longitudinal change patterns of retirees' psychological well-being. *Journal of Applied Psychology*, *92*, 455-474.

Wang, M., & Bodner, T. E. (2007). Growth mixture modeling: Identifying and predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods*, *10*, 635.

Wang, C. P., Brown, C. H., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*, *100*, 1054-1076.

Wang, M., & Chan, D. (in press). Mixture latent Markov modeling: Identifying and predicting unobserved heterogeneity in longitudinal qualitative status change. *Organizational Research Methods*.

Wu, J., & Witkiewitz, K. (2008.). Network support for drinking: An application of multiple groups growth mixture modeling to examine client-treatment matching. *Journal of Studies on Alcohol and Drugs*, *69*, 5.

Ye, Y., & Kishida, K. (2003). Toward an understanding of the motivation of open source software developers. *IEE Proceedings – Software*, 419-429.

Bios

Israr Qureshi is an assistant professor at Hong Kong Polytechnic University. He earned his PhD from University of Western Ontario. He has been interested in understanding various aspects of Information and communication technologies and his research focuses on the impact of information and communication technologies on base of pyramid population.

Yulin Fang is an assistant professor in the Department of Information Systems, City University of Hong Kong. He earned his PhD at Richard Ivey School of Business, University of Western Ontario. His current research is focused on knowledge management, virtual teams, e-commerce, and open source software projects.