World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:9, No:6, 2015

A Quantitative Study of the Evolution of Open Source Software Communities

M. R. Martinez-Torres, S. L. Toral, M. Olmedilla

***Abstract*—**Typically, virtual communities exhibit the well-known phenomenon of participation inequality, which means that only a small percentage of users is responsible of the majority of contributions. However, the sustainability of the community requires that the group of active users must be continuously nurtured with new users that gain expertise through a participation process. This paper analyzes the time evolution of Open Source Software (OSS) communities, considering users that join/abandon the community over time and several topological properties of the network when modeled as a social network. More specifically, the paper analyzes the role of those users rejoining the community and their influence in the global characteristics of the network.

***Keywords*—**Open source communities, social network analysis, time series, virtual communities.

I. Introduction

THE success of OSS projects crucially depends on the contributions of a community of users. Their development is based on the support of virtual communities of individuals spread over the world, which use the software and participate in their development [1]. These advantages result from keeping the source code open to the whole community, so the advances or solutions achieved by a particular developer can be viewed and revised by the rest of the community members [2], [3]. By employing the collective knowledge and diverse experiences of many contributors, the community of users can report software defects, request and inspire new features, reproduce bugs or comment on issues reported by other users [4].

OSS projects do not have a formal hierarchical structure typical of proprietary software. However, their structure is not flat. As is the case in other virtual communities, the phenomenon of participation inequality can also be observed in the OSS case, and different categories of users can be distinguished. In general, OSS communities follow a core-periphery pattern, with a small group of active contributors and a huge group of users with rare or even no contributions [5]. Most peripheral users are also known as free riders, and they are characterized because they take advantage of the community without any contributions. Despite that, they are tolerated. The core-periphery structure is sustained above all by the core group responsible of the majority of contributions [6]. Their role has been highlighted in previous studies, not only as a driving force for OSS projects success [7], but also by their role as knowledge brokers [8]. The core-periphery structure of OSS communities has been mainly studied from the perspective of social network analysis, by modeling the community as social networks, where nodes are users and arcs represent the interactions among them [9], [10]. However, most of these studies are only focused on the static dimension of networks. Social networks are only considered as a snapshot of the network at a given time, since it was created or during a period of time. Several social network features, such as degree, centrality, cohesion, modularity, etc. are calculated and then used either to characterize the core-periphery pattern [11] or to identify certain profile of users. For instance, the identification of the core group of developers is a major issue for the survival of the community, as they have a direct incidence in its successful development [7]. There are only few studies considering the dynamics of OSS communities. However, this is also an important issue for the survival of OSS communities. The core group of developers is not permanent and must be nurtured and reinforced by new members that gain expertise through their interactions with expert members. The dynamic of this process needs to be characterized in order to understand the evolution of communities. Previous studies about the dynamic of OSS projects were focused on the size in lines of code or on bug reports. In [12], the evolutionary behavior of SourceForge projects is analyzed, and the size in lines of code as a function of the time in days is modeled using a quadratic model. In [4], the evolution of structural features of networks modeling user collaborations is analyzed using a bug-tracker of 14 major OSS projects.

This paper is focused on the mailing list of a well-known Debian Linux port, that it is used as a case study. The time evolution of contributors is disaggregated considering new, drop-out and rejoined users, and their changes are then related to several topological properties of the network. The rest of the paper is structured as follows. Next section describes the legitimate peripheral participation as the main mechanism by which newcomers can become experts and reinforce the core group of developers. Section III introduces the case study and the methodology, which is based on the analysis of time series and social networks. Section IV shows the results obtained, highlighting how the community is reinforced by rejoined users. Finally, Section V concludes the paper.

S. L. Toral is with the School of Engineering, University of Seville, Seville, Spain (phone: +34 955481293; fax: +34 9544487373; e-mail: storal@us.es).

M. R. Martinez-Torres and M. Olmedilla are with the Business School, University of Seville, Seville, Spain (e-mail: rmtorres@us.es, mariaolmedilla@hotmail.com).

II. Legitimate Peripheral Participation

The community-based development of OSS has been frequently related to the social learning theory from Wenger, which postulates that learning is a social process, placing learning in the context of the social experience of individuals [13]. The social learning that takes place when people have a common interest in some subject or problem, or collaborate over an extended period to share ideas, solutions, and build innovations leads to the notion of communities of practice, developed in [15]. This is the case of open source communities, where people can freely post their questions related to the underlying software and receive some solutions or alternatives from someone else of the community [14].

The process underlying the construction of communities of practice is called Legitimate Peripheral Participation, LPP [15]. This is the process by which newcomers become full members by learning from more competent practitioners and by being allowed to participate in certain tasks related to the practice of the community [16]. In the case of OSS communities, the newcomer's participation at first is legitimately peripheral. They do not post contributions but questions, and they rarely participate more than once. Some of them are "lurkers" or free riders, that is, observers who exhibit no visible level of activity [17]. However, a percentage of those users start to rejoin the community posting more questions or even some contributions. They browse the community archives, contribute by sporadically reporting bugs, sending patches or ad hoc solutions to problems. As long as they interact with more expert users, they acquire a contextualized learning. Some of these rejoined users can maintain a permanent participation and even be part of the core group of developers.

LPP explains how participation, situated learning, and identity construction interrelate and coevolve as an individual engages in a community of practice [18]. Here participation means to be active in the practices of social communities and constructing identities in relation to these communities. Situated learning is related to the theoretically generative interconnections between persons, actions, knowing, and the surrounding social world [15]. Finally, a member's participation in a community involves the construction of his or her identity and to what extent he or she is legitimized and valued by the other members [18].

Previous studies related to LPP have considered several categories of developers in the OSS community according to the length of their past sustained participation, and then each group was separately studied [19]. In general, LPP has been studied from a social network perspective. Social network analysis reveals not only local topological properties of nodes but also global features of the network [20]. In this paper, a hybrid approach is followed: three groups of users, new, drop-out and rejoined are distinguished and their evolution is jointly analyzed with several global topological features of the network, from the perspective of time series analysis.

III. Case Study and Methodology

This study is focused on the Debian Project, which is an association of individuals who have made common cause to create a free operating system called Debian GNU/Linux, or simply Debian for short [21]. More specifically, the study analyzes Debian port to ARM mailing lists, which can be publicly accessible at https://lists.debian.org/debian-arm/. This website includes people interactions since 1999, and they are organized as threads of discussion. Mailing lists are useful to put in contact information seekers and information providers, and they are a very useful resource for those who need to adapt Debian to a specific processor. Authors posting messages are identified by an email or an alias, which is unique within the community. Using this information, mailing lists are analyzed month by month in order to extract information about new users joining the community, users that abandon the community and users that rejoin the community, which refers to those users who posted messages in the past but not in the previous month.

Mailing lists are also modeled as a social network, considering their evolution month by month. Whereas mailing lists are usually organized by threads of discussion, threads are used as the basic unit of analysis when deciding about the arcs connecting the nodes. More specifically, an author posting to a thread was tied to all the authors who have previously posted to the same thread when constructing the social network. The main assumption to do this is that it is cognitively more complex to answer a thread of discussion than to answer a single message. Answering a thread of discussion usually requires reading all the previous content in the thread to write a coherent answer [22].

Several global topological characteristics of social networks can be extracted once they are modeled as a graph. Table I describes the considered global features and their meaning.

TABLE I
Global Topological Characteristics of the Network

Measure	Description
Density	Number of arcs of the network, as a proportion of the maximum number of possible arcs.
ASP	Average shortest path.
Clustering coefficient	Measure of local cohesiveness through the neighbor interactions of a node. It is defined as twice the ratio between the number of edges which connect the neighbours of a given node and the total number of possible edges among them.
Betweenness centrality	Measure of centrality given by the intermediary role developed by nodes of the network.

IV. Results

The Debian port to ARM mailing lists were extracted and modeled as a graph using a specific crawler developed in R. Each month of the subsequent 13 years after its creation was considered as a separate network, leading to a total of 156 separate networks.

The evolution of the Debian-ARM community was monthly analyzed during the years 1999 to 2013. Fig. 1 shows this evolution in terms of size, new users joining the community, users that abandon the community and how many of the new

users are actually past users that join again the community (therefore, new users include this group). The horizontal axe is the sequence of 155 months corresponding to the period analyzed.

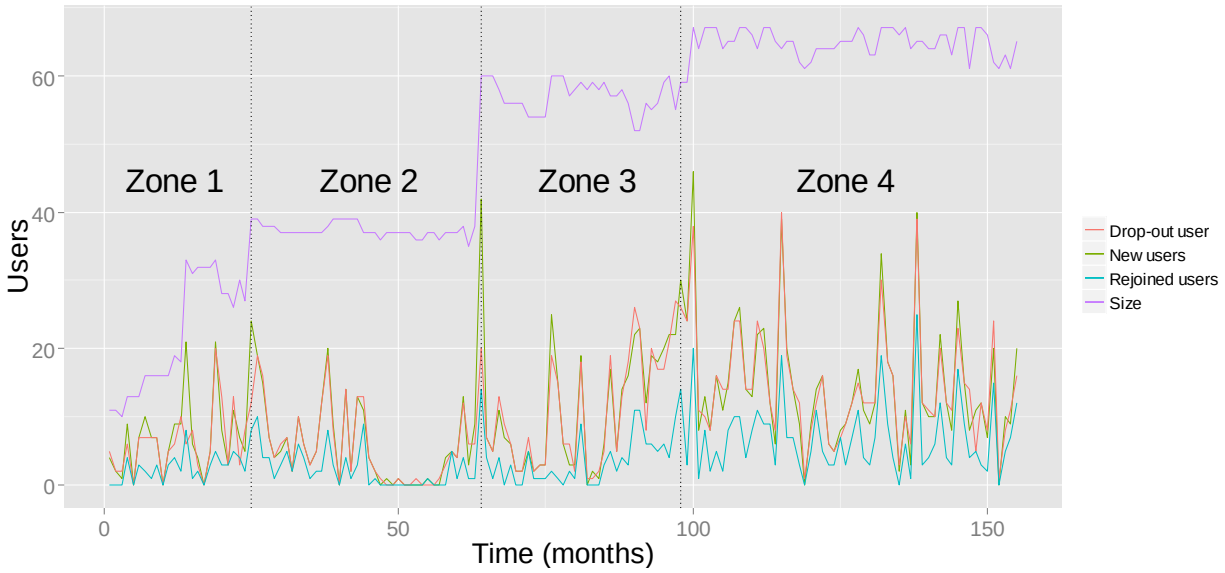


Fig. 1 Evolution of the Debian-ARM community during the period 1999-2013

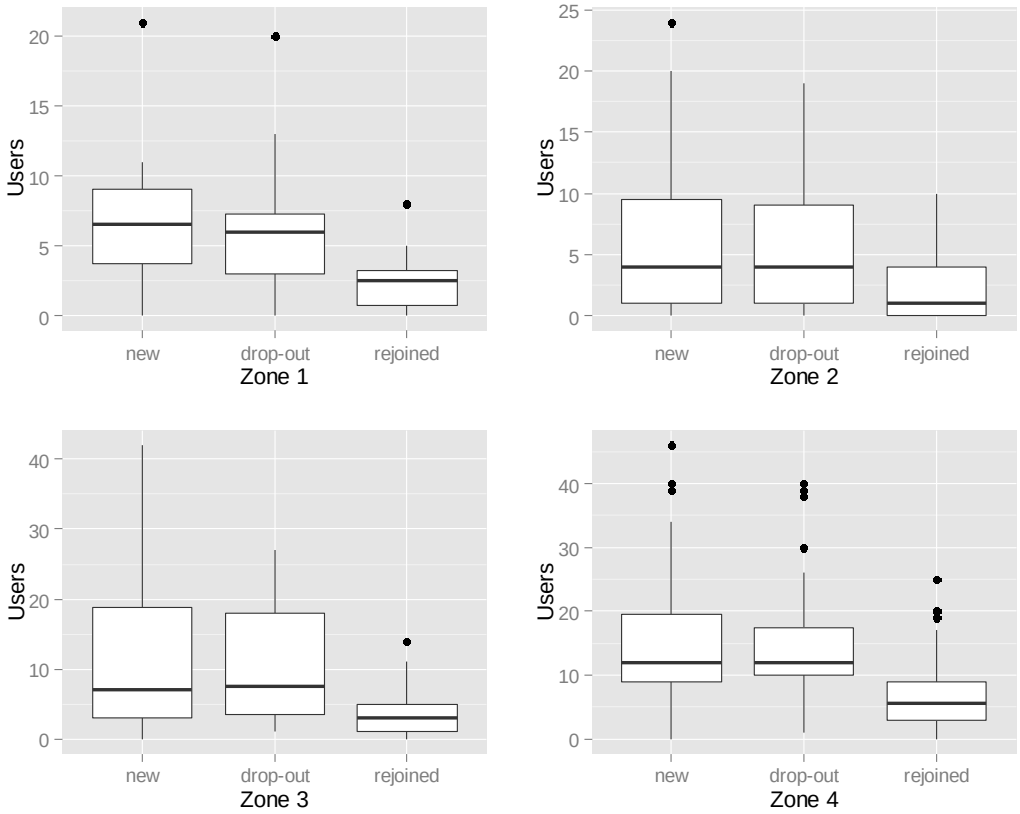


Fig. 2. Boxplot graphs of the four distinguished zones

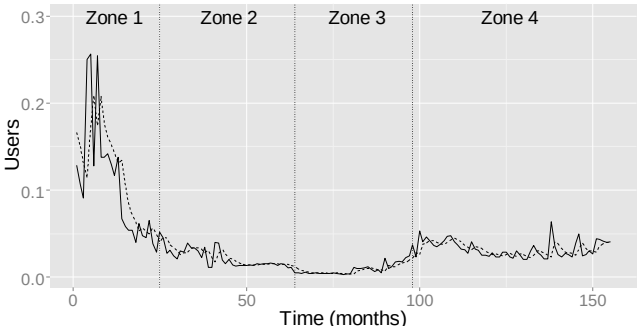
Basically, four stages can be distinguished in Fig. 1: there is a first period in which the size of the community is growing, followed by three stabilization periods in which the size of the community remains almost constant.

The first period goes from month 1 to month 24. The growth during this first stage is mainly explained by a higher number of new users joining the community. The first stabilization period goes from month 25 to month 63. During this period, the size of the community is quite stable, and the

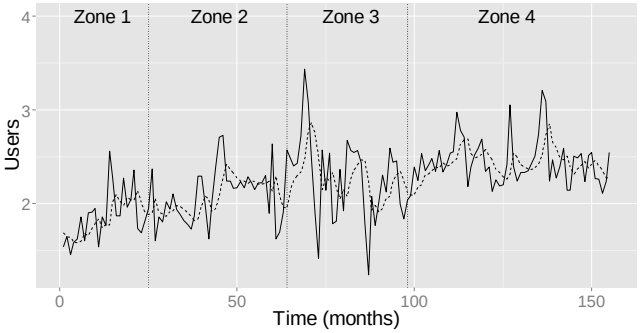
number of new users is almost the same than the number of users that abandon the community. Around month 63 there is a sharp increment of the number of new users, which in turn explains the sharp change in the size of the community, and leads to the third period of Fig. 1. During this period, the community has grown above all due to those new users that have become engaged in the community. Finally, around month 100, there is a new although smaller change in the size of the community. But in this case, the increment is explained by those old users that have rejoined the community.

Fig. 2 illustrates the boxplot graphs for the 4 zones considered in the evolution graph. Except for zone 1, in all other cases the number of new and drop-out users is almost the same. Therefore, the community is sustained by those active users that are permanently engaged in the community. They are the core group of the community, responsible for the majority of contributions and for promoting other users to assume a more active participation. The role of rejoined users is different across the four zones. Intially, rejoined users are only a small percentage of new users (zone 1). As long as the community evolves, rejoined users acquire more importance, so that in zone 4 they can explain the growth of the community. They are actually those users that become experts though the legitimate peripheral participation process described in the literature. Part of rejoined users will become part of the core group with a permanent participation within the community.

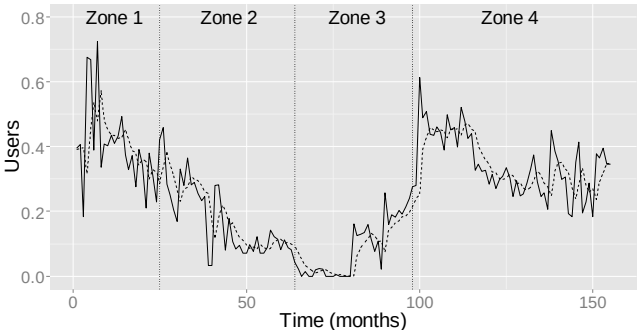
The evolution of the community was also studied in terms of features of the social network representing the community. Fig. 3 shows 5 of these features. In each graph, the solid line is the evolution of the corresponding feature while the dashed line is its Holt-Winters exponential smoothing. Fig. 3 (a) and (b) illustrates the evolution of the density and the ASP of the network, respectively. Initially, the network is a very dense network in the zone 1. However, this is because the network is small and still growing. As soon as the network reaches a stable size, density decays to a low value, as usual in communities reaching a certain size. This is because nodes are connected to only a small part of the overall network. Mathematically, the number of possible edges grows with the number of nodes n as $n(n-1)/2$. Consequently, the number of edges cannot grow so quickly and, as a result, density decays. However, it can be appreciated that density starts to grow by the end of zone 3 and zone 4. This behavior is explained by the increasing number of rejoined users, more active than new users as they have gained experience through interactions during previous months. Fig. 3 (b) illustrates the behavior of ASP. The ASP tends to slightly grow with the size of the community. Again, initially the ASP starts with a low value because the community is small and still growing. As long as the community evolves, the ASP tends to grow because people tend to interact through certain threads of discussion, which facilitates the emergence of subcommunities.



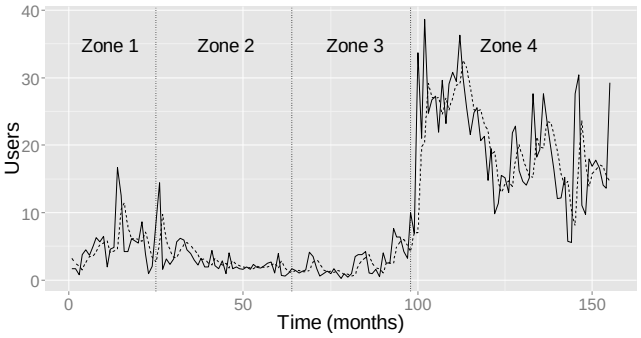
(a) Density



(b) ASP



(c) Clustering coefficient



(d) Betweennes centrality

Fig. 3 Social network features evolution

Fig. 3 (b) shows high temporal fluctuations of the solid line, which seems logical if it is considered that the structure of the network is continuously changing with a lot of users joining and abandoning the network. The dashed line representing a

smoothed version of the ASP illustrates better the general tendency.

The clustering coefficient of Fig. 3 (c) follows a similar pattern to that of the density, with an inverted U shaped curve. The clustering coefficient measures the ratio between the interactions of nodes located in 1-hop neighborhood of a given node and its degree. Initially, it exhibits a high value because the network is small and the average degree is also small. Consequently, the denominator of the formula for the clustering coefficient is low. As the network evolves, its size increases and it becomes less dense. As in the graph of density, there is a turning point when rejoined users acquire more importance, around the central part of zone 3. Rejoined users are more active, increase the density of the network and tend to interact with other active users.

Finally, Fig. 3 (d) shows the betweenness centrality evolution. The betweenness centrality exhibit a sharp change at the beginning of zone 4, coinciding with the increase of the density. That means that in zone 4 the network shows a higher level of cohesion, given by the incorporation of more active users.

In summary, the four graphs of Fig. 3 highlight the importance of past users rejoining the community. Several community features get improved at the moment that rejoined users gain importance. Initially, there is only a small group of core users sustaining the community while the rest of them are just participating occasionally. Zone 1 is the time where the community is forming, increasing its size month by month. During zones 2 and 3 the community reaches two stable sizes, being still most of the users peripheral users only attending occasionally to the community. By the middle part of zone 3, the number of rejoined users starts to increase, and many of them become active users with regular contributions. This effect can be appreciated in the increment of the network's density and clustering coefficient. Zone 4 is the last temporal stage, with also a stable size but with better network features.

Many previous studies highlight the role of the core team. Project success demands the sustained participation of a small number of core developers who possess strong technology skills and proven records to play a vital role in the project [19]. But over time, the core group needs to be reinforced and in some cases replaced by new developers. That is the reason why the periphery group is also important, as they represent potential candidates that can reach the core of the community.

V. Conclusion

This paper analyzes the temporal activity of open source mailing lists considering three types of users and some global topological features of the derived social networks. The main result is that the reincorporation of past users can affect some features related to the structure of the whole network. More specifically, a higher rate of rejoined users can improve the size, density and cohesion of the network.

Acknowledgment

This work was supported by the Consejería de Economía, Innovación, Ciencia y Empleo under the Research Project with reference P12-SEJ-328 and by the Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad under the Research Project with reference ECO2013-43856-R.

References

[1] M. R. Martinez-Torres, M. C. Diaz-Fernandez, Current issues and research trends on open-source software communities, *Technology Analysis & Strategic Management*, Vol. 26, Iss. 1, pp. 55-68, 2014.

[2] L. Dahlander, M. G. Magnusson, Relationships between open source software companies and communities: Observations from Nordic firms, *Research Policy*, Vol. 34, no. 4, pp. 481–493, 2005.

[3] S. L. Toral, M. R. Martinez-Torres, F. Barrero, Modelling Mailing List Behaviour in Open Source Projects: the Case of ARM Embedded Linux, *Journal of Universal Computer Science*, Vol. 15, Iss. 3, pp. 648-664, 2009.

[4] M. S. Zanetti, E. Sarigol, I. Scholtes, C. J. Tessone, F. Schweitzer, A Quantitative Study of Social Organisation in Open Source Software Communities, *Proc. Imperial College Computing Student Workshop ICCSW 2012*, London, 2012, pp. 116–122.

[5] A. Mockus, T. Fielding, and D. Herbsleb, Two Case Studies of Open Source Software Development: Apache and Mozilla, *ACM Trans. Software Eng. and Methodology*, Vol. 11, no. 3, pp. 309-346, 2002.

[6] E. von Hippel, G. von Krogh, Open Source Software and the "Private-Collective" Innovation Model: Issues for Organization Science, *Organization Science*, Vol. 14, Iss. 2, pp. 209-223, 2003.

[7] S. L. Toral, M. R. Martínez-Torres, F. Barrero, F. Cortés, An empirical study of the driving forces behind online communities, *Internet Research*, Vol. 19, Iss. 4, pp. 378-392, 2009.

[8] S. L. Toral, M. R. Martínez-Torres, F. Barrero, Analysis of virtual communities supporting OSS projects using Social Network Analysis, *Information & Software Technology*, Vol. 52, Iss. 3, pp. 296-303, 2010.

[9] G. von Krogh, E. von Hippel, The promise of research on open source software, *Management Science*, Vol. 52, pp. 975-983, 2006.

[10] M. R. Martinez-Torres, A genetic search of patterns of behaviour in OSS communities, *Expert Systems With Applications*, Vol. 39, Iss. 18, pp. 13182-13192, 2012.

[11] M.P. Rombach, M.A. Porter, J.H. Fowler and P.J. Mucha, "Core-periphery structure in networks", *Slam J. Appl. Math*, Vol. 74, no. 1, pp. 187-190, 2014.

[12] S. Koch, Evolution of Open Source Software Systems – A Large-Scale Investigation, *Proceedings of the First International Conference on Open Source Systems*, Genova, 2005, pp. 148-153.

[13] E. Wenger, *Communities of practice: learning, meaning, and identity*, Cambridge: Cambridge University Press, 1998.

[14] S. L. Toral, M. R. Martínez-Torres, F. Barrero, Virtual communities as a resource for the development of OSS projects: the case of Linux ports to embedded processors, *Behaviour & Information Technology*, Vol. 28, Iss. 5, pp. 405-419, 2009.

[15] J. Lave and E. Wenger, *Situated learning: Legitimate peripheral participation*. Cambridge University Press, 1991.

[16] C. Kimble, P. Hildreth, and P. Wright, Communities of practice: Going virtual. In *Hildreth, Paul M. and Kimble, Chris, editors, Knowledge Networks: Innovation through Communities of Practice*, Idea Group Publishing, 220-234, 2000.

[17] P.A. David, F. Rullani, Dynamics of innovation in an open source collaboration environment: Lurking, laboring and launching floss projects on Sourceforge. *Industrial and Corporate Change*, Vol. 17, Iss. 4, pp. 647–710, 2008.

[18] K. Handley, A. Sturdy, R. Fincham, and T. Clark, Within and beyond communities of practice: Making sense of learning through participation, identity and practice, *Journal of Management Studies*, Vol. 43, Iss. 3, pp. 641–653, 2006.

[19] Y. Fang, D. Neufeld, Understanding Sustained Participation in Open Source Software Projects, *Journal of Management Information Systems*, Vol. 25, No. 4, pp. 9–50, 2009.

[20] N. Ducheneaut, Socialization in an Open Source Software Community: A Socio-Technical Analysis, *Computer Supported Cooperative Work*, Vol. 14, pp. 323–368, 2005.

[21] J. Mateos-Garcia & W. E. Steinmueller, "The institutions of open source software: Examining the Debian community", *Information Economics and Policy*, Vol. 20, pp. 333–344, 2008.
[22] N. Knock, "Compensatory adaptation to a lean medium: An action research investigation of electronic communication in process involvement groups", *IEEE Trans. on Professional Communication*, Vol. 44, no. 4, pp. 267-285, 2001.