2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)

Collaboration in Global Software Development: An Investigation on Research Trends and Evolution

Yang Yue*, Iftekhar Ahmed*, Yi Wang†, David Redmiles*
*Department of Informatics, University of California, Irvine, Irvine, CA
†Department of Software Engineering, Rochester Institute of Technology, Rochester, NY
y.yue@uci.edu, iftekha@uci.edu, yi.wang@rit.edu, redmiles@ics.uci.edu

***Abstract*—Global software development (GSD) done by geographically distributed teams of developers is one of the most common ways of developing software nowadays. Though GSD has various benefits, it also introduces challenges that have led to a plethora of research. This paper analyzes research papers published in top software engineering venues in recent years (2009-2018) focusing on team collaboration in order to understand the trend in GSD research. Out of 4,292 papers published in these venues, we found 33 papers that focused on team collaboration in the context of GSD. We study the kinds of data used in these papers and classify them into primary data (i.e., interview and observation data) and secondary data (i.e., repository and communication data) and found that interview data is the dominant type of data in these papers. We also found that the strength of evidence presented in most papers tends to be moderate.**

***Index Terms*—Global software development, collaboration, software engineering, empirical study**

I. Introduction

Software development has evolved from being developed by a small group of co-located developers to large groups of geographically distributed teams [5]. These global teams have to manage various aspects of software development such as conceptualization, development, and maintenance, which makes global software development (GSD) a non-trivial task. GSD also incorporates complexities resulting from temporal, geographical, cultural, and language [3] disparities. Hence, many global projects fail to meet their expectations, [1] making managing global software projects a key concern.

Researchers have been investigating various aspects of GSD including the benefits of using GSD, its limitations, and tools for mitigating the limitations, among others. Since GSD permits the complexity of development to be distributed throughout various organizations it helps to reduce the time-to-market and development costs [4]. However, such distribution of tasks also introduces challenges that have been broadly classified into five categories: team, control and coordination difficulty, loss of communication richness, loss of team spirit and cultural differences [2]. Researchers have looked into various tools and techniques for mitigating and resolving these challenges [10] as well.

Researchers have also conducted meta-studies to analyze GSD research trends [5]. However, the type and source of data used has not been systematically analyzed in the previous studies. The data source and its quality is one of the key foundations of any research and using improper data sources can impact the acceptability of the findings [12]. In this paper, our goal is to investigate the data sources and the strength of evidence provided by the papers in GSD research.

We started by conducting a literature review on the papers that investigated team collaboration in the context of GSD from 2009 to 2018. We categorized the data used in these papers into several categories, i.e. log data, interview data, and observed data. We also analyzed the strength of evidence [5] to answer the following research questions:

- **RQ1**: What are the categories of data sources used in the GSD papers related to team collaboration?
- **RQ2**: What level of strength of evidence do these publications show?

II. Study Design

A. Paper Selection

We selected papers from five venues: International Conference on Software Engineering (ICSE), Foundations on Software Engineering (FSE), International Conference on Global Software Engineering (ICGSE), Conference on Computer Supported Cooperative Work (CSCW), and International Conference on Supporting Group Work (ACM GROUP). We selected these venues due to the interdisciplinary nature of GSD research. We identified 4,292 papers published in these venues from 2009 to 2018. Then we filter full-length research papers with the words "collaboration" and "distributed software engineering" in the abstract or keyword list. We identified 33 papers using this filtering criteria.

B. Dataset analysis

We analyze the datasets used in these 33 papers. First, we categorize the data source into *Primary Data* and *Secondary Data* [11], where *Primary Data* are directly collected by researchers for the research purpose, i.e., *Interview data*, *Observation data*, and *Survey data*. While *Secondary Data* is already existing data and is adopted by researchers, i.e., *Communication data*, and *Repository data*.

Furthermore, we analyze the strength of evidence presented in these papers according to the measurement proposed by [5] which are *Weak (score 1)*, *Average (score 2)*, and *Strong (score 3)*. These strengths of evidence are calculated based on the size of the dataset, the subject of the dataset, the availability of the data, etc.

978-1-5386-9196-0/19/$31.00 ©2019 IEEE
DOI 10.1109/ICGSE.2019.00028

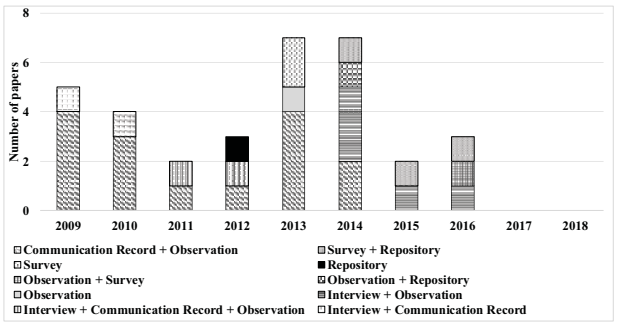

Fig. 1. Source of datasets used.

III. Results

We analyze the source of data used in the papers. Our results show that interviews are the dominant source (Fig. 1). This indicates that researchers prefer interviewing individuals from the development team directly to collect data. However, our results also indicate that in recent years more researchers have been adopting a combination of different sources of data instead of a single source, e.g., interview the individuals, but also observe them in their working context to collect observation data. We also found that only a few researchers are using repository data in their research.

As for the strength of evidence presented in most papers (Fig. 2), it tends to be moderate. For example, some papers adopted online datasets or published their datasets, which makes it easier to validate the conclusion of the research. Researchers also tend to invite more developers instead of studying a small number of developers. These efforts contribute towards strengthening the evidence of their research.

IV. Discussion

The majority of GSD researchers prefer to use *Primary Data* to conduct research [7]–[9]. Surprisingly only a few researchers used *Secondary Data* such as repository data [6] for their research. Since distributed software development teams use various tools to facilitate collaboration (i.e., email for communication, version control system for managing development), these tools generate a large amount of data and the traces of collaboration are reflected in the data. Hence, it is of utmost importance to include *Secondary Data* in the analysis in order to find interesting insights which may be missed if only *Primary Data* is analyzed.

Furthermore, the dataset used in research also affects the strength of evidence. Our findings indicate that the overall strength of evidence of recent GSD related papers is moderate. To improve the situation, researchers need to pay more attention to the data collection process. As an example, data collected by studying professional developers represents first-hand information related to practice, while the data collected by studying students is a kind of simulation. The number of subjects involved in the research also matters. The findings generated by analyzing developers from large teams spread around the world or even across various companies, tend to be more valuable and contributes to the strength of evidence.

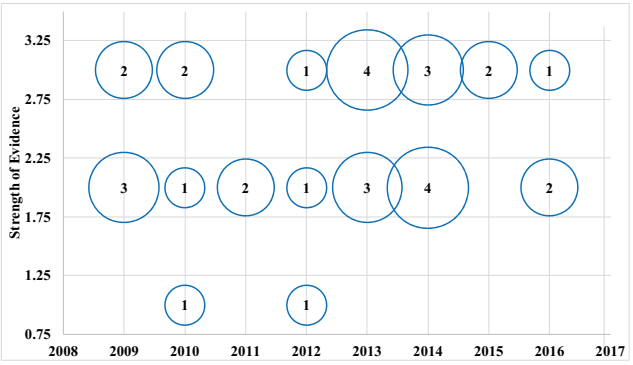

Fig. 2. Trend of strength of evidence.

V. Conclusion

In this work, we analyzed papers published in the top GSD related conferences in the last 10 years focusing on collaboration. Our analysis provides a picture of the research trend of the data used. We also found that the published work has moderate strength of evidence. Moreover, we found that in spite of the availability of rich software repository data, out of the 33 analyzed papers, only 4 used a software repository as a data source. Given our findings, we advocate for incorporating multiple data sources that are available in the software repositories for GSD related research.

References

[1] "Global 100 software leaders," https://www.pwc.com/gx/en/industries/technology/publications/global-100-software-leaders.html.

[2] E. Carmel, "Global software teams: collaborating across borders and time zones," 1999.

[3] V. Casey, "Leveraging or exploiting cultural difference?" in *2009 Fourth IEEE International Conference on Global Software Engineering*. IEEE, 2009, pp. 8–17.

[4] C. Ebert, *Global software and IT: a guide to distributed development, projects, and outsourcing*. John Wiley & Sons, 2011.

[5] C. Ebert, M. Kuhrmann, and R. Prikladnicki, "Global software engineering: evolution and trends," in *2016 IEEE 11th International Conference on Global Software Engineering (ICGSE)*. IEEE, 2016, pp. 144–153.

[6] K. Ehrlich and M. Cataldo, "All-for-one and one-for-all?: a multi-level analysis of communication patterns and individual performance in geographically distributed software development," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* ACM, 2012, pp. 945–954.

[7] H. C. Estler, M. Nordio, C. A. Furia, and B. Meyer, "Collaborative debugging," in *2013 IEEE 8th International Conference on Global Software Engineering*. IEEE, 2013, pp. 110–119.

[8] Z. U. R. Kiani, D. Smite, and A. Riaz, "Measuring awareness in cross-team collaborations–distance matters," in *2013 IEEE 8th International Conference on Global Software Engineering*. IEEE, 2013, pp. 71–79.

[9] A. Piri, T. Niinimaki, and C. Lassenius, "Descriptive analysis of fear and distrust in early phases of gsd projects," in *2009 Fourth IEEE International Conference on Global Software Engineering*. IEEE, 2009, pp. 105–114.

[10] J. Portillo-Rodriguez, A. Vizcaino, C. Ebert, and M. Piattini, "Tools to support global software development processes: a survey," in *2010 5th IEEE International Conference on Global Software Engineering*. IEEE, 2010, pp. 13–22.

[11] N. J. Salkind, *Encyclopedia of research design*. Sage, 2010, vol. 1.

[12] N. Smeeton and D. Goda, "Conducting and presenting social work research: some basic statistical considerations," *British Journal of Social Work*, vol. 33, no. 4, pp. 567–573, 2003.