JOURNAL FIRST PAPER

How are project-specific forums utilized? A study of participation, content, and sentiment in the Eclipse ecosystem

Yusuf Sulistyo Nugroho[1] · Syful Islam[2] · Keitaro Nakasai[3] · Ifraz Rehman[2] · Hideaki Hata[4] · Raula Gaikovina Kula[2] · Meiyappan Nagappan[5] · Kenichi Matsumoto[2]

Abstract

Although many software development projects have moved their developer discussion forums to generic platforms such as Stack Overflow, Eclipse has been steadfast in hosting their self-supported community forums. While recent studies show forums share similarities to generic communication channels, it is unknown how project-specific forums are utilized. In this paper, we analyze 832,058 forum threads and their linkages to four systems with 2,170 connected contributors to understand the participation, content and sentiment. Results show that Seniors are the most active participants to respond bug and non-bug-related threads in the forums (i.e., 66.1% and 45.5%), and sentiment among developers are inconsistent while knowledge sharing within Eclipse. We recommend the users to identify appropriate topics and ask in a positive procedural way when joining forums. For developers, preparing project-specific forums could be an option to bridge the communication between members. Irrespective of the popularity of Stack Overflow, we argue the benefits of using project-specific forum initiatives, such as GitHub Discussions, are needed to cultivate a community and its ecosystem.

Keywords Eclipse · Forum · Participation · Discussion · Sentiment

1 Introduction

As one of the discussion group channels, forums support mass communication and coordination among distributed software development contributors (Storey et al. 2017). Forums are considered to be an improvement over mailing lists in that it provides browse and search functions, which are especially helpful for repetitive questions and answers (Squire 2015).

Recent studies have been conducted that compare forums to communication channels that seem to have similarities to a forum. For instance, there have been studies on

✉ Yusuf Sulistyo Nugroho
   yusuf.nugroho@ums.ac.id

Extended author information available on the last page of the article.

mailing lists (Guzzi et al. 2013; Zagalsky et al. 2018), question and answer sites (Stack Overflow) (Ye et al. 2017; Wang et al. 2018; Zou et al. 2017; Calefato et al. 2018), Microblogs (Guzman et al. 2016; Mezouar et al. 2018), and News aggregators (Aniche et al. 2018). Kahani et al. analyzed discussion topics using a topic modeling technique (Kahani et al. 2016). Squire studied the transition from self-supported forums to Stack Overflow (Squire 2015). It has been reported that generic question and answer platforms such as Stack Overflow have taken over the roles of forums. For instance, many software development projects had closed their self-supported forums and moved to Stack Overflow (Squire 2015). Furthermore, gamification strategies such as awarding of badges or a voting system of Stack Overflow are considered to be incentives designed to participate and improve answer quality (Squire 2015).

From the perspective of forums, the novelty of this work is a comprehensive investigation into how a large forum in the Eclipse ecosystem is utilized. Due to the large amount of information available, we are now able to cover different facets of communication between its members. As a long-living free and libre and open-source software (FLOSS) project, the Eclipse project still maintains an established and active forum, that has also been a targeted data source for three MSR Mining Challenges.[1,2,3] The forum handles its massive ecosystem, centralizing all contributor activities within the Eclipse ecosystem, such as contributed projects, reviews in Gerrit, and topics in forums (i.e., Profile, Forums, Gerrit, Bugzilla, Projects).[4] Furthermore, Eclipse forums support donation badges, which does have a practical impact and signalling on decreasing response time of bug reports (Nakasai et al. 2018). In addition, unlike the other online question and answer platforms such as Stack Overflow, Eclipse manages their users, assigning different contributions based on status (i.e., Junior, Member, and Senior).

In this paper, we conduct an investigation the Eclipse forum in terms of participation, content and sentiment. We first perform an analysis to classify the membership status of the users. Then we set out to empirically investigate forum threads and 2,170 connected contributions to other four systems, (i.e., around 416 thousand profiles, 120 thousand code review submissions, 532 thousand bug reports with 2,883 committers from multiple projects) within the Eclipse ecosystem. To guide our empirical analyses, we formulated the following research questions:

– *RQ₁*: ***How does participation differ according to membership status?***

  Motivation: First, this study is conducted to answer the extent to which users in the ecosystem participate in (1) using forums and (2) using various systems. Second, we would like to analyze the extent to which different types of forum members in the ecosystem take to respond a particular type of post.

  Results: Senior Members are the most active forum users to participate in responding to the bug-related threads (66.1%) and the non-bug-related threads (45.5%), with active contributors in other systems are also active in forums, making it a source of expert knowledge for all systems.

– *RQ₂*: ***How do communication content vary for different membership statuses?***

  Motivation: The key motivation for *RQ₂* is to understand the message contents posted by both the organization and users, the most frequent content of posts in the

---

[1]MSRMiningChallenge2007: http://2007.msrconf.org/challenge/.

[2]MSRMiningChallenge2008: http://2008.msrconf.org/challenge/.

[3]MSRMiningChallenge2011: http://2011.msrconf.org/msr-challenge.html.

[4]Antoine Thomas, The Eclipse User Profile, http://blog.ttoine.net/en/2016/12/01/the-eclipse-user-profile/, December 1, 2016.

forums, and how users reference the information sources to support their answers. In detail, we also investigate different contents that users share in Eclipse forums.

Results: Eclipse forums are dominated by question and answer threads, especially discrepancies, amounting to 28.5%, most of which were posted by Junior Members, with replies and welcome messages being the two most frequent posts by the organization.

– *RQ₃: What are the differences perceived in the sentiment of interaction based on membership status?*

Motivation: The motivation of RQ_3 is to get deep insight on developers sentiment while sharing knowledge. Throughout this RQ, we aim to test the degree of consistency while interacting with each other.

Results: The sentiment among developers is inconsistent while knowledge sharing in the Eclipse forums.

Our results provides arguments for the need to support project-specific forums, especially for large ecosystems of projects. In summary, the main contributions of our work are as follows (i) a comprehensive study of threads in Eclipse forum, covers 832,058 threads with links to profiles, code review, bugs and project systems, (ii) an analysis of Eclipse forum membership using topological analysis, (iii) a manually labelled taxonomy of forum content communicated, (iv) a manual analysis on social interactions within an ecosystem and (v) a set of recommendations for both researchers, and practitioners on the use of forums. The rest of this paper is structured as follows. Section 2 describes our procedure to collect data through five different systems in the Eclipse ecosystem, membership classification, and an online appendix. Section 3 presents our approach to empirically analyze the users' participation, content of forums, and sentiment of interactions. The results of the study are presented in Section 4. Sections 5, 6 and 7 describe our recommendation, threats to validity and provide related works. Finally, we conclude this paper in Section 8.

2 Data Collection

In this section, we describe the procedure of our data collection, and present our replication package. Our goal is to investigate the nature of participation, content, and sentiment of interaction based on the membership in the Eclipse forums. Thus, in this study, we also present how we distinguished the statuses of members.

2.1 Eclipse Community System Collection

Figure 1 illustrates the schema that connects the five software development datasets within the Eclipse ecosystem. This connected data enables the integration of contributors' various activities. Since the Eclipse REST API requests are limited to 1,000 an hour, we collect data not only from API but also from individual systems, that is, Profile, Forums, Gerrit, Bugzilla, and Eclipse Projects.

– **Profile Dataset**. Each contributor in the Eclipse ecosystem is required to register for a profile.[5] The profile contains basic contributor information and can be used to track contributor activities.
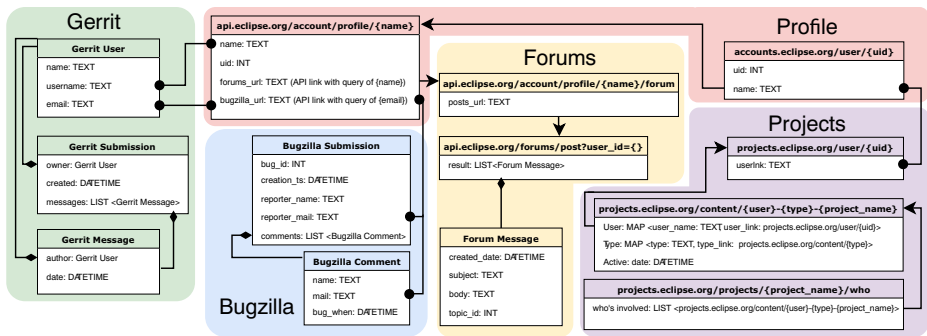
---

[5]https://accounts.eclipse.org/user/register

Fig. 1 Connected data from Gerrit, Bugzilla, Forums, and Projects via REST API in the Eclipse ecosystem

– **Bugzilla Dataset**. The Bugzilla dataset contain the tracking of bug reports and their fixes within the Eclipse ecosystem.[6]
– **Projects Dataset**. To track social roles (committer, review, mentor, etc.) of contributors, we include the project dataset.[7]
– **Gerrit Dataset**. Gerrit is a review tool that facilitates collaboration between committers and contributors within Eclipse.
– **Forum Threads**. Eclipse community forums, a user-to-user interaction site for Eclipse users, have a hierarchical structure. It contains a number of sub-forum categories which may have several topics. Within a forum's problem-related topic, each new initial post by a user can have a response from other users in the community.

The participation of contributors within the Eclipse ecosystem can be analyzed from the connected data of five datasets, as described in Fig. 1. As presented in Table 1, the data used in this analysis are extracted through several steps.

Step 1: In this step, we extracted the dataset from five different sources, that are, Profile, Bugzilla, Projects, Gerrit, and Forums. For Profile dataset extraction, 416,150 *uid*s and *name*s were collected on September 14, 2018 from the profile system.[8] Using the Profile API, 416,126 Profile data were obtained (24 users were not found) from November 5 to 13, 2018. To connect the activities in Bugzilla with the Profile dataset, we used the *email* addresses. We extracted 531,752 bug reports from October 10, 2001 to January 13, 2019. From the Projects data source, we then collected a total of 2,942 committers from the obtained 438 projects (on January 16, 2019). Among them, the information of 59 committers could not be accessed because of the deletion of accounts. Hence, the remaining 2,883 committers can be connected with the Profile dataset through the *uid*. We collected 120,165 Gerrit submissions from October 1, 2009 to October 31, 2018. We link users with the Profile dataset, using the *name* or *email* to match. To download the forum threads, we extract the data from its API. We downloaded 1,097,174 threads (topics) available in the Eclipse community forums until January 9, 2019.[9]

---

[6]https://bugs.eclipse.org/bugs/

[7]A detailed list of Eclipse projects is available from https://projects.eclipse.org/.

[8]https://accounts.eclipse.org/user/

[9]https://www.eclipse.org/forums/index.php/t/

Table 1 Connected data extraction from five data sources within the Eclipse ecosystem

Step	Dataset	Quantity	
Step 1:	Profile	416,126	Profiles
	Bugzilla	531,752	Bug reports
	Projects	2,883	Committers
	Gerrit	120,165	Submissions
	Forums	1,097,174	Threads
Step 2:	Connected Gerrit contributors	2,170	Contributors
	Connected forum messages	467	Messages

Step 2: The connected forum data are collected in this step. On November 23, 2018, we extracted forum message data from Gerrit contributors, via the Eclipse REST API. As part of the data collection process, we applied a pre-processing to detect and remove duplicated accounts. We report 39 pairs of Gerrit and Eclipse forum accounts that have same *name*s but different *username*s. Through manual examination, we verified identities within those pairs; 27 pairs were found to be identical and merged. After the duplicate removal and linking to Profile data, we obtained 2,170 Gerrit contributors and extracted 467 forum messages that connect to these Gerrit contributors.

The collected 2,170 contributors from this extraction are then used to investigate the participation of contributors within the Eclipse ecosystem using a topological data analysis technique (i.e., TDA). TDA is a method to extract valuable information from high-dimensional data that is insensitive to metric, noisy, and incomplete without formulating an initial hypothesis (Lum et al. 2013). To build the topology, we construct the metrics of each contributor from the four data sources, namely, Bugzilla, Gerrit, Forums, and Projects (see the details in Section 3.1). The metrics describe the connected activities of the contributors within the Eclipse ecosystem.

Figure 2 depicts different elements of a thread in the Eclipse forums, including (i) the topic category and subcategory, (ii) the member status, (iii) date of creation, and (iv) the link that was posted in the thread.

To improve the quality of our dataset, as shown in Table 2, the forum threads underwent three stages of filtering.

1. **Step 1**. Raw extraction. In this step, we use all 1,097,174 collected threads yielded from *Step 1* in Table 1.
2. **Step 2**. Remove duplication. In the Eclipse community forums, we found that some threads had been duplicated by the system several times. To avoid redundancies, we removed such threads based on the identity number of initial posts. We were able to reduce the number of threads to 832,058.
3. **Step 3**. Separation. To investigate how the organization communicates with the members and analyze the characteristics of forum usages, contents, etc., we separate those threads posted by the webmasters and the non-webmaster users that include the threads from users that contribute in the other systems. Webmasters deal with the servers and software that runs the eclipse.org site. [10] Webmasters have different roles in the forum

---

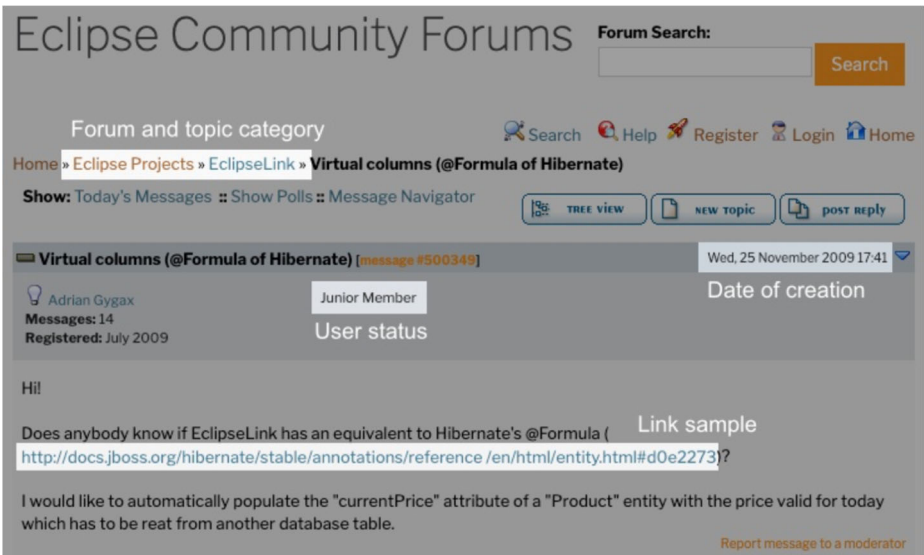[10]https://wiki.eclipse.org/WebMaster

Fig. 2 Example of a forum thread

than other users, so they need to be classified. Also, by analyzing the content of web-masters' posts, it is possible to understand how the organization is engaged in the forum. From the total number of threads resulted in *Step 2*, we were able to separate (i) 542,997 threads posted by the webmasters and (ii) 289,061 threads posted by non-webmaster users.

2.2 Membership Classification

This section describes our techniques to classify the membership of users for each posted message. Unlike the other question and answer online forums such as Stack Overflow, in the Eclipse community forum, all registered users are assigned into three statuses of membership, that are, (1) Junior, (2) Member, and (3) Senior. These user statuses are included in our collected data resulted from *Step 3* in Table 2 which can be seen in every post of a user, as shown in Fig. 2. The status of each user may change from the lowest level (i.e. Junior) into the highest one (i.e. Senior) depending on the contributions of the user in the community. However, in the forum, we could not differentiate which posts were posted by users when they were a Junior, Member or Senior. This is because once the status of a user has changed, it will replace the old status in all posts of a user with the latest status, including

Table 2 Outputs of the pre-processing of the forum dataset

Step		# Threads
Step 1:	Raw extraction	1,097,174
Step 2:	Remove duplication	832,058
Step 3:	Separation:	
	(i) Threads by webmaster	542,997
	(ii) Threads by non-webmaster users	289,061

their first posts. Furthermore, we also did not find any information about the time when the status of a user changed.

To define the member status of each registered user, we attempted to calculate the total number of posts of every user. From this amount of posts, we summarized the quantity of posts per author based on the user identity number. The total number of posts per author varies, from less than ten to more than one thousand posts. In this step, we found the maximum number of posts of each user if we consider the latest status of users as collected in the dataset, as shown in Fig. 3. The maximum number of posts by Juniors and Members are 29 and 106 respectively, while the maximum number of posts by Seniors is 28,476. Based on this finding, we used these maximum numbers as the thresholds to differentiate the user status for each post based on the sequence number of a post. The sequence number of a post depends on its creation date in order. The earliest post is assigned as the first post, then followed by the other posts ordered by date of creation.

Users' classification As shown in Table 3, we are able to assign the status of the users in all posts based on the sequence number of the posts and able to distinguish the posts that were posted by the users when they were Junior, Member and Senior. The posts that are assigned from 1 to 29 are considered as posts that were posted by Juniors, while posts from number 30 to 106 were posted by Members, and Seniors posted the posts from 107 up to the remaining posts.

In the Eclipse forums, we also found that there are a lot of posts posted by authors with the username "Eclipse User", accounting for 30% of the forum users, as described in Table 3. In the threads, this username is automatically assigned by the system and could be from anyone. It does not show the status explicitly whether Junior, Member, or Senior. Furthermore, it does not have a user identity number and a link that connects to the user's profile. Since we only focus to identify the three statuses of the user in the analyses, the messages that were posted by "Eclipse User" were excluded. After the exclusion process, we obtained 202,602 threads that were only posted by Juniors, Members and Seniors.
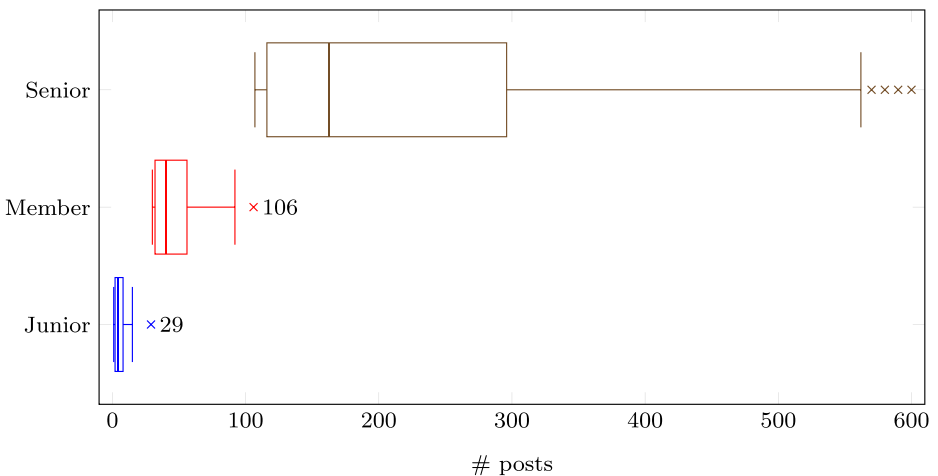

Fig. 3 Frequency of posts per user status. The maximum number of posts for each type of users is used to define the threshold of post-based membership. The threshold for Juniors and Members are 29 and 106, respectively. Although Seniors have posted more than 28 thousands posts, we limit up to 600 in the figure

Table 3 Frequency of posted threads based on the classification results

Status	# Threads	(%)
Junior Member	144,440	(50%)
Member	31,719	(11%)
Senior Member	26,443	(9%)
Eclipse User	86,459	(30%)
Sum	**289,061**	**(100%)**

Message distribution Figure 4 shows the contributions of forum members in the Eclipse forums based on the memberships of our classification. Even though the number of Junior Members is higher, they had posted a smaller quantity of posts, compared with the two other user categories. In contrast, although the total number of Senior Members is lower than the other forum members, most Seniors have contributed in the forum multiple times.

2.3 Online Appendix

Our online appendix contains five data files used in this study: (i) 289,061 non-webmaster threads associated with the information of thread URL, message identity number, forum name and topic, thread title, name of initial-post user, user identity number, and member status; (ii) 216,864 links with the description of thread url, thread title, user's name, extracted link, the link with first directory, domain name, and top level domain; (iii) 2,170 contributors with the metric values described in Table 4; (iv) 1,149 samples of the annotated threads to answer the types of contents; and (v) 1,149 pairs of initial post and first response that are labeled based on the polarity and interaction sentiment. The appendixes
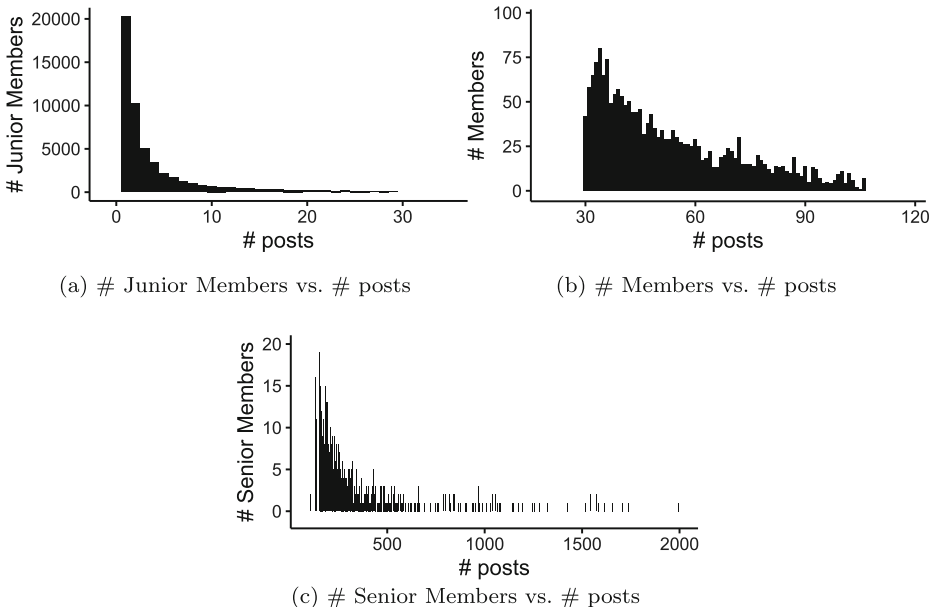

(a) # Junior Members vs. # posts

(b) # Members vs. # posts

(c) # Senior Members vs. # posts

Fig. 4 Distribution of posts per user category after classification

Table 4 Contributor metrics

Metric	Description	System
bug_subm	# of submitted bug reports	Bugzilla
bug_comm	# of bug reports to which only are added comments	Bugzilla
review_subm	# of submitted patches for review	Gerrit
review_comm	# of reviews to which only are added comments	Gerrit
thread	# of threads participated in	Forums
committer	# of committer roles	Projects

are available at https://doi.org/10.5281/zenodo.4451766 and https://github.com/yusufsn/EclipseForumData.

3 Methods

In this section, we describe our approach to answer the three research questions.

3.1 RQ$_1$: How Does Participation Differ According to Membership Status?

Before we study the participation of users' based on their membership statuses, how contributors participate to each system (i.e., gerrit, bugzilla, projects and forums) were analyzed. We would like to understand how many of the contributors are using multiple systems, and whether or not they are forum users.

Shape of contributors' participation To show how contributors use the multiple systems we adopt a topological data analysis technique (i.e., TDA) (Lum et al. 2013). TDA is a topological technique to extract valuable information from high-dimensional data that is insensitive to metric, noisy, and incomplete without formulating an initial hypothesis. This technique applies (algebraic) topology and computational geometry to create a shape of a high-dimensional dataset (Lum et al. 2013).

TDA has been applied in various fields such as an analysis of bacteria survival in soils (Ibekwe et al. 2014) and the performance of NBA players (Lum et al. 2013). Most recently in software engineering, Lertwittayatrai et al. (2017) used TDA to distinguish characteristics of JavaScript npm libraries (such as licence usage, dependency usage and so on) and Tantisuwankul et al. (2019) to analyze the implementation of such communication channels over open source software projects. We use TDA to provide a visual representation that shows (i) the activity levels for each feature and (ii) show the contributor activity across systems, especially with respect to forums.

As shown in Table 4, we first prepare a contributor metrics data that will be used as features for the TDA mapper. This data contains 6 metrics that pertain to the different activities for each system from the 2,170 Gerrit contributors described in Section 2.1. We use the Knotter tool,[11] which is an implementation of mapper algorithm and the t-Distributed Stochastic Neighbor Embedding (t-SNE)(Van der Maaten and Hinton 2008), a technique

---

[11]https://github.com/rosinality/knotter

for dimensional reduction and clustering, and our defined features as the filters for the visualization construction.

For evaluation, we present the typologies of the four systems, highlighting the most active contributors. To show this relationship in terms of their contributions to the forum, we will annotate the most active contributors of forums.

Membership-based participation   To study how users participate in the forums based on their statuses, we divided the type of threads into two categories, (i) bug-related threads, and (ii) non-bug-related threads. In this analysis, we targeted all 202,602 threads that were posted by the registered users (after excluding "Eclipse User" from the dataset), as shown in Table 3. To distinguish the category of threads, we filtered the threads using the Eclipse bug-report URL as the keyword, that is, bugs.eclipse.org. If the threads contain at least one keyword of the bug-report URL, the threads are categorized as bug-related threads, otherwise, non-bug-related threads. Since we investigate how developers respond to bug-related and non-bug-related threads, the threads without any response were excluded. After this process, we subsequently investigated 128,698 bug-related threads, and 25,052 non-bug-related threads.

For both categories of threads, we separated the targeted threads into three classes based on the status of the askers (initial poster), that is, Juniors' threads, Members' threads, and Seniors' threads. In each thread, we then classified the response messages into the same classes as the askers to investigate the user types of the answers. To reduce bias, the answers from the same authors who ask the questions in the threads were excluded. We also removed the response messages with duplicate identification numbers of the answers in each thread. Finally, we quantified the frequency of the answers grouped by their membership statuses.

In addition, to statistically validate our results, we applied Pearson's chi-squared test (χ^2) (Pearson 1992) with the null hypothesis "*bug and non-bug related threads have independent participation by all members of different statuses.*" The Pearson chi-square statistical test is commonly used for testing relationships between one or more categorical variables.

3.2 RQ$_2$: How do Communication Content Vary for Different Membership Statuses?

This analysis includes the investigation of message patterns posted by the organization, the hottest topic of threads discussed in the forum, and the forum linkage to the external sources.

Threads by organization   In the Eclipse community forums, a large number of posts were posted by the organization that deals with the servers and software that runs the eclipse.org site.[12] When the organization posts a message in the forums, they employ the same identity, that is "Eclipse Webmaster". To understand how the Eclipse Webmaster manages its community forums, we used 542,997 messages collected in Section 2.1. Since we found there are a large number of duplicated messages in our collected dataset (e.g. the closure of forum listed in the topic name[13] [14]), we removed them. After the removal of these duplicate messages, there were 934 messages with different contents. For these messages, we found some frequent patterns. Since the messages were unstructured and did not contain explicit keywords, we manually examined the contents of the 934 messages and classified them into 9 patterns.

---

[12]https://wiki.eclipse.org/WebMaster

[13]https://www.eclipse.org/forums/index.php/f/102/

[14]https://www.eclipse.org/forums/index.php/sp/9957/

Forums and topic categories To identify the forums and topics that are mostly discussed by the users, we quantitatively analyzed the messages based on their forum names and topic categories that are embedded in the collected threads. We investigated 289,061 threads that were posted by the non-webmaster users, as presented in Table 2.

Forums linkage To understand how forum links to the external sources in the Eclipse ecosystem, we manually analyzed a representative sample of links. The link targets are extracted from 289,061 threads posted by the users (see Table 2). We first prepared a statistically representative sample from 216,864 collected links by computing a random sample of data with a confidence level of 95% and a confidence interval of 5.[15] The calculation of the sample size yields 383 links.

To extract the usage of the link targets, we performed a manual labeling to all link targets from our sample (similar to Hata et al. (2019)). At this stage, the authors of this paper specified the code for categorizing the usage of the links. The initial codes used to categorize the links were imported from the study by Hata et al. (2019). However, we dropped several labels because they are unrelated to our research. We also combined and adjusted some codes to make them appropriate to our study. Similar to other works in classifying the link targets (Hata et al. 2019), four authors of this paper (the first, third, fifth, and the sixth authors) then coded the first 30 links from the representative sample independently using the designed codes. The kappa agreement from the four raters is 0.81 or 'almost perfect' (Viera and Garrett 2005). Based on this encouraged agreement, the remaining link targets were then coded by the first author.

The codes used to characterize the link targets including the description are as follows:

– *404*: the link target cannot be accessed or missed
– *bug report or Bugzilla*: a specific bug report or a Bugzilla top page
– *other documentation*: documentation of a product or project in general except for API documentation
– *personal or organization homepage*: a web page of an individual or organization
– *product or project homepage*: a web page of a product or project
– *API documentation*: specific documentation of an API component
– *tutorial or article*: a tutorial or technical article without comments
– *thread*: thread in the forums
– *blog post*: informational website that displays postings by one or more individuals and usually has commenting section
– *release*: a web page informing the release of new files, new versions of a software, new packages, etc.
– *code*: a web page of a source code file
– *book or research paper*: a web page of a book or entire book or academic paper
– *licence*: licence of a software project
– *other*: anything that does not fit the other labels, or a web page requiring sign-in. For example, a link that directs to a non-English web page http://www.vsa.de/marvin/dsl/mask/0.1 referenced in a forum content. [16]

Membership-based content Our approach to categorize the content based on the user's statuses is through manual analysis of a representative sample. Since this analysis only focus

---

[15] https://www.surveysystem.com/sscalc.htm

[16] https://www.eclipse.org/forums/index.php/t/627844/

on the threads that were initially posted by the three types of registered users (i.e. Juniors, Members and Seniors), we excluded all threads that were initially posted by "Eclipse Users". Similar to the procedure of forum linkage analysis, the statistically representative samples from the 202,602 threads were prepared. The calculation of the sample size yields 379, 380, and 383 threads for Senior Member, Member, and Junior Member from a total of 202,602 threads.

To extract the content types, we conducted an interactive process of coding. In this process, three authors of this paper (i.e. the first, fourth, and fifth authors) firstly discussed the initial coding guide from previous work (Beyer et al. 2020). To make the coding guide fit with our study, the initial coding guide were adjusted and the other codes were added for labeling the threads. Similar to the approach from prior works (Abdellatif et al. 2020; Hata et al. 2019), the authors subsequently coded 30 initial posts of a thread independently for each type of askers in the sample using the designed labels. The kappa scores for each status of askers were calculated to see the level of agreement between three authors.[17] We gained 0.76 and 0.72 for both Juniors and Members which mean 'acceptable', and 0.82 for Seniors which indicates 'almost perfect' (Viera and Garrett 2005). Based on these motivated agreement scores, the coding task for the remaining main posts from the sample were undertaken only by the first author. The following terms list shows the labels used in our analysis to code the threads discussed including the descriptions:

1. **Question-and-Answer threads**: if the thread is triggered by a question for asking a solution, reason, advice, clarification, or guidance even if the questions are unanswered, then we used the following coding guide:

   – *Errors*: questions about specific errors, including error messages, exceptions, the stack trace, etc. Beyer et al. (2020).
   – *Review*: this category merges the categories Decision Help and Review (Treude et al. 2011), the category Better Solution (Beyer and Pinzger 2014), and What (Rosen and Shihab 2016), as well as How or Why something works (Allamanis and Sutton 2013). Questioners of these posts ask for better solutions or reviewing of their **code snippets**. Often, they also ask for best practice approaches or ask for help to make decisions, for instance, which API to select.
   – *Plans*: asking for the future plans or processes of the development.
   – *Learning*: this category merges the categories Learning a Language or Technology (Allamanis and Sutton 2013) and Tutorials or Documentation (Beyer et al. 2017). In these posts, the **questioners ask for documentation, tutorials, or examples** to learn a tool or language. In contrast to the first category, they do not aim at asking for a solution or instructions on how to do something. Instead, they aim at asking for support to learn on their own.
   – *Usage*: this category subsumes questions of the types **How to implement something** and Way of using something (Allamanis and Sutton 2013), as well as the category How-to (Beyer and Pinzger 2014; Treude et al. 2011), and the Interaction of API classes (Beyer et al. 2017). Specific functionalities are mentioned in the question. The questioner is asking for concrete instructions.
   – *Versions*: this question category is equivalent to the categories Version (Beyer and Pinzger 2014) and API Changes (Beyer et al. 2017). The questions arise due to the

---

[17]http://justusrandolph.net/kappa/

changes in an API or due to compatibility issues between different **versions** of an API. Specific versions are mentioned in the questions.

– *Conceptual*: The posts of this category are equivalent to the category Conceptual (Treude et al. 2011) and subsumes the categories **Why...?** and **Is it possible...?** (Beyer and Pinzger 2014). Furthermore, it merges the categories **What** (Rosen and Shihab 2016) and **How/Why something works** (Allamanis and Sutton 2013). The posts consist of questions about the limitations of an API and API behaviour, as well as about understanding concepts, such as design patterns or architectural styles, and background information about some API functionality.

– *Discrepancy*: the posts that are not questions of particular errors, but **no clue how to solve** questions (Beyer et al. 2020).

2. **Non-Question-and-Answer threads**: if the thread is triggered by a non-question post, we categorized the topics using the following labels:

– *Misuse*: a post is identified out of the scope, unrelated to the community forum, or difficult to understand.

– *Announcement*: a post provides an announcement from system or core developers about specific events (e.g. future updates, file release).

– *Information*: a post provides a general information. It could be from anyone.

– *Recruitment*: indicates an offer of a job vacancy or recruiting people.

– *Test*: a post is used for a test.

– *Other*: anything that does not fit the above labels, including posts that respond to a post from different threads (the post is not an original question from the asker).

To validate the results statistically, we then apply Pearson's chi-squared test (χ^2) (Pearson 1992) with the null hypothesis "*the communicated content and the members of different statuses are independent.*"

3.3 RQ₃: What are the Differences Perceived in the Sentiment of Interaction Based on Membership Status?

Our approach to answer RQ₃ is through manual analysis of a representative sample to understand the sentiment of forum users based on their member statuses while sharing knowledge. To analyze the sentiment, we only focus on the initial posts of any threads that were posted by three types of users (i.e., Juniors, Members and Seniors) and their first replies. This is because the first reply in each thread shows the tangible feeling of a user reaction in responding to the initial posts. Similar to Section 3.2, we excluded all the initial posts from "Eclipse User". We then prepared statistically a representative sample for each users type with 95% confidence interval.[15] The calculation of the sample size yields 379, 380, and 383 pairs of initial post and first response for Senior Member, Member, and Junior Member, respectively.

To find the types of sentiment among users, we conducted an interactive process of coding. In this process, three authors of this paper first discussed the initial coding guide from previous work (Kauffeld and Lehmann-Willenbrock 2012). To make the coding guide fit with our study, the initial coding guides were adjusted. We (first, second, and sixth authors) then independently applied the adjusted coding rules on both initial posts and the first responses of the first 30 samples of thread. We used the kappa score calculator to check the agreement level and find the score 0.78.[17] According to Viera and Garrett (2005), this kappa

agreement score is 'acceptable'. Based on the agreement level, the coding tasks for the remaining threads from samples were undertaken by the three authors, which each author classified 383 different samples of threads.

Our sentiment analysis is composed of two different analyses. First, we investigated the polarity more generally. Second, we performed a deeper analysis of the contents to characterize the social interaction. The following terms list shows the labels used in our analysis to code the type of polarity and interactions while sharing knowledge through forums.

1. **Polarity analysis**: To analyze the polarity of messages, we used the following three labels:

   – *Positive*: It includes posts that has positive feeling while reading and also include positive words (i.e., good, nice, working example, thanks a lot etc). For example:

     Thanks, you have been a great help. It is enough for me.

   – *Neutral*: It includes posts that has neutral feeling (e.g. started with positive then ended with negative talks) while reading or do not include any biasing words (i.e., positive, negative). For example:

     See the Task List's view menu's "Show UI Legend" action for an explanation of the colors and a link to how to change them.

   – *Negative*: It includes posts that has negative feeling while reading and also include negative words (i.e., error, bug, not working etc). For example:

     I've added some custom key binds in M8 and sometimes they don't work. They still show up in the preferences dialog, but they are not indicated on the menu items they are assigned to. Closing Eclipse and restarting fixed the problem. Anyone else seen this? I don't know of any repro steps at this point.

2. **Interaction analysis**: To analyze the interactions between forum members in the threads, we used the following seven statements:

   – *Positive Procedural (PPc)*: It includes procedural statements that are goal oriented (i.e., pointing out or leading back to the topic), procedural suggestion (i.e, suggestions for further procedure), procedural questions (i.e., questions about further procedure), economical thinking (i.e, weighing costs/benefits), etc.
   – *Positive Socioemotional (PS)*: It includes socioemotional talks like encouraging participation, agreeing suggestions, offering praise, etc.
   – *Positive Proactive (PPa)*: It includes proactive statements that discusses interesting ideas, planning actions, agreeing upon tasks to be carried out, etc.
   – *Neutral*: It includes statements that can't fall in the above categories or posts that are neutral in nature.
   – *Negative Procedural (NP)*: It includes procedural statements that talks more about failure of procedure, complaining the procedure (i.e., behaviour of API, code), irrelevant things, etc.
   – *Negative Socioemotional (NS)*: It includes socioemotional talks like criticizing, disparaging comments about others, etc.
   – *Negative Counteractive (NC)*: It includes counteractive statements that cover irrelevant problems, no action plan, terminating discussions, etc.

Similar to the statistical validation of the results in RQ_1 and RQ_2, Pearson's chi-squared test (χ^2) (Pearson 1992) is applied with the null hypothesis "*the sentiment of interactions and the members of different statuses are independent.*"

4 Results

In this section, we present the answer to the research questions and describe the results.

4.1 RQ$_1$: How Does Participation Differ According to Membership Status?

Figure 5 shows the topology of active Eclipse contributors in all other systems. Each node in the visualization represents similar sets of contributors. In general, the map is read as follows:

1. *Topology cluster* - a cluster of nodes represent contributors that shared similar features (i.e., share similar contributions per systems). Hence, closely clustered nodes indicate these contributors share the same attributes.
2. *Topology color activity* - the color represent the density of each feature. Starting from red to blue, the red color (i.e., red=low activity) indicates a low activities of the feature, while green to blue color represented high activities (i.e., blue=high activity). For example, with the review comments feature, contributors that contributed many reviews to the Gerrit system were clustered in the green nodes, while those contributors with almost no activity are assigned red color.

Shape of contributors' participation Figure 5 shows that the Eclipse contributors with high activities in all other systems are more active in the forums. As shown in Fig. 5b and c, Eclipse contributors seem to work on Gerrit and Bugzilla systems to report bug and submit review. This is because permission may be required for submission and committing of patches. Hence it is not easily accessible. Another interesting observation is that forums are used by the more socially active contributors. This is illustrated in Fig. 5d that those who are more active in Eclipse forums also participate more frequently in Eclipse projects as committers, as indicated in the white area of Fig. 5b and c. This provides evidence that the forum is a source where expert knowledge can be shared.

> **Summary**: We find that active contributors are also active in forums, making it a source of expert knowledge for all systems.

Membership-based participation As shown in Table 5, Seniors seem to be the most active forum users in responding any category of threads, 66.1% in bug-related threads and 45.5% in non-bug-related threads, followed by Juniors and Members respectively. In the category of non-bug-related threads, Juniors mostly responded to Juniors' initial posts, and Seniors mostly responded to Seniors' initial posts, while Members frequently responded to Juniors' initial posts. In both categories of threads, Juniors and Members also replied to the Seniors. It indicates that the communication between forum users is not limited by the type of users.

For the statistical evaluation, we find that there is an association between the bug/non-bug related threads and the members of different status. Our null hypothesis on "*bug and*
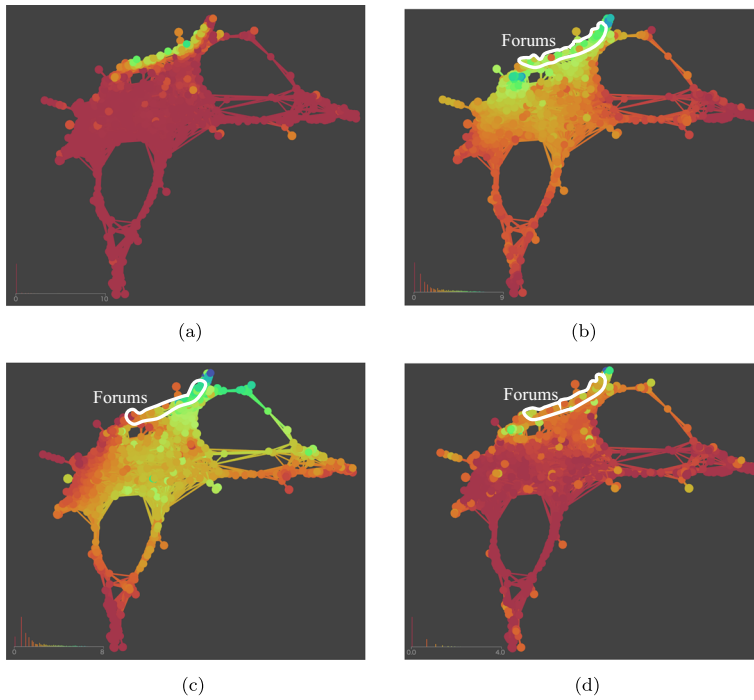
(a)   (b)

(c)   (d)

Fig. 5 The topology shows that Eclipse contributors with high activities in all systems (i.e., (a) Forums, (b) Bugzilla, (c) Gerrit, (d) Projects) are active in forums. Note that the metrics are taken from Table 4. We draw the white area manually to represent the contributors that actively participate in the forums

non-bug related threads have independent participation by all members of different statuses" is rejected (i.e., p-value is < 0.001).

Summary: Compared to the Juniors and Members, Seniors are the most active forum users in responding any category of threads. For bug related threads, Seniors are contributing way more than the Juniors and Members. For non-bug related threads, Junior users responded more to Juniors' initial posts, and Senior users responded more to Seniors' initial posts.

Table 5 Frequency of developers responding bug-related and non-bug-related threads

Answerers	Askers	Bug-related Threads (%)			Sum	Non-bug-related Threads (%)			Sum
		Juniors	Members	Seniors	(%)	Juniors	Members	Seniors	(%)
Juniors		15.4	2.7	1.3	19.4	30.3	3.9	5.2	39.3
Members		11.7	1.7	1.0	14.5	6.9	5.0	3.2	15.2
Seniors		51.1	10.6	4.4	66.1	19.2	5.8	20.5	45.5
Sum		78.8	15.1	6.8	**100.0**	56.4	14.7	28.9	**100.0**

4.2 RQ$_2$: How do Communication Content Vary for Different Membership Statuses?

Threads by organization Table 6 shows the message pattern posted by the Eclipse web-master. We see that the frequency of the reply and welcome messages have surpassed other classes. The Re: messages are the most posted message patterns by the webmaster to indicate the responses to previous messages. The Eclipse webmaster was also frequently sending a welcome message to the newly registered members in a particular community discussions group. This type of message is used to provide a brief statement of the purpose of the newsgroups.

Despite the occurrences of the closure messages are not as frequent as the top 2 patterns, but they were posted by the organization occasionally to inform the contributors that the related forums had been closed. The other pattern of messages that have a number of messages posted in the forum are test and archive, as many as 14 and 13 message contents, followed by guideline, announce and move, with 6, 5 and 4 messages, respectively. Finally, the prevalence of the pattern "other" in the findings is an indicator of the various message patterns in the threads posted by the Eclipse Webmaster.

Forum and topic categories Table 7 shows that the Eclipse platform is the most common topic discussed by Eclipse users. This may imply that the threads are more generic, rather than being very specific. The next topic is followed by BIRT, i.e., an open source software project to create data visualizations and reports. Both topics are in the forum of Eclipse projects. In our statistical records, Newcomers has become the third most topic in the Eclipse forum. This conversation subject is used dominantly by those who are new in the Eclipse community. The other prevalent topics are EMF of Modeling, C/C++ IDE (CDT) and Java Development Tools (JDT) of Language IDEs, to complete the top six topics discussed in the Eclipse community forums.

Forums linkage Table 8 shows the result of our qualitative analysis. From Table 8, we see that a bug-related link target from online issue trackers, such as Bugzilla, is the most links inserted in the messages as a supplement to the answer, accounting for 15%. It indicates that facing problems or finding a software defect was frequently reported by the users in the

Table 6 Webmaster's message patterns

Pattern	Description (keywords)	#
Re:	A reply for a previous message ("*Re:*")	538
Welcome	A greeting to a newcomers ("*Welcome to*")	270
Closure	Notification of closing threads ("*closure*", "*forum closed*")	25
Test	A test message ("*test*", empty message)	14
Archive	Notification of archiving threads ("*been archived*")	13
Guideline	Rules, intention, or suggestions for users ("*posting guidelines*", "*please read before posting*")	6
Announce	Notification of specific events or internal conditions ("*outage*", "*we're hiring!*", "*available*", "*shutting down*")	5
Move	Notification of moving threads ("*this group has moved to*", "*forum move*")	4
Other	A message that does not fit the above	59
Sum		**934**

Table 7 Top 10 forum and topic categories in the Eclipse community forums

Forum	Topic	# threads
Eclipse Projects	Eclipse Platform	31,795
	BIRT	29,019
Newcomers	Newcomers	23,054
Modeling	EMF	16,171
Language IDEs	C / C++ IDE (CDT)	13,449
	Java Development Tools (JDT)	13,193
Eclipse Projects	Standard Widget Toolkit (SWT)	11,307
	Rich Client Platform (RCP)	10,842
Modeling	TMF (Xtext)	9,909
	GMF (Graphical Modeling Framework)	8,484

forum. Other documentations are also prevalent, nearly the same as personal or organizational homepages, accounting for 11% and 10% respectively. The common use of the label 'other' in the link target types represents the heterogeneous of links present in the Eclipse forum. Lastly, we also see that 28% of link targets referenced by the Eclipse forum members are currently inaccessible (i.e. 404).

Frequently linked resources Table 9 shows that the top 10 most prevalent referenced domains from the collected 216,864 links. We see that the Eclipse organizational homepage is the most popular link to refer to, which is separated into four most common directories: Forums, and some Eclipse projects (Modeling, EMF, and BIRT). Bugzilla, a web-based

Table 8 Frequency of link target types in our sample

Availability	Target	# Links	(%)
Available			(72%)
	Bug report/Bugzilla	58	(15%)
	Other documentation	41	(11%)
	Personal/organizational homepage	38	(10%)
	Product/project homepage	28	(7%)
	API documentation	20	(5%)
	Tutorial or article	16	(4%)
	Thread	15	(4%)
	Blog post	10	(3%)
	Release	8	(2%)
	Code	4	(1%)
	Book or research paper	1	(0%)
	Licence	1	(0%)
	Other	34	(9%)
Not available			(28%)
	404	109	(28%)
Sum		**383**	**(100%)**

Table 9 Frequently referenced domains in the Eclipse community forums

Domain		Description	# Links	
www.eclipse.org			39,917	
	/forums/	Eclipse Community Forums		(5,723)
	/modeling/	Eclipse Modeling Project		(4,076)
	/emf/	Eclipse Modeling Framework (EMF) (moved to under the above /modeling/)		(3,119)
	/birt/	Eclipse BIRT (Business Intelligence and Reporting Tools) Project		(2,391)
bugs.eclipse.org		Eclipse Bugzilla	30,135	
wiki.eclipse.org		Eclipse wiki pages	22,920	
download.eclipse.org		download page for Eclipse product	9,343	
dev.eclipse.org		(currently not available or redirected to wiki.eclipse.org/Development_Resources)	9,297	
	/viewcvs/	Eclipse CVS repositories (not available)		(4,528)
	/newslists/	Eclipse news bulletin (not available)		(2,533)
xtext.itemis.com		Xtext framework for programming language development	3,459	
www.w3.org		international standard organization of world wide web	3,323	
github.com		web-based hosting service for version control using Git	3,298	
twitter.com		an online social networking site	3,096	
git.eclipse.org		Eclipse Git repositories	2,123	

Table 10 Frequency of content types in our sample (gray color represents question-and-answer threads, white color represents non-question-and-answer threads)

Topic	Juniors		Members		Seniors		Total	
	#	%	#	%	#	%	#	%
Discrepancy	153	(39.9%)	114	(30.0%)	58	(15.3%)	325	(28.5%)
Review	51	(13.3%)	63	(16.6%)	39	(10.2%)	153	(13.4%)
Conceptual	31	(8.1%)	50	(13.1%)	62	(16.2%)	143	(12.4%)
Usage	45	(11.7%)	36	(9.4%)	44	(11.6%)	125	(10.9%)
Errors	49	(12.8%)	41	(10.7%)	31	(8.2%)	121	(10.6%)
Information	6	(1.6%)	17	(4.4%)	45	(11.9%)	68	(6.0%)
Learning	19	(5.0%)	19	(5.0%)	13	(3.4%)	51	(4.4%)
Versions	9	(2.3%)	22	(5.7%)	18	(4.7%)	49	(4.3%)
Announcement	3	(0.8%)	5	(1.3%)	22	(5.7%)	30	(2.6%)
Plans	8	(2.1%)	4	(1.0%)	9	(2.3%)	21	(1.8%)
Misuse	1	(0.3%)	1	(0.3%)	6	(1.6%)	8	(0.7%)
Recruitment	2	(0.5%)	1	(0.3%)	2	(0.5%)	5	(0.4%)
Test	0	(0.0%)	2	(0.5%)	3	(0.8%)	5	(0.4%)
Other	6	(1.6%)	5	(1.3%)	27	(7.0%)	38	(3.3%)
Sum	**38**	**(100%)**	**380**	**(100%)**	**379**	**(100%)**	**1,142**	**(100%)**

general-purpose bug tracker, is the second most prevalent domain to be referenced, followed by the wiki pages and the download pages of Eclipse. Reporting newly discovered issues, or referencing the existing bugs in the Bugzilla is very common amongst the Eclipse forum discussions. The domain of `dev.eclipse.org`, which is currently not available, was also frequently referenced, especially for CVS repositories and news bulletins. We also found that the link resources from outside the Eclipse organization are preferred by forum users to refer to. The most communal external links posted in the threads are from `xtext.itemis.com`, i.e. the originator of the Xtext framework, `www.w3.org`, i.e. the world wide web standard organization, `github.com`, i.e. distributed version-control platform, and `twitter.com`, i.e. online news and social network. In addition, the Eclipse Git repositories website (`git.eclipse.org`) remains hot in the forum threads to complement the top 10 referenced domains.

Summary: Specific projects of `Eclipse Platfrom` and `BIRT`, and the forums for `Newcomers` are most active categories. We found that referencing to other resources, like bug reports, documentation, etc., is common in forum discussions, which indicates that forums are essential platform for linking various resources in Eclipse ecosystem.

Membership-based content The results of our manual classification show that the Eclipse forum is very much similar to other community-based question-and-answer sites, dominated by the users that share problems through questions to get the answers from the community. As shown in Table 10, discrepancy is the most common type of problems shared by the forum members, especially from Junior Members. Review, conceptual, usage, and errors
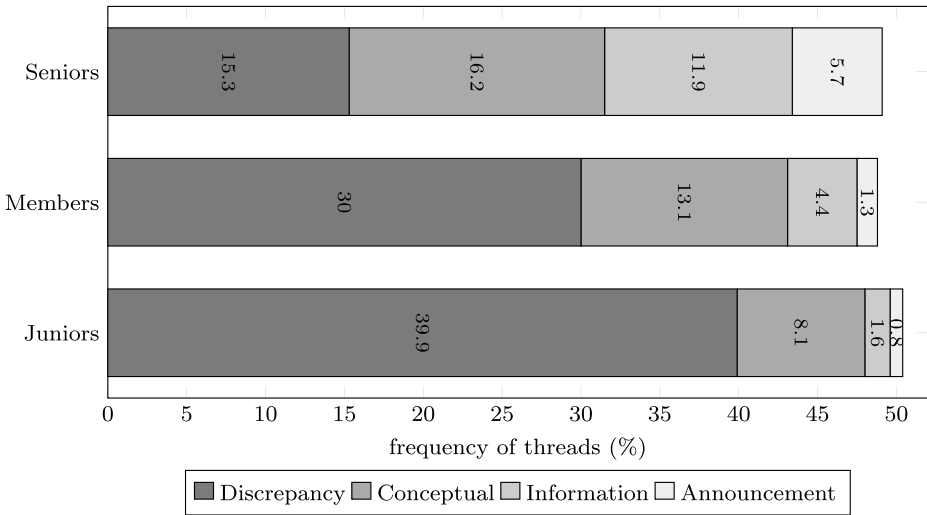
Fig. 6 Four content types communicated in the forums

complete the top five types of content that dominate in the Eclipse forum shared by all forum users, which accounts for more than 10% for each type. For the non-question-and-answer threads, information and announcement are two types of content that were most frequently shared between the forum members.

Our other findings on the content types that shared amongst the forum members at least hint that the forum is not only utilized for asking solutions or responses from other users, but also to announce some specific events or updates, distributing general information, offering some job vacancies, or even just testing the forum as well.

Our findings also highlight that there are several distinguished types of content that were communicated between the members based on their statuses. As illustrated in Fig. 6, discrepancy is the most common type that was shared by Junior Members. However, the number of this type reduces once the status of the user levels up into higher status. Conversely, the quantity of conceptual contents increases inline with the changes of the user statuses, from the lowest level of membership (Juniors) to the highest one (Seniors). This shows that the higher the membership status of a user, the higher the level of the problems discussed in the forums. Furthermore, Senior Members posted general information and announcement more often than the other types of users. The information shared in the forums is commonly general events, such as a seminar, presentation or competition to motivate users' participation. While the announcement describes specific events related to products, for example, file release, future updates, etc. This result at least hints that most of the Senior Members are core members of a software development company who act as sources of information and convey the information to inspire and motivate other people (Yunwen and Kishida 2003).

For the statistical evaluation, we find that there is an association between the communicated content and the members of different status. Our null hypothesis on *"the communicated content and the members of different statuses are independent."* is rejected (i.e., p-value is < 0.001).

Summary: We found that sentiment among developers are inconsistent while knowledge sharing. Interaction among developers shows that positive procedural initial posts overall have more chances to receive positive first responses.

4.3 RQ$_3$: What are the Differences Perceived in the Sentiment of Interaction Based on Membership Status?

Polarity detection Table 11 shows the polarity results of main posts and first responses among developers. We found that Juniors and Members post more positive messages than Seniors. On the other hand, compared with the other types of developers, neutral and negative messages were mostly posted by Senior developers. We also observed that overall positive posts received more positive replies.

Social interaction Figure 7 shows the nature of first responses in the interaction among different type of developers. From Fig. 7a and b, we observed that social interaction by Juniors and Members are overall positive. We found that Juniors and Members posted the initial posts in a more positive procedural way. For this positive procedural interaction, Juniors receive more positive procedural and neutral responses while Members receive both positive proactive and no answer as responses.

In contrast with the other types of developers, Fig. 7c shows that Seniors post positive procedural, neutral and negative procedural initial posts. Complement to RQ$_2$, the neutral posts mostly include an announcement related to a release of a new product, and general information such as advertisement, seminar, etc. In addition, Seniors' main posts did not receive responses in a large number of threads. This hints that Seniors seem to post more complex questions than Juniors and Members that may be difficult to answer. If we combine all types of developers together as illustrated in Fig. 7d, we observed that the social

Table 11 Polarity of communication among developers

| Users' Type	Main post	First response (%)				Sum (%)
		Positive	Neutral	Negative	No response	
Juniors	Positive	24.3	17.5	5.7	13.8	**61.4**
	Neutral	2.1	3.1	0.8	4.4	**10.4**
	Negative	6.0	8.4	7.3	6.5	**28.2**
Sum		**32.4**	**29.0**	**13.8**	**24.8**	**100.0**
Members	Positive	36.6	4.7	3.9	16.6	**61.8**
	Neutral	2.6	0.5	0.5	5.3	**8.9**
	Negative	18.4	2.1	2.1	6.6	**29.2**
Sum		**57.6**	**7.4**	**6.6**	**28.4**	**100.0**
Seniors	Positive	14.8	6.3	3.7	19.8	**44.6**
	Neutral	5.0	4.2	2.4	13.2	**24.8**
	Negative	10.8	4.5	4.5	10.8	**30.6**
Sum		**30.6**	**15.0**	**10.6**	**43.8**	**100.0**
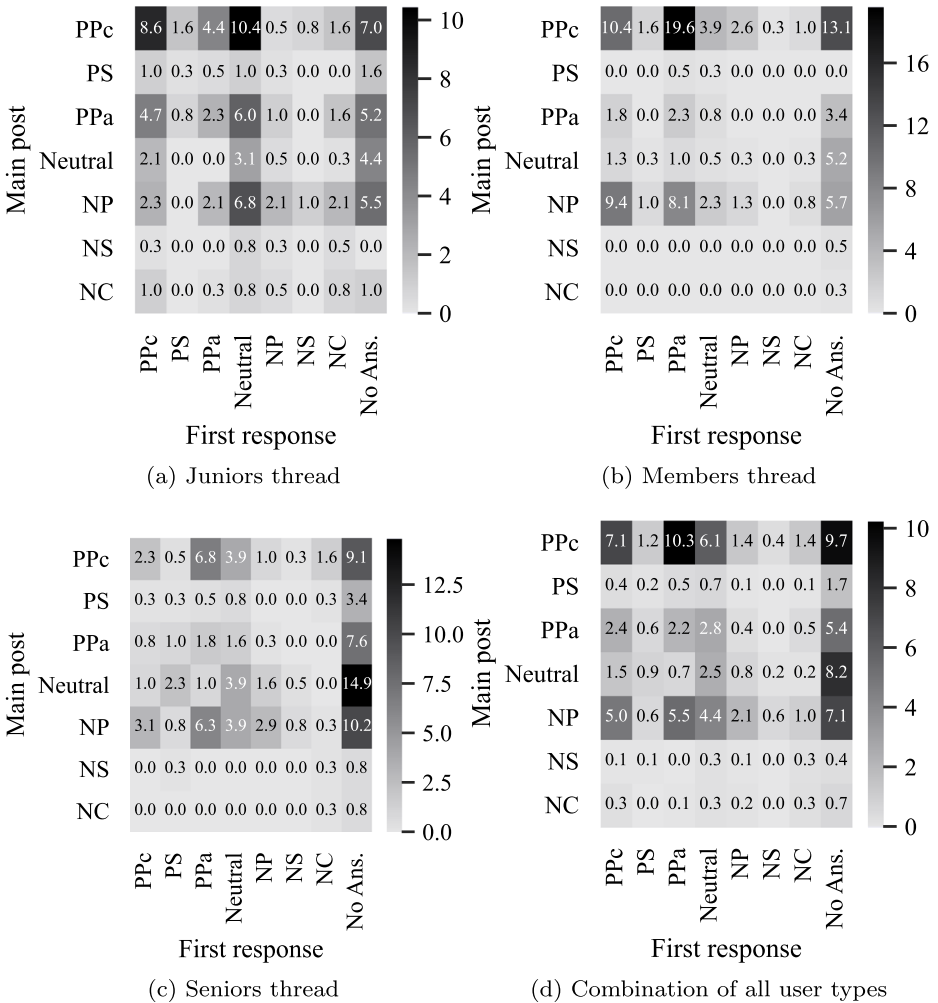
Fig. 7 Nature of first responses in the interaction between forum members. The color scale represents the frequency of the first responses in percentage based on total number of analyzed samples. The darker the area in the heatmap, the more frequent the first responses in the threads

interaction among developers are generally similar to the polarity. The result shows that, positive procedural posts received more positive proactive responses. Furthermore, posting a message in a positive procedural way has higher chances of getting positive replies.

For the statistical evaluation, we find that there is an association between the sentiment of interactions and the members of different status. Our null hypothesis on *"the sentiment of interactions and the members of different statuses are independent."* is rejected (i.e., p-value is < 0.001).

Summary: The Eclipse forums are dominated by question-and-answer threads, especially discrepancies, most of which were posted by Junior Members. Furthermore, the status of users relates to the topics discussed in the forums. The higher the status of users, the more conceptual the content type they communicate in forums.

5 Recommendations

Based on our study results, we provide three sets of recommendation for forum users, software development projects, and researchers.

For forum users, the recommendation regarding the threads and referencing links are as follows:

– *Join forum discussions*, as forums play a role as the center of knowledge sharing platform. The users will not only receive knowledge from community, but also from external sources.
– *Identify appropriate forum topic*. As shown in Table 10, although the percentage of misuse contents is not high, some amount of users posted their messages in the incorrect discussion groups. Sharing problems in a compatible topic category would maximize the probability to get responses and solutions. Otherwise, the questions might be marked out of scope.
– *Share knowledge in positive manner*. As shown in the results, users seem to respond positively to a positive and procedural message. Since the clarity of knowledge shared in the forums might increase the chances to get responses, describing the problems in a procedural way is essential.

One recommendation to software development projects is as follows:

– *Consider preparing project-specific forums*. As observed in the study, Eclipse forums is not only limited to question-and-answer based content, but also enable developers to provide Eclipse community-related information or announcement. Since the information or announcement that specifically relates to the software projects is important to share amongst the community, providing a project-specific forum is necessary. It has been reported that one of the keys to making OSS projects sustainable is to reduce the contribution cost by preparing the development environment (Hata et al. 2015). GitHub Discussions, a new feature of GitHub for asking questions or discussing topics outside of specific Issues or Pull Requests,[18] could be an option for projects hosted on GitHub (Hata et al. 2021).

For researchers, we can consider future works with the following possible challenges:

– *Further studies of activities in multiple software development systems*. Our results show that forums are being actively used by contributors that are also active in the other systems. This is indeed evidence that supports the claim that forums do work as a center of knowledge-sharing. Possible future directions could include tracking the actual

---

[18]https://docs.github.com/en/discussions

threads to understand whether the discussions in forums cross-over into the review and bugs systems, and to what effect.

– *Causal inference on promoting career*. From Eclipse historical data of social roles, we can also collect contributor metrics of a leader. Combined with the metrics of a committer, we can address the impact of the forum on social leadership, that is, the causal effect of forum participation on being promoted from committers to leaders.

6 Threats to Validity

Several threats to the *construct validity* emerge in our study. To connect the dataset between five different sources, we used *uids, email* addresses and *names*. Although the inspection was conducted manually, it is possible that our approach might lead to ambiguity. To mitigate this threat, we cross-reference between those five different systems to ensure that the authors are indeed not multiple aliases. Related to the memberships classification, it is possible to have inaccuracy of threshold definition to identify the member statuses. Although the thresholds are not formally informed by the organization, we specified the thresholds based on the maximum number of posts found in our dataset for each member status. Since the number of user misidentifications is small, we consider to use our defined thresholds in our study. It is also possible that there are identical users that have different identity number and different status in the forums. We also found in several threads, a message posted by a user had been duplicated with a different message identity number. Since we distinguished the users and the messages by their identity number, in our analyses, we recorded them differently even though they are same. In relation with forums linkage, we found that not all links in the threads can be extracted since they are written in plain text by users. However, the number of these issues were small. Thus, we consider that the impact of the missing links is not significant. The other threats to the construct validity is related to our manual labeling in RQ_2 and RQ_3. The label might be affected by annotator misunderstand or mislabeling. Despite annotators resolving disagreements through discussion, the labels might still be incorrect. In addition, there are many potential factors need to be taken into account. To mitigate this threat, we limit the scope of our study into conversational contents of the discussions themselves, to gain the insights of how project-specific forums are utilized.

Threats to the *external validity* appear in our data preparation. Even though we investigated a large scale of discussion forums in Eclipse, the findings could not be generalized to other organizations.

We diminish the threats to *reliability* by preparing the online appendix of our collected threads, links, contributors, and the results of our manual annotation for RQ_2 and RQ_3 (see Section 2.3).

7 Related Work

Forum discussion has been a popular research topic in the software engineering community in recent years. In this section, we introduce the key research related to our study.

7.1 Stack Overflow

There are a number of studies on a web-based question-and-answer communication channel, Stack Overflow (SO). The collaboration between users in different community forums (Zagalsky et al. 2018) and the knowledge sources contained in the discussion forums (Ye et al. 2017) have shaped the characteristics of knowledge sharing in the community.

Zagalsky et al. (2018) reported that SO with R-tag has become one of the two questions and answers communication channels that have a relationship with the users' discussions in R software development community forum. The collaboration between members and the independent individuals work have shaped the knowledge characteristics of the community. Ye et al. (2017) analyzed the reason behind the inclusion of web links in SO forum. It has been reported that the forum users share the URLs in SO for various purposes. Referencing sources to provide the solutions is the most prevalent purposes for the users who posted the web links.

In our study, we analyzed the entire discussion for each topic (not per message) in the Eclipse forums. Thus, we modified the label proposed by Zagalsky et al. (2018) to code the types of a thread (see Section 3.2). Compared to the work by Ye et al. (2017), we found that enclosing link sources to report bug-related problems or find a software defect is common in the forums.

Other works have studied the factors that affect the movement of users to SO and the most common topics discussed.

Squire (2015) showed that the quality measurements, participation of users, and the effectiveness of responding time in the SO forum have been the main factors that induced the developers' movement from self-supported forums and mailing lists. Zou et al. (2017) describes the insights of developers' requirements through SO. The authors found that the usability and reliability are the most common topics discussed in the SO forum and also the most unresolved problems faced by the users.

To complement with the previous work by Squire (2015), our study emphasizes that the users in Eclipse forums also actively participate and contribute over different systems in the Eclipse ecosystem. Expanding the work by Zou et al. (2017), we found that the most common topics discussed in Eclipse forums are Eclipse platforms and BIRT.

Several studies investigated on how to submit a successful question and increase the chance of receiving a successful answer in SO.

Calefato et al. (2018) proposed guidelines for users on how to increase the chances of receiving a successful answer to their questions on Stack Overflow. They found that the factors of potential-to-success questions are conciseness, completeness, and the exactness of the characters usage. In terms of affect, a neutral emotional style may impact successful questions. In line with that, Wang et al. (2018) investigated the factors related to the time to get an accepted answer in four Stack Exchange websites. After controlling for other factors, the accepted answers had been affected strongly by the dimension of the answerers. The more frequent a user is in answering the questions, the faster to receive an accepted answer.

In comparison with the work by Calefato et al. (2018) and Wang et al. (2018), our study shows that it is important to ask a procedural question in a positive manner, since it might increase the chances of receiving responses.

7.2 Twitter and News Aggregators

Communication between programmers are not limited to only mailing lists and question-and-answer channels, but also frequently occur on social media (i.e. Twitter) and news aggregators.

The information hidden in the tweets posted by Twitter users have been attractive for researchers to explore in a number of studies.

Mezouar et al. (2018) investigated the tweets posted by users to relate to the bug reports after removing the noisy tweets. Guzman et al. (2016) analyzed the characteristics of the usage of Twitter, information in the messages, and classified automatically the potential messages about software applications.

In our study, the findings support prior works by Mezouar et al. (2018) and Guzman et al. (2016). We show that bug report are the most dominant link target types referenced in the forums, while Eclipse Projects has become the most popular forums in the Eclipse community. In addition, Eclipse forum is not only used for asking solutions, but also for sharing information or announcement.

7.3 Contributors

Software development could not be separated from users' participation in a forum. Their contributions are not always related to writing code. A number of studies on the contributors in communication channels, has shown that experienced contributors and newcomers play an important role in developing software.

Senior contributors Recent studies show that every individual has an opportunity to become a valuable contributor. Zhou and Mockus (2012) built a model to analyze the users' chances of becoming a senior contributor depend on her competence, passion, and first-time contribution opportunity. Zhou and Mockus (2015) found that the participation of new members in the issue tracking system environment might impact their status of becoming a long term contributor. Expanding the findings from the two prior works, we found that becoming a senior contributor depends on the frequency of posts (i.e., Q&A, general information, or announcement) in the Eclipse forums.

Junior contributors Despite most of the valuable information for improvement to software quality comes from the experienced members in a software project (Li et al. 2015), software developers should not underestimate the newcomers' contributions in a discussion forum.

Steinmacher et al. (2015) identified that the lack of social interaction with the community, having unanswered questions or receiving delayed answers, and their technical experience backgrounds are some difficulties that new members faced when they make contributions to an open source software project. Middleton et al. (2018) also studied the contribution characteristics of new members in OSS projects. The authors identified that the participation forms of the new members, such as pull request and how they comment in the discussion influence the decision to join OSS teams. Our study extends the previous works that posting a message in positive manner might increase the chances of getting positive responses. Thus, it will have an impact on increasing social interaction.

8 Conclusion

To understand the impact of Eclipse community forums related to the linkage of the ecosystem and the social leadership, we conducted (i) a statistical study of 832,058 threads to analyze the characteristics of threads, contributions of Eclipse users, and users' classification; and also (ii) a large-scale study on the membership-based participation, content, and a sentiment investigation on the social interaction.

Our empirical study has shown that forums are an essential platform for linking various resources in Eclipse ecosystem and a source of expert knowledge for other development systems. The results also reveal that forum members actively participate in posting and responding to threads without limitation of membership. The forums are dominated by question and answer threads, with the status of users affecting the contents. Sentiment of interaction among developers are inconsistent. This work shows Eclipse forum is just one example of the utilization of project-specific forums. We recommend the users join forums as they are the center of knowledge sharing. For developers, preparing project-specific forums is important to share community-related information or announcement. Our results also hint at researchers that there are many open issues for future work: understanding and supporting forum discussions, further studies of activities in multiple software development systems, and causal inference for assessment, to name a few.

References

Abdellatif A, Costa D, Badran K, Abdalkareem R, Shihab E (2020) Challenges in chatbot development: A study of stack overflow posts. In: Proceedings of the 17th international conference on mining software repositories, MSR '20, pp 174–185

Allamanis M, Sutton C (2013) Why, when, and what: Analyzing stack overflow questions by topic, type, and code. In: Proceedings of the 10th working conference on mining software repositories, MSR '13, pp 53–56

Aniche M, Treude C, Steinmacher I, Wiese I, Pinto G, Storey MA, Gerosa MA (2018) How modern news aggregators help development communities shape and share knowledge. In: Proceedings of the 40th International conference on software engineering, ICSE '18, pp 499–510

Beyer S, Pinzger M (2014) A manual categorization of android app development issues on stack overflow. In: 2014 IEEE International conference on software maintenance and evolution, ICSME '14, pp 531–535

Beyer S, Macho C, Di Penta M, Pinzger M (2017) Analyzing the relationships between android api classes and their references on stack overflow. Tech.rep. Technical Report, University of Klagenfurt, University of Sannio

Beyer S, Macho C, Di Penta M, Pinzger M (2020) What kind of questions do developers ask on stack overflow? a comparison of automated approaches to classify posts into question categories. Empir Softw Eng 25:2258–2301

Calefato F, Lanubile F, Novielli N (2018) How to ask for technical help? evidence-based guidelines for writing questions on stack overflow. Inf Softw Technol 94(C):186–207

Guzman E, Alkadhi R, Seyff N (2016) A needle in a haystack: What do twitter users say about software? In: Proceedings of IEEE 24th international requirements engineering conference, RE '16, pp 96–105

Guzzi A, Bacchelli A, Lanza M, Pinzger M, Deursen Av (2013) Communication in open source software development mailing lists. In: Proceedings of the 10th working conference on mining software repositories, MSR '13, pp 277–286

Hata H, Todo T, Onoue S, Matsumoto K (2015) Characteristics of sustainable OSS projects: A theoretical and empirical study. In: Proceedings of the IEEE/ACM 8th international workshop on cooperative and human aspects of software engineering, CHASE '15, pp 15–21

Hata H, Treude C, Kula RG, Ishio T (2019) 9.6 million links in source code comments: Purpose, evolution, and decay. In: Proceedings of the 41st international conference on software engineering, ICSE '19, pp 1211–1221

Hata H, Novielli N, Baltes S, Kula RG, Treude C (2021) GitHub discussions: an exploratory study of early adoption. arXiv:2102.05230

Ibekwe AM, Ma J, Crowley DE, Yang CH, Johnson AM, Petrossian TC, Lum PY, Franz E, Flieger A (2014) Topological data analysis of Escherichia coli O157:H7 and non-O157 survival in soils. Front Cell Infect Microbiol 4:1–10

Kahani N, Bagherzadeh M, Dingel J, Cordy JR (2016) The problems with eclipse modeling tools: A topic analysis of eclipse forums. In: Proceedings of the ACM/IEEE 19th international conference on model driven engineering languages and systems, MODELS '16, pp 227–237

Kauffeld S, Lehmann-Willenbrock N (2012) Meetings matter: Effects of team meetings on team and organizational success. Small Group Res 43(2):130–158

Lertwittayatrai N, Kula RG, Onoue S, Hata H, Rungsawang A, Leelaprute P, Matsumoto K (2017) Extracting insights from the topology of the javascript package ecosystem. In: Proceedings of the 24th Asia-Pacific software engineering conference, APSEC '17, pp 298–307

Li PL, Ko AJ, Zhu J (2015) What makes a great software engineer? In: Proceedings of the 37th international conference on software engineering - volume 1, ICSE '15, pp 700–710

Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, Carlsson J, Carlsson G (2013) Extracting insights from the shape of complex data using topology. Sci Rep 3(1)

Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(11)

Mezouar ME, Zhang F, Zou Y (2018) Are tweets useful in the bug fixing process? an empirical study on firefox and chrome. Empir Softw Eng 23(3):1704–1742

Middleton J, Murphy-Hill E, Green D, Meade A, Mayer R, White D, McDonald S (2018) Which contributions predict whether developers are accepted into GitHub teams. In: Proceedings of the 15th international conference on mining software repositories, MSR '18, pp 403–413

Nakasai K, Hata H, Matsumoto K (2018) Are donation badges appealing? a case study of developer responses to eclipse bug reports. IEEE Softw 36:22–27

Pearson K (1992) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: Breakthroughs in statistics. Springer, pp 11–28

Rosen C, Shihab E (2016) What are mobile developers asking about? a large scale study using stack overflow. Empir Softw Eng 21:1192–1223

Squire M (2015) Should we move to stack overflow?: Measuring the utility of social media for developer support. In: Proceedings of the 37th international conference on software engineering - volume 2, ICSE '15, pp 219–228

Steinmacher I, Silva MAG, Gerosa MA, Redmiles DF (2015) A systematic literature review on the barriers faced by newcomers to open source software projects. Inf Softw Technol 59:67–85

Storey MA, Zagalsky A, Filho FF, Singer L, German DM (2017) How social and communication channels shape and challenge a participatory culture in software development. IEEE Trans Softw Eng 43(2):185–204

Tantisuwankul J, Nugroho YS, Kula RG, Hata H, Rungsawang A, Leelaprute P, Matsumoto K (2019) A topological analysis of communication channels for knowledge sharing in contemporary GitHub projects. J Syst Softw 158:110416

Treude C, Barzilay O, Storey MA (2011) How do programmers ask and answer questions on the web? (nier track). In: Proceedings of the 33rd international conference on software engineering, ICSE, vol 11, pp 804-807

Viera A, Garrett J (2005) Understanding interobserver agreement: The kappa statistic. Fam Med 37(5):360–363

Wang S, Chen TH, Hassan AE (2018) Understanding the factors for fast answers in technical Q&A websites. Empir Softw Eng 23(3):1552–1593

Ye D, Xing Z, Kapre N (2017) The structure and dynamics of knowledge network in domain-specific Q&A sites: A case study of stack overflow. Empir Softw Eng 22(1):375–406

Yunwen Ye, Kishida K (2003) Toward an understanding of the motivation of open source software developers. In: Proceedings of the 25th international conference on software engineering, ICSE '03, pp 419–429

Zagalsky A, German DM, Storey MA, Teshima CG, Poo-Caamaño G (2018) How the r community creates and curates knowledge: An extended study of stack overflow and mailing lists. Empir Softw Eng 23(2):953–986

Zhou M, Mockus A (2012) What make long term contributors: Willingness and opportunity in oss community. In: Proceedings of the 34th international conference on software engineering, ICSE '12, pp 518–528

Zhou M, Mockus A (2015) Who will stay in the floss community? modeling participant's initial behavior. IEEE Trans Softw Eng 41(1):82–99

Zou J, Xu L, Yang M, Zhang X, Yang D (2017) Towards comprehending the non-functional requirements through developers eyes. Inf Softw Technol 84(C):19–32

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yusuf Sulistyo Nugroho is a lecturer at the Department of Informatics, Universitas Muhammadiyah Surakarta, Indonesia. He received his Ph.D degree from the Nara Institute of Science and Technology in 2020. He is a member of the IEEE. His research interests include Empirical Software Engineering, Software Documentation, and Mining Software Repositories. Contact him at yusuf.nugroho@ums.ac.id.

Syful Islam received the M.E. degree in Information Science from Nara Institute of Science and Technology, Japan. He is currently working toward the P.hD degree in the same institute. At present, he is on study leave from Noakhali Science and Technology University, Bangladesh. His research interests include software ecosystem, mining Stack Overflow, etc. Contact him at islam.syful.il4@is.naist.jp.

Keitaro Nakasai is an assistant professor at the National Institute of Technology, Kagoshima College, Japan. His research interests include mining software repositories and software ecosystems. Nakasai received a Doctor of Engineering in information science from the Nara Institute of Science and Technology. He is a member of the IEEE and the Information Processing Society of Japan. Contact him at nakasai@kagoshima-ct.ac.jp.

Ifraz Rehman is a Ph.D. student at the Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan. He received his Master of Engineering in 2021 from the Nara Institute of Science and Technology. His research interest includes human aspect of software engineering. Contact him at rehman.ifraz.qy4@is.naist.jp.

Hideaki Hata is an Associate Professor at Shinshu University, Japan. He received his Ph.D. in information science from Osaka University. His research interests include software ecosystems, human capital in software engineering, and software economics. More about Hideaki and his work is available online at https://hideakihata.github.io/.

Raula Gaikovina Kula is an Assistant Professor at Nara Institute of Science and Technology. He received the Ph.D. degree from Nara Institute of Science and Technology in 2013. His interests include Software Libraries, Software Ecosystems, Code Reviews and Mining Software Repositories.

Meiyappan (Mei) Nagappan is an Associate Professor and the Current Ross & Muriel Cheriton Faculty Fellow in the David R. Cheriton School of Computer Science at the University of Waterloo. His research is centred around the use of large-scale Software Engineering (SE) data to address the concerns of the various stakeholders (e.g., developers, operators, and managers). He received a Ph.D. in computer science from North Carolina State University. Dr. Nagappan has published in various top SE venues such as TSE, ICSE, FSE, EMSE, JSS, ICSME and IEEE Software. He is an associate editor for EMSE, JSS and TSE and has served on the PC of several conferences like ICSE, MSR, and ICSME. He has also been the PC Co-Chair of MSR 2021, QRS 2020 and Promise 2020. You can find more at mei-nagappan.com.

Kenichi Matsumoto is a professor in the Graduate School of Science and Technology at Nara Institute Science and Technology, Japan. He has a Ph.D. in information and computer sciences from Osaka University. He is a senior member of IEEE, and a member of the IEICE and the IPSJ. His research interests include software measurement and software process. Contact him at matumoto@is.naist.jp.

Affiliations

Yusuf Sulistyo Nugroho[1] (ID) **· Syful Islam[2] · Keitaro Nakasai[3] · Ifraz Rehman[2] ·
Hideaki Hata[4] · Raula Gaikovina Kula[2] · Meiyappan Nagappan[5] · Kenichi Matsumoto[2]**

Syful Islam
islam.syful.il4@is.naist.jp

Keitaro Nakasai
nakasai@kagoshima-ct.ac.jp

Ifraz Rehman
rehman.ifraz.qy4@is.naist.jp

Hideaki Hata
hata@shinshu-u.ac.jp

Raula Gaikovina Kula
raula-k@is.naist.jp

Meiyappan Nagappan
mei.nagappan@uwaterloo.ca

Kenichi Matsumoto
matumoto@is.naist.jp

[1]   Universitas Muhammadiyah Surakarta, Jawa Tengah, Indonesia

[2]   Nara Institute of Science and Technology, Ikoma, Japan

[3]   National Institute of Technology, Kagoshima College, Kagoshima, Japan

[4]   Shinshu University, Matsumoto, Japan

[5]   University of Waterloo, Waterloo, Canada