Appeared in the Proceeings of APSEC2005, Presented at the Workshop on Supporting Knowledge Collaboration in Software Development, IEEE Computer Society, pp.827-834, December, 2005, Taipei, Taiwan.

# Understanding the Nature of Collaboration in Open-Source Software Development

Kumiyo Nakakoji[1] Kazuaki Yamada[1] Elisa Giaccardi[2]
*[1]Research Center for Advanced Science and Technology, University of Tokyo*
*[2]Department of Computer Science, University of Colorado, Boulder*
*{kumiyo, yamada}@kid.rcast.u-tokyo.ac.jp elisa.giaccardi@colorado.edu*

## Abstract

*Our approach to better understand the nature of collaboration in open-source software (OSS) development is to view it as a participative system, where people and artifacts are inter-connected via a computational infrastructure demonstrating a socio-technical system. This paper presents a framework we have developed to describe a participative system, and discusses our hypothesis that the framework is capable of characterizing the evolution of an OSS community through changing the participants' perceived value and types of engagement. We report a preliminary result of our case study on the GIMP development mailing list as an initial step to test this hypothesis.*

## 1. Introduction

Despite the wide spread of open-source software development styles, we still do not have a clear understanding of how and when open-source software (OSS) development projects work. We are particularly interested in a type of OSS development projects that is defined as "Internet-based communities of software developers who voluntarily collaborate to develop software that they or their organizations need" [6]. By this definition, this paper does not include as the object of our discussion other types of projects that develop open source software, for instance, Jun, which is a relatively small-sized, inhouse open-source software development project [1]. OSS development projects throughout this paper refer to those as defined above unless otherwise noted.

While a few OSS projects demonstrate a huge success in terms of their product quality, the increasing momentum for maintenance and evolution, and the growing size of development and user groups, we could also find a large number of "halted" projects at OSS repository services, such as SourceForge [13].

Social theories suggest that a sustainable community requires community members to be aware of benefits of belonging to the community and incentives to help his/her community [12][3]. Studies of OSS projects, however, have found that many of open-source community members do not necessarily see economical benefits contributing source code and answering questions posted by other community members [7]. While some studies have found that OSS developers are motivated to contribute primarily for self-learning experience [7], other studies have found that those who are given group goals contributed more than those given individual goals [8]. A model of private-collective incentive model seems to play an essential role [6] but it is not clear yet how goals of individuals and that of their project depend on each other. Other factors, such as membership size [2], or the degree of supervision by community owners [4], also seem to strongly affect the community performance.

To better understand the nature of collaboration in OSS development, our approach is to view OSS development as a *participative system* [11]. A participative system does not refer to a computational technology but to an organic socio-technical system "in which the social and technical infrastructures interconnecting users and artifacts" through supporting collaboration both in designing and using artifacts and in framing individual and collective goals [5]. By bringing socio-technical perspective that bonds people, artifacts, and computational environments, we are able to develop a method, or taxonomy, to talk about, analyze, and understand OSS development.

In what follows, we first present a framework we have developed to describe a participative system. We demonstrate how the framework is capable of characterizing the evolution of an OSS community through people's changing roles within the community. We show the result of our preliminary analysis of communication data of an OSS project (GIMP Mailing Lists) as an initial attempt to understand the nature of collaboration in OSS development project using the framework.

## 2. A Conceptual Framework for a Participative System

As described above, we argue that OSS development can be viewed as a *participative system*, where people and artifacts are inter-connected via a computational infrastructure demonstrating a socio-technical system. A participative system is a type of traditional collaborative system, in which people are bonded together through engagement rather than collaboration, demonstrating sustainability. Figure 1 illustrates the conceptual framework we have developed to describe such a participative system.

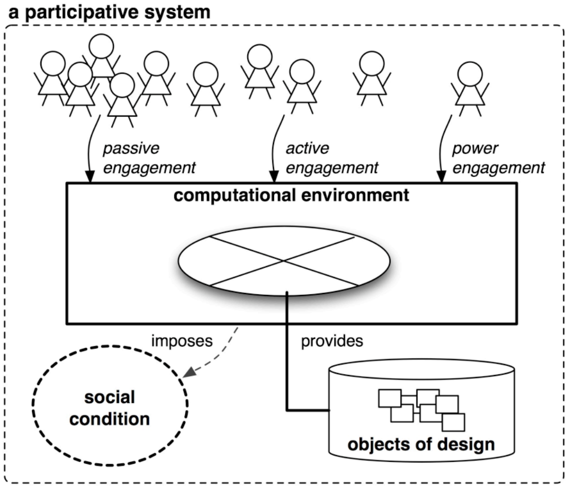

**Figure 1. A Proposed Framework.**

### 2.1 The Three Elements

In the framework, a participative system consists of three elements: (1) *a computational environment* (including networks and artifact repositories), (2) *objects of design* that people collaboratively engage in by using the computational environment, and (3) *social conditions* that are implied by the design and the operation of the computational environment.

The element of *social condition* differentiates a participative system from traditional groupware that supports division of labor. The element of *objects of design* makes a participative system unique from existing communication tools (such as online chat and electronic bulletin boards), and enable people to engage in individual and collective goals. Without a *computational environment*, it is not possible to share and store objects of design among a potentially large number of people, and to allow them to become aware and understand social dynamics demonstrated within the community.

The element of a computational environment of a participative system may consist of existing tools and mechanisms, such as ftp, mailing lists, or distributed databases. In the case of an OSS development project, programming environments also constitute its computational environment. Specialized programs may be developed to produce objects of design.

The element of objects of design in the case of an OSS development project includes source code, CVS repository logs, documentations, or communications among developers. The objects of design in a participative system is typically *open*; that is, open for change and open for evolution, and is associated with the goal of a participative system as a project, for instance, to develop freely available image processing software.

Examples of the element of social condition include social factors that are controlled and encouraged upon a certain group of people through technical elements, for instance, membership control, access control, or identity control.

### 2.2 Three Types of Goals

The concept of *goal* plays an important role in a participative system. We argue that it is important to become aware of the three different types of goals to better understand a participative system:

- *individual goal*: the lower level goal of a single individual
- *collective goal*: the project goal as perceived by a collective of single individuals; and
- *project goal*: the higher level goal imposed by authority, or the power user who designs the technical and organizational structure of the system.

For instance, the goal of a passive user who just uses source code from an OSS project repository may simply want to solve his/her own task; he/she may not be interested in how the project as a digital community evolves over time. In contrast, the goal of a core developer of an OSS project may want to evolve the project by encouraging more and more users to use the software developed by the project. Another user may simply enjoy reading mailing list communications by nurturing the feeling of belonging to a community.

### 2.3 Three Types of Engagement

The framework assumes three different types of engagement of people: *passive*, *active* and *power* (see Figure 1). Those who simply use the objects of design hosted by a participative system *passively engage* in the system. Examples of activities demonstrated by passive engagement include using, reading, viewing, checking-out, or downloading objects of design.

Those who make contributions to the objects of design in a participative system *actively engage* in the system. Example activities of active engagement include writing, creating, drawing, programming, creating, deleting, checking-in, or uploading objects of design.

Those who influence not merely the objects of design but also the participative system itself demonstrate *power engagement.* The initiator of an OSS development project often demonstrates power engagement by designing the three elements of the participative system as described in 2.1.

Nakakoji et al. [9] have identified that participants' roles may evolve over time within an OSS development project, starting from passive users to peripheral developers to core developers. We argue that one's role change over time is a result of his/her changing type of engagement. Such changes of types of engagement demonstrate a *migration path* of a participant engaging in a participative system. In other words, individuals migrate along the spectrum of *passive*, *active*, and *power* engagement.

### 2.4 Three Types of Values and Migration Paths

Value is a motivational factor that encourages people to engage in, or take part in, a participative system. A participant may find value in participating in a participative system by finding *contents* that are useful for his/her external task by *enabling* him/her to create, adapt and/or modify the contents for his/her own goal by being *triggered* to do so [5].

We have identified three factors for participants to perceive values: *efficiency*, *effectiveness*, and *meaningfulness*. A participant values *efficiency* in relation to the use of the system as a computational environment. *Effectiveness* is valued in relation to the importance attributed to his/her own external task. *Meaningfulness* is valued in relation to the personal experience.

Our hypothesis is that as a participant experiences value differently from *efficiency*, to *effectiveness* and to *meaningfulness*, his/her type of engagement changes from *passive*, to *active*, to *power*. Collective of such migration paths would result in the evolution of a participative system. Sustainability and evolvability of an OSS development project as a participative system thus might depend on how much we could encourage participants to experience efficient, effective, and meaningful value through the computational environment that embraces objects of design and social conditions.

The rest of this paper shows a preliminary result of our case study on an OSS development project as an initial step to test this hypothesis.

## 3. A Case Study: Toward Understanding of the Collaboration in the GIMP Project

Following the approach taken by Ye and Kishida [14], we have chosen the GIMP (GNU Image Manipulation Program) project as an example of OSS projects and started looking at the migration paths of community members of the project. The GIMP is freely distributed software for image processing, which has been developed and released since the late 1995. GIMP as an OSS project is interesting because the project has the history of temporal halt for about 20 months when the two original developers have decided to leave the project; the project has then been revitalized by other people taking the role of project owners [14].

GIMP-Developer mailing list has been serving “for interested users and developers to discuss the development and use of the system, to report bugs, and to submit patches for bug fixes and new features” [14]. We have analyzed the mailing list archive during the period between September 1st, 1999 and January 26th, 2005 (about 65 months).

### 3.1 Data Analysis Overview

In this case study, we focused on the email traffics in the GIMP developer mailing list. We were particularly interested in analyzing individual activities within the community as implied by the message-posting to the mailing list over a relatively long period of time, such as who tend to ask questions, who tend to provide answers, and how their roles have or not have changed over time. We looked at the *in-reply-to* relation among the messages. We distinguished two roles of people in communicating within a thread of messages using the relation: those who initiate a discussion by sending *action* messages (with the empty in-rely-to field), and those who respond to the discussion by sending *reaction* messages (with pointers in the in-reply-to field referring to previously posted messages).

In total, 14,031 messages have been posted to the mailing list. Among them, 4,723 messages are *action* messages and 8,343 messages are *reaction* messages. The rest of the messages have invalid in-reply-to field values, and we eliminated those messages from our analyses.

The valid 13,066 messages have been posted by 1,104 different “names,” which we have extracted from the header information of each of the messages. We have decided to use names rather than email addresses because a single person may have more than one email address, and the use of different email addresses may imply the commitment of role changes that the person perceives (for instance, some have started using the "@gimp.org" domain while posting to the mailing list). Note that our analysis has regarded those with identical last and first names as a single person. In the rest of this section, we call names and participants interchangeably.

1,009 participants have posted *action* messages (with no in-reply-to pointers) and 379 participants have posted *reaction* messages (with valid in-reply-to pointers). Among the valid 4,723 action messages, 2,299 messages had reaction messages (i.e., they are replied by other messages) meaning that they are pointed by other messages in their in-reply-to fields. Among the valid 8,343 reaction messages, 2,266 messages are replies to messages posted by other participants; the rest are replied by the sender themselves.

## 3.2 Community Activity Trends

Our initial attempt is to look at how the numbers of posted messages are distributed among the participants. Figure 2 shows a graph of who posted how many messages in total. The vertical axis represents the number of messages and the horizontal axis represents the participants arranged by the descending order of their number of posted messages.

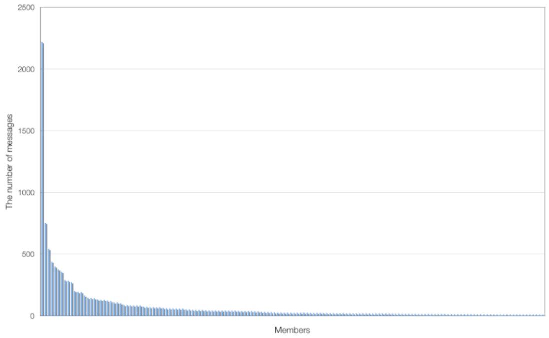

**Figure 2. Total number of messages posted by each participant.**

This distribution of communication among community members corresponds to the results reported by other researchers who have also analyzed other types of OSS development communities [7].

Figures 3, 4, and 5 illustrate the temporal distribution of posted messages by different types of participants according to their degree of activeness in terms of the number of posted messages.

Figure 3 represents the data on relatively active participants. Twenty-five participants have posted more than 100 messages over the 65 months period. The vertical axis represents the time. The bottom point corresponds to September 1st, 1999, when the mailing list has started, and the top point corresponds to January 26, 2005, when the mailing list archival data is collected by us. A point on the horizontal line represents a participant, and each dot on the corresponding vertical line represents a message posted by the participant at the time corresponding to the point on the vertical axis. Figure 3-(a) is depicted with the horizontal axis with the active participants arranged by the ascending order of the time when the *latest* (i.e., newest) message sent by each participant. Figure 3-(b) is depicted with the horizontal axis with the active participants arranged by the descending order of the time when the *first* (i.e., oldest) message sent by each participant.

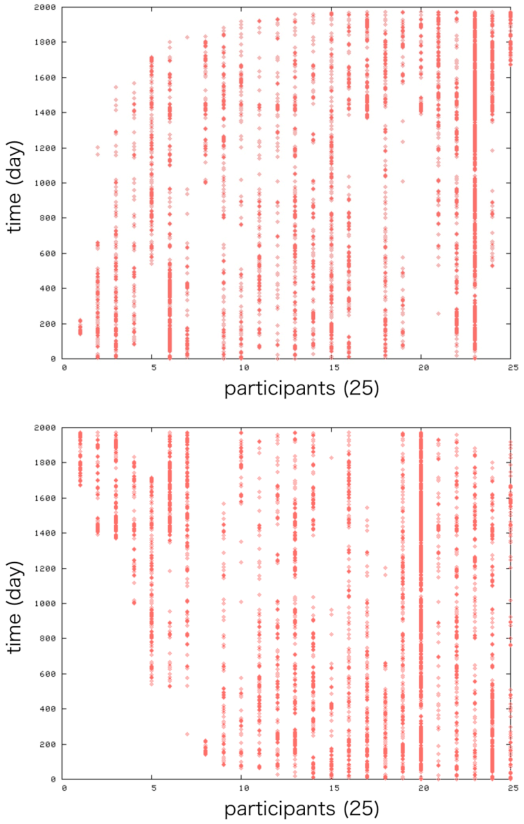

**Figure 3. Two graphs illustrating temporal distribution of messages posted by 25 active participants in different arrangement of participants: ascending order of latest time posted (a; top) and descending order of first time posted (b; bottom).**

Figure 3-(a) shows us that there are a few active participants who have either disappeared from the community or become passive early in the project. Figure 3-(b) tells us that there are several active participants who become active only recently and remain active. The two graphs also show that several participants become active and passive intermittently (indicated by white space between dots within a single vertical line).

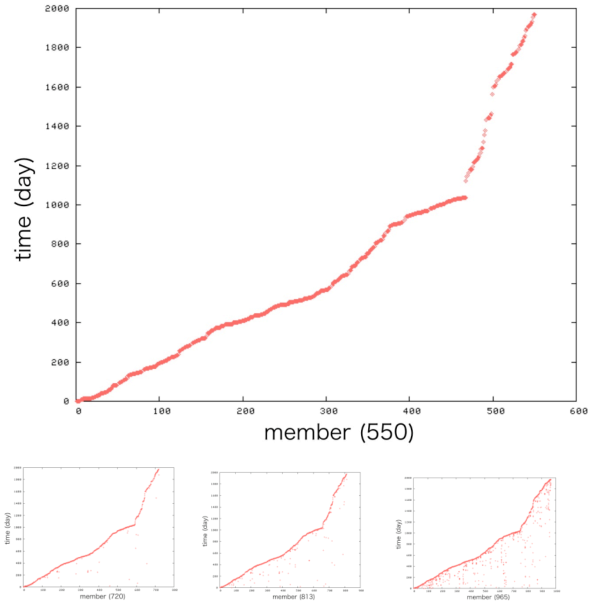

**Figure 4. Graphs illustrating temporal distribution of messages posted by passive and inactive participants: by those who have posted only one message (a; top), by those who have posted two or less messages (b; bottom-left), three or less messages (c; bottom-center) and ten or less messages (d; bottom-right).**

Figure 4 represents the data on new comers or those who remain as peripheral participants indicating passive engagement [9][15]. 550 participants have posted only one message during the 65 months period. They may keep reading messages posted on the mailing list, but we cannot tell from the mailing list data whether they have left the community or they become silent. As the same with the previous graphs, the vertical axis represents the same period of time, and a point on the horizontal axis represents each participant who has posted only one message. Figure 4-(a) is depicted with the horizontal axis with the participants arranged by the ascending order of the time when the message was sent by each of the participants. We have also produced similar graphs for those who have posted two, three, and ten or less messages, respectively (Figures 4-(b)(c)(d)). Their horizontal axes represent the participants arranged by the ascending order of the time when the latest message was sent by each of the participants. There are 720, 813, and 965 participants who have posted two, three, or ten or less messages, respectively.

The four graphs equally demonstrate the same trend. Around mid 2002, there seems to be a changing point in time that changes the frequency of new comers posting messages (see Figure 4-(a)). There might have been a shift of the direction of the project and/or that of the surrounding social and technical situations.

Figure 5 represents the data on regularly active participants; those who have posted more than thirty messages to the mailing list over the 65 months period of time (i.e., more than one message per two months in average). They may include actively engaged participants probably serving as peripheral developer roles but not as core members [9]. 71 participants fall into this category. As the same with the above graphs, the vertical axis of Figure 5 represents the same period of time, and a point on the horizontal axis represents each participant who has posted more than 30 messages in total. This graph is depicted with the horizontal axis with the participants arranged by the ascending order of the time when the latest message was sent by each of the participants.

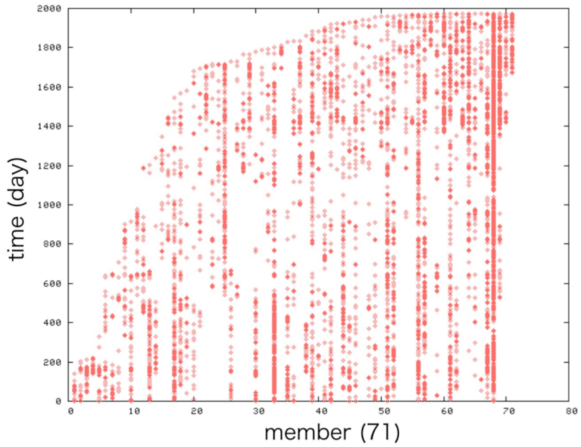

**Figure 5. Graph shows temporal distribution of messages posted by those who have posted more than 30 messages.**

As one can see from this graph, there is an obvious trend change in the curve if one looks at the graph from the vertical axis. The speed of the increase of those who stops posting messages becomes faster around the late spring and early summer of 2004. As with Figure 4, there might have been a shift of the direction of the project around that time.

### 3.3 Roles of Individual Participants

Figure 2 showed a graph of who posted how many messages in total. Figure 6 shows types of messages posted by the top-30 active participants in terms of the number of total messages posted. The vertical axis represents the number of messages. The horizontal axis represents the top-30 participants arranged by the descending order of the total number of their posted messages. For each participant, five values are depicted: the total number of messages sent by this participant (which is used to arrange the participants along the horizontal axis), that of *action* messages sent by this participant (see Section 3.1), that of *action* messages sent by this participant and replied by other participants, that of *reaction* messages sent by this participant, and that of *reaction* messages sent by this participant in reply to action messages sent by other people.

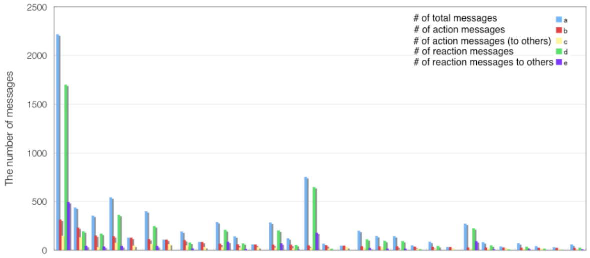

**Figure 6. The numbers of different types of messages posted by each participant.**

Figure 6 illustrates that the five numbers are not necessarily co-related to each other; that is, those who may send a large number of *action* messages may not post many *reaction* messages, and vice versa. In order to understand individual activities such as who tend to take which roles, Figure 7 shows ranking of top-ten actively engaging participants.

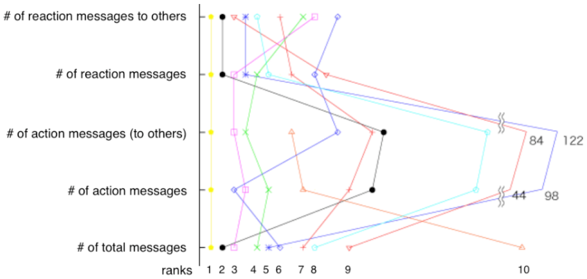

**Figure 7. Participants ranking changes over posting different types of messages.**

In Figure 7, while the most active participant remains the top most across *action* and *reaction* messages, those who ranked 7th, 8th and 9th in the total number of posted messages go down pretty drastically for *action* messages. This indicates that the three participants tend to *reply to others* but infrequently *poses new discussions*. In contrast, the participant who ranked 6th posts more *action* messages than *reaction* messages, indicating that this participant tends to initiate discussions than reacting to those initiated by other participants.

These graphs are based on the accumulated data over 65 months and it would be interesting to see how these ranking have changed over time. For instance, the one ranked 6th might become more "reaction" type participant. This is one of remaining challenge in our project.

### 3.4 Vocabularies Used by Individual Participants

Figure 8 illustrates how the categories of terms and expressions used by messages posted on the mailing list have changed over time. We have collected Subject lines of each document and have applied the principle component analysis method to categorize the topics.

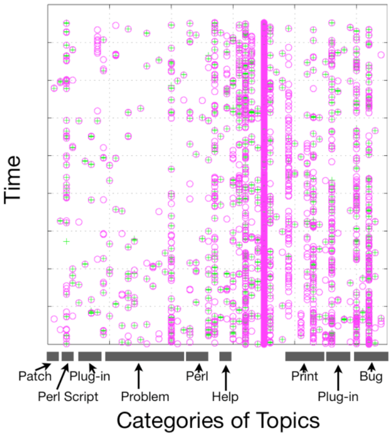

**Figure 8. How topics of message subjects have changed over time.**

The vertical axis of the graph depicted in Figure 8 represents time as the same with the graphs in previous subsections, and the horizontal axis represents categories of topics. While the categorization itself has been performed automatically, category names are manually added by identifying common properties of the subjects grouped within each category. The graph implies that topics among various fields are distributed consistently over the 65 month period of time.

We have examined if differences exist among vocabularies used by individuals In Figure 9, the number in the parentheses is the total number of messages posted by this participant.

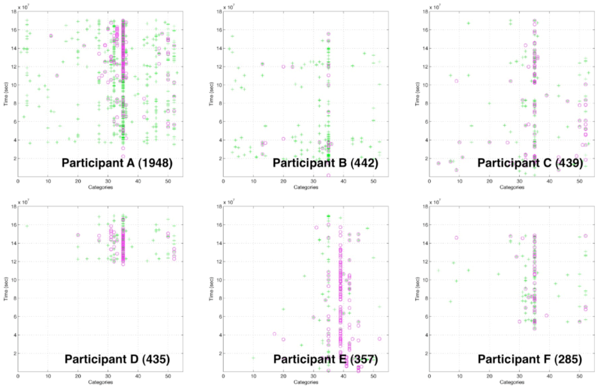

**Figure 9. How topics of message subjects have change over time posted by individuals.**

We have analyzed whether vocabularies typically used by actively engaging participants are different from those of passively engaging participants. Figure 10 shows ranks of how often each term is used in subjects of messages in three groupings: among all the messages, among the messages used by those who have posted two or less messages, and among those used by those who have posted only one message, over the 65 months. As implied by this graph, general terms, such as *error*, *questions*, *window*, and *image processing*, are often used by "new comers" and peripheral, passive users while terms addressing more specific aspect of the system, such as *plug*, *patch*, and *script,* are ranked highly in the overall message group.

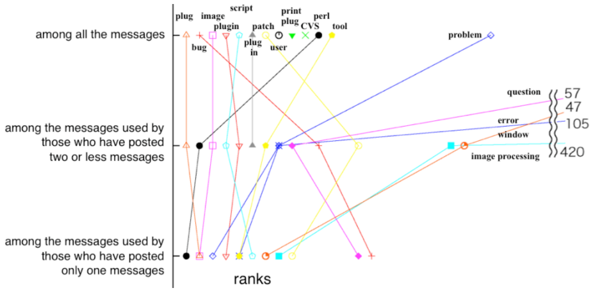

**Figure 10. Vocabulary ranking changes among different groups of users.**

This figure indicates that although we need further investment, it might be possible to identify the role of the participants within a community by looking at what kinds of vocabularies they use. This would be one of our future issues.

### 3.5 Communication Paths among Individuals

Finally, we have analyzed which individual participants communicate with whom in terms of posting an *action* messages and receiving *reaction* messages by other participants. Figure 11 shows data on three participants. In this circular graphic representation, all the members are distributed equally as dots along the circle in the alphabetical order of their last names. An arrow is drawn from one participant (i.e., the corresponding dot) to another participant if the action message sent by the former participant is responded by the latter participant as an reaction message using in-reply-to field.

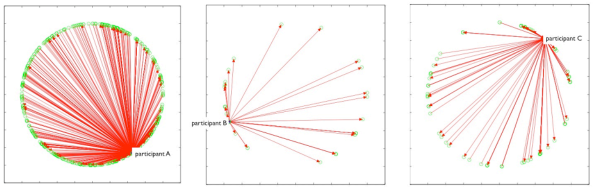

**Figure 11. Communication paths of three different participants.**

As one can see, the most active participant seems to equally communicate with a number of participants within the community. Our future work includes to categorize users and place them in a meaningful manner (rather than the alphabetical order) so that we would be able to identify evolutionary trends of the communication paths among individual participants.

## 4. Discussion

This paper reports our preliminary result of analyzing the GIMP mailing list as communication repositories of an OSS development project. These results are not to draw any robust conclusions on the hypothesis outlined in the end of Section 2. Rather, they are to show potential of future research areas and imply our future directions in testing the hypothesis and understanding the nature of collaboration in software development.

While some of future agenda have been described in the previous section, our next step is to expand our object of analyses and examines the GIMP repositories, as typified with change logs and CVS files. We have currently looked at only Subject text of each message in the mailing list, and we plan to analyze the actual text part in addition to the header information of each message.

To understand the nature of collaboration in OSS, two key factors, which have been little studied in the existing software engineering research framework, are *individuals* and *time.* Artifacts*,* such as source code, test cases, comments or documents, have long been studied in the software engineering research community. Truly understanding the nature of OSS development processes, however, we need to pay attention to how people learn, communicate, and work together to evolve artifacts within a socio-technical framework as a participative system. Analyses and visualization techniques have little studied how one should interact with temporal data [10]. We need to develop an analysis environment that is designed particularly for understanding how relationships among individuals and between individuals and artifacts change over time within a participative system.

Making progress in analyzing and understanding the nature of collaboration in software development would also lead us to develop better understanding of online and digital communities as participative systems. The framework depicted in Section 2 can be applied to other types of digital communities, such as open contents (e.g., http://www.wikipedia.org/) and net arts (e.g., http://www.sito.org/). The kind of work outlined in this paper should be a matter of concern not only within software engineering but also within a much larger research context.

## Acknowledgements

We would like to thank Yoshiyuki Nishinaka, Yunwen Ye, and Yasuhiro Yamamoto for their valuable comments, stimulating ideas, and deep insights for the work presented in the paper. This research is partially supported by the Open Competition for the Development of Innovation Technology (MEXT : No.15103), MEXT Grant-in-Aid for Scientific Research A(1), 15200012, 2003, and MEXT Grant-in-Aid for Exploratory Research, 17650038, 2005.

## References

[1] Aoki, A., Hayashi, K., Kishida, K., Nakakoji, K., Nisinaka, Y., Reeves, B., Takashima, A., Yamamoto, Y., A Case Study of the Evolution of Jun: an Object-Oriented Open-Source 3D Multimedia Library, Proceedings of International Conference on Software Engineering (ICSE2001), Toronto, CA., IEEE Computer Society, Los Alamos, CA., pp.524-533, May, 2001.

[2] Butler, B.S., Membership Size, Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures, Information Systems Research, v.12 n.4, p.346-362, December 2001.

[3] Butler, B., Sproull, L., Kiesler, S., Kraut, R. . Community effort in online groups: Who does the work and why? In Leadership at a distance, Weisband, S., Atwater, L. (Eds.), Laurence Erlbaum, 2005 (forthcoming).

[4] Cosley, D., Frankowski, D., Kiesler, S., Terveen, L, Riedl, J., How Oversight Improves Member-Maintained Communities, Proceedings of CHI 2005, pp.11-20, ACM Press, 2005.

[5] Giaccardi, E., Fogli, D., Beyond Usability Evaluation in Meta-Design: A Socio-Technical Perspective, IJHCS, (submitted).

[6] Hippel, E.v., von Krogh, G.v., Open Source Software and the "Private-Collective" Innovation Model: Issues for Organization Science, Organization Science, Vol.14, No.2, pp.209-223, March-April, 2003.

[7] Lakhani., K.R., Hippel, E.v. , How open source software works free user-to-user assistance. Research Policy, Special Issue on Open Source Software Development, 32, pp. 923-943, 2003.

[8] Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Cosley, D., Frankowski, D., Terveen, L., Rashid, A. M., Resnick, P., and Kraut, R. Using social psychology to motivate contributions to online communities. Journal of Computer-Mediated Communication, Vol.10, No.4, 2005.

[9] Nakakoji, K., Y. Yamamoto, Y. Nishinaka, K. Kishida, and Y. Ye., Evolution Patterns of Open-Source Software Systems and Communities, Proceedings of International Workshop on Principles of Software Evolution (IWPSE 2002), pp.76-85, 2002.

[10] Nakakoji, K., Takashima, A., Yamamoto, Y., Cognitive Effects of Animated Visualization in Exploratory Visual Data Analysis, Information Visualisation 2001, IEEE Computer Society, Los Alamos, CA., pp.77-84, July, 2001.

[11] Pangaro, P., Participative systems. Manuscript, 2000, Available at: http://www.pangaro.com/.

[12] Preece, J and Krichmar, D. M. Online communities. Jacko, J. and Sears, A. (Eds.) The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, pp.596-620, Lawrence-Erlbaum, 2003.

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

[14] Ye, Y., Kishida. K., Toward an Understanding of the Motivation of Open Source Software Developers, Proceedings of 2003 International Conference on Software Engineering (ICSE2003), Portland, Oregon, pp. 419-429, May 3-10, 2003.

[15] Ye, Y., Nakakoji, K., Yamamoto, Y., Kishida, K., The Co-Evolution of Systems and Communities in Free and Open Source Software Development, in Free/Open Source Software Development, S. Koch (Ed.), Chap.3, pp.59-82, Idea Group Publishing, Hershey, PA., 2004.