Collaborative learning in Open Source Software (OSS) communities: The dynamics and challenges in networked learning environments

Raktim Mitra

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Industrial & Systems Engineering

Aditya Johri
Hazhir Rahmandad
G.Alan Wang
Christian Wernz

June 14th 2011
Blacksburg, VA

Keywords: Open source software, online communities, distributed learning, collaborative learning, cyberlearning external jolts, newcomer participation, Sun Microsystems, Oracle, MySQL

Collaborative learning in Open Source Software (OSS) communities: The dynamics and challenges in networked learning environments

ABSTRACT

The proliferation of web based technologies has resulted in new forms of communities and organizations with enormous implications for design of learning and education. This thesis explores learning occurring within open source software (OSS) communities. OSS communities are a dominant form of organizing in software development with implications not only for innovative product development but also for the training of a large number of software developers. The central catalyst of learning within these communities is expert-novice interactions. These interactions between experts and novices or newcomers are critical for the growth and sustenance of a community and therefore it is imperative that experts are able to provide newcomers requisite advice and support as they traverse the community and develop software. Although prior literature has demonstrated the significance of expert-novice interactions, there are two central issues that have not been examined. First, there is no examination of the role of external events on community interaction, particularly as it relates to experts and novices. Second, the exact nature of expert help, particularly, the quantity of help and whether it helps or hinders newcomer participation has not been studied. This thesis studies these two aspects of expert-novice interaction within OSS communities.

The data for this study comes from two OSS communities. The Java newcomer forum was studied as it provided a useful setting for examining external events given the recent changes in Java's ownership. Furthermore, the forum has a rating system which

classifies newcomers and experienced members allowing the analysis of expert-novice interactions. The second set of data comes from the MySQL newcomer forum which has also undergone organizational changes and allows for comparison with data from the Java forum. Data were collected by parsing information from the HTML pages and stored in a relational database.

To analyze the effect of external events, a natural experiment method was used whereby participation levels were studied around significant events that affected the community. To better understand the changes contextually, an extensive study of major news outlets was also undertaken. Findings from the external event study show significant changes in participation patterns, especially among newcomers in response to key external events. The study also revealed that the changes in participation of newcomers were observed even though other internal characteristics (help giving, expert participation) did not change indicating that external events have a strong bearing on community participation.

The effect of expert advice was studied using a logistic regression model to determine how specific participation patterns in discussion threads led to the final response to newcomers. This was supported by social network analysis to visually interpret the participation patterns of experienced members in two different scenarios, one in which the question was answered and the other where it was not. Findings show that higher number of responses from experienced members did not correlate with a response. Therefore, although expert help is essential, non-moderated or unguided help can lead to conflict among experts and inefficient feedback to newcomers.

iv

Finally, thank you Ritika for your love and support, not to mention the late night coffee mugs that kept me awake. Last but not the least thank you to all my friends and colleagues at Virginia Tech for their unparalleled support and help throughout the duration of this project. Working, learning and growing with you has been a unique and lasting part of my experience in graduate school.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER – 1
INTRODUCTION

In this constantly evolving and dynamic global scenario, the citizen today encounters myriad challenges and has to keep himself abreast of the latest technological advances and development. It becomes imperative to strive towards a better and absolute educational and learning system that is as dynamic as the ever changing global challenges of the 21st century. The recent report of the NSF Taskforce on Cyber-enabled learning (NSF, 2008) argues that a fundamental change is needed in the application of technology if education is to keep pace with changes in both skill acquisition and conceptual knowledge and to ensure that educational content is no longer constrained by books or classrooms. The NSF taskforce has identified that cyber-learning has the potential to offer the global audience a mix of diverse content via the combined technological capabilities of the Internet, high performance computing, advanced networking, in-home electronics, and mobile communications. Cyber-enabled learning thus is a sustainable way to achieve a personalized learning system, which according to the NAE is one of the grand challenges for the 21st century.

Open source software development through online communities is an exemplary case of collaborative learning in cyber –learning environments. Open source online development communities showcase learning practices that extend beyond the classroom and even educational institutions and allow learners to engage with like-minded peers, experts, and even machines spread across the globe. The growth of the World Wide Web and its manifestation in the development of open source projects and distributed learning has led to a lot of research and considerable increase in academic interest towards open source software and online development communities (Von Hippel, 2001). Empirical research has led to the development of various

models that dwell on the motivational (Shah, 2006), socialization (Choi et al., 2010) aspects of participation and learning in these collaborative environments. The study here aims at extending the realms of such studies and probe further into some of the unidentified and less discussed areas of open source software development communities.

An examination of these communities, however, shows that while a few such online structures have been highly successful, numerous others have failed to attract the significant participation and then maintain participation levels over time. For instance, on October 11, 2005, proprietary software maker Xara announced its plans to open the source code to its flagship vector graphics package Xara Xtreme, and with the help of community developers port it to Linux. Although the initial use was substantial, the project is since stagnant and on the verge of irrelevance. The inability of the company to work with the open source community and maintain participation after an initial deluge of developers led to its decline. The failure of some open source communities has been extreme and they have eventually been shut down as they could not meet their objectives. For instance, OpenDarwin which shut down in June 2006 failed to gather constructive participation in its 4 years of existence. Therefore, while some open source software development provide an example of successful models to build cyber-learning environments, it is also essential to understand in great detail the socio-technological infrastructure of such domains. Only then would it be possible to establish successful models that can be recreated time and again and restructured to fit the requirements and goals of individual personalized learning environments.

1.1 PROBLEM STATEMENT

For the past century, most educational activities have been organized based on a model of efficiency of delivery (Callahan, 1962). Due to resource constraints, the effective use of teaching resources and building infrastructure served as the primary basis for organizing learning. Since the advent of the Internet and proliferation of digital technology a sweeping change can be observed in how learning occurs. In the age of anywhere, anyplace personalized learning supported by digital media it has become ever so important to make the shift from the classical methods towards these advanced techniques. However the few initiatives undertaken to build digital learning environments have not been entirely successful (Spaulding, 2010). Usage statistics from few such initiatives show the lack of traction and the failure to build on healthy memberships over time. The idea that if learners have access to online environments engagement in learning will follow, has not been observed.

To better understand how to design such environments, researchers have started to study naturally occurring examples of cyber-learning. Over the last decade researchers have closely studied the phenomenon of open source software development which closely mimics the basic ideas of cyber-learning. Open source software communities have led to a lot of innovation, with limited monitoring and involvement by the manufacturer. Such innovation by and for the user is not only limited to creating complex software, but is as diverse as creating windsurfing gear. (Von Hippel, 2001). Many open source efforts have been true success stories. Linux for example is a popular open source computer operating system, which has seen widespread acceptance among users worldwide and is the primary competitor to MS Windows. The Linux family of operating systems developed and maintained by the open source community is a testament to the viability of virtual communities creating valuable products. /

A close inspection of the differences in the usage statistics of the cyber-learning environments vis-a-vie the open source developer communities points towards the low use of the cyber-learning sites and their difficulty in attracting new membership. Since software development is a knowledge-intensive activity that often requires very high levels of domain knowledge, experience, and intensive learning by those contributing to it (Pliskin, Balaila, and Kenigshtein, 1991; Waterson et al., 1997), researchers have argued that a lot can be learned from open source communities in terms of building sustained participation, newcomer integration and learning which are integral to the success of any cyber-learning environment (Ducheneaut, 2005). Newcomer participation since has been studied to a good detail, especially with a sociological light on the open source software development communities. Although such literature has been extremely well cited and provided us with a great platform for understanding online communities, little is known about the exact methods and actual practices that newcomers engage in while participating and integrating in such communities. This study addresses two strongly aligned, but slightly different aspects of newcomer participation dynamics in online communities. The research questions look at newcomer entry and participation in the light of several factors that have not been so well developed in the contemporary literature and aims at shedding some light on these areas. This study expands the understanding of open source development to the next deeper level where we explore the detailed intricacies of practices in these communities. Methodologically, this study proposes a rich quantitative and qualitative of these communities to provide deeper understanding of the knowledge sharing practices and integrating patterns among newcomers. This study aims at contributing to the fundamental theory of newcomer participation in online learning communities, exploring learner – teacher

4

and novice – expert relationships, and understanding and classifying practices within the newcomer population.

1.2 RESEARCH MOTIVATION

1.2.1 EFFECT OF EXTERNAL EVENTS ON COMMUNITY PARTICIPATION

Participation in open source software development has been studied to a great detail including extensive work on the motivational aspects of users, contributing to a larger cause without significant or no monetary gains. Researchers have proposed various theories regarding what motivates individuals to contribute to open source projects. Lawler in his seminal work proposed the expectancy theory which posits that that individuals are motivated by a combination of valence (the intrinsic or extrinsic attractiveness of a reward) and instrumentality (the path to that reward) (Lawler, 1971). Empirical research consistent with expectancy theory describes three motivational aspects among open source contributors. A few well cited research theories (Hars and Ou, 2002; Lerner and Tirole, 2002; Hertel, Niedner, and Herrmann, 2003; Lakhani and von Hippel, 2003) list direct utility, intrinsic benefit (like learning a skill, personal fulfillment etc) and recognition (respect from peers, networking and acknowledgement) as the primary motivators. Further studies have analyzed how newcomers become part of an organization identifying socialization (Ostroff & Kozlowski, 1992; Van Maanen and Schein, 1979) and assimilation (Miller and Jablin, 1991) as the primary aspects. (Johri ,2007) explains these socialization and assimilation in further detail. Based on such work further studies have delved into designing social structure and governance rules, analyzing sustainability of such

models.(Schoberth, Preece, and Heinzl ,2003; Shah, 2006), Community hierarchies, conflict management, and decision making mechanisms (Demil and Lecocq, 1999 ; Joode, 2004). Although this literature serves as a great platform for understanding the online communities, however most of them have taken an internalized approach and believe that these collaborative environments are inclusive and self-sustaining (Kim ,2000) and (Beenen et al., 2004; Ren ,Kraut, and Kiesler, 2008).

What seems to be missing from this body of research is a study about the effects of external or exogenous events – events whose influence is outside the scope of community design, on participation in these cyber-learning communities. The first part of this study takes a step at understanding the implications of such events on participation in online forums and communities. This will provide a holistic picture of the cyber-learning environment that thrives as an OSS development community. This study will add another dimension to the body of knowledge that is available on the reasons for failure of certain communities in sustaining growth, by going beyond the idea of faulty internal community design characteristics to possible ripples caused by external and exogenous events. Finally this study will lead to future research that will address the role of leadership, and strategies in handling such circumstances.

1.2.2 ROLE OF STAR MEMBERS IN OPEN SOURCE SOFTWARE COMMUNITY SUCCESS

The second part of the study looks more into the internal aspects of collaborative learning as it manifests itself in online OSS communities and the study that we propose here will contribute both to the existing cognizance on the importance of newcomer participation in building any collaborative organization and the strategies to sustain continued participation .

Empirical research has revealed that newcomers play an important role in the overall success of the organization as a whole. Not only, is the continuous influx of newcomers necessary for sustained growth (as it also indicates the perception of the community and its marginal benefits to the user), productivity, and innovation (Johri, 2007), it is of paramount importance to integrate these newcomers into the system. A lot of research in this domain points out that socialization has a large impact on the adjustment newcomers to their jobs, groups, and organizations (Ashforth, Sluss, and Harrison, 2007; Bauer, Morrison, and Callister, 1998; Fisher, 1986, Moreland and Levine, 2001, and Saks and Ashforth, 1997). Also Cooper-Thomas & Anderson argue that continued newcomer learning is ''at the heart of any organizational socialization model'' (Cooper-Thomas and Anderson, 2005).

Contemporary studies have researched the behavior of newcomers as they try to join such communities and integrate therein. Morrison explores the social network ties of newcomers and argues that they try to build relationships related to the type of learning they desire (related to different indicators (in depth knowledge and core competency vs. overall organizational knowledge)) (Morrison, 2002). Miller and Jablin look at the newcomer behavior through a slightly different lens, and posit that newcomers on entry into any organizational setup face uncertainty and proactively try to work towards reducing it. One way they adopt is by asking questions and assimilating the information transmitted from the higher levels or 'insiders' in the organization. (Miller and Jablin, 1991).Continued research in these directions have revealed that an important way in which socialization and networking occur is through social interactions between newcomers and "insiders," or more experienced members of their new organization (Feldman, 1981; Louis, 1990; Reichers, 1987). Also the importance of insiders, especially peers and supervisors, for helping newcomers to acquire information and "learn the ropes" (Louis,

Posner and Powell, 1983; Morrison, 1993; Ostroff and Kozlowski, 1992) has been repeatedly highlighted.

Organizations generally present an onion like social structure for the teams. People at the center participate in highest numbers, the next circle is the participants who are posting in medium range, and then the ones in low range and farthest from the core are the peripheral members who contribute very little in these teams (Crowston and Howison, 2004). Open source software development communities present structures similar to this, wherein a few (Knowledgeable/ highly involved/ core) members in communities have a majority of posts and are involved in many discussions, essentially carrying the community on their shoulders sustaining a large developer community (Singh ,Twidale, and Nichols ,2009).

It is evident that the role of these core members is of paramount importance in successful integration of newcomers and their growth – leading essentially to creating a sustainable learning environment and continued innovation in the community. Although a lot of well cited research is available on how these core members work, what roles they play, and how they help newcomers, the approach taken has been to analyze the activities of the core members. Little is known about the quantity, overall quality and method of administration of external help by the core or the experienced members on a case by case basis.

The aim here is to view this agenda from the perspective of a newcomer, and quantify the actual effectiveness of these experienced members. This study shall look at the finer nuances of external help, such as the marginal returns of each individual expert help to a newcomer. This study shall also contribute to the theory of designing successful communities by arguing that just bringing together knowledgeable members is not enough, and a pattern or method should exists

even while these experienced members go about their responsibilities of helping out newcomers integrate into the community .

1.3 RESEARCH QUESTIONS

Based on the review of prior work and our initial investigation of the data, the following research questions were outlined for further study:

1. What is the effect, if any, of external events on the participation in open source software development communities?

   a. What role does the perception of steward/sponsoring firm play in developers' engagement with an open source community?

2. What is the effect of participation by experienced members in providing a useful response to newcomers?

1.4 HYPOTHESES

1. For research question one it was hypothesized that external events would have a bearing on the participation in open source software development communities and significant change in patterns would be observed in and around such exogenous events. This hypothesis was tested using the natural experiment (Meyer, 1982) method and by analyzing key participation metrics of two open source software communities as they experienced external events.

2. For research question 1 (a), it was hypothesized that the perception of the steward/sponsoring firm would affect participation patterns, and would be stronger in the case of newcomers as compared to experienced core members. Such ideological actions are more common among newcomers as compared to committed members who suffer less motivational losses (Hardy and Latane, 1988). Furthermore, it was hypothesized that the change in identity of the sponsoring/steward firm would affect the participation patterns in line with the perceived identity of the firm. If the perception associated of the steward firm was incongruent with the ideals of the community, then it would lead to a decrease in participation.

3. For research question 2, it was hypothesized that an increase in number of experienced members would help sustain newcomer participation, as it would bring in diversity and the higher combined experience as a result of more members. However it was also hypothesized that in inordinate increase in the number of such members may not be beneficial.

CHAPTER – 2
LITERATURE REVIEW

2.1 ENVIRONMENTAL JOLTS AND CHANGES IN ORGANIZATIONAL FIELDS

2.1.1 EXTERNAL EVENT CHANGES AND ITS EFFECT ON ORGANIZATIONS

In the recent decade continuous research on online communities has led to a marked shift in design principles. A more holistic approach has replaced the erstwhile 'technologically sound only' strategy. Community developers now draw immensely from various functional (Kim, 2000), sociological (Shah, 2006) principles and objectives. Socialization and integration of newcomers has been found to be essential for the growth and sustainability of the community, especially to avoid the decay in volunteers' participation, which is an essential recourse for online communities (Schobert, Preece, and Heinzl, 2003). Newcomer participation and integration thus has been the prima facie of a lot of research and have been linked to a number of factors, including governance structures (Shah, 2006), situated learning and identity construction behaviors of participants (Fang and Neufeld, 2009), and ideology (Stewart and Gosain, 2006).

Online communities today, are very similar to offline communities, and are generally embedded in larger organizational and institutional contexts. These institutions and organizations provide immense support to these online communities and very often the largest user or the initial developer of the software. This warrants a wider inspection of design principles for community structure than the fundamentally internalized understanding that we have.

Not all changes and interactions in online communities can be attributed to internal factors. External environment changes can affect a community or organization in a variety of

11

ways. Highly publicized events are critical triggers of institutional transformation (Fligstein, 1990; Sewell, 1995; Hoffman, 1999). Such critical events, are contextually dramatic happenings that gathers sustained public attention and invite the collective definition or redefinition of social problems (Pride, 1995) They have been referred to as shocks (Fligstein,1991), threats, crises, or catastrophes (Billings, Milburn, and Schaalman, 1980), jolts (Meyer,1982), or discontinuities (Lorange, Morton, and Ghosal,1986).As such, critical events have been at the center stage for fostering institutional change and industry evolution (Miles, 1982; Leblebici et al. , 1991).

2.1.2 ENVIRONMENTAL JOLTS

Meyer (1982) examined organizational adaptations to environmental jolts and argued that these events whilst not overtly threatening in nature to the survival of soundly designed organizations do reveal inherent flaws in their architecture. These environmental jolts are transient shocks that cause temporary disruptions that perturb their organizational inhabitants and then they subside (Meyer, Gaba, and Colwell, 2005). Literature suggests that organizations adapt to such environmental forces in different ways and that they are influenced by the strategies they pursue (Miles and Snow, 1978), the structures they adopt (March and Simon, 1958), the ideologies they support (Beyer, 1981), and the slack resources they amass (Bourgeois, 1981). Thus while it is not possible to foresee the exact occurrence of such events, it may be possible to adapt to such transient shocks well.

Online communities as argued earlier are not separate entities in themselves but a part of the environment. Although the importance of external factors and the environment in shaping organizations has been extensively studied, research on distributed learning, cyber learning and

open source communities has been relatively silent in this regard. Many online communities are hosted and run by commercial firms (e.g. Amazon, Usenet by Google, Yahoo Groups, Google Groups, Facebook), by professional associations, educational institutions, and other formal and information organization. The image of a host is especially important for communities associated with software products and community supported software products such as open source software, since these are likely to change hands through mergers or acquisitions. Such instances can lead to 'jolts' to the community that are outside the scope of the internal community design parameters and the response to such events can be critical in the absence of any preparation.

A deeper understanding of the role of external events thus can complement the body of knowledge on community systems design and enhance the understanding of participation changes (if any) that accompany such events. This study takes a first step in this direction by examining the changes in participation in an online community that accompany two significant events. The perceived image of the steward company has been identified as a key factor in determining participation.

2.2 COMMUNITY DYNAMICS

2.2.1 ONLINE SOCIAL STRUCTURES

Open source software development and adoption has inspired a lot of research in the last decade and scholars have investigated several characteristics of an open source software community such as their size, motivation for participation, the economic, legal, and political issues involved in its appearance and growth, use of technology (Yamauchi et al., 2000) and the

success of open source projects (Weber, 2004). Learning and knowledge sharing occurs and newcomers move from peripheral to central or full participants (Kogut and Metiu, 2001).

Starting with Eric Raymond's groundbreaking work 'The Cathedral and the Bazaar' ,OSS has been viewed as the collective effort of a group of developers and has been seen in sharp contrast to proprietary development by emphasizing OSS's distinctive social and communications structures (Raymond,1999). Sacchi pointed out the importance of understanding the coordination process in these software development communities that pan across different settings, and the necessity to unravel the software processes, work practices, and organizational contexts are necessary to their success. (Scacchi, 2002). Studies by (Cox, 1998; Gacek, Lawrie, and Arief, 2001; Moon and Sproull, 2000; Mockus, Feilding, and Herbsleb, 2002) have shed light on the structural composition of such communities although they argue that the coordination processes and work practices differ based on the type of activity undertaken. Krishnamurthy (2002) however argues based on the research on 100 top open source projects on Sourceforge that many such many such communities function as highly centralized development structures that sometimes eludes the idea of a team based structure at all. (Mockus, Feilding, and Herbsleb, 2002) identified the existence of both centralized and decentralized activities in such development structures and different form of participation in different discussions and activities. Crowston and Howison (2008) extend these studies and substantiate the presence of many user levels of expertise, interacting in one discussion and posit that the communication strategies used by help givers and help seekers are adaptive to the situation at hand. Identification of the different roles of users in these communities and the transition of users from newcomers to experts has also been observed (Singh, Twidale, and Nichols, 2009). Experts lie at the center of

the structure and contribute largely, and then the participation level decreases to finally reach the lurkers in the community.

Together this literature points in the direction of a presence of a virtual hierarchy, wherein a handful of experts are able to sustain a large developer community. These group of community members essentially described as 'elders', are active members of the community, regularly posting to share their knowledge and the culture of the community (Kim, 2000).

2.2.1 CRITICAL MASS, GROUP SIZE AND COMMUNITY CRAFT

Attaining a critical mass is very important for learning communities, especially for open source software development communities. A critical mass is necessary for attracting new users, functionality and useful service. This in turn paves way for further participation in the community. (Markus, 1987) pointed out the importance of attaining a critical mass for diffusion of interactive media within online communities and pointed out that such environments are prone to discontinuance. Williams (2000) suggested that building a critical mass is one of the four steps to building successful and sustainable online communities of practice.

While theories on building critical mass aim at building up a sustainable community size that aids the seamless integration of information, services, discussion groups and other offerings (Williams, 2000), Group size theories basically discuss the limit on the size of such communities, and essentially provide the numerical limit to the number of users that may participate without decreasing the combined productivity and effectiveness of the group.

A lot of well cited literature in the organizational science domain point out the importance of ideal group sizes in the success of team oriented projects. (Katzenbach and Smith,

1993) suggest that the ideal group size should be 12 members, while (Schraf, 1989) suggests that it should be kept down to 7 members. Other studies have reported ideal group sizes to be 8 (Nasser, 1988). The basis of these studies seems to stem from medical evidence on the theoretical cognitive limit on human beings to be able to maintain stable social relationships. Referred to as the Dunbar's number, this theory serves as the basis for a lot of research in organizational literature.

Some researchers argue for larger group sizes, (Dennis and Valacich, 1993; Grofman, Feld, and Owen1987) citing the access to more resources, time, energy, money and expertise. Others however support the idea of smaller groups (Deihl and Stroebe, 1987; Stasser and Taylor, 1991). They argue that larger groups are prone to confusion and co-ordination problems such as task clarity, miscommunication. Studies also show that (Hare, 1952; O'Dell 1968) that large groups are prone to conflicts among members. While there seems to be a lot of debate on the idea of an ideal group size, it does highlight the importance application of such knowhow while designing teams and determining the usability of such teams.

Usability and success of online communities have been found to be related to several group factors such as number of participants in a community, the number of messages per unit of time, members' satisfaction, and some less obvious measures such as amount of reciprocity, the number of on-topic messages, trustworthiness, number of errors, productivity, user satisfaction (Preece, 2001) which are all functions of the group size.

While there is little research on the idea of an ideal group size for online communities, it has been pointed out that such group sizing is inherently inbuilt into the design of such communities (Mockus, Feilding, and Herbsleb, 2002). Groysberg (2010) analyzed the role of star members in group success. This study analyzed the marginal returns of adding star members

16

to groups in terms of group effectiveness. Though this study is done with a backdrop of Wall Street sell side equity analysts, it clearly indicated that not only was the overall group size important, but also the diversity of the group. Group effectiveness cannot be achieved by just adding experienced and knowledgeable members, and there exists a tipping point in the number of such members that can be a part of the group, beyond which they tend to decrease the overall group effectiveness.

The aim of this particular study is to look at individual threads (questions) posted on online software forums and determine is such group dynamics operate within individual threads (although they are not defined teams, but the final outcome of a thread/question is dependent on the collaboration of different users with varied expertise).

CHAPTER –3
METHODS AND ANALYSIS

3.1 RESEARCH SETTING

Two very popular open source communities have been studied. The newcomer forums at JAVA (New to Java at forums.sun.com) and MySQL were chosen for this study. The newcomer forum for MySQL developers and users was chosen specifically for the study on external events. While JAVA is a very popular object oriented programming language with diverse applications from application software development to web applications, MySQL is highly popular as Open Source database software. It is a relational database management system that runs as a server providing multi-user access to a number of databases. Both JAVA and MySQL development project source codes are available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements.

The choice of these particular online forums was made for several reasons. First, newcomers indicate participation and product uptake increase or decrease. Second, Java and MySQL are very popular open source programming language and open source database management system respectively, and therefore the findings may be more reflective of general trends in open source software and their related communities. Third, several incidents in the recent past of Java (the open sourcing of Java by Sun (Nov. 19, 2006) and acquisition of Sun (and consequently of Java) by Oracle (April 20, 2009)) and MySQL – (its acquisition by Sun Microsystems (a proponent and major contributor of open source software and systems) and then the subsequent acquisition of the steward company Sun by Oracle (MySQL was among other few Open Source Relational Database Management Systems that were a direct competition to the Oracle Flagship), both offer useful setting for a natural experiment. These changes, being the

result of external events rather than researchers' intervention provided an adequate setting for studying the effect of external events on newcomer participation and behavior. Fourthly, the New to Java community had a built in appraisal system which awarded points based on the quantity and quality of help provided by members, and eventually awarding separate star ratings to users (Platinum, Gold, Silver and Bronze members). This provided with a very good setting for the research question on the effectiveness of experienced/ knowledgeable community members in helping newcomers and overall community success.

3.2 DATA COLLECTION

The data collection procedure involved parsing information using perl scripts from the online web pages and storing them as a relational database. The datasets were created based on the activity in the forum during a three year time window starting 12th February 2007 through 12th February 2010 (A second dataset was created specifically for the study on external events to capture the forum dynamics post 12th February 2010 until 30th October 2010). The activities over these forums were captured in three basic categories. These categories were: 1) User Information, 2) Thread information, and, 3) Individual Post/Message information. The User information dataset captured details about all the registered users posting within the forum during the given time frame. Details such as User Name, User ID, Registration Date and Total number of posts were extracted from the HTML profile pages for each user. The Thread information table details each thread that appears within the forum. At the thread level we have gathered information like the thread ID, thread subject, post date, post time, the post content and the status (answered/unanswered). We created the individual posts/message information table to look into the finer details of the forum activity. This exhaustive table contains information on all

individual posts/messages in the forum. We have the ability to track each topic/thread posted to the forum, the number of replies it received, and the time stamp for each post. All information is stored in a relational database to enable easy data access and manipulation. The final datasets for this research study were created by joining multiple tables from this database. The collected data is taken at a static standpoint, as of 12th February 2010. Essentially the data collection is done in a retrospect manner, and captures a few essential attributes as of the collection date. Figure 1 represents the process on a timeline.
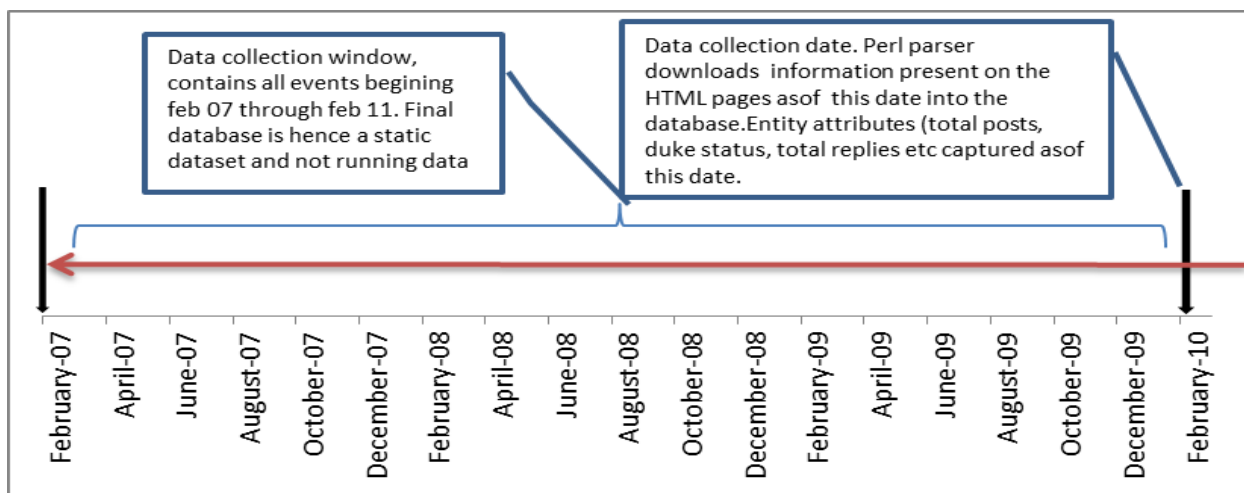
Figure 1: Data collection timeline

3.3 OVERALL RELATIONAL DATABASE DESIGN

Entity –Relationship diagrams are used to represent the relationships between objects and depict their behavior in the relational database design process. An entity is a unique representation of a real world object that is created by using values of its attributes in a computer readable form. The data captured from the forum can be classified into 3 basic entities.

20

Table 1: List of various entities and their primary attributes

Entity	Primary Attribute(Key)
User	User_id
Thread	Thread_id
Message	Message_id

A relational database model following the ER model was developed to represent the relationships between different entities. Figure 2 shows the ER representation of the database design and the data structure. Primary keys uniquely identify each record in the table and are used to link separate tables in one-to- many relationships. Many-to-many relationships can be achieved by concatenating primary keys and the creation of new tables from old tables. A foreign key on the other hand is a field in a relational table that matches the primary key of another table. Figure 3 indicates the primary and the foreign keys for the database.

Figure 2: Entity –Relationship Diagram

User	user_id (PK), user_name(AK),registration_date(AK),total_posts(AK),duke_status(AK),duke_stars(AK)
User_Thread	user_id(FK), thread_id(FK), user_id & thread_id (PK)
Thread	thread_id(PK),thread_desc(AK),thread_post_date(AK),thread_post_time(AK),answered_flag(AK),original_post_content(AK)
Thread_Message	thread_id(FK), message_id(FK), thread_id & message_id (PK)
Message	message_id(PK),message_desc(AK),post_date(AK),post_time(AK),reply_to_message_id(AK),message_content(AK)
User_Message	user_id(FK),message_id(FK), user_id & message_id (PK)

Figure 3: Relational data structure

22
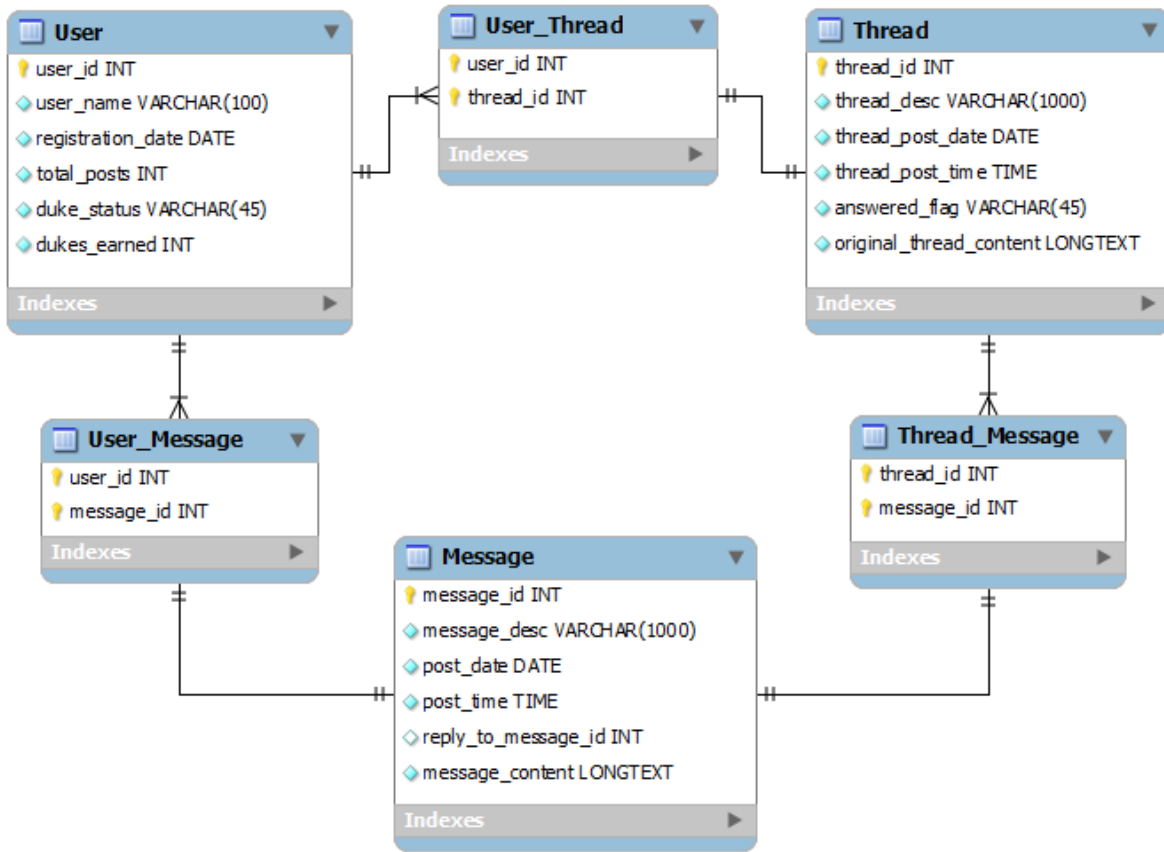
Figure 4: Database class diagram

3.4 BASIC DATA ANALYSIS: COMMUNITY STRUCTURE

A basic data analysis was performed to examine the structure of the New to Java community, understand the basic user roles and behavior and also develop a preliminary guide to the kinds of practices in this community. Since the basis of the methods adopted in this study were at the grass root level, tied to the individual users, their behavior and characteristics, it was but necessary to start with an idea of the basic classification provided by the community itself.

User profiling available on the forum was studied to develop an understanding of the different classes of users present and their characteristics.

3.4.1 USER PROFILING

As a first step to understanding the structure of the forum, users on the forum were profiled based on their existing classification and then a further look into their behavioral characteristics. The method adopted was to proceed based on the classification provided on the community and then explore the subtle differences between the groups. Users were profiled based on their 'Duke Status' and subsequently user groups were compared on several key attributes like number of such users present, overall number of posts, overall number of questions asked, overall number of clarifications requested in each category. Statistical significance of the observed difference on key matrices was established through paired t tests (The paired t-test is a statistical test Used to compare means on the same or related subject over time or in differing circumstances .The paired t-test is actually a test that the differences between the two observations is 0. So, if Δ represents the difference between observations, the hypotheses are: Ho: $\Delta = 0$ (the difference between the two observations is 0) Ha: $\Delta \neq 0$ (the difference is not 0).The test statistic is t with n-1 degrees of freedom. If the p-value associated with t is low ($<$ 0.05), there is evidence to reject the null hypothesis and hence indicating that difference in means across the paired observations.).

3.4.1.1 DUKE STAR RATINGS – DUKE STATUS

The New to Java forum classifies four different user profiles. The profiling is based purely on a feedback system that rates users on the quantity and quality of help provided on the forum. The feedback system functions in a way so as to give the primary question asker or the help seeker the option to rate the users who provide answers to his questions. This is done in terms of 'duke points' which is stored on a cumulative basis against the help providers' name. The 'duke status' of an individual user is determined on the basis of hard cut offs set on the accumulated duke points. The status of a newcomer to the community is set to bronze by default and remains that way until the user garners up 100 duke points. Members with a silver star rating have between 100 to 500 duke points. While gold members have amassed duke points in the 500 to 1000 range, platinum members lie at the top of the rating scheme with more than a 1000 duke points. The rating system is not just inclusive of the activity of members in the 'New to Java' forum but is a cumulative assessment of the individual user across all developer forums housed at forums.sun.com.

3.4.1.2 BRONZE MEMBERS

Bronze members in the community are essentially the newcomers. This is the default duke status set to an individual on joining the community. This remains that way until the individual garners 100 duke points. However since duke points can only be made by providing meaningful answers to questions asked by fellow community users, an average newcomer spends quite some time before transitioning onto the next level. On an average the number of bronze members active on the forum in a month makes up for about 90.6% (89.59% - 91.60% confidence interval) (P < 0.0001) of the forum in terms of their numbers (Averaged over a

25

period of 3 years starting Feb 2007 – Jan 2010). A majority of these users exhibit a 'help seeking' behavior (Miller et al 1991) and ask only questions and clarifications to their own questions. A lesser number of such individuals post replies to others questions (on a monthly basis where about 71% of bronze members ask just questions and clarifications, only about 23% post only replies to others questions without asking a question themselves while the rest ask questions and also reply to others) (P<.0001). The credibility or usability of these replies are often not great as reflected by the average duke star rating of the individuals who engage in such replying behavior (Average duke star rating of 2 among bronze users who exhibited replying behavior). Although these users posts outside the new to java (NTJ) forum, however the difference in total posts (sum of all posts by the user including other forums as well) and posts to new to java is not substantial when compared to the other groups (about 50% percent of these users were only active on New to Java) .Consistent with the onion like social structure model proposed by (Crowston & Howison 2004) this group of members represent the outer layer of the forum structure.

3.4.1.3 SILVER MEMBERS

The silver members represent the next rung of users in the forum structure. Accounting for only about 4% (3.63% - 4.34% confidence interval) (P<.0001) of the number of such active members in the forum on a monthly basis, they represent the 'in transition' members in the forum. While a majority of them (95%) post replies to other questions while not asking a question or clarification themselves, only about 4.5% of these users are engage themselves in asking questions and also replying to others (on a monthly basis) . A marginal 0.5% of these users engage in purely help seeking behavior (i.e. ask only questions while not replying to

26

others). However none of these individuals engage in the purely help seeking behavior over time (over the three year time period, there existed no silver member who had only asked questions and not replied). The average duke star rating for this group was found to be 200 indicating some usability of their replies. Silver members also did not restrict themselves to the new to java forum and posted freely on other forums as well (as indicated by a significant difference in their total post count and post to new to java only) and the presence of no silver member who posted exclusively at new to java.

3.4.1.4 GOLD MEMBERS

According to the design principles of the community structure individuals are awarded the gold duke status on accumulating 500 duke points. They account for about 1.83% of the number of active members on the forum on a monthly basis (1.62% - 2.05% confidence intervals) (P= <.0001). They primarily constitute the outer rung of the core members and are heavily engaged only in posting replies to the questions asked by other members in the forum (94%) (92.22% - 95.93% confidence intervals). They rarely ask questions in the new to java forum. With an average duke star rating of 667 (611 -723 confidence interval) they seem to have gathered substantial credibility. Gold members are involved extensively outside the new to java forum as well as apparent from the difference in total posts and posts to new to java.

3.4.1.5 PLATINUM MEMBERS

Platinum members represent the inner core of the community structure, and participate in the highest numbers. Representing about 3.6% (3.13% - 4.06% confidence interval) (P<0.0001) active members in the forum per month, they represent the third largest active group interestingly more than the number of gold members. They exhibit the same behavior as that of gold members, however in a more pronounced fashion. They outnumber the gold members in terms of the number of replies provided and also in the average duke star rating (1905 duke points) (1550 – 2258 confidence intervals). They like gold members are posts extensively outside new to java.

3.4.1.5 COMPARATIVE SUMMARY

	Bronze	Silver	Gold	Platinum
Average Duke Stars points earned	2	200	667	1905
Average number of such members active on the forums on a monthly basis	831	37	17	33
Average Total Posts	23	1578	4852	15,976
Average number of replies per month per user	4	18	43	80
Average number of questions per month per user	2	1	*2	*1
Average number of users posting replies	245	33	15	33
Average number of users asking questions	628	2	0	0
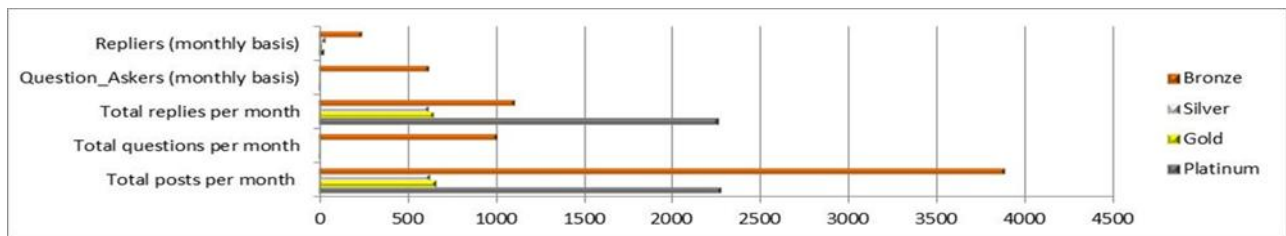
Figure 5: Comparative summary of different groups

3.5 EXTERNAL EVENTS – EFFECT OF ENVIRONMENTAL JOLTS

3.5.1 REVIEW OF CONTEMPORARY METHODS FOR EVENT ANALYSIS

Conforming to the definition of external events and environmental jolts (defined as transient perturbations whose occurrences are difficult to foresee and whose impacts on organizations are disruptive and potentially inimical) (Meyer, Gaba, and Colwell, 2005), several methods were identified as possible alternatives to studying such situations in organizational fields. Ethnographic study is one such qualitative method which has been highly cited, has a long and a distinguished history and has the benefits of the direct involvement of the researcher in the environment (key insights, direct observations) (Sanday, 1979). However the limitation of ethnographic studies is the possible sample bias that may creep into the study or the peril of deciding based on a small sample and lack of comparisons across the entire population (Sanday, 1979).

Another method that has been greatly studied and often used in business and medical studies for event studies and impact analysis is the Test vs. Control methodology. It involves a test of hypotheses that the behavior of a test group (who are exposed to the event being studied) is similar to the behavior of a control group (the control group is generally a smaller sample from among the population that makes up the test, who remain in the same environment as the test, apart from the fact that they are not exposed to the event being studied). If enough evidence is found to reject the null hypotheses, the researcher can claim some impact of the event. Although this looks to be a viable method for this study, but the absence of a defined control discourages the use of this methodology. Also since the type of event being studied - environmental jolts are exogenous events outside the control of the researcher, the test vs. control methodology cannot

be used (Since the test vs. control methodology posits the study of an event administered by the researcher).

A natural experiment method, which is commonly used in a variety of disciplines such as sociology, economics, political science and medicine (Meyer, 1982) is another alternative. In a natural experiment the treatment by definition is not administered by the experimenter, thereby providing the opportunity to examine data as real – world experiments. This method unlike ethnographic studies does not require the direct involvement of the researcher, nor does it necessitate the presence of a defined control group (unlike test vs. control methodology).

3.5.2 NATURAL EXPERIMENT METHOD

This method entails a simple post – vs. –pre analysis of the group and looking for significant changes due to the event on certain key attributes identified beforehand. Hence this method is particularly useful for examining the effects of public announcements on user participation as in the study undertaken here.

In this study, the changes in participation in MySQL newcomer forum before and after two takeover announcements – the January 2008 announcement of MySQL's takeover by Sun Microsystems, and the April 2009 announcement of Sun's takeover by Oracle were examined. Further to support the validity of the findings the New to Java forums (open sourcing of Java by Sun (Nov. 19, 2006) and acquisition of Sun (and consequently of Java) by Oracle (April 20, 2009)) was studied under the same scope.

3.6 ROLE OF STAR MEMBERS – LOGISTIC REGRESSION

Statistical models in empirical research can be widely classified as either explanatory models or predictive models. Explanatory models test hypotheses that specify how and why certain empirical phenomena occur (Gregor, 2006). Predictive models are aimed at predicting the future or new observations with high accuracy.

The goal of explanatory models is to shed light on a hypothesized causal relationship between an outcome and a set of inputs. The idea characteristics of such models should include the easy interpretability as well the ability to provide insights into the importance of each of the inputs. Regression type models are popular in explanatory modeling since they provide for each input a coefficient (with a sign and magnitude) and an associated p-value for ranking their importance (Shmueli, Galit, and Koppius, 2007).

Logistic models belong to the family of regression type models and are useful in problems where the dependent variable takes on only a few discrete values. It has been used extensively in econometrics, biostatistics, and educational testing. This study considers the special case in which the response is dichotomous (binary). Logistic regression is by far the most widely used technique for modeling binary response data, although other GLM procedures are used sparingly. This study uses multivariate logistic regression to model the probability of a question posted on the forum being answered, and explores the significance of the involvement of expert/experienced members in the thread.

3.6.1 DATASET FOR ANALYSIS

The dataset for the study consisted of posts by newcomers from the 'bronze' and 'silver' category. Threads with an 'answered' or 'unanswered' tag were selected. Threads without any

such status were excluded from the study. For each thread, information on the number of replies received, the number of different users involved in the ensuing discussion, the number of clarifications requested by the original poster and the star status of individual users who took part in the discussion was gathered. The final dataset for analysis was essentially a longitudinal data with 15,331 records, and each record indicating a question (with ans/unans tag available) posted by a bronze or a silver 'starred' member between Feb 2007 and Jan 2010.

3.6.2 DEPENDENT OR RESPONSE VARIABLE

The status of the question was modeled as the response or the dependent variable. The question was coded as a 1 if the question status indicated that it was 'ans' and coded as a 0 if it was 'unans'. The New to Java forum allows the original help seeker to classify the question as answered if it obtains a satisfactory answer or the problem posted is solved. However the question asker maintains the status as 'unans' unless the question receives a legitimate reply. In the dataset analyzed, which consisted of 15,331 questions, 6,376 or about 41.59% were answered and coded as 1's. The rest were unanswered and coded as 0's.

3.6.3 INDEPENDENT VARIABLES

The independent variables for the model were based on the quantitative information available on each thread or question. The independent variables were chosen based on the following set of hypotheses.

Hypothesis – I

If everything else is held constant, the probability of a question being answered increases with the higher number of members with high 'star' status involved in the discussion.

This line of thought seems intuitive since, the involvement of higher number of members who are experienced and knowledgeable, and have earned high star status on the basis of help provided previously are likely to lead to a successful answer for a question posted. Also in online Open source developer communities, leaders tend to track each other and the involvement of one 'star' member is likely to bring in more participation into the thread and thus increasing the visibility of the post. This in turn is expected to increase the probability of it being answered. However a second line of thought exists, which points to the cases the participation of 'too many' star members leads to replies that deviate from the main topic. Empirical evidence suggests that sometimes these star members get engaged in knowledge and ego tussles of their own which finally leads to the discussion deviating or digressing from the original question. Sometimes this might also be a discussion centered around the topic but beyond the scope of the newcomers (original help seekers) understanding.

The number of 'star' members in the discussion can be quantified by the two variables:

Dist_platinum_members: Quantifies the number of members with a 'platinum' duke star rating involved in the discussion

Dist_gold_members: Quantifies the number of members with a 'gold' duke star rating involved in the discussion.

Their participation in a given thread have been quantified by the variables

Posted_by_platinum_members: Number of replies to a post by platinum starred members

Posted_by_gold_members: Number of replies to a post by gold starred members

Hypothesis – II

Holding other things constant, an increase in participation of the number of non- star members or newcomers in the discussion is likely to increase the probability that the question is answered.

Newcomers to developer forums are mainly help seekers. Although recent studies have revealed that there exists a group of newcomers who apart from asking questions also tend to answer questions posted by their peers in the forum. A smaller group from among these individuals joins the forum to answer questions rather than pose them. The motivational aspect behind such behavior can be traced back to the same aspects that have been attributed to contribution to open source software development. However the number of such individuals among newcomers is fairly small and the newcomer population by and large remains a "help seeking" entity. With more participation from newcomers a thread/question can come into the limelight thereby attracting experts as well. It might on the other hand lead to a cluttered thread with lot of unrelated discussions, multiple question–answer sub threads and digressing sub topics within the original thread.

The number of 'non-star' members in the discussion can be quantified by the variable:

Non_plat_and_gold_repliers: Quantifies the number of members with a 'Silver' or a 'Bronze' duke star rating involved in the discussion

Their participation in a given thread can been quantified by the variables

Non_plat_and_gold_replies: Number of replies by members with 'Silver' or a 'Bronze' duke star ratings

3.6.4 MODEL SPECIFICATION

A logistic regression technique was used to model the log odds of the outcome (Response variable or the dependent variable) as a linear combination of the predictor or the independent variables. The logistic formula is stated in terms of the probability that Y (Thread is answered) = 1, which is referred to as $\hat{p}$. The probability that Y is 0 (thread is unanswered) is 1- $\hat{p}$. The following model is used to test the hypotheses.

$$
\begin{aligned}
\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = {} & \beta_0 + \beta_1 \times \text{Posted\_by\_platinum\_members} \\
& + \beta_2 \times \text{Posted\_by\_gold\_members} \\
& + \beta_3 \times \text{Dist\_platinum\_members} \\
& + \beta_4 \times \text{Dist\_gold\_members} \\
& + \beta_5 \times \text{Non\_plat\_and\_gold\_repliers} \\
& + \beta_6 \times \text{Non\_plat\_and\_gold\_replies} + \varepsilon
\end{aligned}
$$

3.6.5 CRUCIAL ASSUMPTION – INDEPENDENCE OF OBSERVATIONS

Although logistic regression does not assume that the variance of the independent or the response variable is constant (homoscedasticity), however one of the key assumptions to logistic regression models is the independence of independent observations or the absence of multicollinearity. Multicollinearity in logistic regression models is a result of strong correlations between independent variables.

The existence of multicollinearity inflates the variances of the parameter estimates, which is particularly pronounced for small and moderate sample sizes. Strong multicollinearity may also result in lack of statistical significance of individual independent variables while the overall model may be strongly significant. Multicollinearity may also result in wrong signs and

magnitudes of regression coefficient estimates, and consequently in incorrect conclusions about relationships between independent and dependent variables.

3.6.5.1 MULTICORRELATION DIAGNOSTICS

As a first step to detecting multicollinearity the correlation matrix for the individual variables was observed. Figure 6 represents the correlation matrix for the independent variables.

Correlation Matrix	Posted_by_platinum_members	Posted_by_gold_members	Dist_platinum_members	Dist_gold_members	Non_plat_and_gold_replies	Non_plat_and_gold_repliers
Posted_by_platinum_members	1					
Posted_by_gold_members	0.28935	1				
Dist_platinum_members	**0.71753**	0.14384	1			
Dist_gold_members	0.13473	**0.74816**	0.07015	1		
Non_plat_and_gold_replies	0.38737	0.23326	0.20505	0.13771	1	
Non_plat_and_gold_repliers	0.2112	0.15821	0.16088	0.11452	**0.79476**	1

Fig 6: Correlation among independent variables (p <0.0001).

Although a few variables were found to be highly correlated, but to check interdependencies among variables, multicollinearity diagnostic statistics produced by linear regression analysis were used (PROC REG with options VIF TOL in SAS). The collinearity diagnostic statistics are based on the independent variables only, so the choice of the dependent variables does not matter. An independent variable clarifications_sought (not one of the independent variables in the logistic model) was used as the dependent variable in linear regression. The Tolerance and Variance Inflation Factor (VIF) values for each variable were examined. The VIF indicates the number of times the variance of the corresponding parameter estimate is increased due to multicollinearity as compared to as it would be if there was no multicollinearity. However there is no formal cutoff value to use with VIF for determining presence of multicollinearity. Values of VIF exceeding 10 are often regarded as indicating

multicollinearity, but (Allison P.D., 1999) argues that in weaker models, which is often the case in logistic regression, values above 2.5 may be a cause for concern although it is a strict estimate.

The highest VIF observed was 2.58. However the variable was not directly removed from the analysis, as the VIF value was not found to be very high. Various researchers have argued for the use of correct model selection and nested regression models to deal with such cases (where it is hard to decide on redundant variables and directly drop them). Best subset regression and stepwise regression were used for model selection and a significance level of 0.05 was selected for variable entry while a significance level of 0.10 was required for the variable to stay in the model.

	Tolerance	VIF	Pr >|t|
Posted_by_platinum_members	0.47	2.14	<.0001
Posted_by_gold_members	0.40	2.52	<.0001
Dist_platinum_members	0.46	2.16	<.0066
Dist_gold_members	0.47	2.13	<.0037
Non_plat_and_gold_replies	0.39	2.58	<.0001
Non_plat_and_gold_repliers	0.42	2.38	<.0001

Figure 7: VIF and tolerance values for independent variables in the dataset.

3.6.6 INFLUENTIAL DIAGNOSTICS AND TREATMENT

Influential observations or outliers reduce the accuracy of the maximum likelihood estimation. In particular, remote points potentially have disproportionate impact on the parameter estimates, standard error, predicted values, and model summary statistics.

Influential diagnostics were performed on the dataset. (Hoaglin and Welsh, 1978) argue that points with hat distance $h_{ii} > (2p/n)$ (where p is the number of independent variables and n the number of observations) indicate potential leverage points and warrant close examination. These

elements are extreme points in design space. Plots of Pearson residuals and the deviance residuals were also inspected to identify elements that can potentially be influential points. Index plots of DFBETAS and DFFITS were drawn out to identify elements that cause instability parameter estimates.

While a divided opinion exists on the treatment of influential points, straightaway deleting such observations is not recommended (Bollen and Jackman, 1990). The recommended procedure of radical surgery of influential observations was followed and each observation was examined separately. This included a qualitative analysis of the actual threads which were the basis of each such observation. Although the New to Java forum is a strictly technical discussion forum, and is characterized by technical discussions, question and answer sessions, however non- technical general discussions do creep in. Such non-technical discussions (82% of the leverage points) were excluded from the study as a response to such posts would be meaningless. Other observations were persisted with as there were no substantial arguments available for their removal. Figure 8 summarizes the result of the influential diagnostics performed on the model.
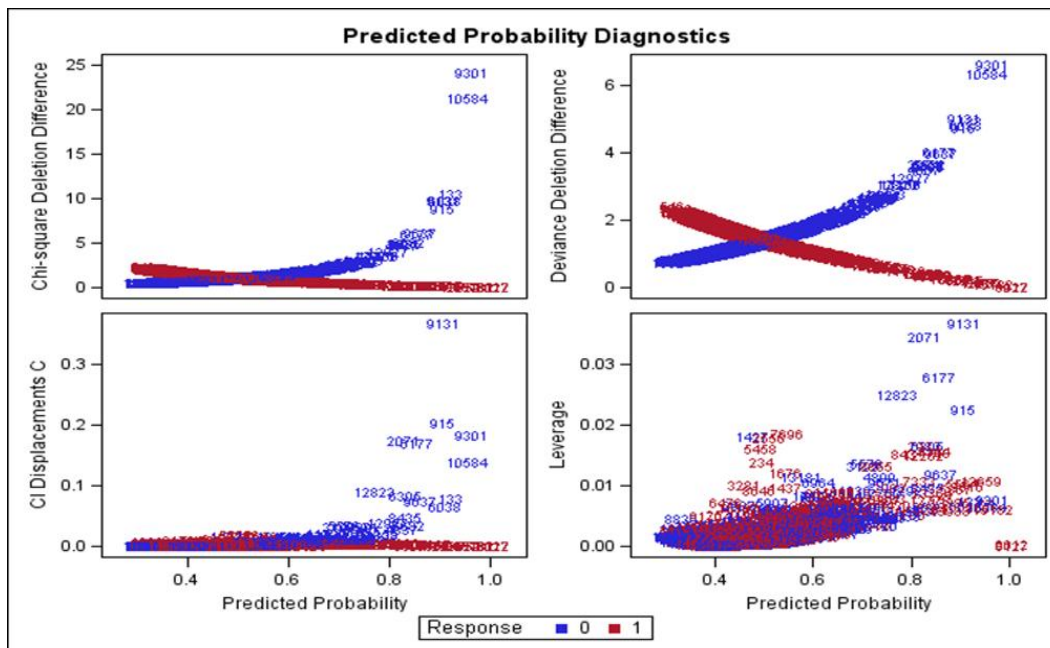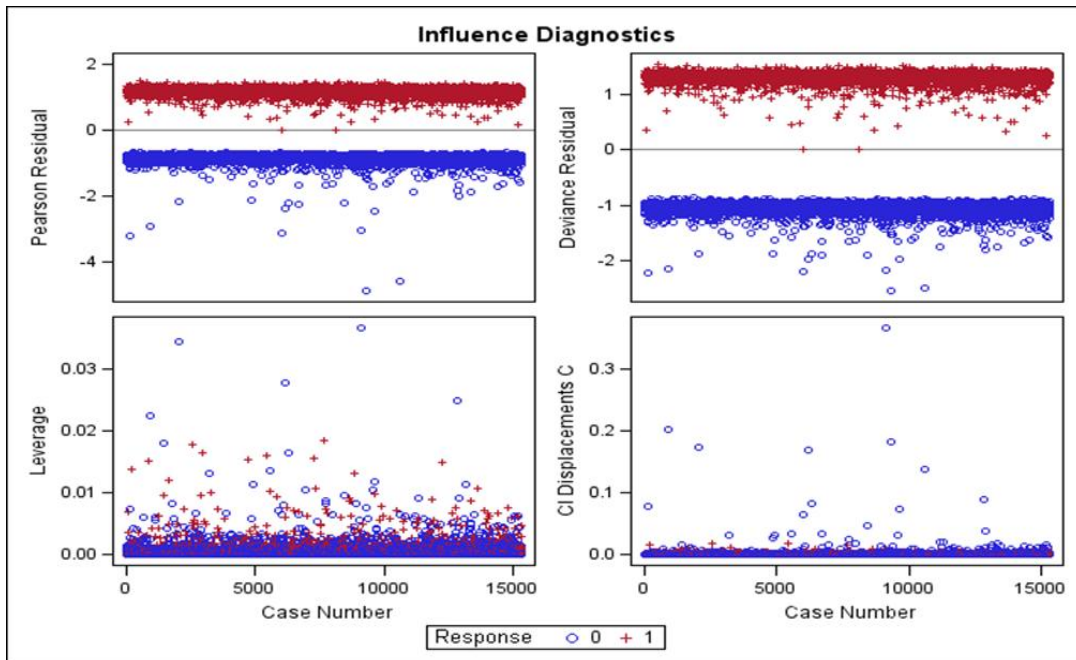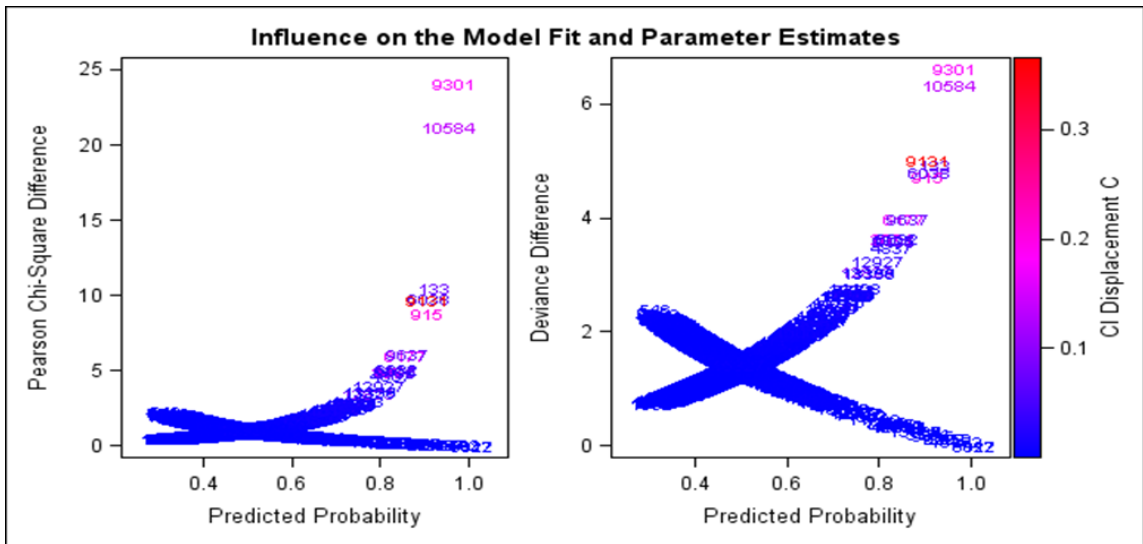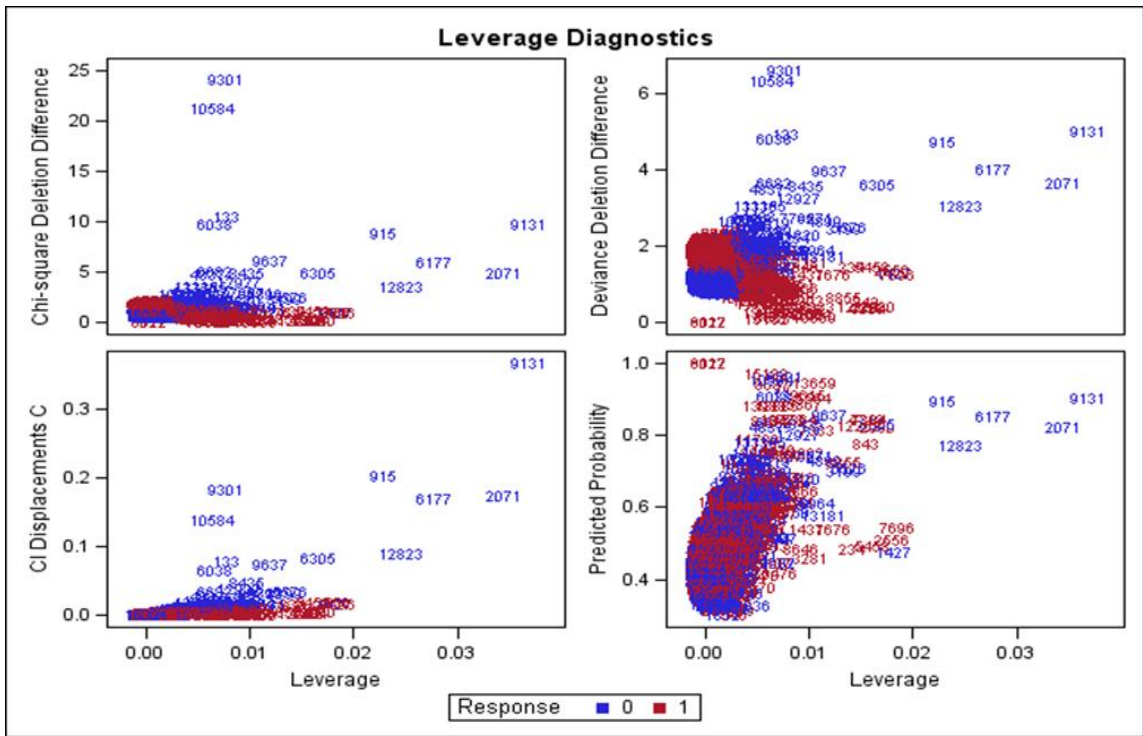
Figure. 8: Influential diagnostic plots

CHAPTER 4: FINDINGS

4.1 EFFECTS OF EXTERNAL EVENTS

As stated earlier in the section on research setting, a few events in the recent past of Sun and MySQL provided an useful setting for a natural experiment in which changes are being the result of external events rather than researchers' intervention. The first step was to examine changes in participation in online developer communities at New to Java (forums.sun.com) triggered by two significant events: (1) open sourcing of Java by Sun (Nov. 19, 2006) and (2) acquisition of Sun (and consequently of Java) by Oracle (April 20, 2009). Both events represented a change either in the nature of product or in the ownership of the product and consequently the stewardship of the associated online community.

Quantitative analysis of the participation parameters of the new to java forum before and after the announcement of the open sourcing of Java by Sun revealed significant changes in participation. The average of new forum users went from 16.54 per day to 20.56 per day before and after the open source announcement. ($t = -2.31$, $P < 0.01$).
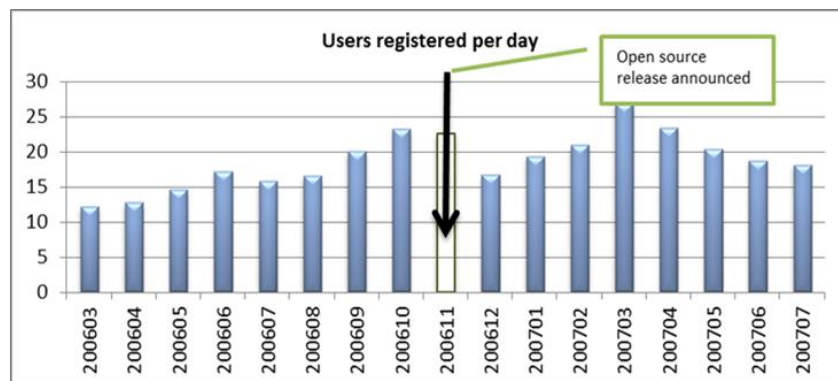
Figure 9: Average number of users registered before and after the open sourcing announcement

While the Average number of questions (threads) posted per day went from 31 to 43.87 (t = -3.01, p <0.01) the average number of messages posted per day increased from 190.18 to 310.47 (t = -3.732, p<.01). The number of messages in the forum was highly correlated to the number of threads posted on the forum with (r = 0.97) before the announcement and (r = 0.866) after the announcement.
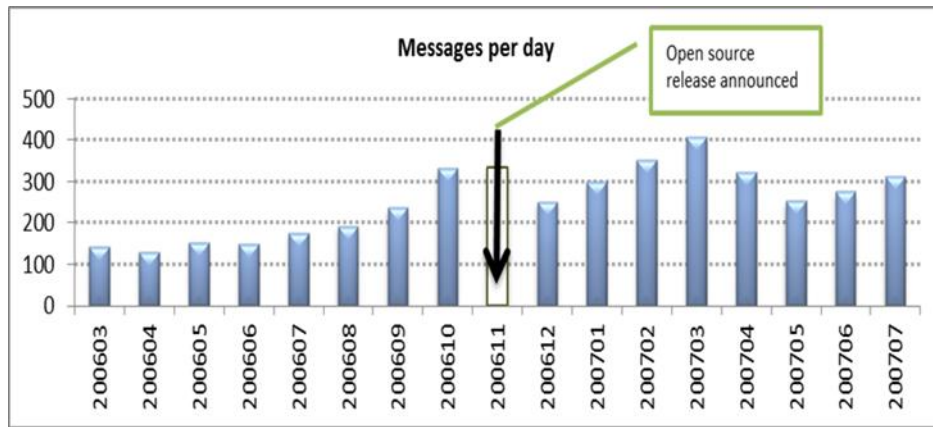
Figure 10: Average number of messages posted (per day) before and after the open sourcing
announcement

An analysis of the number of distinct active users on the forum revealed that there was an increase from an average of 862 per month before the announcement to an average of 1110 users after the announcement. (t = -3.10, P <.01). The number of active users on the forum over the month was highly correlated to the number of messages posted to the forum on the monthly basis (r = 0.95 before the announcement and r = 0.98 after). Also the number of active members in the forum on a monthly basis was highly correlated to the number of new users joining the forum in that given month (0.989 before the announcement and 0.934 after the announcement).
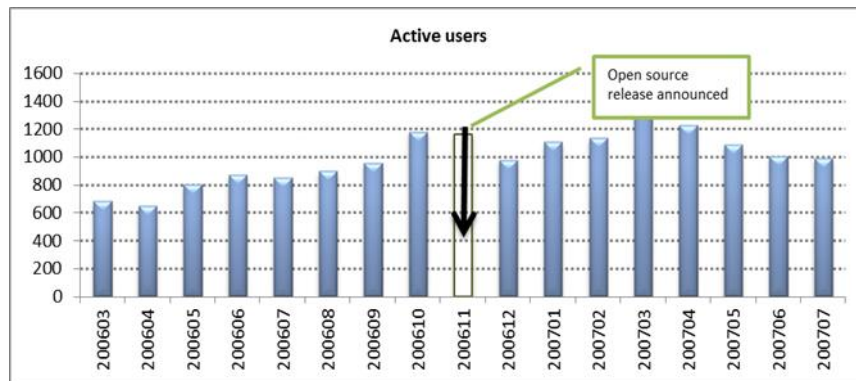
Figure 11: Average number of active users in the forum on a monthly basis

Activity broken down by user status indicated a sharp increase in the number of Bronze users (Average of 776.75 before the announcement to an average of 1020.65 after the announcement (t= -3.10, P<.01). The other groups of users also showed a steady and significant increase in number although the jump in numbers was not large. Table 4.1E provides the statistics broken down by groups.
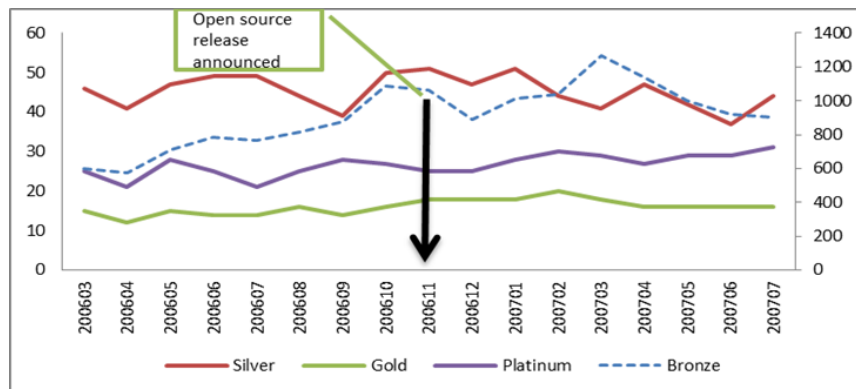
Figure 12: Activity of different groups on a monthly basis

Table 2: Activity statistics for different groups before and after the event.

	Average before	Average after	t statistic	P value
Bronze	776.75	1020.63	-3.11	<.01
Silver	45.63	44.13	0.79	<.01
Gold	14.50	17.25	-3.56	<.01
Platinum	25.00	28.50	-4.14	<.01

To further substantiate the perception of the user community on Sun's open sourcing announcement of Java, an analysis of major news outlets (both online and offline) using online search tools and databases such as ProQuest was performed. User sentiments based on their comments to online news items. Although the views were mixed, they generally leaned towards a positive feeling. Figure 13 showcases a representative sample of comments.

"I really appreciate this. I like the Java Platform, but as a advocate of free software, I never felt quite happy with using Java. Now it is time for me to blow off the dust of all those Java related books I own and start coding in Java once more." Diatha ,Nov 13 2006

"While I'm psyched about the fact that they went whole-hog and went GPL, there are a number of commercial Java Applications out there. The fact that they have this "Classpath Exemption" makes it sound more like the LGPL. But at any rate, it good that Java is now suitable for Linux." Mr. McD,Nov 13 2006 Source : http://developers.slashdot.org/article.pl?sid=06/11/13/0724252

Figure 13: Sample of user comments on popular OSS news sites on the announcement of Sun's open source plans

I think that open sourcing Java will mean very little for developers tomorrow. Very few are really that interested in doing anything with the source. But Short term it will make Java much more acceptable in the open source community. Long term it adds credibility to Java. If something really bad happen to SUN, then Java would be out of reach for predators. http://www.theserverside.com/news/thread.tss?thread_id=41760 (August 15 2006)

I haven't given this a lot of thought, but for such a significant event I'm surprised at the low response on this thread. Compare this with the Anti-Ruby flame baiting and you'll see what I mean. Personally, I'm not sure that the average Java shop wants this type of responsibility. After all one of the percieved benefits of Java is that it's and backed controlled by a "leading" vendor. This is a significant selling point when you consider the FUD factor present in a lot of organisations. It will be interesting to see how the Java community (both developers and other vendors) respond to this news in the comming months. http://developers.sun.com/portalserver/reference/techart/open-src-overview.html (October 19, 2006)

Figure 14: News of an impending open source offering by Sun was available as early as August 2006, User comments on published news even before the official offering

As a second step to the analysis process, the Oracle acquisition of Sun (April 2009) was studied. Key participation indicators of the New to Java forum were studied before and after the event. The average number of registered users (new users to the forum) per day dropped from 14.74 to 6.49 before and after the event (t = 3.94, p<.01). While the average number of discussion threads dropped from 31.76 to 15.19 (t = 4.20, p <.01), the number of messages posted on the forum per day decreased from 256.69 to 127.80 per day (t = 5.069, p<.01). The number of discussion threads and the number of messages posted are highly correlated (r=0.98 before the acquisition announcement, and r=0.93 after the announcement), as they are both facets of user interaction
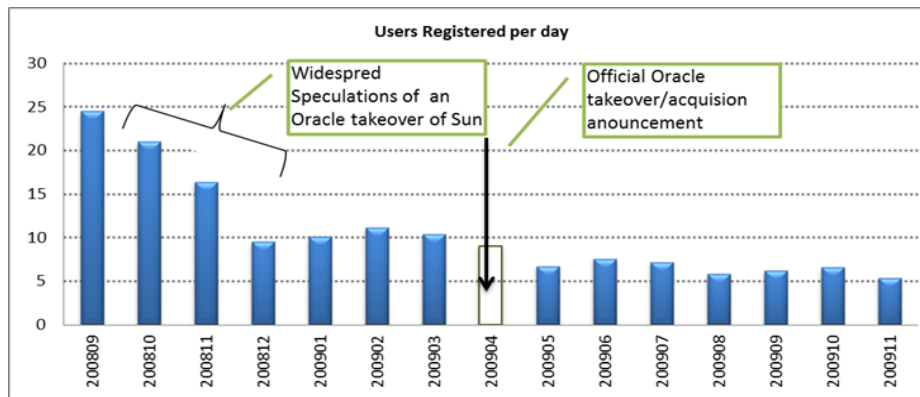
Figure 15: Average number of users registered in the forum before and after the official Oracle buyout of Sun announcement
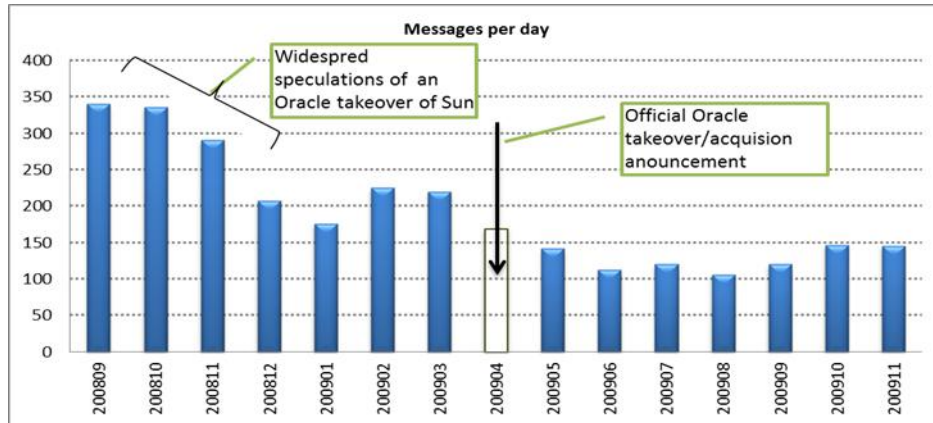
Figure 16: Average number of messages posted (per day) before and after the Oracle buyout of Sun announcement

However, the number of discussions per thread, or the number of messages posted to each questions posted on the forum, did not see any drop and actually increased slightly although the increase was not significant (average of 8.27 messages per thread before the acquisition announcement and an average of 8.41 messages per thread after the announcement ($t = -0.72$, $p<.3$) . This indicates that while there was little change in the help providing feature of the forum, there was a sharp decrease in both new user registration and the number of questions being asked. A sharp decline in the activity in the forum was observed in conjunction with the rumors of an impending buyout of SUN. The average number of distinct members active on the forum saw a sharp drop in the months leading to the announcement. The average number of users active on the forum on a monthly basis before the announcement was 907 as compared to a paltry 477.287 users after the announcement ($t = 4.84$, $p<.01$). However this decline was dominated primarily by the drop in participation among new users, while the more experienced members continued to participate in almost a similar fashion both before and after the event.

Table 3: Activity statistics for different groups before and after the event (Oracle takeover)

	Average before	Average after	t statistic	P value	Correlations	Total Activity
Bronze	832.43	413.86	4.65	0.00	Bronze Member Activity	0.99977
Silver	31.71	23.86	9.48	0.00	Silver Member Activity	0.64023
Gold	13.43	13.00	0.63	0.28	Gold Member Activity	0.29736
Platinum	29.43	26.57	3.33	0.01	Platinum Member Activity	0.49647
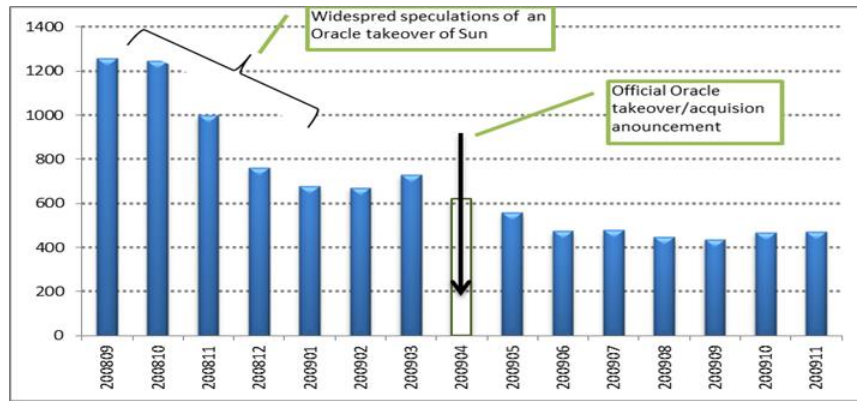
Figure 17: Average monthly activity before and after the Oracle buyout of Sun
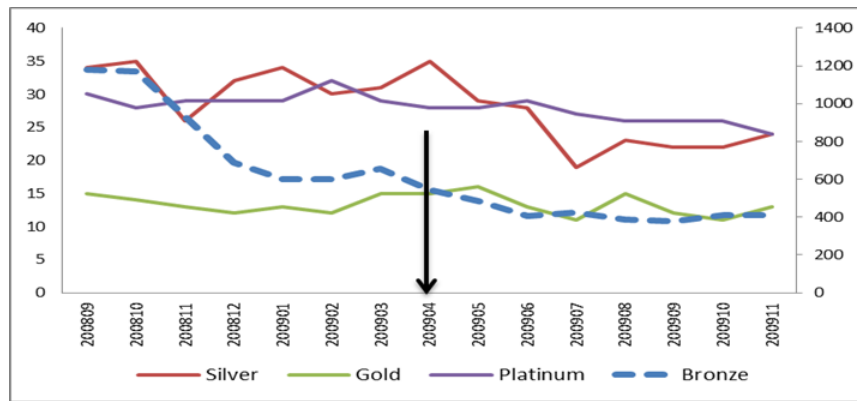
Figure 18: Activity of different groups on a monthly basis

A similar analysis of news sources and was performed to investigate the perception of the user community on the Oracle takeover of Sun:

> *"I find this all very sad. Sun was one of the Unix innovators from the earliest days. Even when they grow large, they still seemed like a "cool company." … Oracle seems to be the antithesis of this; major, corporate, gouging, monster... One can only hope that some of Sun's culture and products will survive.*
>
> *PhotoGuy, April 20 2009.*

Figure 19: User comments from Slashdot after the official announcement

> *"Just hope that Oracle doesn't make all that hard work the FOSS community go to waste by making it proprietary. In that case, expect to see protests by both FOSS developers and leaders within the FOSS protest, like Richard Stallman*
>
> *." Ungood++ Apr. 20th 2009*
>
> *"Oracle doesn't care about their paying customers, so why would they care about anything Open Source?"*
>
> *eldonyo Apr 20th 2009*

Figure 20: User comments from Arstechnica after the official announcement

> *Oracle (the organization) is the root of all evil. Anyone aquired by them will be changed for the worse.*
> *ctgman Jan 20th 2009*
>
> *While Sun may not be the strongest FOSS advocate, they've made many adjustments over the past few years to open up several products. Is Oracle likely to have the same philosophies when it comes to this stuff? I don't know Oracle as an organization too well, but I have a feeling they'd go into 'lockdown' mode of Sun's projects if they bought 'em.Thoughts?*
> *Anonymous Cowardon Jan 20th 2009*

Figure 21: User comments from Slashdot before the official announcement

A review of available news and discussions available both online and offline showed that users and developers were aware of a potential buyout of sun before the official announcement. Initial rumors were rife in the development community as early as late 2008. However the rumors subsided with IBM closing in on Sun for a buyout. This was observed in a sharp drop and a subsequent stabilization of participation in the forum.

An analysis of the participation of heavy contributors during the period studied around both the events (Sun open sourcing and the Oracle acquisition), revealed no significant change in

contribution. Heavy users in the forum are primarily veteran developers who provide help to newcomers (Platinum and Gold members as highlighted in the user activity graphs). Qualitative analysis of the threads revealed that the content was similar to that before the jolts: The questions were still related to technical issues and the responses were similar to those before the jolt. There were no discussions on forum regarding the open source announcement. A qualitative analysis of the forum revealed that there were no messages that referred to the open source announcement.

To validate the trends observed in the New to Java forum, the MySQL newbie forums were also tested similarly. Two significant events (1) Acquisition of MySQL by Sun (January, 2008) and (2) acquisition of Sun (and consequently of MySQL) by Oracle (April 20, 2009) were studied under the same scope as before.

The number of new users, who registered into the forum, was analyzed for the eight months before and after the acquisition announcement of MySQL by Sun Microsystems. The number of new users registered in the forum decreased slightly from an average of 15.56 before the Sun Acquisition announcement of MySQL to a 14.478 after the announcement, although the drop was not significant ($t = 1.13$, $p < 0.2$).
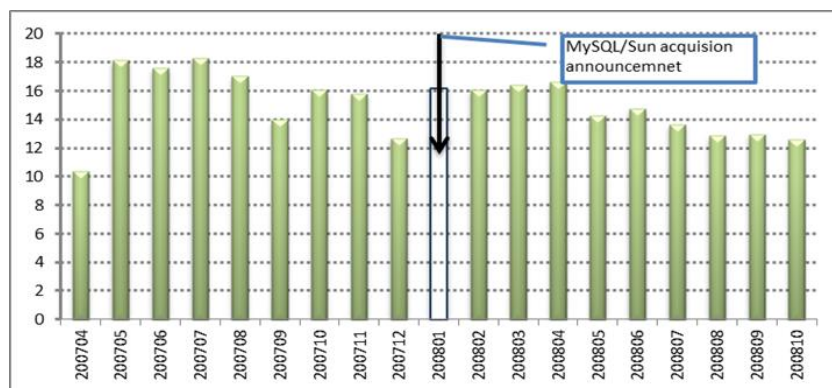
Figure 22: Average number of users registered per day before and after the announcement of MySQL acquisition by Sun

Analysis of online and offline media, developer forum discussions revealed mixed views on the Sun takeover of MySQL. User sentiment based on their comments posted on forums such as Slashdot, etc was found to be positive although unsure of the effect of the acquisition.

"There are many of us who have been working on MySQL for many years (my efforts with MySQL begin a decade ago). None of us are willing to move away from our open source roots. I've seen nothing that makes me think that Sun had any interest in doing anything foolish. They understand the value of MySQL being open source."

Krow, January 16 2008

"Are these great news? It's hard to know in which direction will big companies move. But if Sun keeps it's current track, I would say these are great news."

Slashidiot, January 16 2008

Figure 23: Positive user sentiment from Slashdot discussions on Sun's takeover of MySQL

It will stay open source

by devnull01 January 16, 2008 6:59 AM PST

With a quite good insight in Sun I can say that it's a sure bet that MySQL will stay open source. The commitment to open source in Sun is no gimmick it's acually the real thing.

Oh Really!

by maverick_nick January 17, 2008 12:15 AM PST

Well I guess that you don't know much about Sun, do ya?

Sun is a great company that is committed to open source, and good products. I think that this is great news for the MySQL community, because now finally MySQL can really begin competing with the proprietary systems like MS SQL Server which is brilliant, and would be pretty difficult to beat.

Tim O'Reilly | @timoreilly | Comments: 26 | 16 January 2008

I think Sun makes a great OS and has made good contributions to the standards community, but they aren't really all that great (at least not yet, despite trying pretty hard) at the "Open Source thing", and saying they are and backing it with self-proclaimed inaccurate studies don't make it so.

Robert [16 January 2008 10:33 AM]

I don't have any worries at all. I think Sun is pretty much going to

let it run as is.

Marco [16 January 2008 10:44 PM]

Figure 24: Mixed user sentiment from Slashdot discussions on Sun's takeover of MySQL

However a different trend was observed following Sun's acquisition by Oracle and a significant drop in the number of new user registration was observed. The average number of newcomers registering into the forum per day dropped from an average of 13.57 to 10.34 (t = 7.39, p<.01). The pattern observed in the new to java forum was also observed at the MySQL forums. A drop in participation was observed even before the official announcement of the takeover. There was also a significant drop in the number of threads (questions) per month on the forum (656.6 to 609, t = 2.11, p<.05).
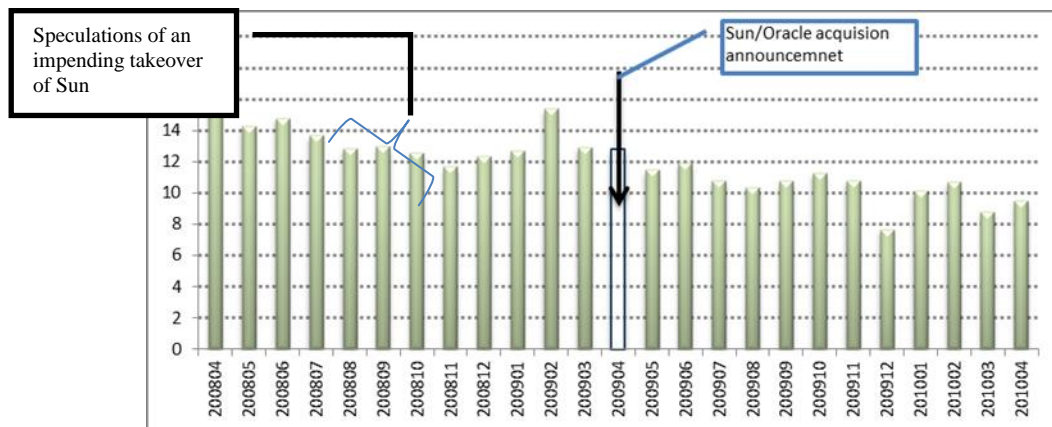
Figure 25: Average number of users registered per day before and after the announcement Sun's acquisition by Oracle.

A similar sentiment analysis performed over major news outlets revealed a very negative outlook about the future of MySQL after the Oracle takeover of Sun.

"I find this all very sad. Sun was one of the Unix innovators from the earliest days. Even when they grow large, they still seemed like a "cool company." … Oracle seems to be the antithesis of this; major, corporate, gouging, monster... One can only hope that some of Sun's culture and products will survive."

PhotoGuy, April 20th 2009.

> "…As for MySQL, the Oracle benefactors will say: do not worry, my dear people, we will keep it with true love, and gradually let it become deprecatingly obsolete."
>
> Linhares, Apr. 20th 2009

Figure 26: Negative user sentiment from Slashdot discussions on Oracle's takeover of Sun and thereby MySQL

A qualitative analysis of the threads revealed no change in content before and after the announcement and the drop in participation can be attributed to the negative perception of Oracle as the acquiring company.

4.2 TESTING FOR TRENDS IN EXTERNAL EVENTS DATA

The next step in the process was to determine if the changes observed in participation parameters in and around the events was in anyway biased due to the presence of any inherent trends in the participation data. To determine this, the key participation metrics were modeled as dependent variables (separately) with the timeline and the pre and the post event as (i.e. a binary dummy which is 0 before the event and 1 after) as the independent ones. Linear regression analysis on the New to Java forum data (against both the number of new users registered and the number of messages posted as independent variables) revealed that while there was a trend in the data, the events (i. Open Source announcement by Sun, and ii. Oracle takeover of Sun) had significant effect on the participation indicators.
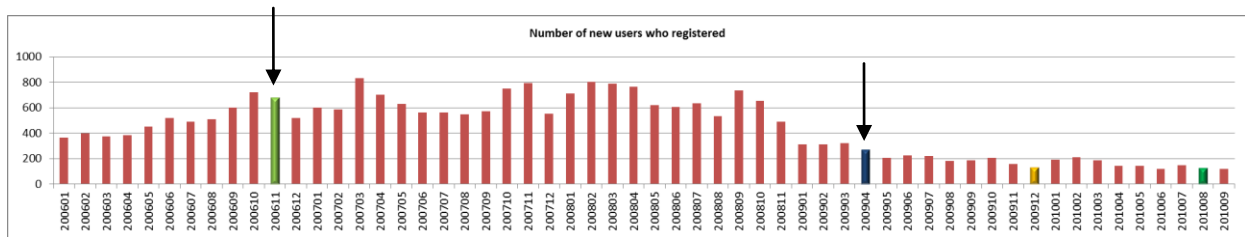
Figure 27 : Number of new users registered (per month figures) in the New to Java forum – Entire timeline
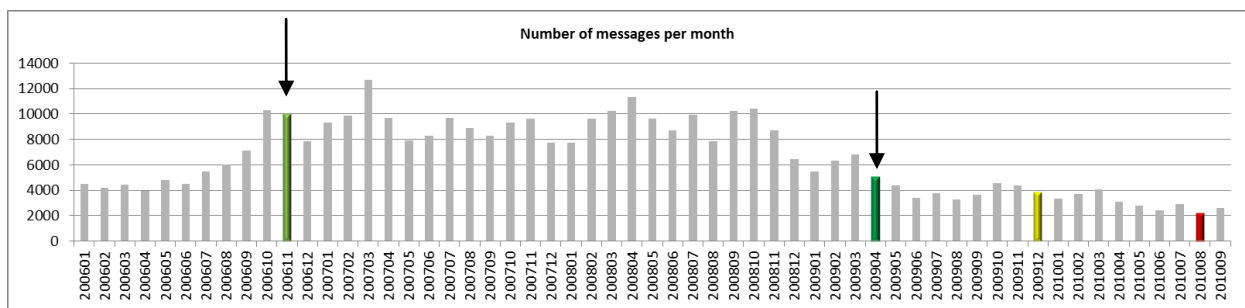
Figure 28: Number of messages posted on the New to Java forum on a monthly basis – Entire timeline

53

```
                                    The REG Procedure
                                    Model: MODEL1
                            Dependent Variable: Users_Registered

                          Number of Observations Read          57
                          Number of Observations Used          57


                                  Analysis of Variance

                                      Sum of         Mean
           Source                DF   Squares       Square    F Value    Pr > F

           Model                  3   2071455       690485      48.36    <.0001
           Error                 53    756683        14277
           Corrected Total       56   2828139


                      Root MSE           119.48659    R-Square     0.7324
                      Dependent Mean     446.66667    Adj R-Sq     0.7173
                      Coeff Var           26.75073


                                 Parameter Estimates

                          Parameter     Standard                              Variance
      Variable     DF      Estimate        Error    t Value   Pr > |t|   Tolerance    Inflation

      Intercept     1        177263        53675       3.30     0.0017          .            0
      timeline      1      -0.88114      0.26756      -3.29     0.0018    0.18662      5.35861
      Event_1       1     234.92851     61.27629       3.83     0.0003    0.42833      2.33465
      Event_2       1    -253.03467     62.13396      -4.07     0.0002    0.30999      3.22592
```

> No Multicollinearity problems as VIF < 10

Figure 29 : Regression model results highlighting the significance of both events, in terms of change in the number of new users registered

The overall model is significant (F test is significant). Individual parameter estimates show that both the events are significant. The effect of any underlining trend is also significant. A similar analysis was performed on the number of messages posted on the forum (on a monthly basis) as the dependent variable and the results re-iterate the significance of the events on the participation parameters

```
                          The REG Procedure
                            Model: MODEL2
                   Dependent Variable: Num_Messages

                 Number of Observations Read        57
                 Number of Observations Used        57


                         Analysis of Variance

                               Sum of         Mean
     Source              DF    Squares       Square    F Value    Pr > F

     Model                3   327359389    109119796      46.89    <.0001
     Error               53   123340371      2327177
     Corrected Total     56   450699760


            Root MSE            1525.50871   R-Square     0.7263
            Dependent Mean      6620.43860   Adj R-Sq     0.7108
            Coeff Var             23.04241
```
<div style="background:black;color:white;">No Multicollinearity problems as VIF < 10</div>

```
                         Parameter Estimates

                  Parameter     Standard                         Variance
Variable    DF     Estimate        Error   t Value   Pr > |t|   Tolerance   Inflation

Intercept    1      1908681       685278      2.79     0.0074          .           0
timeline     1     -9.48496      3.41604     -2.78     0.0076    0.18662     5.35861
Event_1      1   4378.57730    782.32645      5.60     <.0001    0.42833     2.33465
Event_2      1  -3535.68984    793.27649     -4.46     <.0001    0.30999     3.22592
```

Figure 30 : Regression model results highlighting the significance of both events, in terms of change in the number of messages posted

4.3 ROLE OF STAR MEMBERS

As identified in the section on methods, Platinum and Gold (Experienced and core) members essentially play the 'Help provider' role in the forum. It was also identified in the previous section on jolts that these groups of users did not show any change in participation or attrition although significant changes in the participation of Bronze (absolute newcomers) and (to some extent) Silver members was observed.

It was identified previously that newcomers, on joining a forum ask questions and look for answers (Miller and Jablin, 1991). The first step of the analysis was to validate this statement in the context of the forum under study. Quantitative analysis was performed to determine the percentage of users who came back after their first post (question/thread) to post a second question against the one who never posted another question (after the first post). Analysis revealed that about 39% of users who received a correct answer to their first question came back to post another question. However among users who did not get their first question answered, only 29% came back to post another question again.
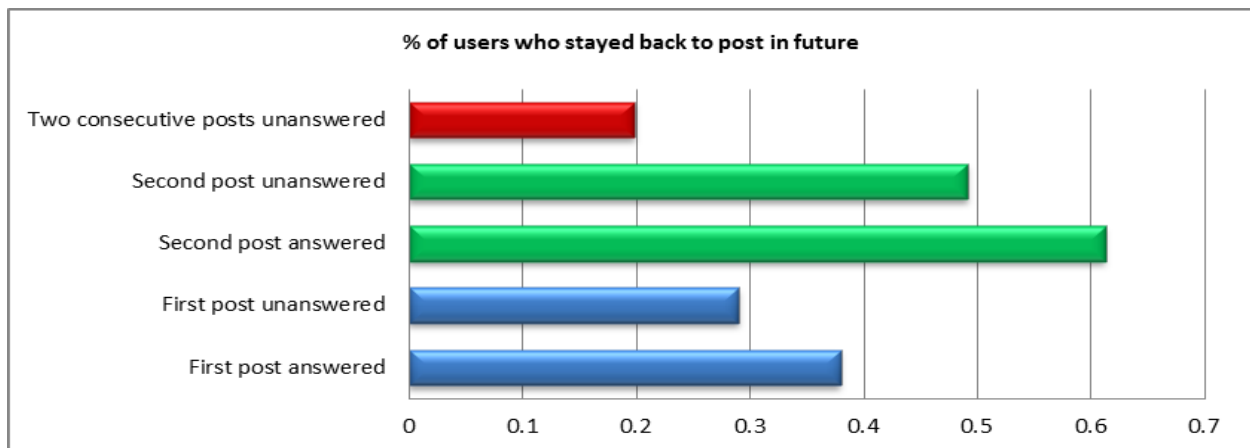
Figure 31: Percentage of returning newcomers after depending on the fate of their questions in the community

Among users who received a correct answer to their second post on the community, almost 62% returned to post another question, while among users who did not receive a correct answer to their second question only 49% returned to post a question again. However for users who did not receive a correct answer to the first two questions on the community, the dropout rates were substantial and only about 20% returned to post again (Analysis excludes users who posted their last question in Feb 2010, for allowing them some time to come back and post again – data was captured as of Feb 12th 2010)

The higher percentage of returning newcomers indicated that receiving correct answers to questions posted is an integral part of their participation characteristics. This in turn indicated that the onus is on the experienced members of the community to provide correct or acceptable answers to question posted by the newcomers. While the sheer number of replies provided by experienced members (platinum and gold) seem to indicate that they are thoroughly engaged in doing so, further analysis on the actual outcome of such participation warrants a closer look.

As discussed in the section on methods, this study incorporated an analysis of the participation of experienced members in individual threads and their role in a successful outcome of the discussion that ensues. Also the idea was to look at individual threads from the perspective of a team activity and investigate the critical mass, group size theories that are associated with actual groups. Consistent with the hypotheses (as discussed in the methods section), preliminary analysis was done and percentage of answered/unanswered questions were tiered with the independent variables.
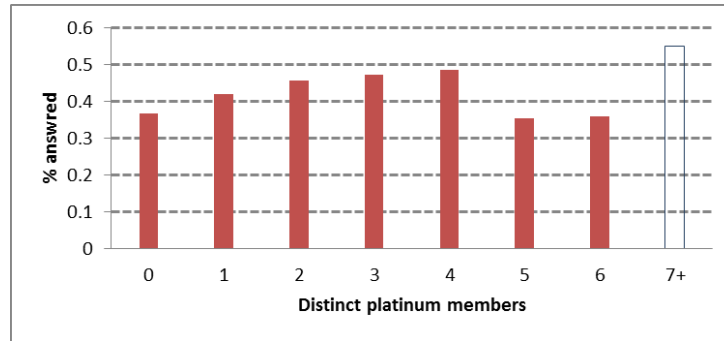
Figure 32: Percentage questions answered tiered with the number of distinct platinum members involved in a thread
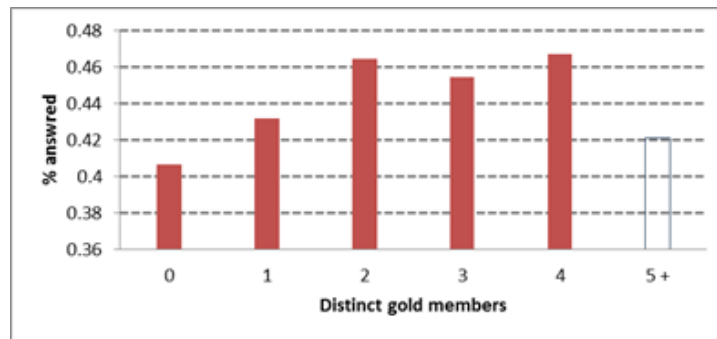
Figure 33: Percentage questions answered tiered with the number of distinct gold members involved in a thread
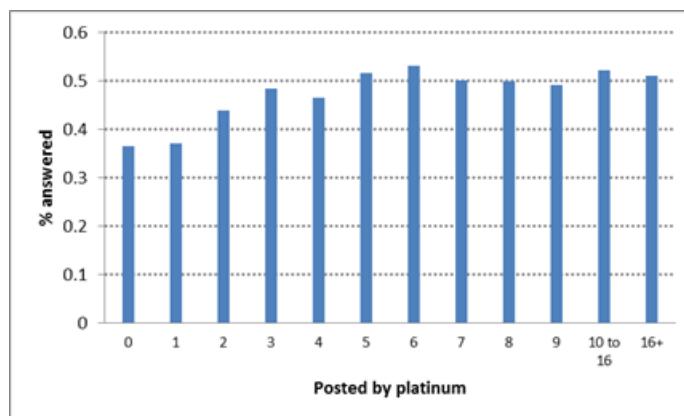
Figure 34: Percentage questions answered tiered with the number of posts by platinum members in a thread

The rough plots of the percentage of answered questions did reveal an interesting trend, and indicated that while the percentage of questions answered increased with the increase in number of distinct platinum or gold members participating in the discussion (question), questions with large number of such individual members had a lower answered rate. A logistic regression model was used to identify the significance of these independent predictors of response to questions posted on the forum by a newcomer

As detailed in the section on methods, the necessary pre - processing (correlation issues, influential diagnostics) were performed before running the final model. Best subset logistic regression and stepwise logistic regression methods were used for model selection. The variables selected by the STEPWISE method agreed well with the set of variables selected by the Best subset method. Five out of the six variables were chosen to represent the final model. The variable dist_gold_members was removed as it was not significant (P < .2).

Results of Best Subset Logistic Regression (Higher Chi - Square Score better)							
		Variables					
Number of variables	Chi-Square Score	posted_by_platinum_members	posted_by_gold_members	dist_platinum_members	dist_gold_members	non_plat_and_gold_replies	non_plat_and_gold_repliers
3	123.4476	•				•	•
3	110.89			•		•	•
3	108.0476	•	•				•
4	129.9422	•	•			•	•
4	129.4384	•			•	•	•
4	126.3852	•		•		•	•
5	133.6563	•	•	•		•	•
5	132.6605	•		•	•	•	•
5	130.6173	•	•		•	•	•

Figure 35: Results of Best subset Regression - Model selection procedure

The logistic regression model was run without any interaction terms and included the five independent variables identified above. The Convergence criterion (GCONV=1E-8) was satisfied indicating that the model was significant. Also the model fit statistics indicated that the intercept and covariates had a lower AIC, SC and -2 Log L than only the intercept indicating that the variables had some significant effect on the model. The global null hypotheses test was significant again indicating that the model was significant.

```
            Model Convergence Status Convergence criterion (GCONV=1E-8) satisfied.


                            Model Fit Statistics

                                              Intercept
                               Intercept          and
                Criterion           Only     Covariates

                AIC            19711.654      19468.612
                SC             19719.239      19514.122
                -2 Log L       19709.654      19456.612


                     Testing Global Null Hypothesis: BETA=0

            Test                  Chi-Square       DF      Pr > ChiSq

            Likelihood Ratio       253.0422         5         <.0001
            Score                  160.7130         5         <.0001
            Wald                   241.7274         5         <.0001
```

Figure 36: Model convergence status, fit statistics and global fit tests

Analysis of the parameter estimates revealed that all the remaining variables were significant in the model. A Detailed model summary (Analysis of Maximum Likelihood Estimates table summarizes information regarding the independent variable including parameter estimates, variability and significance while the Odds Ratio Estimates table summarizes the significant independent variables and indicates their associated odds ratio and confidence limits) have been provided in Figure 37. Odds ratio estimates indicated that the probability of a question

60

getting answered increased with the increase in number of posts by platinum members (odds ratio point estimate 1.133), posts by gold members (odds ratio points estimate 1.084), and posts by silver and bronze members (odds ratio point estimate 1.147). However the probability that a question would be answered decreased with the increase in the number of distinct platinum members involved in the discussion (odds ratio point estimate 0.921). The number of non-platinum and gold members involved in the discussion also led to a decrease in the probability of the question getting answered (odds ratio point estimate 0.815).

```
                    Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
     Parameter           DF    Estimate    Error   Chi-Square   Pr > ChiSq

     Intercept            1     -0.5300   0.0318    278.2029      <.0001
     posted_by_platinum_m 1      0.1252   0.0135     86.5281      <.0001
     posted_by_gold_membe 1      0.0808   0.0174     21.6032      <.0001
     dist_platinum_member 1     -0.0827   0.0297      7.7632      0.0053
     non_plat_and__gold_r 1      0.1372   0.0213     41.4232      <.0001
     non_plat_and_gold_re 1     -0.2050   0.0335     37.5122      <.0001




                           Odds Ratio Estimates

                            Point          95% Wald
          Effect          Estimate     Confidence Limits

          posted_by_platinum_m   1.133      1.104      1.164
          posted_by_gold_membe   1.084      1.048      1.122
          dist_platinum_member   0.921      0.869      0.976
          non_plat_and__gold_r   1.147      1.100      1.196
          non_plat_and_gold_re   0.815      0.763      0.870

     Association of Predicted Probabilities and Observed Responses

          Percent Concordant     57.3    Somers' D    0.177
          Percent Discordant     39.6    Gamma        0.182
          Percent Tied            3.1    Tau-a        0.086
          Pairs              51247014    c            0.588
```

Figure 37: Parameter estimates, Odds ratio estimates and Association statistics

The results indicated that while high participation of a limited number of experienced members in a thread was desirable, the involvement of many at a time was not. The model also indicated that the participation of non-platinum and gold members was also significant, and influenced the threads in a similar fashion as the participation of the experienced members. While these findings did support the initial set of hypotheses (outlined in the section on methods), it was confounding at the same time to capture the actual import of the dynamics without investigating the interaction among the independent variables.

As a next step to the analysis, a logistic regression model was developed to include two-way interaction among the independent variables. The interaction model was found to be significant with the satisfactory convergence of the GCONV=1E-8 criterion. The AIC, SC , and - 2Log L fit statistics for the interaction model were found to be lower than the model including only the intercept indicating that the variables in the final model had some effect. The global null hypotheses test was also found to be significant.

Inclusion of potential two-way interaction terms in the model suggested that the interaction between the number of posts by platinum members and number of distinct platinum members involved in a discussion was significant (P <.0001) and lead to a decrease in the probability of the question being answered (odds ratio 0.975407448). Among other significant terms that influenced the probability of the question getting answered were the number of posts by platinum members (odds ratio = 1.101860374) (P = <.0001) and distinct number of platinum members (odds ratio PE = 1.059714996) (P = <.0001). This clearly indicated that threads (questions) which involved high number of posts and high number of experienced members at the same time were less likely to be answered. This may be attributed to the possible digression from the original topic with too many knowledgeable members posting too many messages. This

might also be the case where in the knowledgeable members get into ego tussles among themselves trying to prove a point and the thread (question) is never answered or the help seeker is unable to figure out the correct reply.

Also the interaction term between the number of distinct platinum members and number of (relatively inexperienced) non platinum and gold members (silver and bronze members) was found to be significant in the model and led to a decrease in the probability of the question getting answered (odd ratio = 0.916768768) (P = <.0001). This indicated that too many eyes looking at a thread at the same time was undesirable.

The other significant variables in the two way interaction model included the number of posts by gold members (odds ratio = 1.101860374) (P = <.0001), number of posts by non-platinum and gold members (silver and bronze members) (odds ratio PE = 1.151654956) (P = <.0001), and the distinct number of non- platinum and gold members (odds ratio PE = 0.913382991) (P = 0.0176) involved in the discussion.

The effect of the involvement of the lesser experienced members in technical discussions was evident in both the models. While an increase in the number of posts by these lesser experienced members increased the probability of the question getting answered, the involvement of larger number of such members lead to a decrease in the probability. It can be argued that though posts from lesser experienced members might help bring a question/thread into the limelight, however a thread which is cluttered with too many digressive posts (more likely to happen with higher number of inexperienced members involved) lead to the question or thread not receiving correct replies. As is often observed in such threads, when many such inexperienced members are involved, the thread tends to digress as even experienced members

start commenting on, clarifying or even correcting the replies posted by the lesser experienced

members and eventually forgetting the main subject.

```
                    Analysis of Maximum Likelihood Estimates

 Parameter                                         DF    Estimate    Pr > ChiSq
Intercept                                          1     -0.7325       <.0001
posted_by_platinum_m                               1      0.1828       <.0001
posted_by_gold_membe                               1      0.097        <.0001
dist_platinum_member                               1      0.058        0.0833
posted_by_platinum_m * dist_platinum_member        1     -0.0249       <.0001
non_plat_and_gold_replies                          1      0.1412       <.0001
non_plat_and_gold_repliers                         1     -0.0906       0.0176
dist_platinum_member*non_plat_and_gold_repliers    1     -0.0869       <.0001


                          Odds ratio Estimates

          Parameter                                       Estimate
          Intercept                                       0.480706
          posted_by_platinum_m                            1.200574
          posted_by_gold_membe                             1.10186
          dist_platinum_member                            1.059715
          posted_by_platinum_m * dist_platinum_member     0.975407
          non_plat_and_gold_replies                       1.151655
          non_plat_and_gold_repliers                      0.913383
          dist_platinum_member*non_plat_and_gold_repliers 0.916769




       Association of Predicted Probabilities and Observed Responses

          Percent Concordant      57.7    Somers' D    0.185
          Percent Discordant      39.2    Gamma        0.191
          Percent Tied             3.1    Tau-a        0.090
          Pairs              51247014    c            0.593
```

Figure 38: Parameter estimates, Odds ratio estimates and Association statistics for the two way interaction model

CHAPTER 5: DISCUSSION

5.1 EFFECT OF EXTERNAL EVENTS ON ONLINE COMMUNITY PARTICIPATION

The motivation for this study was to investigate the effect of external changes and events on online communities, especially because this has not been studied to a great detail. OSS communities which are central to software development and innovation are a very important cog in the OSS development paradigm. A holistic understanding (not only internal design principles and community structure but an understanding of the environment in which these communities thrive) is necessary before such knowledge can be used to shape cyber-learning entities.

Although much has been explored about the internal design principles in such communities, very little is known about what role the environment may play, if any in the success or failure of such communities. This investigation becomes ever so important as it talks not only about OSS communities but the cognizance will be important for any community as such (many online communities are hosted and run by commercial firms e.g. Amazon, Usenet by Google, Yahoo Groups, Google Groups, Facebook, by professional associations, educational institutions, and other formal and information organization)

The analysis on the online communities for two open source products as undertaken in this study indicated that there were significant and substantial effects of external events on the participation. Participation studies in and around the open source announcement by Sun and the buyout of MySQL by Sun, indicated an increase in participation parameters while the acquisition of Sun by Oracle lead to a decrease in community participation. Meyer (1982) suggested that ideological issues are a major factor in how people respond to environmental jolts. The same can

be argued here in terms of the perception of the steward company and its values (open source values in this case). The users of the community may change their participation patterns although they may not be directly affected. Another line of thought that can be argued with is that the users reassess the relevance and potential benefit participation carries for them, and as a result, join, leave or change the intensity of participation. Both arguments seem valid and if reasoned with, the latter may just be the reason why significant increase in participation patterns were seen when the users assessed the event to have a positive connotation (Suns open sourcing Java, MySQL buyout by Sun). Users may expect the open sourcing to lead to an extensive use of Java and its expansion into new domains, thereby creating more opportunities for them. Career advantages, increased personal exchange with other software developers, and the chance to work on new applications can be possible pragmatic motives behind an increased participation. The review of news journals and developer forums undertaken as a part of the study indicated such motives.

However the drop in participation metrics in an around the Oracle acquisition of Sun can be better explained by the ideological theory as put forward by Meyer (1982). The perception and the values (open source values) associated with the new steward firm may not be the same as that of the old steward firm. This was seen as a part of the analysis of the online forums and news portals wherein the major discussions were centered on the image of Sun and Oracle as proponents of open source, and the difference in their business models. The motives of the users in this case can be seen to be more ideological in nature as the participation decreased although the other things associated with the community remained the same (the help giving, involvement of experienced members and the design parameters). Such ideological mechanisms in users may be triggered by speculations and rumors as well, and can result in significant change in

participation patterns even before the actual event. Speculations of a probable buyout of Sun were seen to effect participation in the analysis.

Another revelation from the study was that changes in newcomer participation patterns were more prone to such events, and even speculations or rumors of such events could have momentous effects on their behavior. While the experienced/core members in the community were found to be relatively unwavering to such environmental jolts (and particularly not to rumors), there participation patterns may be perturbed if the external event results in punctuated equilibrium or perpetuated disequilibrium scenarios as opposed to the 'seeking equilibrium' scenario in the case of environmental jolts being studied here (Meyer, Gaba, and Colwell, 2005).

The present study thus explores and sheds light on the effect of external events (environmental jolts) on community participation. Significant changes were found in community participation parameters, indicating the imminent need to extend our understanding of such events. While this is a first step in this process, further research is necessary to determine the steps that can be taken and the strategies that can be adopted to avoid the destructive effects of such events or leverage the positive effects and ride on the ripples created to new heights. This study argues in favor of the necessity to integrate learning's from organizational science into open source software community design principles although literature including Raymond's seminal work (Raymond, 1999) defines such development as a bazaar and where the definition of an organization is still fuzzy.

5.2 EXPERIENCED MEMBERS AND THEIR ROLES IN COMMUNITY SUCCESS

This study was undertaken to understand the effectiveness of the help provided by experienced members in developer communities. The study explored the typical discussion patterns in online developer communities, where experienced members post bulk of the replies while newcomers show a predominantly help seeking behavior. Drawing analogies on participation in groups and teams, the study tested certain hypotheses on the participation patterns of experienced members and non-experienced members and the subsequent effect of such patterns on the outcome. A quantitative analysis of newcomer forums at forums.sun.com (New to Java) forums was performed to test these hypotheses. While this may not have been the best method to tackle the problem (as compared to ethnographic studies), this study provided empirical evidence that individual discussion threads behaved as virtual groups and the presence of certain participatory patterns and order were more beneficial to the final thread (group) objective.

Empirical evidence was found, that supported pre-existing theories on the necessity of help, guidance and feedback from experienced members. The final logistic regression model found the number of posts by experienced members and the number of such distinct members involved in the discussion to be significant. Both these attributes contributed positively to the discussion making in more likely to be answered. However the interaction of these two variables had a negative bearing on the final outcome in the thread (while being significant in the model), thereby indicating that although the involvement of experienced members was necessary excessive crowding of such members and their inputs in the discussion may not beneficial.

Studies in organization science have already discussed this phenomenon and state that such results are likely because there tends to be more conflict in larger groups (Hare, 1952;

68

O'Dell, 1968), as they are less likely to co-operate with each another (Brewer and Kramer, 1986), failing to help people in need. Such dynamics were also observed in the community studied, where on a few occasions (out of a random sample of threads chosen for qualitative analysis) wherein experienced members got into lengthy debates on possible solutions to a problem posed by a newcomer, while the newcomer was seemingly lost. On a few occasions this debate ensued after a successful answer to the primary question posted was available, however follow-up questions by the newcomer went unanswered among the debate that ensued.

CHAPTER 6: CONCLUSION

6.1 FINDINGS AND RECOMMENDATIONS

The findings from the study on the effects of external events are prima facie, important for companies interested to involve users in activities related to their products, companies which are involved in the development and stewardship of open source software. The findings suggest that it is essential for such companies to actively address concerns that might arise due to changes such as acquisitions, even if such changes involve no change in the product itself.

The findings presented here also add to the growing body of knowledge on community, and open source software development participation, by providing evidence of the effect of external events on participation. Whereas the majority of prior work on online communities has largely focused on users' motivations for participation, community governance structure and the design of communities to foster participation, the present study sheds light on the role of external events.

Another issue the findings raise is the importance of a change of perspective, that is, in cases where community activity is a critical aspect of a firm's product, response to external events should be taken into account by company leadership before decisions are made. Strategies can be adopted that facilitate the monitoring of ongoing activities in an online community. Future research is needed to develop new ways of understanding the effect of external events on community participation.

Cyber-learning environments may also start to look in this direction and mull about the possible design parameters and strategies they might want to incorporate. Participation in online

learning environments can be influenced in the same way as was found during the study. This may be as a result of specific announcements regarding the scope of the coursework, or the tools and techniques taught.

The study on the role of experienced members in online communities pointed out the necessity to follow certain norms when helping newcomers. Interesting findings from the study indicated that the mere presence of large number of experienced members in online discussions may not enough, and actually at times may be detrimental for the final outcome of the discussion. The implication of this study in design of online learning communities is that it emphasizes the role of moderation in such communities. Effective moderation and management will be necessary to keep an eye on the erstwhile 'eyes' in the discussion. This may mean that moderators be separated from the experienced members in the forum and only perform the administrative task to avoid conflicts of interest.

6.2 LIMITATIONS AND ASSUMPTIONS

As with any research study, this study has certain limitations. For the study on external events the primary exogenous effect has been attributed to Oracle's and Sun's relative perception by the community. While this seems plausible, it may probably not be the only source of difference between the companies. Therefore, the findings should be interpreted cautiously. An analysis of Java developer communities, and online news sources (Slashdot) was performed to determine the perception in the general community about these events. However a deeper analysis of community sentiment, through surveys for instance, would make for a more convincing case.

The study on the role of experienced members made several assumptions, including the very basic assumption of tying the only positive outcome of a discussion to the benefit it provides to the original help seeker. The study has not captured the benefits of having healthy technical debates in such discussions which may be attractive to other members of the forum. Also, it is quite possible for some questions asked on the forum to be open ended, so as to not have a definite answer (However in technical software development forums such questions are far and wide).

REFERENCES

Allison, P. D. (1999), Logistic Regression Using the SAS System: *Theory and Application, Cary, NC: SAS Institute Inc*.

Ashforth, B. E., Sluss, D. M., & Harrison, S. H. (2007). Socialization in organizational contexts. *International Review of Industrial and Organizational Psychology*, 22, 1–70

Mockus, A., Fielding, R. T., and Herbsleb, J. D.(2002). Two Case Studies Of Open Source Software Development: Apache and Mozilla*, ACM Transactions on Software Engineering and Methodology, volume 11, number 3*, pp. 309–346.

Bauer, T. N., Morrison, E. W., & Callister, R. R. (1998). Organizational socialization: A review and directions for future research. *In G. R. Ferris (Ed.), Research in personnel and human resource management, vol. 16:* 149-214. Greenwich, CT: JAI Press

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P. and Kraut, R. E. (2004). Using social psychology to motivate contributions to online communities. *Proc. CSCW* 2004, 212-221

Beyer, J. M, (1981). Ideologies, values, and decision making in organizations. In P. Nystrom & W. H. Starbuck (Eds*.), Handbook of organizational design, (Vol. 2). London: Oxford University Press*, 166-197

Billings, R. S., Milburn, T. W. and Schaalman, M. L. (1980). 'A model of crisis perception: a theoretical analysis'. *Administrative Science Quarterly, 2 5, 300- 16*.

Bollen, K.A. and Jackman, R.W. (1990). Regression diagnostic: An expository treatment of outliers and influential cases. *In: Fox, J. and Long, J.S., Editors, 1990. Modern Data Analysis, Sage, Newbury Park, CA*, pp. 257–291.

Bourgeois, L. J. (1981). On the measurement of organizational slack. *Academy of Management Review*, 6: 29-39.

Callahan (1962). Education and the Cult of Efficiency. *Chicago: University of Chicago Press*.

Choi, B., Alexander, K., Kraut, R.E., and Levine, J.M. (2010). Socialization tactics in Wikipedia and their effects. *Proc. CSCW 2010*, 107-11

Cooper-Thomas, H. D., and Anderson, N. (2005). Organizational socialization: a field study into socialization success and rate. *International Journal of Selection and Assessment*, 13, 116–128

Cox A. (1998). "Cathedrals, Bazaars and the Town Council," *Slashdot*

Crowston, K. and Howison, J. (2004). The Social Structure of Free and Open Source Software. *First Monday*, 10(2).

Demil, B., Lecocq, X. (1999). Neither market nor hierarchy or network: the emerging bazaar governance. *Research in Organizational Behavior* 12, 295–336

Dennis, Alan R.; Valacich, Joseph S. (1993). Computer brainstorms: More heads are better than one. *Journal of Applied Psychology*, 531-537.

Diehl, Michael, Stroebe, Wolfgang (1987), *Journal of Personality and Social Psychology*, Vol 53(3), 497-509

Ducheneaut N. (2005) Socialization in an Open Source Software Community: A Socio-Technical Analysis. *Computer Supported Cooperative Work (CSCW)*, 14(4):323 368

Fang, Y. and Neufeld, D. (2009). Understanding sustained participation in open source software projects. *Journal of Management Information Systems*, 25(4):950

Feldman, D. C. (1981). The multiple socialization of organization members. *Academy of Management Review*, 6: 309-318.

Fisher, C. D. (1986). Organizational socialization: An integrative review. In K. M. Rowland & G. K. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 4, pp. 101 - 145).

Fligstein, N. (1990). The Transformation of Corporate Control. *Harvard University Press, Cambridge, MA*.

Fligstein, N. (1991).The structural transformation of American industry: *An institutional account of the causes of diversification in the largest firms*

Gacek C., Lawrie T., and Arief B. (2001). The many meanings of open source, Technical Report 1, DIRC—*Interdisciplinary Research Collaboration in Dependability*

Gregor, S. (2006).The Nature of Theory in Information Systems, *MIS Quarterly*, (3)30, pp. 611-642

Grofman B., Feld S. L., and Owen G. 1983 Thirteen Theorems in Search of the Truth. *Theory and Decisions* 15:261-278.

Groysberg, B., J.T. Polzer and H.A. Elfenbein. (2008).Too many cooks spoil the broth: How High Status Individuals Decrease Group Effectiveness. *Harvard Business School, Cambridge, MA*

Hardy, C. J., and Latane, B. (1988). Social loafing in cheerleaders: Effects of team membership and competition. *Journal of Sport and Exercise Psychology,* 10, 109-114.

Hare, A. Paul. (1952). Interaction and consensus in different sized groups." *American Sociological Review*, 17: 261-267.

Hars, A., Ou, S. (2002). Working for free? Motivations for participating in Open-Source projects. *International Journal of Electronic Commerce* 6, 25–3

Hertel, G., Niedner, S. and Herrmann, S. (2003). Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel. *Research Policy*, 32, 7, 1159–1177

Hoaglin D.C., Welsch R.E. (1978) The hat matrix in regression and ANOVA. *The American Statistician 32* (1978), pp. 17–22

Hoffman, A. (1999). Institutional evolution and change: Environmentalism and the US chemical industry. *Acad. Management J*. 42(4) 351-371.

Howison J, Inoue K, Crowston K (2006) Social dynamics of free and open source team communications. *Proceedings of the IFIP 2nd International Conference on Open Source Software (Lake Como, Italy), IFIP International Federation for Information Processing, vol 203/2006, Springer, Boston*, USA, pp 319–330

Vernon D. Miller, Fredric M. Jablin. (1991), Information Seeking during Organizational Entry: Influences, Tactics, and a Model of the Process. *The Academy of Management Review Vol. 16*, No. 1 pp. 92-120

Jae Yun Moon and Lee Sproull, (2000). Essence of Distributed Work: The Case of the Linux Kernel, *First Monday, volume 5, number 11* (November)

Johri, A. (2007). The Socio-Technical Process of Newcomer Participation: Lessons from a Field Study *Proceedings of the 8th international conference on Computer supported collaborative learning* .978-0-6151-5436-7 New Brunswick, New Jersey, USA317-319

Katzenbach, J.R., and Smith, D.K. (1993).The rules for managing cross-functional reengineering teams. *Planning Review*, 21, 2 12-13.

Kim, A.J. (2000). Community building on the web: Secret strategies for successful online communities. *Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA*

Kogut, B., A. Metiu. (2001). Open source software development and distributed innovation. *Oxford Rev. Econom. Policy* 17(2) 248-264.

Krishnamurthy S., (2002). Cave or Community? An Empirical Examination of 100 Mature Open Source Projects, *First Monday, volume 7, number 6.*

Lakhani, K.R. and von Hippel, E. (2003) How open source software works: ''free'' user-to-user assistance. *Research Policy*, 32, 6, 923–943

Lawler, E.E. (1971) Pay and Organizational Effectiveness: *A Psychological View. New York, NY: McGraw-Hill.*

Leblenbici, H., G. R. Salancik, A. Kopay, T. King. (1991). Institutional change and the transformation of interorganizational fields-An organizational history of the United-States radio broadcasting industry. *Admin. Sci. Quart*. 36 333-363.

Lerner, J. and Tirole, J. (2002). Some Simple Economics of Open Source. *Journal of Industrial Economics,* 50, 2, 197–234

Lorange, P., M. Scott Morton, S. Ghoshal. (1986). Strategic Control Systems. *West Publishing Co., St. Paul, MN.*

Louis, M. R., Posner, B. Z., and Powell, G. N. (1983). The availability and helpfulness of socialization practices. *Personnel Psychology*, 36: 857-866.

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

March, J. C., and Simon, H. A. (1958). *Organizations, New York: Wiley.*

Markus, M.L. (1987).Toward a Critical Mass Theory of Interactive Media: Universal Access, Interdependence and Diffusion, *Communication Research (14:5)*, pp. 491-511.

Meyer, A. (1982). Adapting to environmental jolts. *Administrative Science Quarterly*, 27:515-537.

Meyer, A.D. and Gaba, V. and Colwell, K.A. (2005). Organizing Far from Equilibrium: Nonlinear Change in Organizational Fields. *Organization Science*, 16(5): 456-473.

Miles, R. E. and C. C. Snow (1978), Organizational Strategy, Structure, and Process. *New York: McGraw-Hill Book Company*

Miles, R. H. (1982). Coffin Nails and Corporate Strategies. *Prentice-Hall, Edgewood Cliffs, NJ*

Miller, V. D., and Jablin, F. M. (1991). Information seeking during organizational entry: Influences, tactics, and a model of the process. *Academy of Management Review*, 16: 92-120

Moreland, R. L., and Levine, J. M. (2001). Socialization in organizations and work groups. In M. E. Turner (Ed.), Groups at work: *Theory and research* (pp. 69-112). *Mahwah, NJ: Lawrence Erlbaum Associates, Inc.*

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. *Acad. Management J.* 451149-1160

Nasser, D. (1988). How to run focus group. *Public Relations Journal* 44:33-34

O'Dell, Jerry W. (1968).Groups size and emotional interaction. *Journal of Personality and Social Psychology, Vol 8*(1, Pt.1), 75-78

Ostroff, C., & Kozlowski, S. W. J. (1992). Organizational socialization as a learning process: The role of information acquisition. *Personnel Psychology*, 45: 849-874.

Pliskin, N., Balaila, I., Kenigshtein, I. (1991). The knowledge contribution of engineers to software development: a case study. *IEEE Transactions on Engineering Management* 38 (4), 344–34

Preece J. (2001), Sociability and usability in on-line communities: determining and measuring success. *Behaviour and Information Technology 20 5* pp. 347–356

Pride, R. A. 1995. How activists frame social problems: Critical events versus performance trends for schools. *Political Comm.* 12 5-26

Raymond, (1999). The Cathedral and the Bazaar: Musings on Linux and Open Source from an Accidental Revolutionary. *O'Reilly and Associates, Sebastapol, CA*.

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

Ren, Y., Kraut, R., and Kiesler, S. (2008). Applying common identity and bond theory to design of online communities*. Organization Studies*, 28(3):377-408

De Joode R.V.W.(2004). Managing conflicts in open source communities. *Electronic Markets*, 14(2): 104–113.

Saks, A. M., & Ashforth, B. E. (1997). Socialization tactics and newcomer information acquisition. *International Journal of Selection and Assessment*, 5, 48–61.

Sanday, P. R. (1979). The ethnographic paradigm(s). *Administrative Science Quarterly*, 24: 527-538.

Scacchi, W. (2002). Understanding the Social, Technological, and Policy Implications of Open Source Software Development. *Position paper presented at the NSF Workshop on Open Source Software*.

Scacchi, W. (2002), Understanding the Requirements for Developing Open Source Software Systems. *IEE Proceedings: Software*, 149(1), pp 24-39.

Schoberth, T., J. Preece, A. Heinzl. (2003). Online communities longitudinal analysis of communication activities. *Proc. 36th Hawaii Internat. Conf. System Sci., Big Island, Hawaii*.

Sewell, W. H. (1995). Historical events as transformations of structures: Inventing revolution at the Bastille. *Theory and Society* 25 841- 881

Shah S.K. (2006). Motivation, governance, and the viability of hybrid forms in open source software development. *Management Science* 52 (7), 1000–1014

Shmueli, G. Koppius O. R (2007). Available at http://www.citi.uconn.edu/cist07/5c.pdf

Shmueli, G. and Koppius, O. R. (2010). Predictive analytics in information systems research. *MIS Quart.*

Singh V., Twidale M., and Nichols D. (*2009).* Users of open source software – how do they get help? *In HICSS'09.*

Spaulding, Trent J (2010). How can virtual communities create value for business? *Electronic Commerce Research and Applications*, 9, 38–49

Stasser, G., and Taylor, L. A. (1991). Speaking turns in face-to-face discussions. *Journal of Personality and Social Psychology*, 60, 6754.

Stewart, K., & Gosain, S. (2006). The impact of ideology on effectiveness in open source software development teams. *MIS Quarterly*, 30(2), 291–314

Van Maanen, J., and Schein, E. H. (1979).Toward a theory of organizational socialization. In B. M. Staw (Ed.), *Research in organizational behavior* (Vol. 1, pp. 209-264).

Von Hippel, E., (2001). Open source shows the way: innovation by and for users—no manufacturer required. *Sloan Working Paper* #4163.

Waterson, P.E., Clegg, C.W., Axtell, C.M., (1997). The dynamics of work organization, knowledge and technology during software development. *International Journal of Human-Computer Studies* 46 (1), 81–103

Weber, S. (2004). The success of open source. *Harvard University Press. Cambridge, MA*

Williams, R. L., and Cothrel, J. (2000). Four smart ways to run online communities. *Sloan Management Review*, 81 – 91.

Yamauchi, Y., Yokozawa, M., Shinohara, T., T. Ishida. (2000). Collaboration with Lean Media: How Open-Source Software Succeeds. *ACM. CSCW '00*.