Why Do Newcomers Abandon Open Source Software Projects?

Igor Steinmacher, Igor Wiese, Ana Paula Chaves
Computer Sciences Coordination
Federal University of Technology – Paraná (UTFPR)
Campo Mourão, PR, Brazil
{igorfs, igor, anachaves}@utfpr.edu.br

Marco Aurélio Gerosa
Computer Sciences Department
University of São Paulo
São Paulo, SP, Brazil
gerosa@ime.usp.br

***Abstract*—Open source software projects, are based on volunteers collaboration and require a continuous influx of newcomers for their continuity. Newcomers face difficulties and obstacles when starting their contributions, resulting in a low retention rate. This paper presents an analysis of the first interactions of newcomers on a project, checking if the dropout may have been influenced by lack of answer, answers politeness and helpfulness, and the answer author. We have collected five years data from the developers' mailing list communication and issue manager (Jira) discussions of the Hadoop Common project. We observed developers' communication, identifying newcomers and classifying questions and answers content. In the analyzed period, less than 20% of newcomers became long-term contributors. There are evidences that the newcomers decision to abandon the project was influenced by the authors of the answers and by the type of answer received. However, the lack of answer was not evidenced as a factor that influences newcomers' decision to remain or abandon the project.**

***Index Terms*—Newcomer, communication, collaboration, open source software, retention**

I. Introduction

A continuous influx of newcomers and their active engagement with development activities are crucial to the success of Open Source Software (OSS) projects [1]. However, the first steps in open software projects require overcoming many obstacles. Degenais et al. [2] compare newcomers in software projects to explorers who need to orient themselves in a hostile environment. On the one hand, newcomers need to learn social and technical aspects alone, exploiting existing information in mailing lists, source code repositories, and issue managers [3]. On the other hand, it is not easy to access this information due to the large volume, lack of tools to navigate the repository, and the difficulty of making connections between logically related items in different sources [4].

In a previous study [5], we presented reports from developers who tried to initiate their participation in two well-known open source projects. Developers indicated that the lack of awareness and guidance during the course of their first steps discouraged further contributions. To reduce this problem, newcomers generally post their questions and request help to choose their tasks in forums and mailing list or send emails to specific developers who have central roles in the project (e.g. owners, project leaders) [1, 6]. As mailing lists and forums are public communication channels, people often use such means to start their interaction in the project. However, receiving replies that do not offer guidance or unpolished answers can result in newcomers dropout.

Given this scenario, it is important to observe different open source software communities to understand the way they interact and what are the newcomers' needs when they start their participation in such projects. This understanding enables the creation of mechanisms and tools to better support the retention of newcomers in open source software projects, by means of, for example, defining specific awareness mechanisms for them. This understanding and tools may also be extended to other communities that depend on collective production by volunteer work, such as virtual encyclopedias and other social media systems.

This paper presents a study that aims to verify whether the lack of response, politeness, and usefulness of the answers, or the authors of the replies received by newcomers in the mailing list and in the issue manager influence the decision to remain in the project. We seek to understand the reasons why newcomers do not stay based on their first interactions in the project. To reach this, we examine the research question:

> *Does the absence of response, politeness, usefulness or the author of answers influence the retention of newcomers in an open source project?*

To answer this research question, we defined three specific objectives, namely:

- check if the newcomers receive answers;
- observe who are the authors of the answers to newcomers' questions;
- classify the answers received by the newcomers.

For this study we chose to observe the Hadoop Common project, hosted by the Apache Software Foundation. For this analysis, we used data from the developers' mailing list, issue manager (Jira), and the users' mailing list.

The rest of the paper is organized as follows: in Section II, we present some related work. In Section III, the research method. In Section IV, the results. In Section V, we present the threats to validity. Finally, in Section VI, the conclusions and future work.

978-1-4673-6290-0/13/$31.00 © 2013 IEEE

II. Related Work

Many studies in the literature deal with newcomers joining process in collective production communities, including studies on Wikipedia [7, 8, 9] and on open software projects [5, 6, 10, 11, 12, 13]. Degenais et al. [2] and Begel and Simon [14] also present studies regarding newcomers joining process in software projects, but their focus is in proprietary software.

Studying newcomers in open source projects is important because, according to Jensen et al. [10], they are potential contributors that are vital to projects growth and survival. This paper is aligned to the problem addressed by other previously published works, studying the initial steps and the difficulties faced by newcomers in open source software projects [6, 10].

Von Krogh et al. [6] conducted a study on the project FreeNet, using interviews with developers, analysis of emails, source code repository, and project documents. The authors proposed a joining script for developers who want to take part in the project. One of their contributions indicates that newcomers often lurk the project before starting their participation, and then start interacting. Although they studied the joining process in an open source software project, they did not analyze the reasons why newcomers leave, checking only the behavior of those who became project developers.

Nakakoji et al. [15] studied four open source software projects to analyze the evolution of these communities. They presented eight possible roles for the members of an OSS project and structured them into a model composed of concentric layers, like the layers of an onion. This structure was later called the onion patch, and other authors conducted studies based on this model [10, 16, 17]. According to this model, newcomers usually start by outer layers and go toward the center. Although these papers deal with the joining and evolution of members' participation in open source communities, none of them concerned with the reasons for newcomers leaving the community.

Some researchers worked on giving support to newcomers. Zhou and Mockus [18] and Schilling et al. [19] worked on identifying the newcomers who are more likely to remain contributing to the project in order to offer active support for them to become long-term contributors. Čubranić et al. [4] presented Hipikat, a tool that supports newcomers by building a group memory consisting of four types of artifacts: source code, email discussions, change tasks (issues), and other project documents (e.g., design documents). Users proactively request recommendation based on existent artifacts. Hipikat returns a list of source code, mails messages, and bug reports related to the queried artifact. Sarma and Wang [13] present a Tesseract extension to enable newcomers to identify bugs of interest, resources related to that bug, and visually explore the appropriate socio-technical dependencies for the bug in an interactive manner. Malheiros et al. [12] present Mentor, a tool intended to help newcomers by recommending potentially relevant source code pieces that can be used by a developer when working on an issue. Park and Jensen [1] show that visualization tools support the first steps of newcomers in an open source project, helping them to find information more quickly. Canfora et al. [11] proposed an approach aimed at identifying and recommending mentors to newcomers of open source projects by mining data from mailing lists and source code versioning systems. They evaluated the approach using data from mailing lists and surveying some developers to understand mentoring in their projects. These articles deal with the admission and evolution of the participation of members in open source software, but they do not analyze the factors that influence the newcomers' decision to abandon the project.

Jensen et al. [10] made an analysis on four mailing lists from open source software projects in order to verify if the e-mails sent by newcomers are quickly answered, if gender and nationality influence the kind of answer received, and finally, if the reception is different in users lists and developers lists. Our study is close to Jensen's study in terms of goals and methodology. However, our main goal is to delve into the reasons why the newcomers drop out from the project.

III. Research Method

To conduct this study, we used data from the Hadoop Common project. This project was chosen because it is a successful project, already consolidated, and has an active and well organized community. Furthermore, data from the issue manager (Jira) and mailing lists were publicly available.

We gathered data from the developers' mailing list (common-dev) and from the comments posted to the issue manager (Jira) from January 2006 to December 2010. We analyzed the emails and Jira comments separately. The following sections detail the data collection.

A. *Collecting Data from Jira*

To collect the data from the issue manager (Jira), we built a tool to extract the data from the issues and store them in a local relational database. The extraction was performed by accessing the web page that displays the issue and varying the URL: "https://issues.apache.org/jira/browse/ <proj_name> - <issue_number>. The system receives the name of the project as a parameter and varies the issue number sequentially.

The extractor parses every HTML page and collects the following information for each reported issue: description, issue reporter, assignee, creation date, closing date, priority, status, and comments (with author, date, and message). For the analysis, we considered the users that appear as reporters, assignees, or those who have commented any issue.

B. *Collecting Data from Mailing List*

To extract data from the mailing list, first we downloaded the mbox files (storage format for email collections) for each month of the investigated period. These files contain the header and body of emails sent to the list in the corresponding month.

The information was extracted from the mbox files by analyzing message headers and body to obtain the message content, subject, message ID, sender, and thread identifier (In-Reply-To). The threads were reconstructed by checking the field in-reply-to header as well as the email subject (examining the prefixes "Re:", "Fwd:") and the header field references, to lessen the chances of losing messages related to a discussion.

Finally, the messages details were stored in a database. For the analysis, we disregarded messages sent automatically in the creation, review, or change of status of an issue in Jira. Such messages were identified by checking if it was sent by jira@apache.org or if it started with the identifier "[jira]."

C. *Data Analysis*

We collected 60 months of mailing list discussions, containing 7,891 threads with 37,095 responses, resulting in 45,076 posts. We also collected 60 months of issues created in Jira, resulting in 6,793 issues with 53,664 comments.

As described in Figure 1, the analysis period was divided into 4 intervals. Initially, we carried out a query in interval 1 to identify the people who contributed to the project during the first 3 years. In our analysis, these users are considered members that already contribute to the project.

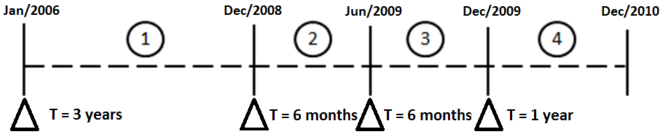

Figure 1. Timeline of data collection

In the second interval, we identified persons who started their contributions during this period, not appearing in interval 1. These persons were considered newcomers in the context of this study.

In the third interval, we verified which newcomers remained within 6 months and those that did not return. Newcomers who have not appeared in interval 3 were considered dropouts.

Finally, in interval 4, we identified users who kept contributing. Users who appeared in intervals 2, 3, and 4 were considered newcomers who were retained by the project.

We conducted a manual analysis of the messages sent by the dropouts and the answers received in order to classify them according to the absence of response, the authors of the answers, and the type of answers sent to the newcomers. The model used to classify the messages was defined by adapting the method used by Qu et al. [20]. Figure 2 illustrates the method used to create the classification.

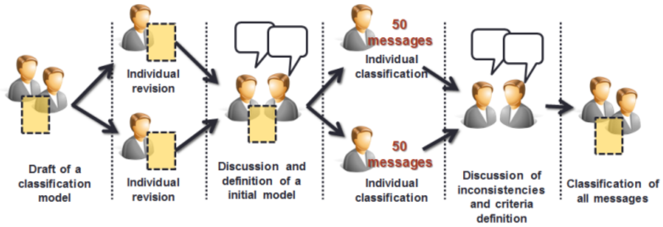
Draft of a classification model
Individual revision
Individual revision
Discussion and definition of a initial model
50 messages
Individual classification
50 messages
Individual classification
Discussion of inconsistencies and criteria definition
Classification of all messages

Figure 2. Method applied to define the message classification model

According to the method, two researchers create jointly a draft of the model, after a brainstorming session. After that, they individually review the draft suggestions of both revisions are discussed, resulting in an initial model. Then, the two researchers categorize independently a random sample of messages, according to the initial model. The two researchers discuss the inconsistencies until they came to a consensus regarding the classification and the criteria. A new model based on the resulting criteria is then defined and applied to the full set.

The following subsections describe some details for each method defined.

1) Check if newcomers receive answers: Once newcomers were identified, we checked whether they received answers or not, and the time elapsed until receiving the first answer. For the issue manager, it was necessary to check if the newcomer was the reporter or just joined the discussion by posting comments. If the newcomer was a reporter, the comments made by other users in the task were considered answers.

In the mailing list, we queried the messages sent by newcomers that were starting threads, and classified them as questions. If they were part of a discussion in progress, we classified the message as an answer. For messages classified as questions, we checked for answers in the same thread.

2) *Who answer newcomers' questions:* For messages that received answers, we analyzed the email address and username of the members who answered to the messages. Then, we classified the addresses and usernames according to the interval that the member started in the project and to the amount of previously sent messages, dividing them into three categories:

- Core members: appeared in interval 1 and were among the 10% more participative members;
- Newcomers: have not appeared in interval 1 and appeared in interval 2;
- Other members: appeared in interval 1 and were not among the 10% most participative members.

This classification was defined to check if the previous 'experience' of the member who answered the question influences the retention of a newcomer

3) *Type of answer received by the newcomers:* We classified the answers received by newcomers according to the types of answers defined by the method shown in Figure 2. Therefore, each message was classified into the following types:

- In Topic / Help: when the answer addresses the problem raised;
- Indifferent: when the message is not informative, show no receptive tone and usually indicates an external link to answer the question;
- Not Useful / Off Topic: when the answer is off topic and does not contribute to address the problem;
- Not Useful / Another question: when the answer is a new problem, creating a different discussion;
- Other: when it was not possible to classify as one of the previous types, for example, product announcements or not understandable messages.

4) *Questionnaire sent to the dropouts:* After analyzing the data, we conducted a survey via email with dropout newcomers, to complement the understanding of the reasons that led them to first interact and then abandon the project. To do it, we sent a questionnaire to the dropouts. The questions submitted and the results can be verified in Section IV.D.

IV. Results

In this section, we present and discuss the results. First, we present a general analysis on the information obtained, and then we present a discussion for each specific goal.

Table I presents the results of the analysis made on developers' mailing list data. For each interval, we present the number of users found, the percentage of users in each interval, and the percentage of newcomers who remained in the project in the next period and, subsequently, those who continued contributing during interval 4. In the second interval, 67 newcomers joined the community and 20 of them (29.85%) remained in the next interval. After one year, only 12 (17.91%) were still active.

TABLE I. USERS THAT SENT EMAIL TO THE MAILING LIST

	# users	*% of existing*	*% of newcomers*
Existing members (interval 1)	677		
Newcomers (interval 2)	67	9.90%	
Remaining (interval 3)	20	2.95%	29.85%
Retention (interval 4)	12	1.77%	17.91%

A similar analysis was performed on the data obtained from Jira. Table II presents the results. We considered those users who reported or commented some task. One can verify that the number of newcomers is greater when compared to the mailing list.

TABLE II. USERS THAT REPORTED OR COMMENTED AN ISSUE IN THE ISSUE MANAGER

	# users	*% of existing*	*% of newcomers*
Existing members (interval 1)	483		
Newcomers (interval 2)	127	26.29%	
Remaining (interval 3)	30	6.21%	23.62%
Retention (interval 4)	17	3.52%	13.39%

Table III presents an overview of the evolution of the participation of newcomers in mailing list. The table shows the number of newcomers who kept contributing to the project and the amount of messages sent by these newcomers in each interval. The first observation we can make is that, although there was a great newcomers' dropout rate in interval 2, the amount of messages sent by members who remained in the project increased in interval 3.

TABLE III. EVOLUTION OF NEWCOMERS' PARTICIPATION IN MAILING LIST

	Interval 2			***Interval 3***			***Interval 4***		
	Msgs	*New-comers*	*Msgs/ New-comer*	*Msgs*	*New-comers*	*Msgs/ new-comer*	*Msgs*	*New-comers*	*Msgs/ new-comer*
Questions	68	47	1.45	18	9	2.00	6	5	1.20
Answers to others' discussion	56	24	2.33	160	15	10.67	55	12	4.58
Replies to own thread	56	20	2.80	17	6	2.83	12	4	3.00
TOTAL	180	67	2.69	195	20	9.75	73	12	6.08

We also observed that the second interval has a greater number of questions than intervals 3 and 4, both in absolute terms and when compared with the amount of newcomers. The reason is that the first interactions with the list are made to clarify doubts, set environment or request help to take the initial steps in the project.

Regarding the answers to the discussions initiated by other members, in interval 2, only 56 messages were sent in 32 different discussions by 24 newcomers. Few newcomers (9) participated in more than one thread and only eight newcomers wrote more than one answer in discussions of other members. In the third interval, 15 newcomers answered to 160 messages sent by third parties in 94 discussions. This shows that, after an initial period in the list, newcomers who have continued in the project began to contribute more in discussions initiated by others, and assist in troubleshooting. In interval 4, he responses sent by the newcomers who remained decreased.

The decrease of messages sent by newcomers who continued (interval 4) can be observed in all table rows. No reason for such reduction was found during manual analysis. However, there may be a relation with the members' evolution process within the project, so they start contributing in other ways, such as answering questions or fixing bugs. One thing to observe is that, although newcomers who continued sent only 55 messages answering to others, all 12 newcomers appeared in these answers.

TABLE IV. EVOLUTION OF NEWCOMERS' PARTICIPATION IN JIRA

	Interval 2			***Interval 3***			***Interval 4***		
	Msgs	*New-comers*	*Msgs/ New-comer*	*Msgs*	*New-comers*	*Msgs/ New-comer*	*Msgs*	*New-comers*	*Msgs/ New-comer*
Issues reported	154	78	1.97	61	15	4.07	76	10	7.60
Comments to others' issues	420	107	3.93	308	17	18.12	356	14	25.43
Comments to own issues	421	55	7.65	260	18	14.44	331	11	30.09
TOTAL	995	127	7.83	629	30	20.97	763	17	44.88

We also analyzed the evolution of the newcomers found in Jira. The results are shown in Table IV. Differently from what occurred in the mailing list, we can see that the average number of posts per newcomer (msgs / newcomer) increased in every interval analyzed. This increase is observed in the number of issues reported and comments sent in discussions. This growth is due to the visibility and confidence that users increasingly acquire when contributing.

The initial analysis showed a small proportion of newcomers who remained in both environments. To further investigate the possible reasons for dropping out, the following Sections present discussions related to the specific objectives of this study.

A. Are the newcomers answered?

In Table V we present the data regarding the answers received by newcomers in the mailing list and their decision to give up or remain in the project. The observation shows that 47 newcomers sent messages during interval 2. These newcomers sent 68 questions to the mailing list. We found that 34 newcomers (72.34%) were answered by other members in 40 different threads. Among those who were answered, the average time to receive the answer was 1.32 days, and 19 of them were answered on the same day.

TABLE V. ANSWERS VERSUS DROPOUT IN INTERVALS 2, 3 AND 4 CONSIDERING MAILING LIST PARTICIPATION

	# people	*Dropout*	*Remained in intervals 3 and 4*
No answers	13	11	2
Received answer	34	30	4

Among the other 13 newcomers who have not received any answers to their questions, eleven gave up (84.62%) and the other two (15.38%) continued contributing to the project. Among those who have had some question answered, 30 gave up (88.24%) and four continued (11.76%) in the project. Therefore, there is no evidence that the lack of answer influences on the dropout decision.

The manual analysis of the messages enabled us to verify that some messages are not answered because the questions were off topic. For example, we found some questions regarding Hadoop installation and configuration in the developers list, when they should be sent to the users list.

We have also noted that most of the questions are promptly answered, and the authors thank for the answer. Even after getting correct and useful answers, some newcomers left the community after receiving the answer. In this case, it is clear that the people who sent the e-mails did not intend to contribute to the project, but to solve a problem that they were facing momentarily.

Table VI presents the data related to the receipt of comments on issues reported by newcomers in Jira and the newcomers' decision to remain or to abandon the project. We found that 78 newcomers reported tasks during interval 2, among them, 71 (91%) received comments. Only seven newcomers had not received any comments in eight issues posted. By analyzing these issues manually, we note that six of them were redirected to the MapReduce project, whose activity level is lower than the project Hadoop Common. So, we considered that only two issues had not received feedback.

We can see that the receptivity in Jira is very good. Even issues reporting something that is out of the scope of that tool or reporting problems already reported previously, were commented, guiding the users.

TABLE VI. REPORTS, COMMENTS AND DROPOUTS IN INTERVALS 2, 3 AND 4 CONSIDERING ISSUE MANAGER PARTICIPATION

	# people	*Dropout*	*Remained in intervals 3 and 4*
Not commented	7	6	1
Received comments	71	55	16

Thus, we can say that Jira is an environment in which new members are well received, and that receiving comments on this tool does not influence the retention of newcomers on Hadoop Common Project.

B. *Who answered the newcomers?*

Figure 3 presents a Venn diagram showing the relation between the questions asked by newcomers and the members who answered in the context of the developers' mailing list during interval 2. In the figure, each set represents the type of author who answered questions triggered by newcomers. The respondents were categorized according to the types presented in Section III.C.3. The values shown within the sets represent the number of threads that a particular type of member participated.

It can be noticed that most part of the questions sent by newcomers are answered by core members. Considering the discussions initiated by dropouts, 21 (63.63%) had the involvement of core members, seven of them were answered only by core members. Eleven discussions (34.38%) initiated by dropouts had no answers sent by core members. While reading the messages sent by email, we found that, in some cases, newcomers answering newcomers bring negative influence, which could result in a dropout. For example, in a discussion in which a newcomer requested assistance in choosing a bug to start his contribution, another newcomer replied. His answer said that only committers could work on bugs. In another case, the newcomer asked about the architecture and what would be the simplest way to start his contribution in the project. Two other newcomers also sent messages in the thread saying they also wanted to contribute to the project.

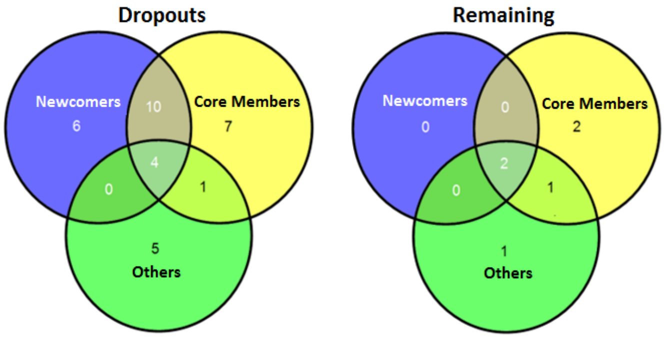

Figure 3. Analysis of respondents of emails sent by newcomers to the developers mailing list

Observing Figure 3, it is also possible to see that most part of the discussions initiated by newcomers who quit the project received answers from core members. By analyzing the text of emails, it is clear that much of this discussion relates to specific technical questions of a user or configuration of a specific environment, including some messages replicated to the user list. These newcomers sometimes use the mailing list to solve their own problems, without any intention to keep contributing.

Observing the answers received by newcomers who continued contributing, we realize that there is no discussion in which only newcomers sent messages. In addition, newcomers sent answers to only 2 threads. These discussions were manually analyzed, and they present six and nine messages sent by five different persons in each discussion. The newcomers that appeared in these threads continued their contribution to the project during intervals 3 and 4.

The results obtained from the analysis of Jira are depicted in Figure 4. It can be noticed that there is a greater amount of answers to newcomers, both for those who remained, and for those who dropped out. In Jira, 39 reports (45.34%) of dropout newcomers had no comments from core members, while for those who continued there were 11 (28.20%) comments from core members.

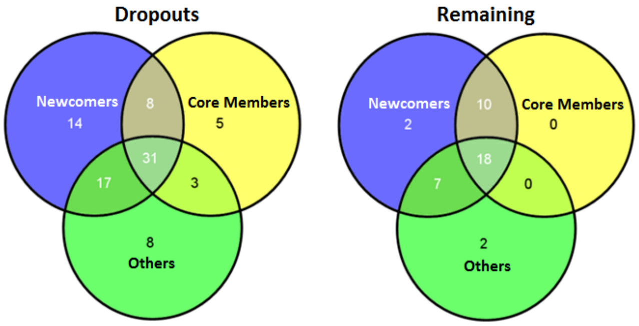

Figure 4. Analysis of the authors of the comments to issues reported by newcomers in Jira

In general, we noticed a greater contribution of core members, and more answers from newcomers in Jira than we found in the mailing list. This occurs because the issue management tool stimulates more contextualized and focused discussions, which will have an effective result in the project. We also noticed that, during the discussions, newcomers receive some guidance from the core members when they submit some contribution (bug report, enhancement or patch).

C. *What kind of answer the newcomers receive?*

While reading the threads started by newcomers, the answers to each of them were classified in order to verify their impact in the withdrawal. Table VII presents the results of the classification according to the model defined according to the method presented in Section III.C.2. It is possible to see that the answers "Not Helpful" or "Indifferent" were received only by dropouts. There are nine answers classified into one of these types. This may be an indication that the type of answer can influence the decision to leave the project. Section IV.E presents some other evidences on this.

The data presented in Table VII show that, even when receiving helpful answers, some newcomers left the project. The manual inspection identified that the messages sent by newcomers were asking questions not related to "how to contribute" or technical questions related to a contribution. The messages were concerning specific user needs, for example, the integration of a proprietary technology to Hadoop, asking for a library in a release, reporting results of a test using computer grids, incompatibility with a Java virtual machine and setup questions and. Thus, it is clear that some users had specific intention of clarifying their individual doubts, and were not necessarily interested in contributing.

TABLE VII. TYPE OF ANSWERS RECEIVED BY THE NEWCOMERS

Answer Type	*Left the project*	*Remained*
In topic / Help	20	7
Not Useful / Another Question	5	0
Not Useful / Off topic	3	0
Indifferent	1	0
Other	4	0

The manual analysis conducted over Jira issues and comments showed that the comments posted were on topic, contextualized and provided useful information. Because it is a controlled environment, there are no patterns or discrepancies to discuss. Some exceptions appear when, for example, an issue was reporting a setup or installation problem. In these cases, users answered by redirecting the beginner to correct forum,

D. *Survey conducted with the dropouts*

In this section, we present details of the questionnaire sent via email to the newcomers who quit the project. After analyzing the data extracted from the mailing list, we sent a short questionnaire to 55 newcomers who left the project, with the following questions:

1. Do you remember sending an email to hadoop-common-dev mailing list? [Y/N]
2. At that time, were you interested to keep contributing to Hadoop project?[Y/N]
2a. In case you answered YES to question 2, why did you give up?
2b. In case you answered NO to question 2, what was the goal of the messages sent to developers list?
3. Have you contributed to the project after June 2009? [Y/N]
4. Have you contributed to other Open Source project BEFORE 2009? [Y/N]

From 55 e-mails sent, 10 deliveries failed and 13 were answered. The answers received are summarized in Table VIII. Other 31 persons did not respond to e-mail and one user responded to the email with no answers to the questionnaire. He just said he was still in the project and had become committer recently, but had changed his email address.

TABLE VIII. ANSWERS TO THE SURVEY SENT TO THE DROPOUTS OF PROJECT HADOOP COMMON

	Question 1	*Question 2*	*Question 3*	*Question 4*
Yes	13	11	1	7
No	0	2	12	6

We can notice that the 13 respondents recalled having sent the email to the list and said their intention was to keep contributing to the project (11 answered "yes" to question 2). This question was specialized in two others, according to the option chosen, to understand the reasons why they left. The answers were analyzed and the result is shown in Table IX.

TABLE IX. TYPE OF ANSWERS TO THE QUESTIONS SENT TO THE DROPOUTS OF PROJECT HADOOP COMMON

Type of answer	*Answer to Question 2*	
	Yes	*No*
i. The user just wanted to clarify some doubt	0	2
ii. Question not answered or answer did not help	2	0
iii. Lack of help to choose a task	3	0
iv. Not accepted by the Project	1	0
v. Changed focus or company	4	0
vi. Resumed the contributions later	1	0

The answers classified as types *ii, iii* and *iv* show that a possible reason for quitting the project was the receptiveness. From the 13 respondents who intended to contribute to the project, six sent answers related to reception. We highlight two answers that clearly show the dissatisfaction of dropouts:

"*My issue was how to start contributing. Hadoop looked so vast, even If I wanted to start fix some defect I don't know where to start from. If I could have got some hand holding that might have helped ...*," said one of them;

"*I got no answer for my question*," complained the other.

From the people who were keen to contribute we also found a case of a user that resumed his contributions to the project. He said that since 2011 he is back to the project answering questions and discussing them in the mailing list.

V. Threats To Validity

This section discusses the threats to validity that may have influenced the study. The next three subsections present threats to internal, external and construct validity. The risks related to internal validity are concerned with factors that may affect the dependent variables without the researcher's knowledge [21]. The risks to external validity are related to the ability to generalize the results of the experiments to a wider population [21]. The risks to construct validity concern the relation between the concepts and theories behind the experiment and what is measured and affected [21].

A. *Internal Validity*

Despite the relatively large collection period, we identified just a few newcomers who remained in the project. This small number of newcomers and messages sent by them can influence the outcome due to the low data density.

Other factors could have been considered as reasons why newcomers leave. To reduce this threat, a survey was conducted with the dropouts. However, the low response rate did not allow a more complete evaluation.

B. *External Validity*

The validity of this study is limited to the project Hadoop Common. The conclusions and discussions presented are specific to the project. For more generic results, it is necessary to analyze a representative sample of projects and different analysis periods.

C. *Construct Validity*

The measures used in this paper may not be the best way to show the results, and can have different interpretations. We did not find other studies providing other means that enabled different ways to measure or to compare/confirm our results.

The manual classification of newcomers' questions and answers is subject to errors, as they were performed by humans. We cannot guarantee that the classification has covered all questions and answers types. To reduce this threat questions and answers were analyzed independently by two researchers who discussed until a consensus on the classification.

The timeframes chosen may have affected the observations. Changing intervals size or their start and end date may produce different result. Users who made punctual contributions at different intervals and may have been misclassified as newcomers, may have caused some bias in the analysis.

Some questions asked by newcomers may have not being classified as "questions" in the mailing list. This can occur because we classified only the first message in a thread as a question. However, there may be cases in which a question appears as a reply to an existing discussion.

Users can have two or more usernames in issue manager or join the list of developers with two different email addresses. There may also be cases where the same user is registered in the issue manager and the list of developers with different emails. To reduce this threat, the email collector uses some heuristics to combine different addresses used by a single user. A manual analysis was conducted to combine emails used in the mailing list and Jira usernames.

VI. Conclusion

This paper presented an analysis of newcomers' dropout in open source software projects, observing the Hadoop Common project. We could point out some possible reasons for leaving the project, based on first interactions in the project's mailing list and issue manager.

The results showed that the newcomers' retention rate is small – 18% in the mailing list and 13% in the issue manager. By reading the discussions initiated by dropouts, we realized that part of the newcomers had no intention to join the project. Many of them sent only one message to the mailing list to clarify a specific doubt, received correct answers, sometimes they expressed thanks and did not return. To deeper analyze the reasons why newcomers leave the project we contacted the dropouts and sent to them a questionnaire.

We classified the 13 answers received by the newcomers and could perceive that negative messages or message directing to external links may influence dropouts. Six of the 13 responses (46.15%) revealed newcomers unhappy with the answers received, because they could not find the support needed to start.

Among the factors that influence the decision to abandon, we found evidence that receiving inadequate answers and the experience of the respondent affect the decision of newcomers. In contrast, we concluded that the lack of answer does not have much influence.

After reading the mailing list discussions, we can conclude that newcomers that are interested in contributing to the project but have questions answered by other newcomers, are more likely to quit. We noticed that some newcomers answer questions with wrong information or merely replicate the intention to join the project. However, for a more realistic analysis it would be necessary to interview each member who left the project in order to understand their reasons.

Regarding the newcomers who remained in the project, a deeper analysis on their activities showed that they tend to contribute more and diversify their actions. Thirteen of the 24 newcomers who remained (54.17%) were active until May 2012 in Jira and three of them have become committers. Their participation in mailing lists (even for newcomers who initiated the participation by this means) decreased over time.

This work was the first step to understand how developers collaborate on open source projects, and how the newcomers behave in these communities. Understanding this behavior is important to create recommendation engines to better receive the newcomers and increase their retention in open source software projects.

Our future work includes conducting a qualitative study to analyze the factors that lead newcomers to leave or to remain in a project. We will conduct interviews with core members, dropouts, and short-term contributors to understand the joining process, the interaction patterns, and the demands and needs of newcomers when they wish to start contributing to such projects. This understanding will enable us to define on mechanisms and tools to better support the newcomers' first steps.

ACKNOWLEDGMENT

The authors of this paper thank Fundação Araucária for the financial support. Marco Aurélio Gerosa receives individual grant from Brazilian National Research Council (CNPq).

REFERENCES

[1] Y. Park and C. Jensen, "Beyond pretty pictures: Examining the benefits of code visualization for Open Source newcomers", in 5th IEEE Workshop on Visualizing Software for Understanding and Analysis (VISSOFT), pp. 3-10, 2009.

[2] B. Dagenais, H. Ossher, R. K. E Bellamy, M. P. Robillard and J. P. de Vries, "Moving into a new software project landscape", in Proc. of ACM/IEEE 32nd International Conference on Software Engineering, pp. 275-284, 2010.

[3] W. Scacchi, "Understanding the requirements for developing open source software systems", in IEE Proceedings Software, v. 149, no. 1, pp. 24-39, 2002.

[4] G.C. Čubranić, C. Murphy, K. Singer and K. S. Booth, "Hipikat: a project memory for software development", IEEE Trans. Softw. Eng., v. 31, no. 6, pp. 446-465, 2005.

[5] I. Steinmacher, I. S. Wiese and M.A. Gerosa, "Recommending Mentors to Software Project Newcomers". In Proc. of 3rd International Workshop on Recommendation Systems for Software Engineering (RSSE '12). IEEE CS, pp. 63-67, 2012.

[6] G. von Krogh, S. Spaeth and K .R. Lakhani, "Community, joining, and specialization in open source software innovation: a case study", Res. Policy, v. 32, no. 7, pp. 1217-1241, 2003.

[7] P. Vora and N. Komura, "The n00b Wikipedia Editing Experience". In *Proc. of the 6th International Symposium on Wikis and Open Collaboration*, Article No 36 , 3 pp., 2010.

[8] J. Thom-Santelli, D. Cosley and G. Gay, "What do you know?: experts, novices and territoriality in collaborative systems". In *Proc. of the 28th international conference on Human factors in computing systems* (CHI '10), pp. 1685-1694, 2010.

[9] A. Halfaker, A. Kittur and J. Riedl, "Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work". In *Proc. of the 7th International Symposium on Wikis and Open Collaboration* (WikiSym '11), pp. 163-172, 2011.

[10] C. Jensen, S. King and V. Kuechler, "Joining Free/Open Source Software Communities: An Analysis of Newbies' First Interactions on Project Mailing Lists", in Proc. of the 44th Hawaii International Conference on System Sciences, pp. 1-10, 2011.

[11] G. Canfora, M. Di Penta, R. Oliveto and S. Panichella. "Who is going to mentor newcomers in open source projects?" In Proc. of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering . ACM, 11 pp, 2012.

[12] Y. Malheiros, A. Moraes, C. Trindade and S. Meira, "A Source Code Recommender System to Support Newcomers," In IEEE 36th Annual Computer Software and Applications Conference (COMPSAC), IEEE, pp.19-24, 2012

[13] J. Wang; A. Sarma and "Which bug should I fix: helping new developers onboard a new project." In Proc. of the 4th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE '11). ACM, pp. 76-79, 2011.

[14] A. Begel and B. Simon. "Novice software developers, all over again." In Proc. of the 4th International Workshop on Computing Education Research (ICER '08). ACM. pp. 3-14, 2008

[15] K. Nakakoji, Y. Yamamoto, Y. Nishinaka, K. Kishida and Y. Ye, "Evolution patterns of open-source software systems and communities". In *Proc. of the International Workshop on Principles of Software Evolution* (IWPSE '02). ACM, pp. 76-85, 2002.

[16] C. Jensen and W. Scacchi, "Role Migration and Advancement Processes in OSSD Projects: A Comparative Case Study". Proc. of the 29th International Conference on Software Engineering, pp. 364-374, 2007.

[17] C. Jergensen, A. Sarma and P. Wagstrom, "The onion patch: migration in open source ecosystems". In *Proc. of the 19th ACM Symposium and the 13th European conference on Foundations of Software Engineering*, pp. 70-80, 2011.

[18] M. Zhou and A. Mockus, "What make long term contributors: Willingness and opportunity in OSS community," In 34th International Conference on Software Engineering (ICSE), pp.518-528, 2012

[19] A. Schilling, S. Laumer and T. Weitzel, "Who Will Remain? An Evaluation of Actual Person-Job and Person-Team Fit to Predict Developer Retention in FLOSS Projects," In 5th Hawaii International Conf. on System Sciences, pp.3446-3455, 2012.

[20] Y. Qu, C. Huang, P. Zhang and J. Zhang, "Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake". In *Proc. of the Conference on Computer supported cooperative work* (CSCW '11), 2011, pp. 25-34, 2011.

[21] C. Wohlin, P. Runeson and M. Höst, "Experimentation in Software Engineering: An Introduction". Kluwer Academic Publisher, 2000.