Effects of Communication Variety on Performance in Collaborative Software Development

Mehmet Gençer[1]

[1]mgencer@cs.bilgi.edu.tr
Istanbul Bilgi University, Department of Computer Science

Abstract

In this paper we present the findings of an exploratory study about the effects of communication variety on performance in collaborative software development communities. Performance was defined separately for the two generic activities that cover most of participants in a software development community: 'knowledge creation' and 'knowledge brokerage'. This study attempts to combine different, and conflicting theoretical lenses to understand effects of communication on these intertwined processes. In this attempt we introduce measures to assess variety of communication sources within and across sub-communities, which represents knowledge domains within a larger community, and build hypothesis in line with theoretical synthesis proposed. We have tested these hypotheses using electronic repositories from one large and one relatively smaller virtual software development community. Our results confirm the hypothesised effects for the most part. Deviations suggest that the community scale matters in terms of how communication variety corresponds to information variety, hence effects differing with respect to community size in knowledge creation process.

Introduction

Virtual communities are becoming increasingly commonplace with the advent of information and communication technologies (ICTs), and knowledge management systems (KMSs) built upon them. Besides well known cases like Facebook where the virtual replaces and/or complements the old ways we are connected in our private lives, many areas of professional life is effected by this change. Certain areas of work, where the subject of work is a digital good, is quite amenable to penetration of ICTs. An extreme case is software production, particularly open source software (OSS) communities, where not only all the subjects of work are digital but also the objects, programmers, are extremely comfortable with ICTs.

The volume, variety, and quality of information experienced by the participants of a virtual community varies with respect to the scale of the community and structural design of information exchange mechanisms in a KMS. However, the effect of such features of communication on outcomes of knowledge processes, such as productivity of individuals, is not well understood. For example, information variety to which an individual is subjected to may be thought of as a useful thing generally, but software engineers consider it as harmful, and instead subscribe to the principle of hiding information within the boundaries of software modules Parnas (1976). In general, the problem relates to several research areas such as knowledge management, informetrics, innovation, and organisation research, but its proper empirical treatment is missing in most, if not all.
In this paper, we report results of an empirical study concerning the effects of communication variety in virtual software development communities. We have tested our hypotheses on two separate communities of different sizes. While designing the research, we adopt a perspective which acknowledges heterogeneity of activities in such communities. In the next section, we develop a theoretical position to build hypotheses concerning the effects of communication variety in two particular types of knowledge processes (while there may be more): knowledge brokerage/dissemination, and knowledge creation. Then we present research data set and analysis methods, followed by a summary of findings. Finally we discuss the findings in relation to theory and further work, followed by conclusions.

Theoretical Background and Hypothesis

With the increasing emphasis on innovation, rather than production, in the business world, organisation and management research started to engage on the studies concerning knowledge processes in –work– organisations. Being a research thread in the early period of its development, organisational knowledge and learning research fails to resonate with theory and findings in related fields such as social networks and informetrics, and it is particularly underdeveloped beyond the level of business organisation (i.e. teams and individuals level) (Kostopoulos and Bozionelos, 2011). Furthermore, as Styrhre notes (2003) the uni-dimensional understanding of knowledge in the rationalist tradition of management studies is a stumbling block for development and refinement of theoretical constructs which is suitable to analyse and interpret the great variety of knowledge processes around us.
A relatively consistent finding in organisational knowledge research is differentiation of knowledge exploration and knowledge exploitation processes. Separate treatment of these two types of processes goes back to March (1991). Powell (1998) also notes knowledge-seeking and knowledge-creation as two generic types in inter-organisational knowledge processes. However, these and consequent approaches does not address the individual level dynamics of knowledge processes. Recent research literature on open source communities particularly address the individual level (i.e. the developers), but is largely detached from the theoretical concepts of organisational knowledge field.
In investigating the effects of communication on performance in open collaborative open source software development, we focus on two stereotypes among the most active participants in such a software community: (1) those who are focused on coding and improving particular parts of software, and (2) those who assist others in finding information. Here

we label the corresponding activities as 'knowledge creation' and 'knowledge brokerage', respectively. It is worth noting that these labels reflect a different cut regarding actor versus process oriented views of the two generic processes noted earlier by March (1991) and Powell (1998).

Few perspectives in the broader field of organisational research are suitable to approach the problem, each with their own merits to examine certain aspects. One group of approaches subscribe to the value of information variety in knowledge processes. One of these approaches is the social networks literature which point to advantages of variety of ties for performance in networks. This advantage stems from access to scarce information and is articulated in various forms in the literature under terms such as 'strength of weak ties' (Granovetter, 1973) and 'structural holes' (Burt, 1992). The notion of absorptive capacity (Cohen and Levinthal, 1990), although commonly applied at business organization level, similarly emphasizes importance of previous exposure to variety of information for adaptation and problem solving. Taken broadly, these approaches reflect a perspective which considers knowledge as a commodity which is passed around. As a consequence, variety of knowledge sources to which an individual or an organisation is exposed to is generally considered as a stimulator of learning or innovation performance in these perspectives.

A different approach to knowledge processes comes from the communities of practice framework (Orr, 1986, 2006; Lave and Wenger, 1991; Wenger, 1998). Contrary to a commodified view of knowledge, these studies emphasise that collective knowledge processes in expertise groups require socialisation with the group, establishment of common meanings and language, etc. By focusing on the life-cycle of participation and integration of individuals to a community, this approach points to usefulness of information coming from familiar, rather than novel, sources as it is more likely to adhere to a common language, and hence better facilitating knowledge transfer between individuals (Brown and Duguid, 2001). Brown and Duguid (2001) suggest that in knowledge processes which require combination of in-depth expertise of various individuals, knowledge transfer across team boundaries are more problematic.

Hypotheses

Our study attempts to combine these two different lenses in understanding the effects of communication variety on individuals' performance where separate types of knowledge processes (with their corresponding performance definitions) take place in the same community. For this purpose we have used two different performance outcome measures corresponding to the knowledge creation and knowledge brokerage processes in the context of software development. For those individuals engaged in knowledge creation, performance is associated with their contribution to changing the software. For those engaged in knowledge brokerage, on the other hand, performance is associated with how much others care about information disseminated by these individuals.

Organisation of large scale virtual software development communities are based on separation of work domains for communication efficiency. This generally means that there are several mail-groups formed to accommodate communication related to different parts or aspects of software development or use. The basis of our approach for combining two different theoretical approaches concerns the information exchange across sub-community

boundaries. Sub-communities are loci of relatively separated domain knowledge. In other words each sub-community is a community of practice which is relatively separated from others. For this reason we hypothesise that communication variety within, but not across sub-community boundaries increases performance in knowledge creation process. These dual qualities of communication ensure that an individual is exposed to the wealth of information needed to participate in a sub-field of practice, but not distracted with non-relevant information coming from other sub-fields. In using the communities of practice perspective in such a manner, we implicitly consider the amount of expertise of an individual as a controlling factor:

Hypothesis 1 Knowledge creation process requires access to knowledge variety within a focused domain, and thus its performance is associated with higher peer-diversity and lower domain-diversity in communication, in addition to experience in the domain.

For the knowledge brokerage process, the line of reasoning is reversed. We suggest that this process is more appropriate to apply a commodified view of knowledge. Individuals who are able to match and deliver a higher variety of knowledge across domain boundaries would be more successful in brokerage. As a consequence, better performance in knowledge brokerage is associated with communication variety across sub-communities. On the other hand, as social networks research points to, different sources within the boundary of a sub-community are unlikely to provide novel information as the information they provide is somewhat similar:

Hypothesis 2 Knowledge brokerage process requires knowledge variety across different domains, thus its performance is associated with lower peer-diversity in presence of higher domain-diversity.

Research data and methods

There are various measures of contribution to software development used in software engineering (Ghezzi et al., 2002). In this study we have preferred a less controversial and easily accessible measure of bug-solving performance. Thus performance in knowledge creation process is operationalised as the number of software bug-fixes by an individual, i, denoted by b_i.

As a measure of performance in knowledge brokerage, we consider how responsive is the community to information disseminated by an individual. Thus we resort to measures of centrality in the communication network emerging from e-mail exchanges. Although various network centrality measures are available (Wasserman and Faust, 1994), in-degree centrality measure (i.e. the number of replies to e-mails sent by the individual) is the proper metric with a straightforward interpretation in this case. The knowledge brokerage performance operationalized as the in-degree centrality of an individual, i, is denoted by c_i^{in}.

Apart from these two performance measures which are taken as the dependent variables in our hypotheses, we have developed measures for communication variety to operationalize the independent variables. Our hypotheses required assessment of communication

diversity of an individual both in terms of its heterogeneity across peers, and also in terms of its heterogeneity across different domains (sub-communities, and corresponding mail-groups in the software community with different subject focus). Following are the measures we define on per individual basis:

in-degree, out-degree centrality : Counts of e-mails addressed-to or originated-from the individual, denoted as c_i^{in} and c_i^{out}.

peer-diversity : To assess heterogeneity of one's communication with different peers, we have used Herfindahl-Hirschmann Index (HHI) (Hirschmann, 1964) commonly used in economics to compute homogeneity. The peer-diversity is computed using the counts of e-mails sent by an individual to each of his/her peers, and denoted as d_i^p. Assuming that an individual have a set of communication peers, communication strength with each of whom are denoted as $x_1^i, x_2^i, \cdots, x_{n_i}^i$, and whose sum is $\mathbf{x^i} = \sum_{k=1}^{k=n_i} x_k^i$, HHI gives us the homogeneity of this distribution of communication among peers as follows:

$$HHI_i = \sum_{k=1}^{k=n_i} \left(\frac{x_k^i}{\mathbf{x^i}} \right)^2$$

Thus the HHI index is bounded between 0 and 1. Simply computing $1 - HHI$, gives us a bounded measure of heterogeneity/diversity. Thus one can compute the d_i^p measure defined above as:

$$d_i^p = 1 - HHI_i = 1 - \sum_{k=1}^{k=n_i} \left(\frac{x_k^i}{\mathbf{x^i}} \right)^2$$

domain-diversity : Obtained by collating e-mails originated from the individual with respect to the sub-communities/mail-groups, and using HHI as described above for peer diversity. Similar to peer-diversity, this measure is also computed from counts of e-mails sent by an individual in each mail-group, and denoted as d_i^d.

time in community is the amount of time passed since an individual's participation to the community to time of performance measurement, denoted as t_i.

We have constructed data sets to test our hypothesis using publicly available electronic repositories of two separate open source software communities of different scales:

Eclipse is a project started in 2001. This data set covers the bug solving and e-mail exchange records from the beginning of the project until the end of 2010. There were 75.000 people in this large data set who have participated in the Eclipse community, and exchanged e-mails in 149 mail-groups/sub-communities. There were 1.2 million bug solving records for this project.

Jakarta has started in 2002, and the data set contains records from the beginning of the project until the end of 2010. There were 8.500 people in this relatively smaller data set, with 19 sub-communities, and 48.000 bug solving records for this project.

To construct the data sets we have used custom software to both retrieve mail-group e-mail records and bug-fix records, and also to produce the necessary measures. In each

of the two cases e-mail traffic was divided into multiple mail-groups which correspond to different sub-communities, each with its own focus on a different aspect or component of the software being developed. To produce the social network stemming from e-mail communication we have used the e-mail headers indicating 'who replies to who'. Although messages in mail-groups reach all registered users, it is reasonable to think that the immediate reference indicated in messages capture the actual correspondence that takes place in these mail-groups.
Since all dependent and independent variables are measured as bounded scalars We have used generalised linear models to test our hypotheses, using R statistics software (R Development Core Team, 2009).

Findings

In testing hypothesis 1, we have normalized an individual's bug solving performance with respect to the size of his/her time-frame of participation to the community, because this performance measure is cumulative. Concisely the linear model used in testing this hypothesis is:

$$b_i \sim d_i^p + d_i^d + t_i$$

The results of applying this generalised linear model to data are shown in Table 1. These

Table 1: **Effect of communication peer and sub-community diversity on bug-solving performance in knowledge creation process, normalized with respect to time spent in the community.**

```
ECLIPSE DATA-SET Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6.029e+01  1.529e+01  -3.942 8.15e-05 ***
peer-diversity  1.632e+02  2.803e+01   5.822 6.01e-09 ***
domain-divers. -8.174e+01  3.759e+01  -2.174   0.0297 *
time-in-comm.   3.313e-09  2.007e-10  16.509  < 2e-16 ***

JAKARTA DATA SET Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8.989e+00  2.070e+01  -0.434   0.6642
peer-diversity  1.906e+01  4.076e+01   0.468   0.6402
domain-divers.  1.301e+02  7.856e+01   1.655   0.0984 .
time-in-comm.   1.641e-09  3.389e-10   4.843 1.62e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

results support hypothesis 1 in the larger (Eclipse) data set because model coefficient for peer-diversity is positive and for domain-diversity it is negative as hypothesised, in

addition to being significant at $p < 0.001$ $p < 0.05$ levels, respectively. But the model is not affirmative of the hypotheses in the smaller (Jakarta) data set.
In consideration of testing hypothesis 2, it was apparent that the performance measure of in-degree was related to out-degree. In other words, those who work harder in terms of communication with others, would be responded back. In order to normalise the data in terms of this effect we have built a linear model accordingly:

$$\frac{c_i^{in}}{c_i^{out}} \sim d_i^p + d_i^d$$

The results of applying this generalised linear model to data are given in Table 2. These

Table 2: **Effect of communication peer and sub-community diversity on information value performance in knowledge brokerage process.**

```
ECLIPSE DATA-SET Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.73583    0.01580  109.89   <2e-16 ***
peer-diversity -1.16924    0.02773  -42.16   <2e-16 ***
domain-divers.  0.83635    0.03552   23.55   <2e-16 ***

JAKARTA DATA-SET Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.9776     0.1517  13.039  < 2e-16 ***
peer-diversity  -1.9071     0.2753  -6.926 1.08e-11 ***
domain-divers.   1.9045     0.5520   3.450 0.000598 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

results confirm hypothesis 2 for both large and small data sets, with high significance levels. For both data sets, the model gives a negative coefficient for peer-diversity and a positive coefficient for domain diversity, as hypothesised, and all are significant at $p < 0.05$ level.

Discussion

Perhaps the most substantial implication of the empirical evidence presented here is confirmation of the applicability of ostensibly conflicting lenses to collaborative knowledge processes. We have developed an approach which considers collaborative software development communities as the loci of a multitude of processes. In articulation of this approach we have considered two generic knowledge processes: knowledge creation and knowledge brokerage. In developing our hypotheses we have combined social network theory, which generally considers knowledge as a transferable commodity, and communities of practice theory which suggests that knowledge is sticky and its transfer is possible

between members of a community with a history, a common language and understanding of the affairs specific to a domain. Our combination of these theories was based on the assumption that a commodified view of knowledge was applicable for knowledge brokerage, but a sticky view of knowledge is appropriate for the deeper process of knowledge creation. As a result of this approach we have hypothesised that diversity of communication in across and withing knowledge domains has adverse effects, and the effect on performance of individuals depend on the process type in question. Despite relatively long history of acknowledgement of generic knowledge process types, application of such critical synthesis of different approaches is rather rare in organisational knowledge research. Further development of understanding knowledge processes will require novel empirical designs as well as critical evaluation of existing concepts and theories about organisational knowledge and learning (Styhre, 2003).

As part of our approach we have used commonly known performance measures for the two generic knowledge process types, but proposed rather novel measures to operationalise knowledge variety within and across domains. Although our application of these measures may seem specific to e-mail communities, it is quite generic. The only underlying assumption here is that the knowledge community in perspective is divided into multiple, recognisable sub-communities corresponding to knowledge domains. This is indeed the case for most communities of considerable size. Such communities (virtual or otherwise, concerned with software or some other professional subject) are almost always structured into groups with respect to work focus due to reasons of organisational efficiency.

An important deviance in the empirical analysis in this study is the difference in the nature of knowledge creation process between large and small community cases, where the hypotheses failed for the small community case. These results indicate the community size (and structure) as a mediator of how knowledge variation effect performance. A possible cause of the findings may lie in the fact that knowledge landscape in the larger community is more 'compartmentalized' compared to the smaller community case. As a consequence domain diversity in the smaller community may essentially be equivalent to peer diversity within the same field. In other words, small worlds phenomena (Watts, 1999), may be more involved in larger communities in terms of how novel information reaches different places of the communication network. However, we content that substantiating any possible reasons for this difference requires a different research design.

Conclusion

This paper presented results of an empirical study on the effects of information source variety in communication on performance of individuals, separately for knowledge creation and knowledge brokerage processes as two generic knowledge process types. We have developed a theoretical framework combining alternative theories, and produced hypothesis considering counter-veiling effects of peer and domain diversity. The hypothesis were confirmed on two data-sets from different sizes of open source software development communities, with some deviations which bear on the importance of scale in terms of interaction between the community structure and knowledge processes.

The theoretical articulation, measures proposed, and empirical evidence presented in this study contribute to somewhat disparate literature on organisational knowledge and

learning processes at the individuals level. We suggest that differing perspectives on organisational knowledge (or rather knowledge organisation) such as social networks theory or communities of practice theory may well be combined with a critical synthesis to understand and empirically validate the nature of collective knowledge processes in large communities.

References

Brown, J. S. and P. Duguid (2001). Knowledge and organization: A social-practice perspective. *Organiation Science 12*(2), 198–213.

Burt, R. (1992). *Structural holes: The social structure of competition*. Cambridge: Harvard University Press.

Cohen, W. M. and D. A. Levinthal (1990). Absorbtive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly 35*(1), 128–152.

Ghezzi, C., M. Jazayeri, and D. Mandrioli (2002). *Fundamentals of Software Engineering*. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Granovetter, M. S. (1973, May). The strength of weak ties. *The American Journal of Sociology 78*(6), 1360–1380.

Hirschmann, A. O. (1964). The paternity of an index. *The American Economic Review 54*(5), 761.

Kostopoulos, K. C. and N. Bozionelos (2011). Team exploratory and exploitative learning: Psychological safety, task conflict, and team performance. *Group & Organization Management 36*(3), 385–415.

Lave, J. and E. Wenger (1991). *Situated learning: legitimate peripheral participation*. Cambridge University Press.

March, J. G. (1991, feb). Exploration and exploitation in organizational learning. *Organization Science 2*, 71–87.

Orr, J. E. (1986). Narratives at work: story telling as cooperative diagnostic activity. In *CSCW '86: Proceedings of the 1986 ACM conference on Computer-supported cooperative work*, New York, NY, USA, pp. 62–72. ACM.

Orr, J. E. (2006). Ten years of talking about machines. *Organization Studies 27*(12), 1805–1820.

Parnas, D. (1976). On the design and development of program families. *IEEE Transactions on Software Engineering 2*, 1–9.

Powell, W. W. (1998). Learing from collaboration: Knowledge and networks in the biotechnology and pharmaceutical industries. *California Management Review 40*(3), 228 – 240.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Styhre, A. (2003). *Understanding Knowledge Management*. Copenhagen Business School Press.

Wasserman, S. and K. Faust (1994). *Social Network Analysis*. Cambridge.

Watts, D. J. (1999). Networks, dynamics, and the Small-World phenomenon. *The American journal of sociology. 105*(2), 493+.

Wenger, E. (1998, June). Communities of practice: Learning as a social system. *Systems Thinker*.