Takeshi Kakimoto, Yasutaka Kamei, Masao Ohira, and Ken-ichi Matsumoto, ``Social Network Analysis on Communications for Knowledge Collaboration in Oss Communities,'' In Proceedings of the 2nd International Worksho on Suporting Knowlege Collaboration in Software Development (KCSD2006), pp.35-41, September 2006. (Tokyo,Japan)

Social Network Analysis on Communications for Knowledge Collaboration in OSS Communities

Takeshi Kakimoto Yasutaka Kamei Masao Ohira Ken-ichi Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama Ikoma Nara Japan 630-0192
{takesi-k, yasuta-k, masao, matumoto}@is.naist.jp

Abstract

Knowledge collaboration is the key for success of open source software (OSS) communities, because not all members have knowledge and skills necessary for software development. Generally, members in OSS communities communicate for knowledge collaboration using communication tools (e.g. mailing lists, discussion forums, bug tracking systems, and so on) so that geographically distributed members collaborate and coordinate their work. In this paper, we apply social network analysis to the data accumulated in communication tools. We analyzed relationships between the density of social networks and OSS releases by time series analysis of 4 OSS communities in SourceForge.net in order to investigate the quality of communications for knowledge collaboration. The analysis results showed that communications among community members with a variety of roles are active before/after OSS released in communities where knowledge collaboration is going well.

1. Introduction

Nowadays software developers continuously require a considerable amount of new and diverse knowledge about technologies for software development such as programming languages and components libraries, since such technologies have been evolving from day to day and the past knowledge about them cannot be used soon. In this situation, an individual developer cannot possess every kind of knowledge about latest technologies needed for software development. Knowledge collaboration [10] is not desirable but necessary for modern software development.

Especially, open source software (OSS) development communities rely on knowledge collaboration among community members who have a variety of roles such as community leaders, developers, bug reporters, passive users and so forth [6, 11], because OSS communities, differently from traditional software development organizations, cannot recruit members who have sufficient skills and knowledge required for building software systems in advance.

In typical OSS communities where community members are geographically distributed, knowledge collaboration takes place through using collaboration tools such as version control systems, bug tracking systems, and mailing lists. Based on the data stored in the collaboration tools, prior studies discussed the model of collaboration processes in distributed environments [9], the efficiency of communication and coordination in distributed software development [4], the benefits of OSS style software development [5], communications metrics for knowing the quality of group work [2] and so forth.

In this paper, we would like to investigate the quality of communications for knowledge collaboration by analyzing the data from communication tools used for distributed software development and the data denoting the success and failure of knowledge collaboration (e.g. number of software releases and number of software downloads). In OSS development, community members rarely meet to discuss but communicate heavily using electronic media (e.g. mailing lists and forums). So, we supposed that we might comprehend the success and failure of knowledge collaboration from the quality of communications among community members through collaborative communication media.

As an approach to inspecting the quality of communications for knowledge collaboration, we use social network analysis (SNA) [7, 8], especially the density of social networks which is a measure to know the quality of social relationships among people (e.g. intimacy or solidarity among people). In this paper, we applied SNA methods to the communication data stored in forums for OSS communities in SourceForge.net (SF.net) [1].

[1] SourceForge.net, http://sourceforge.net/

In what follows, in Section 2 we hypothesize on communications for knowledge collaboration, more specifically, how knowledge collaboration in OSS communities is conducted using electronic communication media. Section 3 describes density of social networks, which is a measure for SNA. In section 4 we analyze 4 OSS communities in SF.net. Section 5 is the results of our analysis. We discuss the results and our hypothesis in Section 6. Section 7 concludes the paper.

2. Communications for Knowledge Collaboration in OSS Communities

In this section, we discuss communications for knowledge collaboration in OSS communities. Typical OSS communities where community members are geographically distributed and rarely meet to discuss together, heavily relies on collaboration tools such as version control systems and bug tracking systems and electronic communication media such as mailing lists and forums to precede their knowledge collaboration. Yamauchi et al. [9] had conducted two case studies to investigate how OSS development communities achieve smooth coordination and effective collaboration. One of the findings of the case studies was that collaboration and communication tools (e.g. CVS, TODO lists and Mailing lists) were used in a good balance between centralization and spontaneity [9].

In this paper we would like to focus on the quality of communications for knowledge collaboration through communication media. In OSS development, communications for knowledge collaboration involve a variety of people. For instance, software developers discuss technological problems, bug reporters point out bugs of released software, end-users request developers to add new features and so forth. It is important for knowledge collaboration to involve such a variety of community members because ”voice” from bug reporters and end users often makes OSS reliable and innovative, and motivates OSS developers to create further OSS development [3].

Figure 1 shows a simple model on a cycle of knowledge collaboration in OSS development. Before OSS released, OSS developers discuss their products and related problems (development period). After OSS released, users ask questions on usage of the products to other users or developers and also report bugs or requests of a new features to developers (feedback period). Again, developers discuss the reported bugs and requested features, and then modify and refine their products. This would be a simple view of a cycle of OSS development but an important aspect of knowledge collaboration, because an end user would not use the products if s/he can get help from other community members, a bug reporter would not report bugs if developers do not modify reported bugs, and developers would not continue

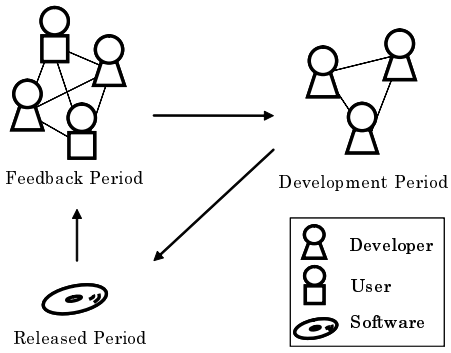

Figure 1. Cycle of Knowledge Collaboration in OSS development

to create software products if no one use them. Here we can make a hypothesis on communications for knowledge collaboration in OSS development communities as follows.

Hypothesis: Communications are actively encouraged before/after OSS released, especially among community members with a variety of roles but not among particular members.

We thought that we might be able to know the success and failure of knowledge collaboration or “health condition” in OSS communities by analyzing the quality of communications among community members before/after OSS released. The next section describes use of the density of social networks which is our approach to investigating the quality of communications.

3. Density of Social Networks

Using the density of social networks in social network analysis (SNA) is a simple way to know the quality of social relationships among people [7, 8]. Social relationships can be graphed as social networks, which consist of persons (nodes) and their relationships (edges).

The density of social networks is defined as the number of lines (edges) in social networks, expressed as a proportion of the maximum possible number of lines [7, 8]. The formula for the density of social networks is

$$ND = \frac{2l}{n(n-1)} \quad (1)$$

where l is the number of lines (edges) in the networks and n is the number of nodes in the networks. The values of ND (network density) can be from 0 to 1.

If social networks show low density, the social relationships tend to be “open” which means *a large, open, diverse, and externally focused relationships* [1]. If social

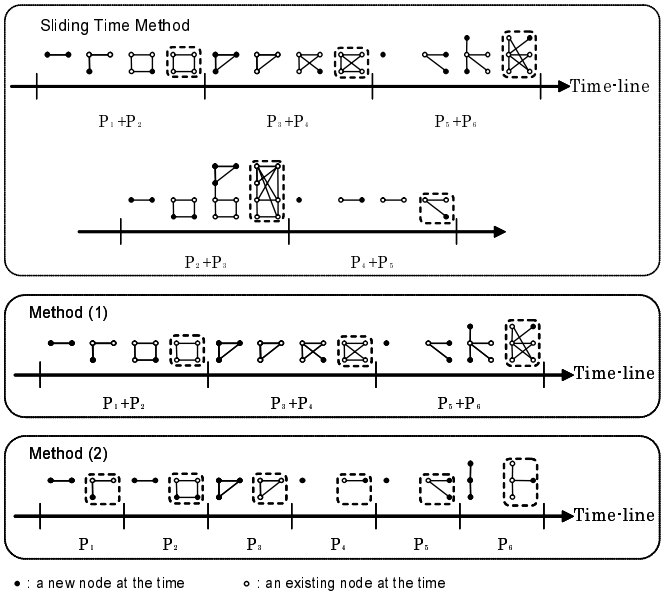

Figure 2. Calculation methods for the density of social netowrks

networks indicate high density, the social relationships often have characteristics of "closed" which means *a small, closed, homogeneous, and internally focused network* [1].

In this paper we apply SNA to the communication data stored in communication tools such as mailing lists and forums (bulletin board systems) to know the quality of communications for knowledge collaboration in OSS development. In this case, social relationships can be defined by posts and replies. Community members (e.g., developers, end-users, bug reporters, and so on) discuss issues related to OSS development. If a member (A) posts a message to a forum for a community (Pi) and another member (B) replies the message, then it assumes that there is a social relation between A and B in Pi. As described before, we anticipate that the density of social networks would be high if comunications for knowledge collaboration go well.

4. Analysis on The Quality of Communications for Knowledge Collaborations

4.1. Dataset

We collected the data involving public forums and released OSSs for 4 OSS communities for the time interval between December 1, 1999 and December 31, 2005. These communities ware selected as target communities for analysis because they indicated characteristic measurements results (e.g. a large number of developers, downloads, or posts). We did not collect the data of mailing lists because the mailing lists were not used for communications among community members but for announcements of OSS releases or archives of CVS logs. The data on public forums includes ID of each posted message, user 's name who posted messages, the date of messages posted, ID of each replied message, and ID of each OSS community. The data on released OSS includes the number of developers in each community, the start date of each community, the number of downloads, the number of average downloads per a day, version numbers of released OSS, the release date of OSS, and ID of each OSS community.

4.2. Analysis Procedure

The followings show the procedure of our analysis using social network analysis (SNA) [7, 8].

Preparation Before calculating the density of social networks, firstly we need to define social networks in the context of our analysis. As described before, our aim of using the density of social networks is to know the quality of communications for knowledge collaboration. We use the communication data made from discussions (messages) in forums.

From messages in forums for a target community[2], we identify who posted a message to the forums (node A) and who replied to the message (node B). Then we regard the relation between the poster (node A) and

[2]A community can have several forums for different purposes of discussions

the respondent (node B) as an edge, by threading relationships between posts and replies as social networks. Repeating this for all messages in forums of a target community, we can graph the relationships as social networks and calculate the density of the social networks.

Calculations of network density by a certain period

Calculating the density of social networks from all the data is inadequate, because structures of social networks change over time. Therefore, time series analysis is necessary to know changes of the quality of communications among community members, that is, changes of the density of social networks. In order to see temporal changes of the density of social networks, we have to fix a particular time interval.

We calculate the density of social networks from social networks for a period P sliding a $\frac{P}{2}$ interval (sliding time method) in this paper. Figure 2 shows calculation methods for the density of social networks. The density of a social network for a certain period is calculated from the structure of the network at the end of the period.

The sliding time method in this paper is sensitive to changes of network structures than method (1) and (2) which not overlap neighboring periods. For example, communications are active in the period of $P_2 + P_3$. However, method (1) can not reflect such the activeness. Method (2) which divides the period in half also can not reflect the activeness because it can only show small changes.

In this paper, the density of social networks is calculated by one and a half month ($P =$ three months. The reason why we fix 3 months is we considered that one topic in a forum is finished about 3 months. We need further consideration for this period or a way to fix an appropriate period.

Time series analysis We analyze relationships between the density of social networks and OSS releases in order to verify our hypothesis. Changes of the density of social networks in time series are used in the analysis. The number of users who posted messages (posters), links among posters (links), and posted messages (posts) are also used.

4.3. Target Communities

In this paper, we analyze 4 characteristic communities. For instance, a community has forums which are posted by only users, a community has a number of developers, a forum has a large number of posted messages, and so forth.

Table 1 shows the measurement results of each community. In what follows, we describe an overview of each community, which consists of characteristic measurement results, developing software, and usages of forums.

Community A Community A has a number of developers. This community has been developing an operating system for controlling small electronic devices. The posted messages in the forum consist of questions on implementation from developers. This community is currently working on own web site but not on SF.net.

Community B Community B has only one developer but provides a tool downloaded by a large number of users. This community provides windows installers for image manipulation software which is originally developed for UNIX. The posted massages are only from users.

Community C The characteristic measurement results of Community C is that there has been a large number of downloads. Community C has been providing a CD ripping tool. The posted messages in the forums of the community consist of posts regarding the implementation of software, questions on released software, bug reports, and requests for new features. Both developers and users often post to the forums. Anonymous users who do not have user ID of SF.net also use them.

Community D The characteristic measurement results of Community D are that the network density is very high and the number of downloads and posters is very small. Community D creates an OpenGL viewer with command line tools. The forums of this community are used only by developers excepting one post by a user.

5. Analysis Results

Figure 3 shows the time series graph of 4 target communities. The horizontal axis shows time period, the vertical axis at the right side is values of the density of social networks, and the vertical axis at the left side is the number of posters, links among posters and posts for each period. Each arrow shows the date of OSS release excepting beta versions released by Community C. The analysis results of each community as follows.

Community A The following pattern of the changes of the density in community A is repeated. At the initial phase of the community started, values of the density become high. Then, values of the density are decreased as the community progressed. Finally, values of the density become zero. Version 0.6.0 and version

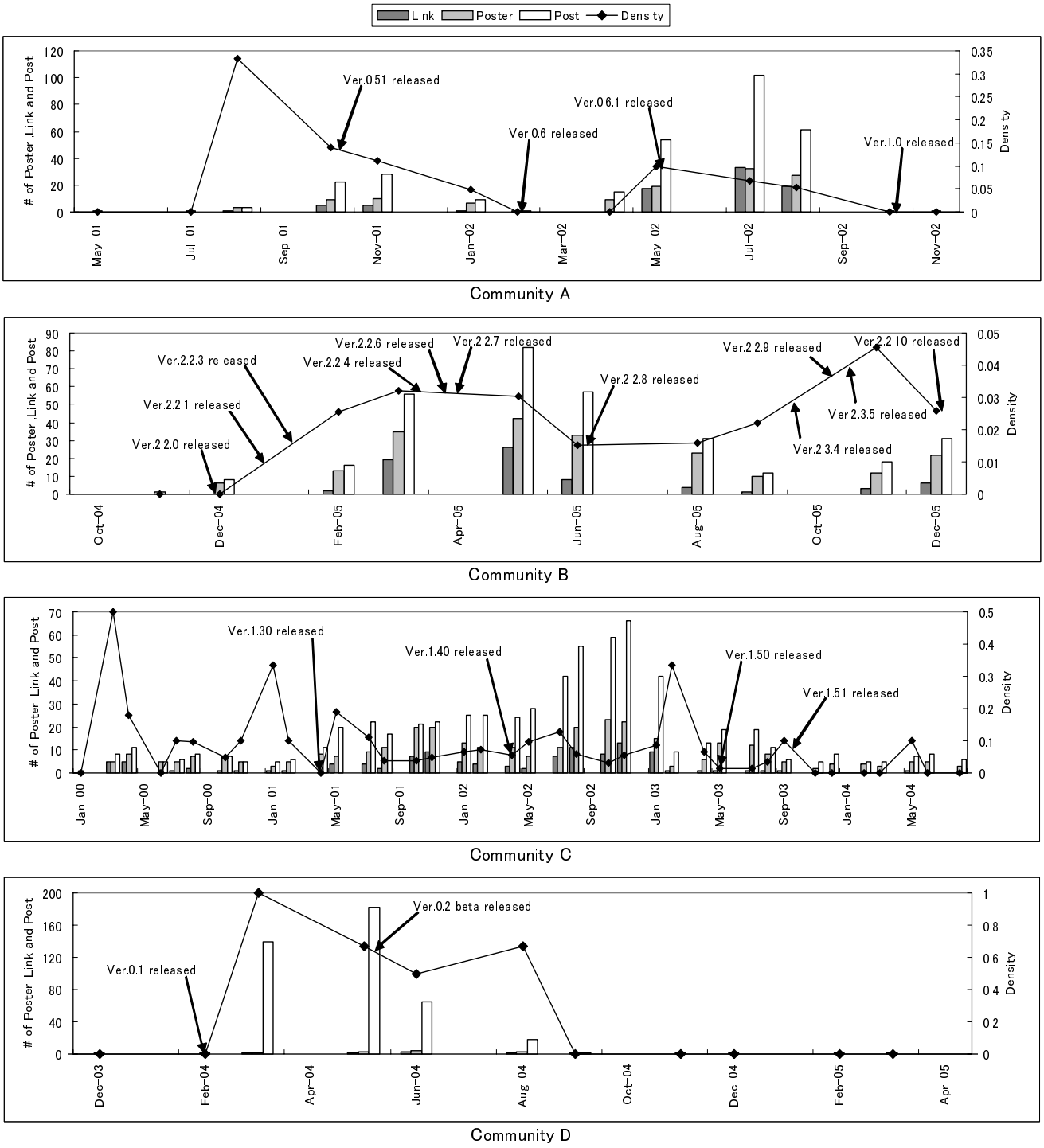

Figure 3. Analysis Results

Table 1. Characteristics of target communities

Community	Num. of developers	Density of all periods	Num. of posts	Date of communities started	Num. of downloads	Num. of average downloads per a day
Community A	138	0.022	174	04-Jun-01	28,265	16.92
Community B	1	0.013	165	07-Oct-04	7,734,629	17188.06
Community C	11	0.007	766	05-Dec-99	26,000,000	11878.12
Community D	3	0.500	203	29-Dec-03	156	0.21

Table 2. Topics in local peak periods

Contents of topic	Mar 2005 peak of density	Apr 2005 peak of post	Nov 2005 peak of density	Dec 2005 peak post
Question	4.8 %	5.3 %	20.0 %	50.0 %
Help	23.8 %	26.3 %	50.0 %	21.4 %
Bug report	71.4 %	57.9 %	20.0 %	7.1 %
Request for new features	–	5.3 %	10.0 %	14.3 %
Others	–	5.3 %	–	7.1 %

1.0 were released when values of the density became zero. Values of the density before releases were higher than that for a released period excepting version 0.6.1. The number of posts attains the local peak (Nevember 2001, July 2002). It is after a few months when the value of the density indicates the maximum locally (local peak) (August 2001, May 2002). When the second time of the pattern is compared with the first time, the second time of the patterns was larger in the number of posters, links among posters and posts.

Community B In the community B, when the density was increasing or high, new versions were released in a short interval. On the other hand, when the density was decreasing, new versions were released in a long interval. Values of the density after releases were higher than that for released periods in most cases. The number of posts attains local peak in the next period of local peak in the density (March 2005 – May 2005, November 2005 – December 2005). When the number of posts is small, the number of posters was near the number of posts. Therefore, when the number of posts is small, only few people posted several messages to the forum.

Community C In the community C, values of the density before OSS releases were higher than that for a released period in all releases. And, values of the density after OSS releases were higher than that for a released period excepting version 1.51. The degree of incenses of density values after releases is decreasing as the community progressed. The number of posts is the maximum after a few months when the value of the density was the maximum.

Community D The number of posters is small against the number of posts in the community D. In the version 0.1 release, values of the density after the release were higher than that for the released period. And, values of the density before the release were higher than that for the released period in the version 0.2 beta release.

6. Discussion

The analysis results excepting community D showed that values of the density before OSS releases are high in the community that has a number of posts from developers. And, values of the density after OSS releases are high in the community that has a number of posts from users.

In the community C that meets both conditions, values of the density before and after OSS releases are higher than values of the density for released periods. In other words, communications are active before and after released periods in community C. On the other hand, Community D that is not the case with these conditions seems be stagnant as the number of downloads and posters is very small and the last release is a beta version. Therefore, we consider that our hypotheses are true for communities where knowledge collaboration among community members with a variety of roles is going well.

Table 2 shows the rate of contents of topics in the local peaks of the density and posts in community B. The results in table 2 indicate community members in the local peak of the density discussed various topics more than in the local peak of posts.

The analysis results showed that a few months after the local peak of the density, the number of posts attained the local peak. Members discussed focusing on topics about bug reports in March 2005. In this period, the density was high. On the other hand, members discussed various topics when the number of the posts was large (April 2005). The value of the density was lower than that in March 2005. We can consider this that communications among members are centered toward a particular topic if the density is higher and communications are dispersed by various topics if the density is lower.

7. Conclusions

In this paper, we investigated the quality of communications for knowledge collaboration by time series analysis using the density of social networks. From the results of analyzing changes of the density in 4 OSS communities our hypothesis (communications are actively encouraged before/after OSS released, especially among community members with a variety of roles but not among particular members.) was verified.

In the future, we will analyze separating developers from end users to distinguish between development periods and feedback periods in more detail.

Acknowledgments We would like to thank Shinsuke Matsumoto for helping us analyze OSS communities. This work is supported by the EASE (Empirical Approach to Software Engineering) community in the Comprehensive Development of e-Society Foundation Software program and Grant-in-aid for Scientific Research (B) 17300007, 2006 and for Young Scientists (B), 17700111, 2006, by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] W. E. Baker. *Achieving Success Through Social Capital*. John Wiley & Sons Inc., 2000.

[2] A. H. Dutoit and B. Bruegge. Communication metrics for software development. *IEEE Transactions on Software Engineering (TSE)*, 24(8):615–628, 1998.

[3] J. Feller and B. Fitzgerald. *Understanding Open Source Software Development*. Addison-Wesley, 2002.

[4] J. D. Herbsleb and A. Mockus. An empirical study of speed and communication in globally distributed software development. *IEEE Transactions on Software Engineering (TSE)*, 29(6):481–494, June 2003.

[5] A. Mockus, R. T. Fielding, and J. D. Herbsleb. Two case studies of open source software development: Apache and mozilla. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 11(3):309–346, 2002.

[6] K. Nakakoji, Y. Yamamoto, Y. Nishinaka, K. Kishida, and Y. Ye. Evolution patterns of open-source software systems and communities. In *Proceedings of the International Workshop on Principles of Software Evolution (IWPSE'02)*, pages 76–85, New York, NY, USA, 2002. ACM Press.

[7] J. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, 2000.

[8] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[9] Y. Yamauchi, M. Yokozawa, T. Shinohara, and T. Ishida. Collaboration with lean media: how open-source software succeeds. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work (CSCW'00)*, pages 329–338, New York, NY, USA, 2000. ACM Press.

[10] Y. Ye. Dimensions and forms of knowledge collaboration in software development. In *Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05)*, pages 805–812, Taipei, Taiwan, December 2005. IEEE Computer Society.

[11] Y. Ye and K. Kishida. Toward an understanding of the motivation open source software developers. In *Proceedings of the 25th International Conference on Software Engineering (ICSE'03)*, pages 419–429, Washington, DC, USA, 2003. IEEE Computer Society.