Does Distance Still Matter?

Timo Wolf,* † Thanh Nguyen and Daniela Damian

Software Engineering interAction Lab (SEGA), Department of Computer
Science, University of Victoria, Victoria, BC Canada

Nowadays, distributed development is common in software development. Besides many advantages, research in the last decade has consistently found that distribution has a negative impact on collaboration in general, and communication delay and time to complete tasks in particular. Adapted processes, practices, and tools are demanded to overcome these challenges.

We report on an empirical study of communication structures and delay in IBM’s distributed development project Jazz. The Jazz project explicitly focuses on distributed collaboration and has adapted processes and tools to overcome known challenges. We explore the effect of distance on communication and task completion time and use social network analysis to obtain insights about the collaboration in the Jazz project. We discuss our findings in the light of existing literature on distributed collaboration and delays. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: Global software development; distributed collaboration and delay; Jazz; social network analysis

1. INTRODUCTION

Global software development promises many benefits such as decreased development cost, access to a larger skill pool, proximity to customer, and 24-hour development by following the Sun. Unfortunately, many challenges have to be addressed for global teams to truly benefit from global development. These challenges cover many facets of software engineering, including processes, infrastructure, or business strategies (Prikladnicki et al. 2007). A recurring challenge, widely recognized by research and industry, is cross-site communication and collaboration (Perry et al. 1994, Battin et al. 2001, Herbsleb and Moitra 2001, Damian and Zowghi 2003, Holmstrom et al. 2006, Damian 2007, Herbsleb 2007).

A significant amount of research has reported about delays in the communication of distributed teams, since the acknowledgment that distance matters (Olson and Olson 2000) and the early reports from the field by Herbsleb and colleagues (Grinter et al. 1999, Herbsleb and Grinter 1999b, Herbsleb and Mockus 2003). In this paper we report from an empirical study of communication and work delay in global software development involving 16 distributed sites around the world. We study a distributed project at IBM, the Jazz development project, over a period of 23 months. The project uses the Jazz platform, which is a state-of-the-practice development environment aiming at providing collaborative support to distributed teams (Frost 2007). Our goal was to investigate whether the effects of distance reported in the literature in the last decade are still significant in affecting the project communication and time to complete tasks when a state-of-the-art collaborative tool and adapted processes and practices support the work of distributed software teams.
In Section 2 we motivate our work by discussing related work in collaborative software development, as well as current reports about challenges in distributed communication. Our research questions are provided in Section 3. Specifically, we focus on the effects of distribution on communication delay, communication structures, and task completion times. The study settings, the data constructs, measures, and the data collection are described in Section 4. We report and discuss our results in Sections 5 and 6 respectively. Possible threats to validity are discussed in Section 7. We conclude our paper in Section 8.

2. RELATED WORK

We review here the results of empirical studies of distance, communication, and delay, and motivate the design of our investigation into Jazz.

One of the notable series of empirical investigations of distance and delay in software development is the work of Herbsleb and colleagues. Early reports of communication problems and lengthened cycle time to resolve system issues (e.g. Herbsleb and Grinter 1999a,b) are followed by systematic studies into distance and speed at large distributed organizations (e.g. Herbsleb and Mockus 2003, Herbsleb et al. 2005). The 2003 study at Lucent (Herbsleb and Mockus 2003) provides systematic evidence about significant communication delay and task completion delay for modification requests that involved cross-site work. A comparison of data from same-site and cross-site projects indicates that tasks involving distributed participants take about 2.5-times longer to complete than similar collocated tasks. The result is explained by the perceived communication delay as reported in interview data. Other factors influencing task completion time included the number of people involved in the task, as well as the size of the task. Not surprising to this study, there were significant differences in the size of distributed versus same-site communication networks. This negative impact of distance on the properties of distributed social networks has recently been found by Ehrlich and colleagues (Ehrlich et al. 2007) as well.

Most of these findings are confirmed in later studies. An experience report (Herbsleb et al. 2005) of nine distributed projects at Siemens brings insights about benefits and challenges in distributed work, identified through interviews conducted at three different sites. Besides reported benefits, communication and collaboration across sites continued to be identified as a significant problem, leading to reduced development pace. The interviewees clearly stated that face-to-face communication is the most important and fastest communication. Frustrations arose when the pace of interaction went down.

In contrast to work in industry commercial environments, open source projects do not seem to have problems with distributed communication, collaboration and development following the sun. In 2006, Spinellis (2006) analyzed the source code from the CVS repository, the bug reporting system and the geographical developer locations of the FreeBSD operating system. He found that round-the-clock development takes place and the amount of distance does not negatively affect performance, code quality or the density of defects. Although these findings are promising, it is difficult to derive lessons learned for industrial projects, as the characteristics of open source projects are very different. First, the analysis has no collocated control group so that everyone is distributed and face-to-face communication is not existent. Thus, it is impossible to conclude that distributed development works as well as collocated development. Second, project characteristics such as developing on time and on budget are not existent. Third, developers are volunteers and are highly experienced and motivated.

Besides speed, the quality of team communication has also been of interest to research and recently linked to quality of developer and team performance. Research studies often use methods of social network analysis in studying the characteristics of communication networks of these teams. Studies of software teams support Conway’s law (Conway 1968) by showing that the organizational structure has a direct influence on the communication structure as indicated by the social networks (Gloor et al. 2003, Hinds and McGrath 2006, Hosain et al. 2006). Using social network analysis has the major advantage that one can draw from its extensive knowledge of analysis and implications (Freeman 1979, Burt 1995). For example the density of a social network reflects the ability to distribute knowledge (Rulke and Galaskiewicz 2000). In studies of communication and coordination specifically, social network measures such as density and centralization have been used to study the properties of communication networks. Communication
structure – the topology of a communication network – has been studied in relation to coordination ease (Hinds and McGrath 2006) and coordination capability (Hossain et al. 2006).

In summary, the literature in the last decade contains a body of empirical evidence about increased communication delay and work completion time in distributed projects, generally due to lack of informal communication and difficulties in expertise finding.

In this paper we describe an empirical study that aims at evaluating the impact of distance on communication and task completion time after almost a decade since the problems of distance have been reported. We specifically investigate distance and delay in a project that uses the Jazz development tool and which was developed specifically to support the collaboration of large distributed teams. The outline of our research questions in the next section is followed by the description of our research methodology and discussion of our empirical findings.

3. RESEARCH QUESTIONS

Our research questions center around communication delay, task completion time, and communication structures in distributed collaboration, as well as other factors that could introduce delays in communication besides geographical distance. Our investigation is in the context of Jazz development, of which characteristics we describe in detail in Section 4.1.

RQ1: Does geographical distribution introduce delay in project communication? Previous research (Grinter et al. 1999, Herbsleb and Mockus 2003, Sarker and Sahay 2004) identified communication delay by increased response times in distributed projects. We examine quantitative data from the Jazz project and compare response times between same- and multiple-site communication.

RQ2: Do distributed tasks take longer to complete than same-site tasks? Previous research (Herbsleb and Mockus 2003, Sarker and Sahay 2004) found that tasks involving people from distributed sites take longer than tasks involving people from only one site. We investigate if these findings still hold by comparing completion times between tasks involving different numbers of geographical distributed sites.

RQ3: What other factors influence communication delay and time to complete a task? Factors besides distance may influence the communication and task delay (Herbsleb and Mockus 2003). We investigate the relationship between communication and task delay and other factors such as the number of participants in the communication, the severity, and priority of tasks.

RQ4: What is the communication structure of the distributed, project-wide Jazz team? In the study of teams, social networks are recognized as important in fostering relationships, trust, and knowledge management. In distributed software development teams in particular, examining communication structures proved useful in studying coordination in global teams (Hinds and McGrath 2006, Ehrlich et al. 2007). We examine the structure of communication network of the entire distributed project using social network analysis techniques and seek to identify characteristics that explain observations about communication speed and task completion.

RQ5: Do collocated and multi-site communication-based social networks have different structures in Jazz? While RQ4 focuses on the project-wide communication structure in Jazz, we are also interested in identifying if characteristics of the communication in collocated or multi-site (two, three, or more sites) differ. We define and compare social networks measures across networks involved in collocated and multi-site communication.

4. METHODOLOGY

This section describes the design of our study. We provide details about the study setting before we describe our constructs and data collection methods.

4.1. Study Settings

We studied the communication and collaboration in the Jazz development project, over a period of 23 months. Jazz is a development environment that tightly integrates programming, communication, and project management. With collaboration
support as one of its main goals, Jazz provides integrated support for work planning and tracking, continuous builds, static code analysis, and reporting tools to measure the quality of the work products and the development process.

Jazz development involves 16 different sites that are located in the USA, Canada, and Europe. The development and testing takes place at 7 different sites and involves 151 active participants working in 47 functional teams. From now on we will only refer to the seven sites in our study that are Hawthorne, Raleigh, Lexington, and Beaverton in the USA, Ottawa and Toronto in Canada, and Zürich in Europe. Jazz members belong to multiple functional teams (referred to as teams henceforth), each managed by a team lead. The team leads report to a project management committee, which is responsible for the project-wide coordination. The team size ranges from 1 to 20 and has an average of 5.7 members. The number of developers per geographical site ranges from 7 to 24 and is 14.8 in average.

The Jazz development uses the Eclipse Way process. It defines 6-week iteration cycles, which are separated into planning, development, and stabilizing activities. Every iteration releases a new milestone build of the product. The project management committee formulates the goals and features for each release. Every team has an iteration plan that consists of unstructured text and task descriptions, which are captured in work items.

Work items represent single assignable and traceable tasks. Different types of work items are created to represent defects, enhancements, and general tasks. Commenting on the work items is the main task-related communication channel. A second broadcast communication channel is email, which is mostly used for announcements. Chat is used for synchronous communication among the developers.

In this paper, we focus on work items as they represent the most identifiable unit of work, i.e. tasks, and which provide the context for communication and collaboration. Task related communication is established by commenting on the work items. Depending on the location and team memberships of the authors, the communication can be across teams and different geographical sites.

4.2. Data Constructs

The main variables of interest in our study are distance, communication, and task completion time. To study communication, we focus on communication speed and the structure of the communication network formed around the resolutions of work items. As such, our unit of analysis is the work item and the associated collaboration that involves members from same and different-sites. We define the following constructs:

**Number of sites** involved in the communication and completion of a work item is our conceptualization of distance. For each work item, we use the location of each author of a comment to obtain the number of sites that were involved in communication. For example, if a work item had two commenting authors from two different sites like Zürich and Ottawa, then the work item falls into the two-site category.

**Response time** is our conceptualization of communication speed and it relates to the possible communication delay because of distance. We assume that the comment thread on a work item represents a conversation about the work item, and that a comment is a reply to the previously created comment. We measure the time intervals between the creation dates of sequential comments. For each work item, the response time is the average of all comment time intervals of a work item.

**Resolution time** is our conceptualization of task completion time. For each work item, we compute the difference between the completion date and the creation date.

**Developer communication network** is a construct we define to study the collaboration aspects in the distributed Jazz project team. Using information about work item-centered communication, we build communication networks (a) of the entire Jazz project and (b) of developers involved in different types of multi-site interaction (e.g. two-site, three-site). To construct the networks, we first select a set of work items and obtain all contributors who commented on these work items. For the project-wide communication network, all work items are considered; for the multi-site communication, we choose work items that involve communication from one, two, three, or more sites. The networks then includes these contributors as nodes, with edges between nodes being added if the contributors represented...
by the nodes comment on at least one common work item.

Previous literature (Herbsleb and Mockus 2003) suggests that also other factors may affect communication in distributed development. We study the possible relationships between response and task completion time with the following variables:

**Number of comments** on a work item, as a measure of how controversy the work item was discussed.

**Number of authors** as the number of different authors that created the comments of each work item.

**Severity** of a work item as an indication of how critical the work item is (could be unclassified, minor, normal, major, critical, and blocker).

**Priority** of a work item, as a measure of importance during planning (could be unassigned, low, medium, or high).

4.3. Data Collection

The main artifacts, including work items, associated comments and contributors are located in the server component of Jazz. A query plug-in was implemented to extract the required artifacts from the Jazz server. We imported the resulting data into a relational database management system.

We extracted a total number of 18,618 work items. However, as some geographically locations of Jazz participants were unknown, we decided to remove 4876 work items in which any missing contributor location or any unknown contributor was involved. In the resulting data, we could identify all authors, locations, and the creation dates for all work items and comments.

Our analysis thus involved data on a total of 13,742 work items that were created between October 2005 and November 2007. About 0.73% of these work items were created in 2005, 29.46% in 2006, and 69.82% in 2007. The work items have 43,967 different comments in total. The number of comments per work item ranges from 0 to 58 and is 2.75 in average. As much as 2669 work items had no comments and when considering only work items with comments, the average number of comments became 3.41. We identified 148 contributors including their locations.

5. DATA ANALYSIS AND RESULTS

In this section we describe the data analysis methods and findings in relation to each of our research question in part. As our variables do not follow a normal distribution (e.g. see Figures 7 and 8), we use nonparametric statistics for the tests reported below.

**RQ1: Does geographical distribution introduce delay in project communication?**

To examine response time differences for work items involving communication from different numbers of sites, we categorized the work items by the number of involved sites and computed the average response times for each of the work items. Table 1 shows the different categories (up to five sites involved in a work item communication), the number of work items in each category, and the mean and median response time for each category. For example, a work item with only comments from Ottawa and Zürich is in the two-site category. We also computed the mean of the first two comments in each site category as a comparison with the mean response time of all comments.

One can see that the means and medians behave differently when the number of sites increases. Although the mean response time decreases, the median response time increases when more sites are involved. With data that is not normally distributed, the median values become a better indicator of variations in the variable. Note that the number of work items is much larger when communication involved up to three sites than when communication involved four or five sites.

The box plot in Figure 1 shows the response times of all work items categorized by their number of sites. The horizontal bar in the middle shows the median value. The bottom/top line of the box
Research Section

Figure 1. Box plot of the work item response times categorized by the number of sites involved in the work item communication shows the 25th/75th percentile. The complete box shows where 50% of the data lie. The whiskers show the minimum and maximum respond time values, excluding outliers.

The Kruskal-Wallis test of difference in response times across the five different number of site categories yields a statistically significant difference ($K = 26.31, p < 0.001$).

To examine the size of this effect, we applied correlation tests. In other words, to investigate the strength of the relationship between distribution and delay in communication, we tested for a correlation between the number of sites and the average work item response time. We ran the non-parametric correlation test Kendall Tau (Siegel 1956). The correlation result yielded a very low correlation value ($\tau = 0.05, p < 0.001$), which means that distributed communication does not appear to introduce a significant amount of delay compared with same-site communication.

RQ2: Do distributed tasks take longer to complete than same-site tasks?

Similar to the previous analysis, in order to compare resolution times of work items completed at the same site or distributed sites, we categorized the work items by their number of sites and calculated the mean and median resolution times for all work items (Table 2).

In contrast to the effects on response time, the trend in the mean and median resolution times shows an increase in both means and medians when the communication involved contributors from one to many sites.

Table 2. Mean and median resolution times in days for the different numbers of involved sites

<table>
<thead>
<tr>
<th>Number of work items</th>
<th>Mean (days)</th>
<th>Median (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One site</td>
<td>3975</td>
<td>43.17</td>
</tr>
<tr>
<td>Two sites</td>
<td>2019</td>
<td>54.10</td>
</tr>
<tr>
<td>Three sites</td>
<td>329</td>
<td>72.86</td>
</tr>
<tr>
<td>Four sites</td>
<td>51</td>
<td>85.90</td>
</tr>
<tr>
<td>Five sites</td>
<td>5</td>
<td>53.13</td>
</tr>
</tbody>
</table>

Figure 2. Box plot of the work item resolution times categorized by the number of sites involved in the work item communication

The box plot in Figure 2 also shows that the maximum and the variance of resolution time increase when the number of involved sites increases. Although the Kruskal-Wallis test of difference in resolution times across the five different number of site categories yields statistically significance ($K = 101.37, p < 0.001$), the Kendall Tau correlation test of association yields an extremely low correlation value ($\tau = 0.09, p < 0.01$). This indicates that distributed communication does not appear to introduce a significant amount of delay compared with same-site task completion time.

Table 3. Kendall Tau correlation results

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response time</td>
<td>1.0</td>
<td>0.54</td>
<td>-0.10</td>
<td>0.08</td>
<td>0.05</td>
<td>0.05</td>
<td>-0.04</td>
</tr>
<tr>
<td>Resolution time</td>
<td>1.0</td>
<td>-0.13</td>
<td>0.11</td>
<td>0.15</td>
<td>0.09</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>Severity</td>
<td>1.0</td>
<td>0.10</td>
<td>0.12</td>
<td>0.10</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of authors</td>
<td>1.0</td>
<td>0.56</td>
<td>0.53</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of comments</td>
<td>1.0</td>
<td>0.35</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of sites</td>
<td>1.0</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Priority</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The column header relates to the row numbers. Medium and high correlation values are shown in bold font.

RQ3: What other factors influence communication delay and time to complete a task?

To investigate whether other factors influence the communication delay and task completion time, we investigated the relationship between response and resolution time and the other factors defined in Section 4.2: severity, number of commenting authors, number of comments, number of sites, and priority. Table 3 shows the results of the Kendall Tau correlation tests for each of these relationships.

Despite these correlations being statistically significant at $p < 0.001$, the actual correlation strengths are very low, except for the association between response time and resolution time, the association between number of authors and number of comments, and the association between number of authors and number of sites. We believe that these correlations, the very low included, were found statistically significant because of our very large sample size.

RQ4: What is the communication structure of the distributed, project-wide Jazz team?

To obtain more insights about the project-wide collaboration patterns in Jazz, we constructed a communication-based social network that includes all Jazz contributors and their communication behavior. We create an edge between two nodes in the network (Jazz contributors) if the two contributors commented on at least one common work item. The Jazz-wide communication network is shown in Figure 3 and consists of 112 nodes and 3296 edges. The contributors’ geographical location is shown by the shape of the node as well as delineated by grey line ovals. To examine the structure of the Jazz team communication network, we compute the $k$-core (Seidman 1983), core-periphery structure (Borgatti and Everett 2000), degree centrality, and group efficiency tests (Everett and Borgatti 2005). The analysis for each of these tests and results are described below.

To investigate how cohesive the Jazz team was with respect to communication, we used the $k$-core and the core-periphery structure tests. First, the $k$-core test identifies whether there is a cohesive subgroup in the network (referred to as the core) on the basis of a defined nodal degree. The $k$-core value specifies the minimum number of ties that must be present from each node of the core to the other core members. More details on this measure and an example is provided in the Appendix. Second, the core-periphery structure test indicates whether the structure of a network resembles a star-like structure that has one core and a surrounding periphery (rather than multiple-core or spark structure in which there is no or many cores). The core-periphery test results in values between 0 and 1. A high value indicates a strong core-periphery structure and is independent of any $k$ value.

For the Jazz team, the $k$-core test with $k = 25$ results in a core subgroup with 60 out of 112 members of the entire Jazz project; 2118 out of the total of 3296 ties are among the participants of the core. This large core consists of about 50% of the members and about 75% of the communication of the project. The core-periphery test yields a value of 0.758, a high value that indicates the structure of one core and a surrounding periphery. In Figure 3, the $k$-core with $k = 25$ is highlighted with a circle in the center.

To examine the level of connectedness of each geographical location to other project members and to examine how central each geographical location is compared with other locations, we computed, for contributors in the network of each geographical location, the group degree centrality and the group efficiency indexes (Everett and Borgatti 2005), respectively, as follows.
The Group Degree Centrality index provides a centrality measure for a group of nodes in a network. It is defined as the number of nodes outside the group that are connected to the members of the group. Unlike the group degree centralization measure (used later in RQ5), the Group Degree Centrality does not focus on individual nodes, but identifies how central a group of nodes is in relation to the rest of the network. To compare the geographically different groups in Jazz, we normalized the index by the number of actors outside the group and thus report percentages as values of this index. More details on this measure and an example is provided in the Appendix. The Group Degree Centrality index for each of the seven Jazz development locations are shown in Table 4.

The table also shows the number of participants at each location. For example, 12 project participants are located in Zürich and have communicated to 97% of all project members that are at a different location other than Zürich. As almost all locations have a very high group degree centrality index and the indices of the top four locations (>90) are almost equal, we cannot identify geographical locations that are more central than the other locations in terms of number of external communication peers.

The Group Efficiency index provides a measure of how redundant are the communication ties from members inside the group to members outside the respective group. It is defined as the fraction of the size of the minimum subgroup, which has the same group degree centrality index as the whole group, and the number of group members. More details on this measure and an example is provided in the Appendix. This index ranges from 0 to 1, where 1 indicates that the group has no redundant communication ties to project members outside the group. In Jazz, the efficiencies of the seven geographical locations range from 0.2 to 0.57 (Table 4) and have an average of 0.37. The overall low-efficiency index indicates redundant communication ties across project participants from different geographical locations.

RQ5: Do collocated and multi-site communication-based social networks have different structures in Jazz?

To explore the characteristics of single- and multi-site communication, we computed a number of network measures and examined any differences across networks that involve single- or multi-site communication. We constructed distinct networks that reflect communication about work items that involve contributors from single- and multiple (two, three, etc.) sites respectively. Specifically, we categorized each work item by the number of locations of the commenting contributors. The nodes in the communication networks were contributors that commented on those work items, whereas the ties between two contributors indicate that they comment on at least a common work item. This categorization results in communication networks that involve only participants from one location (e.g. Zürich, Raleigh, Lexington), as well as all combinations of two to five locations. For example a two-site communication network would involve only participants from Zürich and Raleigh or only participants from Zürich and Lexington. Networks involving participants from Zürich, Raleigh, and Lexington, or Zürich, Raleigh, and Ottawa are examples of three-site communication networks. Table 5 shows the number of constructed communication networks, the total number of work items for each location category, and the average number of contributors in

<table>
<thead>
<tr>
<th>Location</th>
<th>Group degree centrality (%)</th>
<th>Group efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zürich (12)</td>
<td>97</td>
<td>0.25</td>
</tr>
<tr>
<td>Raleigh (17)</td>
<td>97</td>
<td>0.41</td>
</tr>
<tr>
<td>Lexington (23)</td>
<td>94</td>
<td>0.26</td>
</tr>
<tr>
<td>Ottawa (25)</td>
<td>93</td>
<td>0.20</td>
</tr>
<tr>
<td>Beaverton (10)</td>
<td>83</td>
<td>0.40</td>
</tr>
<tr>
<td>Toronto (10)</td>
<td>75</td>
<td>0.50</td>
</tr>
<tr>
<td>Hawthorne (7)</td>
<td>55</td>
<td>0.57</td>
</tr>
<tr>
<td>Average</td>
<td>85</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Table 5. Number of constructed communication networks, the total number of work items, and the average number of contributors in the networks for each number of locations

<table>
<thead>
<tr>
<th>No. of locations</th>
<th>No. of networks</th>
<th>No. of work items</th>
<th>Avg. no. of contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>14,498</td>
<td>13.14</td>
</tr>
<tr>
<td>2</td>
<td>52</td>
<td>4,593</td>
<td>11.13</td>
</tr>
<tr>
<td>3</td>
<td>66</td>
<td>850</td>
<td>10.02</td>
</tr>
<tr>
<td>4</td>
<td>37</td>
<td>122</td>
<td>8.80</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>31</td>
<td>11.00</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>5</td>
<td>8.50</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>9.00</td>
</tr>
</tbody>
</table>
the networks. For example, we constructed 52 communication networks that involve participants from two locations and are based on the total number of 4593 work items.

To test for any differences in communication structures in single- and multi-site communication, we computed the following network measures: density, centrality (group degree centralization and group betweenness centralization), and structural holes (Wasserman and Faust 1994, Burt 1995). We briefly introduce each measure before describing the results of the tests for difference across communication networks.

**Density** is a measure of the degree to which all members in a team are connected to one another. It is calculated as the percentage of the existing connections to all possible connections in the network. A fully connected network has a density of 1, while a network without any connections has the density of 0. An example is shown in the Appendix. Density measures have been studied, for example, in relation to enhanced group identification (Reagans and Zuckerman 2001) and to problems in coordination (Hinds and McGrath 2006).

**Centrality** measures indicate activity, importance, or prominence of actors in a social network. The most commonly used measurements for centrality are degree centrality, betweenness, and closeness (Wasserman and Faust 1994). Each has a different social implication. Gloop et al. (2003) built communication networks from the email correspondence of three collaborating groups. They use the centrality measures to characterize and compare the different networks. Hossain et al. (2006) also built communication networks on the basis of email correspondence and explores the correlation between network centrality and coordination, which is measured using content analysis of the email content. They found betweenness to be the best measure for coordination.

**Degree centrality** can be calculated at both the node and network level. We compute the network level centrality measure, the **Group Degree Centralization** index, which indicates the extent to which the nodes in the network have more or less an equal number of communicating partners. Thus, it identifies if some nodes have more communication partners than others and are more active and important in the network, or if the number of communication partners are equally balanced and all network nodes are equally active. The actual formula for the group degree centralization is provided in the Appendix. The group degree centralization ranges from 0 to 1, with the index reaching its maximum value of 1 when one node is connected to all other nodes and all other nodes are only connected to the central one, like a star. The index attains its minimum value of 0 if all nodes have the same centrality degree.

Betweenness measures the extent to which a team member is positioned on the shortest path in between other two members. People in between are considered to be ‘actors in the middle’ and to have more ‘interpersonal influence’ (Freeman 1979) in the network. **Betweenness Centrality** is a probability index for the betweenness of the node in the complete network. Similar to degree centrality, the betweenness centrality can be calculated at both the node and network level. We compute the network level measure, the **Group Betweenness Centralization** index, and which relies on the measures of node betweenness (for actual formulas see Appendix). Similar to the group degree centralization, its minimum value of 0 is reached when all network nodes have the same betweenness index. Its maximum value of 1 is reached in a star network, in which the central node is between all other nodes.

**Structural Holes** measures are concerned with the notion of missing connections and connection redundancies among nodes (Burt 1995). At the node level, structural holes are gaps between nodes in a social network. For an example of a structural hole and the measures outlined below, see the Appendix. We use the following structural hole measures in our study:

- **The SH effective size** of a node \( c_i \) is the number of its neighbours minus the average degree of those in \( c_i \)'s ego network, not counting their connections to \( c_i \).
- **The SH efficiency** normalizes the effective size of a node \( c_i \) by dividing its effective size with the number of its neighbors.
- **SH constraint** is a summary measure that relates the connections of a node \( c_i \) to the connections of \( c_i \)'s neighbours. If \( c_i \)'s neighbors and potential communication partners all have one another as potential communication partners, \( c_i \) is highly constrained. If \( c_i \)'s neighbours do not have other alternatives in the neighborhood, they cannot constrain \( c_i \)'s behavior.
To extend the structural hole measures from the nodes to the complete network, we compute the sum of the measures for each node of a network. As the measures are based on network connections, we normalize the sum by computing the fraction of the sum and the number of possible network connections.

Having defined these measures, we performed two sets of tests: (i) the test of differences in these measures across networks involving one, two, three, and more sites and (ii) the test of differences in these measures between single- and multi site networks, i.e. collocated versus distributed communication.

To test for differences in measures across networks involving any number of sites, we used the non-parametric Kruskal-Wallis test (Siegel 1956) for significant differences between the communication networks involving one, two, and three different locations, for each of these network measures. We did not include the networks involving four, five, six, and seven locations, because of the low number of involved work items (See Table 5). Kruskal-Wallis test yields statistical significant differences for the measures group degree centralization ($K = 15.55, p < 0.001$) and group betweenness centralization ($K = 12.83, p < 0.01$). No significant differences were found for the other measures. Figures 4 and 5 show the box plots for each location category.

To test for differences in networks involving collocated versus distributed communication, we separated all networks in two categories: collocated and distributed communication networks. The collocated category includes all single-site networks as defined above, while the distributed category contains all networks involving two and three locations as defined above. Having only two test groups, we used Mann-Whitney non-parametric test (Siegel 1956). The test yielded statistically significant differences for the group degree centralization measure ($U = 177, p < 0.05$). Figure 6 shows the group degree centralization measures for each category. Table 6 shows all the test results of all measures in detail. Significant $p$-values are shown in bold font.

6. DISCUSSION

Motivated by reports of challenges of distributed development and advances in collaborative software development environments in the last decade, our study sought to bring yet another piece of
Research Section

Table 6. Statistical test results of the SNA measure differences between one, two and three locations (Kruskal-Wallis test) and between the collocated and distributed categories (Mann-Whitney test)

<table>
<thead>
<tr>
<th>Measure</th>
<th>1-3 locations</th>
<th>Coll. versus distr.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kruskal-Wallis</td>
<td>Mann-Whitney</td>
</tr>
<tr>
<td>Density</td>
<td>5.77</td>
<td>0.056</td>
</tr>
<tr>
<td>Gr degree centr</td>
<td>15.55</td>
<td>0.0004</td>
</tr>
<tr>
<td>Gr betw centr</td>
<td>12.83</td>
<td>0.002</td>
</tr>
<tr>
<td>SH effective size</td>
<td>4.84</td>
<td>0.089</td>
</tr>
<tr>
<td>SH efficiency</td>
<td>3.19</td>
<td>0.202</td>
</tr>
<tr>
<td>SH constraint</td>
<td>2.75</td>
<td>0.253</td>
</tr>
<tr>
<td>p-value</td>
<td>272</td>
<td>0.496</td>
</tr>
<tr>
<td>U</td>
<td>177</td>
<td>0.047</td>
</tr>
<tr>
<td>p-value</td>
<td>267</td>
<td>0.453</td>
</tr>
<tr>
<td>p-value</td>
<td>227</td>
<td>0.197</td>
</tr>
<tr>
<td>p-value</td>
<td>204</td>
<td>0.108</td>
</tr>
<tr>
<td>p-value</td>
<td>223</td>
<td>0.178</td>
</tr>
</tbody>
</table>

Significant p-values are shown in bold font.

evidence about coordination, communication, and delay in software engineering.

Although delays in communication have been a recurring theme in published reports of industrial practice, our investigation of communication in Jazz does not show the same effect. In what follows we interpret the results of the several analyses we conducted on response time and task completion times in Jazz, along with the findings of communication structure in the Jazz. It is our intention to reveal a rich picture of collaboration supported by one of the more recent development tools, focused on supporting collaboration in distributed teams: Jazz.

Communication and Task Completion in Distributed Teams

In contrast to previous studies, we did not find that geographical distance introduces significant delays in communication and task completion. On the basis of previous literature, we expected that responses on and resolution of work assignments would take longer as the number of sites involved in the communication increased. Thus we analyzed the response and resolution times in relation to the number of sites involved in each work item. Our findings show no clear impact of distance. Although the test of difference Kruskal-Wallis tests showed statistical significance (indicating that the variation in one of the distributed communication settings is different from the others), the correlation results that tested the size of this effect were extremely low. This indicates that variations in communication and task completion times cannot be associated at all with the variation in geographical distance of collaborators in the project. Thus, the expectation that larger distribution, as found in collaboration involved a higher number of sites, is associated with an increase in both response and work item resolution times was not confirmed in our study.

One possible explanation for this difference in our findings is that previous studies used qualitative, self-reported interview or survey data, which might not have provided an objective measure of communication delay. The interviewed participants might have remembered and reported only about high delayed communication instances, possibly a small and unrepresentative sample to generalize from. Our data is less prone to participant recollection problems.

Another possible explanation might come from the characteristics of the Jazz project participants. We can expect that developers developing a tool that focuses on collaboration support reflect on how they communicate and collaborate within the project and optimize these activities. Thus, they may perform better than the developers in other projects, which focus on a different domain and deal with distribution issues as lower priority items. In addition, demographic information such as age and familiarity with text-based communication specific to commenting on work items or chat might influence the ability to collaborate across sites. Unfortunately, we do not have demographic information of the Jazz team.

The descriptive statistics in our data set also allows us to do a more in-depth analysis of the trends in communication and resolution times in same-site and distributed communication:

First we discuss the distribution of response times for the five different categories of work items as shown in Figure 7. With the exception of the five-site category, the single-, two-, three- and four-site categories show a consistent pattern: about one-third of work items have an average response time of 1 day; about 10% of work items have an average response time of 2 days (exception is in the four-site category), and the remaining work items exhibit a long-tail distribution pattern indicating the presence of some comments with very long response times. This distribution together with information shown in the box plot may explain the inverse trend in the pattern of mean versus median shown in Table 1: the number of long responses decreases as the number of involved sites increases (mean value decreases as number of sites increases).
whereas there are fewer quick responses as the number of sites increases (median value increases as number of sites increases).

Further, as our data is not normally distributed, the median is an important measure for the trend in response time (as opposed to means). As such, one can observe that the median response time in the two-site distributed communication is only 0.61 days larger than the median response time in the same-site communication category. Without claiming that this is the difference in average times between same-site and two-site response times (as such computation is not permissible on median values), we believe it still indicates that collaborators in two-site communication wait for a response in average about half a day longer than those in collocated communication. When this is considered in the light of 6-week long iteration cycles, one may understandably not find it disturbing. A similar trend can be observed as the number of sites increases. The five-site case is somewhat difficult to analyze. With a small sample size of ten work items, the response time distribution shows an almost even distribution of the response times from 1 to 13 days. However, this is too small for a sample to allow us to regard it as representative for this case.

In summary, we believe that the geographical distribution did not have a practical effect on the response times in the Jazz environment.

Second, the evidence on the resolution times in the several distributed communication settings is not as clear and rather mixed. The descriptive statistics provide interesting insights about the distribution of resolution times when the number of sites increases. As shown in Table 2, both mean and median values increase from single- to five-site communications. The histograms in Figure 8 show a similar trend for the first three single-, two-, and three-site categories: in each of the categories there are many work items with short resolution times, followed by the remaining in a long-tail-like distribution; about 20% of resolution times are within 1 day for both same- and two-site communication, while the percentage of same resolution times is 10% for three-site communication. The fact that both median and mean values increase as the number of sites increases is indicating that probably the number of very long resolution time work items (at the end of the long tail) is increasing as the number of sites increases. The data in the four- and five-site categories, however, are rather small in size to allow us to draw conclusions about any pattern. As seen in the histograms, the resolution times...
Does Distance Still Matter?

Research Section

times for the 51 work items in the 4four-site cate-
gory are almost evenly distributed between 0 and 27-day wait time, with some outliers with over 30-
day wait time. Similarly, the data for the five-site category shows rather long resolution times. Again, the very few data points in our set, 5, are not suf-

cient for drawing any conclusions for the trend in the five-site communication category. In summary, we believe that the difference in resolution times when the number of sites increases is not practically significant, given the rather similar distribution of response time values in the same site, 2 two- and three-site distributed communication.

Interestingly enough, none of the investigations into correlations between response and resolution times and other factors as suggested by literature revealed possible alternative explanations for vari-

ations in communication and task completion in the different distributed settings. We thus turn our attention to the findings about the communication structure in the Jazz social network, to explain our findings on communication and task completion.

Communication Structure in Distributed Social Networks

The lack of significant communication and resolution time delay does not come as a surprise when the characteristics of the Jazz collaboration envi-

ronment are being considered. Drawing on the knowledge of the Jazz process, the Eclipse Way (Frost 2007), as well as on our private conversations with the Jazz project managers, we provide further explanation to our findings: The applied process of the Jazz project, with its collection of best practices and the supporting Jazz tool, contributes positively to reducing communication delay. As indicated to us, Jazz project members are encouraged to imme-

diately respond to requests and comments from the members of remote teams. This practice is con-

ducive to cross-site communication and, apart from reducing the communication delay across sites, contributes to an enhanced familiarity and closeness of participants located at different geographic sites. As a result, we believe that increased familiarity reduces communication problems such as misunder-

standings and mistrust across sites (Grinker et al. 1999, Jarvenpaa and Leidner 1999) and contributes to a denser social network in the project.

A result of our analysis of the communication structure in Jazz is that the project-wide

communication-based social network has a large core of active contributors, suggesting that many project members were actively communicating across functional teams as well as geographical distance.

This finding contradicts expectations formed by previous studies of subjects that were not using Jazz. Since maintaining interpersonal relationships and awareness of work and expertise have been found challenging in other distributed teams (e.g. Grinker et al. 1999, Herbsleb and Grinker 1999b, Ehrlich et al. 2007), a less dense network would be expected in the large distributed Jazz team. The social networks of former distributed projects are typically characterized by a more restricted flow of information across sites (Herbsleb and Mockus 2003, Ehrlich et al. 2007). The presence of particular team members acting as point people through whom much of the team communication flows to the distanced teams was found to contribute to the efficiency of coordination across teams (Hinds and McGrath 2006).

The Jazz network, in contrast, appears to be very connected across distances. We find that the distributed network has a large core with mem-

bers from all geographical locations, and none of the locations has a dominant or central role in the overall project communication. Instead of having multiple geography-dependent clusters of connected cores (as in a fractionated network), the entire project communicates through one large core. Studies (e.g. Hinds and McGrath 2006) found that communication networks with a strong core-

periphery structure lead to less coordination prob-

tems than fractionated networks with multiple cores that are loosely connected. In a core-periphery structure, many core members are available to broker information from one end of the periphery to the other. In fractionated networks the task of brokerage between two groups depends on several members. It is that these brokers can become communication bottlenecks if their work load is too high and mem-

bers outside the core get disconnected from the group when these brokers are not available.

With a core comprising about half of the entire project team, the Jazz network in Figure 3 also shows that about half of the members from each geographical location are in the core. Having multiple members of each team in the core, connecting their team to members of other teams,
reduces possible communication bottleneck problems and introduces redundant communication channels, enabling fast communication. While these core members may still act as information brokers to the rest of the team, the network as a whole exhibits a rather informal hierarchical organization, since the other peripheral members in each geographical location appear to be fairly connected to the other members outside the core area. These redundancies reduce the risk of communication bottlenecks and increase the ease of communication.

Our analysis of properties of collocated versus distributed networks in Jazz brings additional evidence to explain our hypothesis that the Jazz team acted as a single cohesive team, despite the distance between its several worldwide locations. The two network structure measures found to significantly differ across teams that were collocated or multi-site were the measures of group degree centralization and group betweenness centralization.

The degree centrality of a node in a communication network is a measure for activity and prominence (Wasserman and Faust 1994) and emphasizes the most visible actors of the network. At a network level, the measure of group degree centralization indicates the extent to which there are members with a higher degree centrality than other members. The low group degree centralization index for the single-site networks (Figure 4) suggests that members in single-site networks have almost the same centrality degree, and thus roughly the same number of communication peers at the particular location. For tasks that are carried out at only one location, the members are well balanced in terms of activity and prominence. In contrast, the higher group degree centralization index for the communication networks involving multiple locations (in both Figures 4 and 6) indicates that some members are more active and prominent in the network. We expect that these are the members of the core, which actively communicate to members of other locations.

According to Freeman (1979) and Hossain et al. (2006) the betweenness centrality measure is an indicator for control and coordination. Actors with a high betweenness index are responsible for controlling and coordinating information between actors that do not have an otherwise direct connection. The low group betweenness centrality values for communication networks involving tasks at only one location (Figure 5) suggest that all members in single-site communication have an almost equal betweenness centrality index. Thus, the controlling and coordination activities at one location are more evenly distributed among all members of one location. At the same time, the group betweenness test for communication networks involving multiple locations results in a statistically higher index for two-site networks, while the index for three-site networks is low again (Figure 5). We expect that the reason for the high index at two locations is that the core members control and coordinate the information between the two locations. When three locations are involved, a larger number of members from all locations are in the core and the need for coordination and control by each individual core member is decreased. These activities are shared among more participants of the core, each core member being less often in a position to mediate communication between other members, resulting in the decrease in the group betweenness centrality index.

In summary, we believe that this project-wide communication structure in Jazz was enabled and continually supported by the above-mentioned communication practices, substantially contributing to reducing possible communication delay caused by distance.

7. THREATS TO VALIDITY

We recognize a number of threats in the construct and external validity of our study results.

7.1. Construct Validity

Our communication is based on comments on work items. Besides commenting on work items, the Jazz project members also use other sources of information, which we did not consider. These include email lists, chat, web-based information, and face-to-face meetings. On the basis of our conversations with the Jazz team and observation of the work item online commenting behavior, we believe that comments are most commonly used to communicate about work items. This is largely because the comments were work-item-specific information immediately available with the work item and easily accessible to project members irrespective of their geographical location.

Our conceptualization of response time assumes that every comment on a work item is a response.
Research Section

Does Distance Still Matter?

to the previous comment on that respective work item. The assumption might not be correct for all comments, since a comment might be also independent from the previous comments. We investigated some work items manually and assume that our assumption is correct for the majority of comments. Further, because the second comment is more likely a response to a first comment, we also computed the means of only the first two comments (see last column of Table 1). Since the means in both cases are similar, we are more confident that this threat is not as important.

The resolution time construct is based on the creation and resolution date of work items to measure the task completion times. This interval also includes possible idle time, i.e. the time between the creation and the actual time when someone starts working on the task described by the work item. For example, an unimportant work item may be created and then ignored for a long time, until someone resolves it in a short time. As we have a very large sample size, we believe that these rare cases lose their impact on our results.

7.2. External Validity

We studied IBM’s distributed project Jazz, of which characteristics are very specific. The Jazz team members uses an integrated development environment, while self-hosting the development of the tool they are developing. Similar results may be expected if other distributed teams use an integrated development environment similar to Jazz, and follow a process similar to the Eclipse Way.

8. CONCLUSION

Does distance still matter? In our study distance does not have as strong an effect on distributed communication delay and task completion as we have seen in past research. The effect of distance is being mitigated in collaborative environments such as Jazz, in which communication in large distributed work teams is facilitated by an ability to asynchronously comment on and tracking activity of work items. The mechanism of contributing to a shared repository containing work items and associated comments from members that contributed to their implementation facilitates knowledge exchange and aids expertise seeking. These mechanisms together with processes and practices that particularly account for distributed interactions enable effective cross-site collaboration.

However, we recognize that we have made only a first step in the study of development environments that enable complex and richer than before distributed collaboration. Perhaps more important than answers are questions that our research is raising about today’s distributed collaboration:

- What are the specific practices enabled by rich collaboration environments that allow distributed groups to engage in cross-site collaboration that is as timely and effective as same-site collaboration?
- What is the interplay between asynchronous commenting on work items of interest and other communication channels that contribute to effective cross-site collaboration?
- What are the main technical features in today’s collaborative environments such as Jazz that are successfully supporting distributed collaboration?
- What brings people together in successful distributed collaboration? Is it the characteristics of the product being developed, history of working relationships or the affordances of advanced collaborative development environments?

The research community as well as the software practitioners will benefit from immediate answers to these questions.

ACKNOWLEDGEMENTS

We thank all members of the IBM Jazz project who helped us, particularly Harold Ossher of IBM TJ Watson. Many thanks to the members of the SEGAL lab. Funding was provided by NSERC and IBM Jazz Technology Fellowship Award. IBM and Jazz are trademarks of IBM Corporation in the USA, other countries, or both.

APPENDIX: NETWORK MEASURES EXPLAINED

In the following we use Figure 9 in a running example.
Figure 9. Example network to illustrate the used measures

**k-core Test**

The k-core specifies the minimum number of ties that must be present from each node of the core subgroup to the members of the core. For example, if the k is set to 5, each node of the k-core must have at least five connections to other members of the core. In Figure 9, the k-core for k = 2 consists of the nodes c_1, c_2, c_3, and c_7, as all have at least two connections to these nodes of the k-core subgroup. Note that this example has only one core. Other networks may also consist of multiple k-cores. Identifying the k-cores with k = 3 in Figure 9 will result in an empty group. There is no subgroup in which all nodes have at least three connections to the members of that group.

**Group Degree Centrality**

A centrality measure for a group of nodes in a network, Group Degree Centrality, is defined as the number of nodes outside the group that are connected to the members of the group. Note the difference to the group degree centralization, which focuses on single nodes and not on a group of nodes. The group degree centralization focuses on the degree differences of all nodes in a network, while the group degree centrality identifies how central a group of nodes is in relation to the rest of the network.

In our study we compared different groups and as such we normalized the index by the number of actors outside the group. For example, the group degree centrality of the group a) consisting of the nodes c_1, c_2, and c_7 is 2, as the group is connected to the two nodes c_3 and c_6. The normalized index is 2/4 = 0.5. Group a) is connected to 50% of the nodes outside the group.

**Density**

Density is calculated as the percentage of the existing connections to all possible connections in the network. For example, the density of the network in Figure 9 is 9/21 = 0.43.

**Group Efficiency**

The Group Efficiency index for a group provides a measure of how redundant the communication ties from members inside the group to outer members are and is defined as the fraction of the size of the minimum subgroup that has the same group degree centrality index as the whole group and the number of group members. For example, the efficiency for group a) in Figure 9 is 1/3 = 0.333, as the nodes c_1 and c_2 can be removed from the group without decreasing the group degree centrality index.

**Group Degree Centralization**

To compute the Group Degree Centralization index for the complete network we use formula (1) as proposed by Freeman (1979), in which g is the number of nodes in a network, and C_D(c_i) is the degree centrality measure of a node c_i. The degree centrality measure provides the degree for each node in a network. The degree of a node c in our communication networks is the number of its connections C_D(c). For example, the degree of c_1 in Figure 9 is C_D(c_1) = 2. C_D(c^*) is the largest node degree index for the set of contributors in the network.

\[
C_D = \frac{\sum_{i=1}^{g} [C_D(c^*) - C_D(c_i)]}{(g-1)(g-2)}
\]  

**Group Betweenness Centralization**

This measure relies on the measures of betweenness and betweenness centrality, introduced here first:

Betweenness measures the extent to which a team member is positioned on the shortest path in between other two members. For example, the interaction between nonadjacent nodes c_1 and c_4 might depend on other nodes in the communication network such as c_2 and c_3, which are on the shortest path from c_1 to c_4.

Betweenness Centrality is a probability index for the betweenness of the node in the complete network.
There is an assumption that a ‘communication’ takes the shortest path from a contributor \( c_i \) to contributor \( c_k \) and if the network has more shortest paths, all of them have the same probability to be chosen. If \( g_{kj} \) is the number of shortest paths linking two contributors, \( 1/g_{kj} \) is the probability of using one of the shortest paths for communication. Let \( g_{kj}(c_i) \) be the number of shortest paths linking two contributors that contain the contributors \( c_j \) and \( c_k \). Freeman (1979) estimates the probability that contributor \( c_i \) is between \( c_j \) and \( c_k \) by \( g_{kj}(c_i)/g_{kj} \). The betweenness index for \( c_i \) is the sum of all probabilities over all pairs of actors excluding the \( i \)th contributor. Formula (2) shows the normalized betweenness index for undirected networks.

\[
C_B(c_i) = \frac{\sum_{j<k} g_{kj}(c_i)/g_{kj}}{(g-1)(g-2)/2}
\]

(2)

The Group Betweenness Centralization index, shown in Equation (3) computes a betweenness index for a complete network instead of a single node. \( C_B(c^*) \) is the largest betweenness index of all nodes in the network. Similar to the group degree centralization, its minimum value of 0 is reached when all network nodes have the same betweenness index. Its maximum value of 1 is reached in a star network, in which the centered node is between all other nodes.

\[
C_B = \frac{\sum_{i=1}^{g} [C_B(c^*) - C_B(c_i)]}{(g-1)}
\]

(3)

**Structural Holes**

At the node level, structural holes are gaps between nodes in a social network. For example, the network in Figure 9a consisting of \( c_1, c_2, \) and \( c_3 \) has no structural holes, as all nodes are connected, while the network in Figure 9b has a structural hole because \( c_5 \) and \( c_6 \) are not connected. The structural hole measures used in our study are:

- **SH effective size** of node \( c_1 \) in Figure 9a is \( 2 - 1 = 1 \). Note that only direct neighbors of \( c_1 \) are considered. The effective size of node \( c_4 \) in Figure 9b is \( 2 - 0 = 2 \).
- **SH efficiency** of node \( c_1 \) in Figure 9a is \((2-1)/2 = 0.5 \). The efficiency of node \( c_4 \) in Figure 9b is \((2-0)/2 = 1 \).

- **SH constraint** of node \( c_1 \) and \( c_4 \) in Figure 9a is 1.125 and 0.5, respectively. The logic of SH Constraint is that if node \( c_i \)’s neighbours and potential communication partners all have one another as potential communication partners, \( c_i \) is highly constrained. Otherwise the neighbors cannot constrain \( c_i \). Constraint is higher than \( c_6 \) because its neighbors, \( c_2 \) and \( c_7 \), are connected to almost the same set of nodes, whereas \( c_6 \)’s neighbors, \( c_5 \) and \( c_7 \), are connected to a very different set of nodes. The computation, however, is not as simple to show and can be obtained in (Burt 1992).

**REFERENCES**


