What Community Contribution Pattern says about Stability of Software Project?

Ayushi Rastogi  
IIIT-Delhi  
Email: ayushir@iiitd.ac.in

Ashish Sureka  
IIIT-Delhi  
Email: ashish@iiitd.ac.in

Abstract—Free/Libre Open Source Software (FLOSS) community management is an important issue. Contributor churn (joining or leaving a project) causes failure of the majority of software projects. In this paper, we present a framework to characterize stability of the community in software maintenance projects by mining Issue Tracking System (ITS). We identify key stability indicators and propose metrics to measure them. We conduct time series analysis on metrics data to examine the stability of the community. We model community participation patterns and forecast future behavior to help plan and support informed decision making. We present a case study of four years data of Google Chromium Project and investigate the inferential ability of the framework.

I. RESEARCH MOTIVATION AND AIM

Research shows that a majority of Free/Libre Open Source Software (FLOSS) projects fail due to lack of sustained contributors [1][2][3]. Lack of sustained contributors in FLOSS projects causes schedule overrun and team regeneration [3], which adversely affects the existence and quality of the projects [2]. In FLOSS projects, contributors join and leave projects (contributor churn) at will [3][4]. The voluntary participation in FLOSS projects makes it hard to understand contributor churn [1] thereby escalating the challenge of estimating its effects. The inevitable and non-trivial nature of estimating the effect of contributor churn makes it hard to characterize the stability of software maintenance projects. The research aim of the work presented in this paper is:

1) To investigate metrics to objectively characterize community stability and key stability indicators by mining Issue Tracking System.

2) To demonstrate the inferential ability of time series data on key stability indicators for investigating the stability of the community and estimate future contribution to support informed decision making.

II. STABILITY CHARACTERIZATION METRICS

We conduct experiments on four years data of Google Chromium Issue Tracking System (GC-ITS) extracted from January 1, 2009 to December 31, 2013 and measure it quarterly (Table I). Attrition, Regeneration, and Retention are three KSIs (Key Stability Indicators). In this section, we propose metrics, and compute results. Each measure of a metric uses data from two consecutive time intervals as input. Thus, we generate time-series data for 15 time intervals (out of 16 data points for each metric measured quarterly) for analysis. Attrition Rate (AR) for time interval $T_t$ measures in percentage the fraction of contributors who left the project in time interval $T_t$ to the total number of contributors who participated in preceding time interval $T_{t-1}$. In set notations, the attrition rate is the cardinality of set difference of contributors in two consecutive time intervals $T_{t-1}$ and $T_t$ divided by the cardinality of all participant contributors in the time interval $T_{t-1}$ multiplied by 100 (refer Equation 1). The unit of Attrition Rate is quarter$^{-1}$.

$$AR_{T_t} = \frac{|Con_{T_{t-1}} \setminus Con_{T_t}|}{|Con_{T_{t-1}}|} \times 100 \quad where \quad t > 1$$

(1)

Regeneration Rate (RgR) for time interval $T_t$ measures the rate (in percentage) at which new contributors start participating in ITS in two consecutive time intervals under analysis. In set notations, we define regeneration rate for time interval $T_t$ as fraction of cardinality of set difference of contributors in $T_t$ from contributors in time interval $T_{t-1}$ to the cardinality of contributors in time interval $T_t$ multiplied by 100. The unit of measurement is quarter$^{-1}$.

$$RgR_{T_t} = \frac{|Con_{T_t} \setminus Con_{T_{t-1}}|}{|Con_{T_t}|} \times 100 \quad where \quad t > 1$$

(2)

Retention Rate (RtR) for time interval $T_t$ measures percentage of contributors retained out of all participants in time interval $T_{t-1}$. In set notations, Retention Rate for time interval $T_t$ is the ratio of cardinality of intersection of contributors (Con) in time interval $T_t$ and preceding time interval $T_{t-1}$ to the cardinality of contributors (Con) in time interval $T_{t-1}$. The unit of Retention Rate is quarter$^{-1}$.

$$RtR_{T_t} = \frac{|Con_{T_t} \cap Con_{T_{t-1}}|}{|Con_{T_{t-1}}|} \times 100 \quad where \quad t > 1$$

(3)

The three metrics proposed in the study cannot be compared. To help cross comparison, we normalize the metrics on the union of contributor count in two consecutive time intervals $T_t$ and $T_{t-1}$. The normalization facilitates comparison at the cost of affecting the scores. We conduct analysis of modified

\footnote{Set difference is set of all contributors who worked in one time interval ($T_{t-1}$) but did not continue participation in next time interval ($T_t$).}
Regeneration Rate (%)

<table>
<thead>
<tr>
<th>Contributor</th>
<th>Reporter</th>
<th>Owner</th>
<th>CC’ed</th>
<th>Commenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009/g23704:06</td>
<td>2009/g23707:09</td>
<td>2009/g23710:12</td>
<td>2010/g23701:03</td>
<td>2010/g23704:06</td>
</tr>
<tr>
<td>2010/g23707:09</td>
<td>2010/g23710:12</td>
<td>2011/g23701:03</td>
<td>2011/g23704:06</td>
<td>2011/g23707:09</td>
</tr>
<tr>
<td>2011/g23710:12</td>
<td>2012/g23701:03</td>
<td>2012/g23704:06</td>
<td>2012/g23707:09</td>
<td>2012/g23710:12</td>
</tr>
</tbody>
</table>

Regeneration Rate of contributors across four years

The three metrics can be restated as:

\[
|AR| = \frac{|\text{Con}_{T_t} \setminus \text{Con}_{T_{t-1}}|}{|\text{Con}_{T_{t-1}} \cup \text{Con}_{T_t}|} \times 100 \quad \text{where } t > 1
\]

\[
|RgR| = \frac{|\text{Con}_{T_t} \setminus \text{Con}_{T_{t-1}}|}{|\text{Con}_{T_{t-1}} \cup \text{Con}_{T_t}|} \times 100 \quad \text{where } t > 1
\]

\[
|RtR| = \frac{|\text{Con}_{T_t} \cap \text{Con}_{T_{t-1}}|}{|\text{Con}_{T_{t-1}} \cup \text{Con}_{T_t}|} \times 100 \quad \text{where } t > 1
\]

Together the three metrics sum to 100 as shown below:

\[AR + RgR + RtR = 100\]

Figures 1, 2, and 3 show Attrition Rate, Regeneration Rate, and Retention Rate respectively for 15 time intervals. The horizontal axis of the plot represents consecutive time intervals measured quarterly and the vertical axis shows metric score in percentage. Colored lines (refer to the legend) present metric scores for contributors and four roles of the contributor (reporter, owner, cc’ed, and commenter).

RQ1: Do we observe high Attrition Rate in Google Chromium Project?

In Figure 1, contributor Attrition Rate in Google Chromium Project (shown in black) ranges from 27% to 47% with 34.7% mean and 4.07% standard deviations. It implies that every three months one-third of the contributors discontinue to participate in ITS. The high attrition rate in GC-ITS raises concerns regarding the stability of the project.

RQ2: Do we observe comparable Attrition Rates for all roles? If no, how does it vary with role relevance?

In Figure 1, we observe a marked difference in Attrition Rates for four roles where role relevance follows the order: owner ≥ cc’ed ≥ commenter ≥ reporter. We observe that Attrition Rate increases with decreasing relevance of the role. We see minimum Attrition Rate for owner (shown in blue) and maximum for reporter (shown in red). This observation follows the intuition and research results of previous research and validates the insight. Attrition Rate for reporter varies from 28.9% to 44.2% with mean 37.5% and standard deviation 4.3%. A high Attrition Rate for the role of reporter presents an open question that is, whether the project can compensate for the resources lost and generate resources to meet the increase in requirements.

RQ3: Is high Attrition Rate accompanied by high Regeneration Rate?

In Figure 2, we see that in GC-ITS contributor Regeneration Rate fluctuates from 30.0% to 45.4% with mean 38.7% and 4.2% standard deviation. This indicates that the resource lost is regenerated. However, like RQ2, we observe a marked difference in regeneration rates for four roles and how it may impact the stability. Our next RQ attends anomalous behavior, if any, in Attrition Rate and Regeneration Rate patterns for four roles.

RQ4: Do we observe anomaly, if any, in the temporal behavior of Attrition Rate and Regeneration Rate for four roles of contributors?

We observe that, for the owner, Regeneration Rate is always higher than Attrition Rate while the same does not hold true for the other three roles. For reporter, commenter, and cc’ed, we observe cross overs and fluctuations. Cross overs and fluctuations may point to concerns.

RQ5: Do we observe changes in lag between Attrition Rate and Regeneration Rate?

We observe that for owners, Attrition Rate remains low and consistent for most parts while a constant dip in Regen-
eration Rate is observed. Similarly for other roles we observe decrease in lag between Attrition Rate and Regeneration Rate accompanied by crossovers and fluctuations. Thus, in last few time intervals contributors lost due to high Attrition Rate is not replenished with high Regeneration Rate. Also, though we do not observe this trend for owner their participation pattern is also heading on similar lines.

**RQ6: Do we observe increasing or decreasing trend in Retention Rate over time?**

Figure 3 shows that contributor Retention Rate ranges from 23% to 32% with wide variations across four activities. Retention Rate is high for owner and cc’ed compared to retention rate for reporter and commenter. One observation from this graph is that commenters stay in projects longer than the contributors who report bugs.

**RQ7: Do we observe changes in the relationship between Attrition Rate, Regeneration Rate, and Retention Rate? What implications can we draw from such observations?**

To answer this RQ, we approximate the trend by fitting a linear model to the temporal data of Attrition Rate, Regeneration Rate, and Retention Rate for 15 time intervals. The simple linear regression model creates an approximation of the trend with correlation coefficients \( r = 0.7314258, -0.7198569, \) and \(-0.7001004\) respectively. The fitted model is statistically significant with p-values 0.004866, 0.008281 and 0.004246 respectively. The fitted model provides an approximation of the trend. Moreover, the residuals of the model have mean close of zero and a finite variance indicating that the residuals do not follow any pattern.

In Figure 4, we observe that Regeneration Rate is decreasing, Attrition Rate is increasing, and Retention Rate is decreasing. The interpretations that we draw from this plot are that over time contributors leaving the project are not regenerated and the retention of contributors is decreasing over time. In a nutshell, the contributors are slowly moving away from the project. This observation may influence the stability of the project depending on whether the project is in initial phase or has attained maturity. Decreasing contributor participation during initial phase may indicate decrease in interest and hence a decrease in popularity of the project.

**RQ8: Do we observe an increase in contribution overhead on each contributor with time?**

In Table I, we observe an increase in the contributor count with increase in issue count. However, the increase is not proportional. One inference that we draw from the trend is that over time there is an increase in workload on contributors. The increase in workload may be welcoming or may further lead to decrease in contributor participation.

### III. Estimating Future Participation

Planning is the response of forecasting. Estimating future participation helps project managers plan ahead and facilitate informed decision making. Research in FLOSS projects shows that contributor participation is a volunteer activity where contributors join or leave project at will. Individual participation or contribution at microscopic level cannot be estimated. We investigate the correlation between the community participation pattern and stability of the community. Because we model historical participation data, the study does not capture changes in participation pattern due to changes in external factors, like similar project gaining popularity etc. Although these changes are reflected in subsequent time intervals.

We propose statistical forecasting models that along with judgmental forecasting of decision makers (to capture environmental factors), will present a complete understanding of factors that influences the stability of the project. We conduct experiments on GC-ITS dataset for 15 time intervals. We enumerate three statistical prediction models, justify the choice, implement, compute accuracy, compare the models for goodness of fit and forecast accuracy and present visualizations and analysis. We present results for the three metrics and visualize one metric. Given the small data size we choose 12 (out of 15) data points for training and the rest 3 for testing. The baseline model (Model-I) for prediction assumes (also conducted experimentally) that the contributor participation pattern follows normal distribution. We compute the mean and standard deviation of the time series data with 99% and 95% confidence interval to estimate participation patterns in following three time intervals (refer Table II). In Table II we see that the simple statistical model computes future participation with 99% confidence interval. Figure 5 shows the simple statistical model and the confidence interval for Attrition Rate, Regeneration Rate, and Retention Rate metrics respectively. From statistical significance testing, a p-value less than 0.01 indicates that the linear fit provides a good...
TABLE II: Model-I: Simple Statistical Model [SD=Standard Deviations; Conf=Confidence Interval; Obs=Observations for next three time intervals]

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Mean</th>
<th>SD</th>
<th>99% Conf</th>
<th>95% Conf</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>34.02</td>
<td>3.68</td>
<td>24.56-43.47</td>
<td>26.80-41.23</td>
<td>36.54, 32.65, 43.04</td>
</tr>
<tr>
<td>RgR</td>
<td>39.38</td>
<td>3.73</td>
<td>29.78-48.97</td>
<td>32.06-46.70</td>
<td>36.54, 32.65, 43.04</td>
</tr>
<tr>
<td>RIR</td>
<td>28.27</td>
<td>2.40</td>
<td>22.10-34.44</td>
<td>23.57-32.97</td>
<td>26.64, 28.29, 23.84</td>
</tr>
</tbody>
</table>

Fig. 6: Model II: Linear Regression Model to analyze trends in time series data of Attrition Rate

For statistical time series modeling, exponential smoothing and ARIMA models are two popular techniques. The small dataset available for prediction makes estimating parameters for ARIMA modeling difficult without overfitting. Thus ruling out the choice of ARIMA modeling for the study. Exponential smoothing, on the other hand, uses recent observations for prediction. Therefore, for third model, we choose exponential smoothing to predict contributor participation patterns. To implement Model III, we use HoltWinters function in R to compute additive exponential smoothing with trend and level and without seasonal components. We forecast participation in three upcoming time intervals with 80% and 95% confidence. Figure 7 shows the observed and expected contribution patterns. A manual inspection of Model-II and Model-III shows slight improvement in prediction accuracy of first time interval (Figures 6 and 7) of Attrition Rate metric. However, not much difference is observed for second, and third time intervals. We observe similar trends for Regeneration Rate and Retention Rate metrics. In the code snippet that follows, we compare the prediction accuracies of Model-II and Model-III for different horizons (1, 2, and 3). We conduct Daibald Mariano Test to compare the predictive accuracies of models (implemented in R). We hypothesize that the predictive accuracy of linear regression model is less than the predictive accuracy of exponential smoothing model. Tables III and IV displays accuracy of prediction of Exponential Smoothing and Linear Regression Model

IV. CONCLUSION

We presented a generalized framework to characterize the stability of software maintenance projects on community participation patterns by mining Issue Tracking System. We proposed metrics for three key stability indicators, investigated temporal trends, and presented our analysis. We modeled participation patterns, and predicted future participation. We proposed three statistical prediction models, compared the models for prediction accuracy, and stated the inferences.

REFERENCES