Mining Email Social Networks in Postgres

Christian Bird, Alex Gourley,
Prem Devanbu, Michael Gertz
Dept. of Computer Science, Kemper Hall,
University of California, Davis,
Davis, California Republic.
cabird,devanbu@ucdavis.edu

Anand Swaminathan
Graduate School of Management,
University of California, Davis,
Davis, California Republic.
aswaminathan@ucdavis.edu

ABSTRACT

Open Source Software (OSS) projects provide a unique opportunity to gather and analyze publicly available historical data. The Postgres SQL server, for example, has over seven years of recorded development and communication activity. We mined data from both the source code repository and the mailing list archives to examine the relationship between communication and development in Postgres. Along the way, we had to deal with the difficult challenge of resolving email aliases. We used a number of social network analysis measures and statistical techniques to analyze this data. We present our findings in this paper.

Categories and Subject Descriptors

D.2.8 [**Software Engineering**]: Metrics—*Empirical, Open Source*

General Terms

Human Factors, Measurement

Keywords

Open Source, Social Networks

1. INTRODUCTION

We have created a framework for mining publicly available OSS project data and using the results to answer questions about the activity in OSS projects. In an effort to test and validate our hypotheses based on earlier results from the Apache HTTP Server project, we have performed the same mining and analysis process on the Postgres SQL Server project[1]. We have mined source code repository activity and used mailing list archives to create a social network of developers and contributors to Postgres. We are hoping to answer the following questions:

- *Are the distributions of email activity, and the social network measures (such as in-degree and out-degree) similar in both projects?*
- *Is there a correlation between mailing list activity and development activity?*
- *Do the developers have significantly higher status than non-developers in the email social network?*

[1] http://www.postgresql.org

MSR'06, May 22–23, 2006, Shanghai, China.
ACM 1-59593-085-X/06/0005.

2. DATAMINING

The Postgres project is a stable and widely used piece of open source software with archives dating back to 1996. In order to mine social data from mailing list archives, we need various forms of information about each message sent on the list. Specifically, we need to know who sent a message, when the message was sent and if the message was sent in reply to a previous message. Mailing lists accomplish this "message linking" by assigning each message a unique message ID. Message a is a reply to message b if there is an *In-Reference-To* or *In-Reply-To* header in a's headers that has b's message ID in it. Unfortunately, although the mailing list archives for Postgres began in January of 1997, this method of using message ID's did not begin until January, 1998. We therefore restricted our mining effort to the time period from January, 1998 to February, 2006.

For the period in question, we found that there were 111,020 messages sent on the mailing lists (over 1,100 per month or 35 per day on average). We were able to parse 110,260 messages (approximately 99.3%). The remaining 760 messages were unparseable mostly due to malformed headers that lacked the *Message-ID* header crucial to our social network reconstruction. However, we believe that our results would not be significantly affected by the small proportion of unparseable messages

A serious hurdle to data collection was email aliasing. We found that during this time period, messages were sent to the list from 4,075 unique email addresses. Mailing list participants often use multiple email addresses, so for our analysis to be a valid, we need to remove the aliasing from the data. Each message sent on a mailing list has a name and an address of the sender. We have constructed an algorithm that uses a number of heuristics (such as address similarity, edit distance between names, etc.) and clustering to detect sets of email aliases that belong to one person. The results of this process are manually verified and editted for better results. Although it is not possible to completely remove aliasing based on name and address heuristics, (it's possible that the name, email pair ⟨shiby thomas, sthomas@cise.ufl.edu⟩ is the same person as ⟨david wetzel, dave@turbocat.de⟩, in which case our algorithm would miss it) we believe that our process is fairly accurate. Details of the aliasing algorithm are presented in the companion MSR paper[2]. After removing aliases we found 3,293 unique "identities" that we believe each correspond to one person. We used a similar technique in conjunction with online research (most OSS projects have a credits file or a developer info page[3]) to match CVS accounts to mailing list identities.

[2] http://wwwcsif.cs.ucdavis.edu/~bird/papers/msr06.pdf
[3] The email addresses of many Postgres developers can be found at http://www.postgresql.org/developer/bios

	changes	srcChanges	docChanges	outdegree	indegree	betweenness	mean	min	max
changes	1	0.974	0.936	0.768	0.782	0.765	3247	0	35883
srcChanges	0.974	1	0.885	0.769	0.785	0.769	2016	0	23345
docChanges	0.936	0.885	1	0.747	0.759	0.767	1231	0	12538
outdegree	0.768	0.769	0.747	1	0.992	0.948	0.0115	0	0.0679
indegree	0.782	0.785	0.759	0.992	1	0.956	0.0092	0.0001	0.0506
betweenness	0.765	0.769	0.767	0.948	0.956	1	.0246	0	0.2634

Figure 1: Cross-correlation table, (using Spearman's rank correlation) showing the relationship between the total number of changes, the changes to source, changes to documents, relative in-degree, relative out-degree, and betweenness. Average, min, and max are also shown. n=25

In addition to mining mailing list data, we also gathered data from the source code repository of Postgres (which uses CVS as its version control mechanism). During the period of interest, 26 CVS accounts were used. We were able to match email addresses to all but one of these. According to the developers[4], the *pgsql* account is used only to tag and package releases, and is not represented on the mailing list so we do not include it in our analysis. We tracked development by counting the number of changes to files over time and found 83,359 changes made to 4,108 files over the course of the time studied.

3. RESULTS

We constructed a social network based on the messages that were sent and replied to on the mailing lists. Three commonly accepted social network metrics were run on the resulting network on a per node basis; in-degree, out-degree, and betweeness. In general, developers had higher levels of all three metrics by at least an order of magnitude over non-developers. This indicates that developers hold positions of high status in the social network of contributors by multiple measures. A Student's t-test shows a significant statistical difference in the in-degree, out-degree and betweenness values for the population of developers and the population of non-developers. Figure 2 shows the social network of highly active Postgres mailing list participants (ties represent at least 150 messages between participants). The two most central participants, Bruce Momjian and Tom Lane, are also the most active CVS committers. The majority of the other participants in this network are also CVS committers. There are, however, nodes in this network that are not CVS committers and not all committers are in the network.

In addition, Figure 1 shows high levels of correlation between the social network measures and CVS activity. Similar to the results of our study of the Apache HTTP Server project, The social network metrics are higly correlated with source file changes. Unlike Apache, however, document file changes correlate to an equal degree. This may be due to the lower number of CVS developers (25 versus 78) and the fact that in this project, many developers work on both source code and documentation. Another possibility may be the number of document translations and how they are dealt with. We plan to mine other OSS projects to investigate this phenomenon futher.

We also examined the distribution of people with in-degree, outdegree, number of sent messages and number of replies. Consistent with data from the Apache project, each distribution exhibits a power-law character. This gives us confidence in our mining methodology and analysis as social processes tend to be characterized by power-laws.

[4]Marc Fournier and Tom Lane both explained this in responses to our inquiries regarding this account

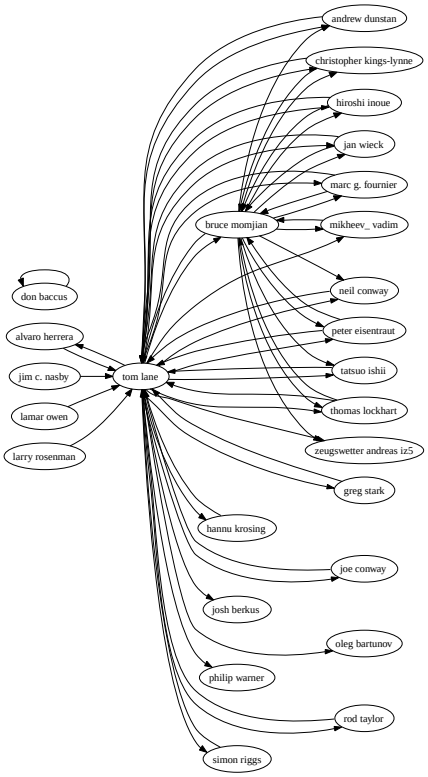

Figure 2: Social network of highly active Postgres mailing list participants

4. CONCLUSION

After mining and analyzing mailing list and source code repository data for the Postgres project we found that the distributions of email activity and social network measures were similar to those found in the Apache project. Our results indicate that developers hold higher levels of status in the social network than non-developers. We also found high correlations between various social network status metrics and source code development. This is consistent with our findings from the Apache project and gives us confidence in our hypotheses and methods. The discrepancy in correlation of document changes with social network status between projects indicates an area that requires further investigation

There is a significant body of related work, which is omitted from this summary for brevity. We refer the reader to our companion paper, "Mining Email Social Networks" accepted to MSR 2006 (located at http://wwwcsif.cs.ucdavis.edu/~bird/papers/msr06.pdf) for details.