Correlating Temporal Communication Patterns of the Eclipse Open Source Community with Performance and Creativity

Yared H. Kidane
MIT
yaredo@mit.edu

Peter A. Gloor
MIT
pgloor@mit.edu

Abstract

This paper studies the temporal communication patterns of online communities of developers and users of the open source Eclipse Java development environment. It measures the productivity of each community and seeks to identify correlations that exist between group communication characteristics and productivity attributes. The study uses the TeCFlow (Temporal Communication Flow) visualizer to create movie maps of the knowledge flow by analyzing the publicly accessible Eclipse developer mailing lists as an approximation of the social networks of developers and users. Thirty-three different Eclipse communities discussing development and use of components of Eclipse such as the Java Development Tools, the different platform components, the C/C++ Development Tools and the AspectJ extension have been analyzed over a period of six months. The temporal evolution of social network variables such as betweenness centrality, density, contribution index, and degree have been computed and plotted. Productivity of each development group is measured in terms of two indices, namely performance and creativity. Performance of a group is defined as the ratio of new bugs submitted compared with bugs fixed within the same period of time. Creativity is calculated as a function of new features proposed and implemented. Preliminary results indicate that there is a correlation between attributes of social networks such as density and betweenness centrality and group productivity measures in an open source development community. We also find a positive correlation between changes over time in betweenness centrality and creativity, and a negative correlation between changes in betweenness centrality and performance.

Contact:
Peter A. Gloor
Center for Coordination Science
MIT Sloan School of Management
Cambridge, MA02142

Tel: 1-617-253-7018
Fax: 1-617-253-4424
Email: pgloor@mit.edu

Key Words: Temporal social network analysis, open source, TeCFlow, creativity, performance, virtual community.

Acknowledgements: We thank Yan Zhao as the main developer of the TeCFlow and online process tools, and Thomas Malone, MIT Center for Coordination Science, and Paul Johannesson, Systems and Computer Science Department Royal Institute of Technology, Stockholm, for their support.

Correlating Temporal Communication Patterns of the Eclipse Open Source Community with Performance and Creativity

Yared H. Kidane and Peter A. Gloor

Self-initiated groups of people who share the same passion and strive to achieve a specific goal began to exist in organizations long ago. These groups have been called by different names meaning the same thing. Some examples are “communities of practice”, “learning communities”, “family groups”, “thematic groups”, “peer groups”, “collaborative knowledge networks” etc…[Gloor, 2005]. Such groups have been creating value through developing and spreading new knowledge and capabilities, fostering innovations, building and testing trust in working relations [Communities of Intelligence, 2005]. The advent of modern Information and Communication Technologies (ICT) has leveraged the performance of these groups by providing instantaneous global accessibility, even though they were active before these technologies were in place. ICT has enabled collaboration of geographically distributed, functionally and/or culturally diverse entities by linking and creating lateral, dynamic relationships for coordination. Hence, such groups have acquired the name “Virtual Teams” or “Cyber Teams”.

The work of virtual teams ranges from IT outsourcing, software and distributed product development to running political campaigns and online charities [Gloor, 2005]. The focus of this paper is on open source software development, which is one of the prime manifestations of cyber teams collaborating over the Internet to produce software that can be used by people all over the globe for free. The role of the Internet is central for open source software development success. Without it, the boundary spanning among virtual entities would not be possible. The Internet has enabled virtual teams to link across distance, time, culture, departments and organizations, thereby creating "anyone/anytime/anyplace" alternatives to the traditional same-time, same-place, functionally centered, in-house forms of organizational experience [Lueg & Fisher, 2003].

Differences in productivity observed in open source development teams have created increasing interest to study their pattern of communication and further uncover the existence of association between communication pattern characteristics (social network attributes) and teams’ productivity measures such as performance and creativity. In this paper a time series analysis of the communication patterns of thirty-three different component development groups of the Eclipse open source developers’ community [Eclipse project, 2004] has been conducted. Assessment of their productivity is carried out in a bid to identify the correlation between temporal social network structures and the two variables performance and creativity.

Objectives of this Study

Currently there are thousands of open source projects in progress; despite this not all are equally successful. Linux, Apache web server, Internet news server, Mozilla web browser, and Eclipse are some that are notable for their influence, size, and success [Cubranic, 2005]. This difference in productivity is observed not only among different open source projects, but also between groups and subgroups in the same project.

There are different streams of active research trying to better understand the reasons for the growing success of open source movements, which not only produces software at lower cost, but also of higher quality [O’Mahony, 2003] [Moon & Sproull, 2000] [Weber, 2004]. Some are aimed at developing and testing a model of the influences of team, project, and user factors on their success. Other focus on social factors that enhance commitment, attract and retain developers, motivate developers’ contribution, facilitate the coordination and combination of developers’ contributions, etc… [Anonymous, 2005].

It is clear that productivity in open source development can be affected by many factors like: member’s skill difference, their commitment to carry out their specific task, nature of the problem they are dealing with, etc. This paper seeks to find out if the communication patterns of developers can explain variability in their productivity, keeping all other factors constant. We are trying to answer the question of what defines the most creative and efficient open source teams with respect to communication pattern characteristics. We would like to come up with recommendations for communication in virtual teams based on insights obtained through temporal social network analysis of open source developer teams. The result of the study will help to determine appropriate means for fostering innovation in software development teams. Identifying supportive attributes of communication patterns should lead to better performance and creativity. Our overall goal is to come up with a method to optimize

knowledge flow that will result in a more creative, innovative and responsive type of setting for knowledge intensive work [Gloor, 2005].

Communication Attributes and Productivity Measures

Software development is a complex socio-technical activity. It requires people to interact with each other, with the technical methods and computing technologies they use to perform their work [Sawyer, 2004]. The social aspect of software development deals with how people interact, behave, and organize. It helps to study patterns of communication among individuals and groups of developers to describe networks of relationships as fully as possible. It helps to tease out the prominent patterns in such networks, trace the flow of information (and other resources) through them, and discover what effects these relations and networks have on the final product i.e. the software system.

There are many attributes used to measure the characteristics of social networks. Two of the most prominent measures that are considered in this paper are betweenness centrality and density. Betweenness centrality measures how an actor controls information and relative access to network resources and information. It can also be interpreted as measuring the degree of independence from others in the network. The TeCFlow tool calculates it in a range from zero to one. The two possible extremes are a 100% centralization of the entire network, which is a case where one actor is bridge to all others, (star networks with betweenness centrality equal to one), and the other extreme is 0% centralization, which is a case where no actor is a bridge of another actor (circle network with betweenness centrality measure equal to zero). Density on the other hand measures the readiness of the group to respond to changes, its complexity and solidarity. It is defined as the percentage of ties that exist in a network out of all possible ties. A density of 1 implies that every actor is connected to every other actor. A density of 0 implies that no actor knows any other actor.

Productivity in software development is not only a measure of the time and/or cost required in delivering and maintaining a software system; but it is also a measure of the usefulness of the software system in satisfying the customers' and users' needs and expectation. Hence there are two major dimensions of software engineering productivity, the change in the quantity of software produced for a given period of time at a given cost (the process dimension) and the quality of the resultant software system (the software system dimension) [Duncan, 2003].

This paper limits itself to productivity measures from the software system perspective. It considers two attributes, size and defect. "Size" measures how the software system evolves in time, enhancements and additional features incorporated. "Defect" is the amount of bugs made during a given period of time. These two attributes give rise to productivity indices used in this research i.e. creativity and performance:

$$\text{Creativity Index} = \frac{\text{\# Of enhancements integrated in time period}}{\text{\# Of bugs resolved in the same time period}}$$

And

$$\text{Performance Index} = \frac{\text{\# Of bugs resolved in time period}}{\text{\# Of bugs reported in the same time period}}$$

Hypotheses

[Cross & Cummings, 2003, 2004] have demonstrated that the extent to which an individual is on the shortest information path connecting individuals who themselves are not connected (betweenness centrality), can be associated with one's ability to obtain and apply relevant information to solve problems [Weber, 2004]. This paper will try to make a similar case at the group-level based on communication and productivity data obtained from the Eclipse open source software development group. The following five hypotheses are defined.

Hypothesis-1: -Decentralized software development groups promote performance and creativity by enabling members to share knowledge in a more efficient and effective manner than centralized ones. In centralized social

networks, dissemination of knowledge takes longer as it needs to travel through the extended hierarchies. Therefore, the following hypothesis should hold *"Group betweenness centrality is negatively correlated with group performance and creativity"*.

Hypothesis-2: - As the number of ties between actors in a social network grows; density, performance and creativity of a group increase. Alternative ways for knowledge to flow through the network will grow which facilitate collaboration. Therefore performance and creativity of software developers should improve. Therefore the following hypothesis follows *"Group density is positively correlated with group performance and creativity"*.

Hypothesis-3: -This hypothesis follows from the economic concept of "opportunity cost": If a given developer group has to carry out competing tasks with the same resources (fixing bugs and plan, design and implement new features or enhance existing ones), there will be a tradeoff. As a result, an increase in one activity will entail a decrease in the other. Hence, hypothesis three follows "*Group performance and creativity are negatively correlated to one another"*.

Hypothesis-4: - Teams that have a constant communication structure perform better. For high performance, it is therefore better for a group to have similar communication patterns over the lifetime of the group. *Group performance is negatively correlated with the number of changes, i.e. fluctuations in betweenness centrality over time.*

Hypothesis-5: - Teams that have changing communication structures are more creative. Changes in communication structure between hierarchical communication (high group betweenness centrality), and dense, balanced, more democratic many-to-many communications (low betweenness centrality) are indications of high creativity. *Group creativity is positively correlated with the number of changes, i.e. fluctuations in betweenness centrality over time*.

Chosen Approach

This research seeks to better understand the characteristics of high performing open source developer teams by applying principles of social network analysis to the analysis of e-mail networks [Weber, Steven, 2004]. It follows the approach of Leenders et. al. [Leenders, R.Th.A.J. van Engelen, J.M.L. Kratzer, J., 2003]. To better understand the correlation between temporal communication patterns and performance of open source developer communities, mailing list archives of thirty-three Eclipse communities [Eclipse mailing lists, 2004] [Eclipse project, 2004] are analyzed. These communities are working on different modules of Eclipse such as the Java Development Tools, the different platform components, the C/C++ Development Tools and the AspectJ extension. Six months of communication data is analyzed. Mailing lists are considered as an approximation of the social network for the fact that they are main means of communication by developers actually working on or otherwise contributing to day-to-day development to discuss and make decision on design and implementation issues.

Social network data was collected from the Eclipse component development groups' online mailing lists by using the online process tool [Gloor & Zhao, 2004]. Data on bugs and enhancements per each group was collected from the Eclipse bugizilla database by using built-in queries [Eclipse bugizilla, 2004]. The social network data was analyzed with the TeCFlow tool [Gloor & Zhao, 2004]. The graph structure, group betweenness centrality, group density and similar attributes were calculated for each software development group. Based on the data found from the bugizilla bugs database performance and creativity indices were computed. Following these, statistical analysis, correlation and regression were used to test the hypothesis formulated above. Finally, on the basis of the results obtained from the analysis and hypothesis-testing phases a conclusion of the study is drawn.

Data Collection

The study is based on data from the three main projects of the Eclipse open source development community, namely "eclipse", "tools" and "technology". We have chosen thirty-three different component development groups for analysis. Selection of these project groups is random to avoid bias in the study.

Communication data was parsed for each of these component groups over a period of six months. The online process tool [on line process tool, 2004] was utilized to collect communication data from their mailing list archives. The online process tool runs a robot that searches for URLs in the projects' mailing list archives to compile a list of the possible URL links. It then extracts communication data as tuples in the form of (sender, receiver,

communication_type, timestamp, communication_contents) and stores it in the database. Further, bugs and enhancement data were collected from the Eclipse bugzilla database. These statistics form the basis for the calculation of performance and creativity indices

Analysis and Interpretation

In the data collection phase communication data from mailing lists was parsed and imported into MySQL databases. Then, the TecFlow tool was used to merge multiple datasets for the same "component" project groups into one. Data cleansing was performed to avoid duplicate presence of the same entity in the database and mass or group mails. The TeCFlow tool was used to automatically calculate group betweenness centrality and density from the communication data stored in the database. There is neither manual step nor subjective interpretation involved in this process. The algorithms for calculating these social network measures are pre-built in the tool. By using the bugs data obtained from the bugzilla bugs database and formula described above creativity and performance indices were computed.

A correlation analysis between group betweenness centrality and creativity reveals an important relation between these two attributes: when the groups get more central the creativity drops, and vise versa. Groups, which tend to be decentralized, make proportionally more enhancements within the analyzed time period than centralized ones. On the other hand, group betweenness centrality shows an insignificant influence on the performance of a software development group (see table 1).

A bi-variate correlation analysis supports hypothesis-2 that there is a strong positive correlation between density and performance of a group. The group with the highest density has resolved the highest number of bugs reported in the analyzed time period and the lowest–density group has the lowest performance index. Even though an inverse relation between density and creativity is observed, the relation is not statistically significant enough. Therefore, in this study, density was not an important social network attribute to affect creativity of a software development group (see table 1).

	Performance Index		Creativity Index	
	Pearson Correlation	P-Value	Pearson Correlation	P-Value
Group Betweenness Centrality	0.292	0.105	-0.169	0.356
Group Density	0.369	0.037 *	-0.224	0.218

Table 1. Correlation between Creativity/Performance and Density/Centrality

Our research indicates that performance and creativity are negatively correlated to one another with a correlation coefficient of B=-0.224, supporting hypothesis-3. This illustrates that the economic law of "opportunity cost" prevails even in open source software development teams. Stated differently, adding more enhancements comes at the price of slower bug fixing.

The relation discovered among group betweenness centrality, density, performance and creativity is illustrated by the social network graphs of the Eclipse open source developer teams in figure 1 below. Figure 1 demonstrates our hypothesis; it displays snapshots of the network graphs of communication networks, temporal evolution of betweeness centrality, degree centrality, density, and creativity and performance numbers of two different Eclipse developer communities. As seen in the figure, the group with high communication density exhibits a higher productivity index than the group with lower density. Further, the lower creativity is observed in the group with relatively centralized communication pattern.

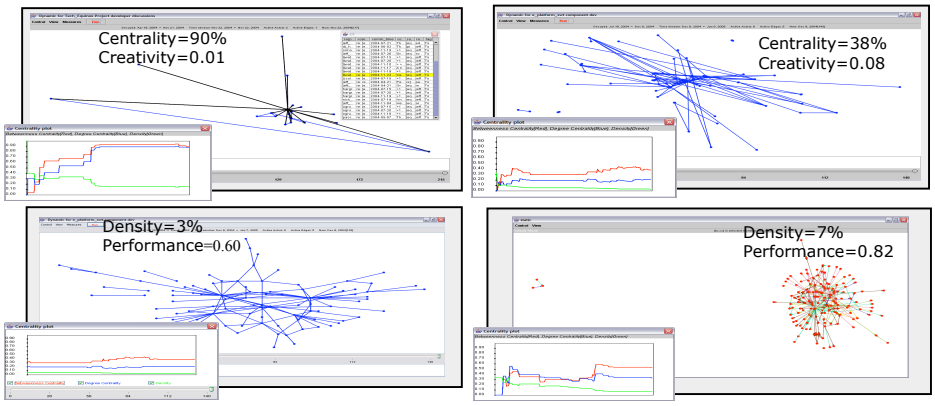

Figure 1: Network graphs of communication networks and temporal evolution of betweeness centrality, degree centrality and density

Temporal Analysis

We also aimed to better understand the impact of changes over times in social network structure on productivity and creativity of the software development groups. Towards that goal, data on the number of changes in the groups' density, betweenness and degree centrality has been collected. We counted the number of peaks and troughs in the temporal density, betweenness, and degree centrality plots, i.e. the number of their local minima and maxima. Correlation analysis between these datasets and measures of performance and creativity has revealed interesting results as shown in table 2.

Change In	Performance Index		Creativity Index	
	Correlation Coefficient	P-Value	Correlation Coefficient	P-Value
Group Centrality	-0.534	0.003*	0.601	0.01*
Group Density	-0.48	0.008*	0.188	0.329

Table 2. Correlation between Creativity/Performance and changes in Group Centrality and Density
*Correlation is significant at the 0.01 level (2-tailed).

Statistically significant relationship between changes in groups' centrality, performance and creativity has been identified. In support of hypothesis-4, we found that groups, which are steady in their density over time, are more performing. On the other hand, groups, which exhibit more change in their centrality, are found to be more creative and less performing. This finding supports hypothesis-5, illustrating that oscillation between hierarchical and decentralized communication structure is a strong indicator for creativity. Similar analysis on density versus creativity has shown no statistically significant correlation between these two variables.

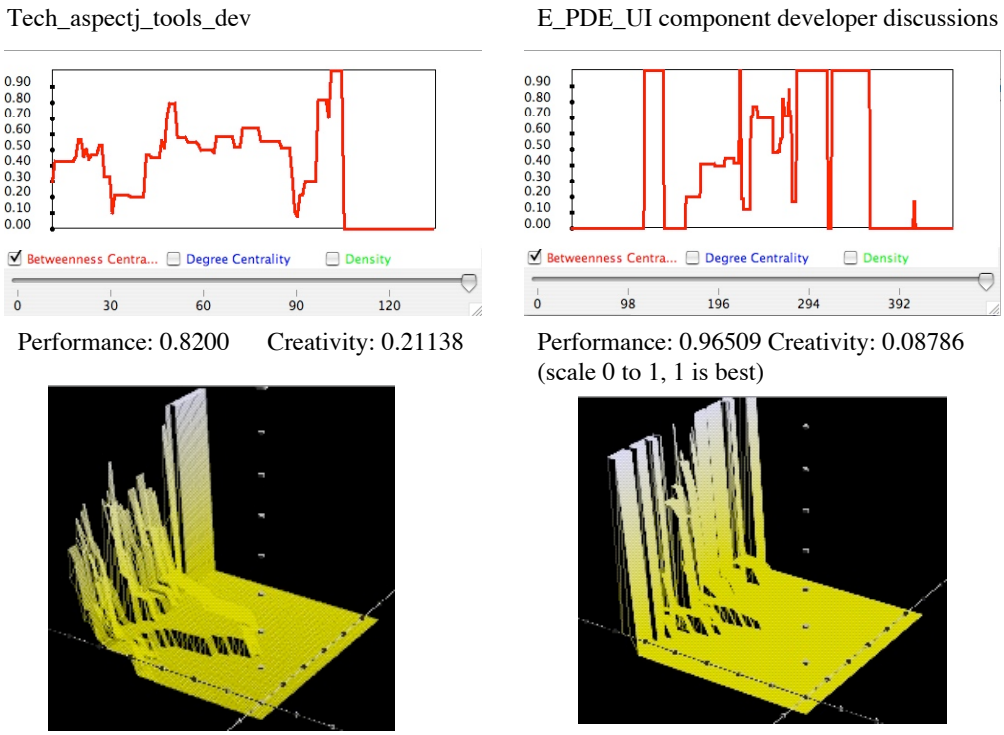

Figure 2: Correlation between temporal evolution of Betweenness Centrality and Performance and Creativity

Figure 2 illustrates our findings graphically. The left group with a less steady dynamic evolution over time of group betweenness centrality (top left of figure 2), and a changing social temporal surface has higher creativity and smaller performance than the group at right, the E-PDE_UI component developer discussion.

Conclusion

The eclipse open source development groups with more centralized communication structure are found to be less creative than decentralized ones. In this context, creativity is the defined as the number of new feature enhancements carried out by eclipse developers. We speculate that as communication between development teams gets more centralized, it prevents innovative ideas from coming up to the floor. On the other hand, we found that centrality of a software development group is not a major factor affecting performance.

While conclusive evidence is not yet found that lower group betwenness centrality is correlated with higher creativity, the study has shown that higher density is an indicator of higher performance. Eclipse development groups with high communication density seem to be better performers than those with low density. As development members get more connected, they become more efficient in dealing with bug resolution, fixing a higher percentage of reported bugs.

Whereas oscillations over time in group betweenness centrality and density are indicative of high group creativity, more steady temporal evolution of group betweenness centrality and density is a potential indicator for high performance.

Our continuing goal is to come up with recommendations for communication in virtual teams based on insights obtained through temporal social network analysis of open source developer teams. This research presents preliminary results on the relationship between communication structure and productivity of open source developer teams. Future work should focus on studying larger numbers of open source software development communities to come up with more insights that will help generalize the kind of correlation between group communication characteristics and productivity variables in open source development communities as a whole. It would also be useful to consider additional social network characteristic in the study. One should also consider how social network characteristics are related to the "process dimension" of software development groups' productivity measures in

order to see how cost of the development (in human resources, hardware resources, and calendar time), is affected by communication, collaboration and coordination variability.

References

A.S. Duncan, 2003, "Software development productivity, tools and matrices" *Proceedings of the 10th international conference on Software Engineering, 2003*

Anonymous, 2005 "A study of open source software project success" *project summary retrieved March 7, 2005 at URL: http://www.smith.umd.edu/faculty/kstewart/ResearchInfo/NSFProjectSummary.pdf*

Cross, Rob, and Jonathon N. Cummings, 2004, "Tie and network correlates of individual performance in knowledge-intensive work" *retrieved December 2004 at URL: http://ccs.mit.edu/fow/cross_cummings.pdf*

Cummings, J., & Cross, R, 2003 "Structural properties of work groups and their consequences for performance" *Social Networks, 25(3), 2003, 197-210*

Davor Cubranic, 2005 "Open-Source Software Development" *retrieved on March 7,2005 at URL: http://sern.ucalgary.ca/~maurer/ICSE99WS/Submissions/Cubranic/Cubranic.html*

Eclipse bugizilla bugs database. *Retrieved on December 1, 2004 at URL: https://bugs.eclipse.org/bugs/reports.cgi*

Eclipse mailing lists *Retrieved on September 1, 2004 at URL http://www.eclipse.org/mail/index.html*

Eclipse project, 2004. *Retrieved on September 1, 2004 at URL: http://www.eclipse.org/eclipse/*

Communities of Intelligence, 2005. *Retrieved on January 20, 2005 at URL: http://www.communityintelligence.co.uk/resources/collaboration_tools.htm#_ftn1*

Gloor, P. Laubacher, R. Dynes, S. Zhao, Y., 2003, "Visualization of Communication Patterns in Collaborative Innovation Networks: Analysis of some W3C working groups". *Proc. ACM CKIM International Conference on Information and Knowledge Management, New Orleans, Nov 3-8, 2003.*

Gloor, Peter A. 2005, forth coming book, "Swarm Creativity, Competitive advantage through collaborative innovation networks" *To appear at Oxford University Press, fall 2005, also available at http://www.swarmcreativity.net*

Gloor, P. Zhao, Y., 2004, "A Temporal Communication Flow Visualizer for Social Networks Analysis*", ACM CSCW Workshop on Social Networks. ACM CSCW Conference, Chicago, Nov. 6. 2004.*

Leenders, R.Th.A.J. Van Engelen, J.M.L. Kratzer, J., 2003 "Virtuality, Communication, and New Product Team Creativity: A Social Network Perspective", *Journal of Engineering and Technology Management, 20, 2003, pp. 69-92.*

Lueg, C. Fisher, D., 2003, "From Usenet to CoWebs, Interacting with Social Information Spaces", *Springer, 2003.*

Moon, J.Y. Sproull, L., 2000, "The essence of distributed work: the case of the Linux kernel." *First Monday, vol. 5, no. 11 (November 2000), URL: http://www.firstmonday.dk/issues/issue5_11/moon/*

O'Mahony, S., "Guarding the commons: how community managed software projects protect their work" *Research Policy 32, 2003. 1179-1198.*

Online process tool, 2004, at URL: http://www.ickn.org/ickndemo/TeCFlow_HToMailingListDia.html

Sawyer, Steve. 2004 "Software Development Teams", *Communications of the ACM, 47(12), December 2004*

Tyler, Josh, Wilkinson, Dennis, Huberman, and Bernardo A., 2003 "Email as Spectroscopy: Automated Discovery of Community Structure within Organizations" *HP Laboratories, 2003. Retrieved February 2005 at URL http://www.hpl.hp.com/shl/papers/email/index.html*

Wasserman, S., Faust, K, 1994, "Social Network Analysis, Methods and Applications", *Cambridge University Press. 1994.*

Weber, Steven, 2004, "The Success of Open source" *Harvard University Press, 2004.*