What Community Contribution Pattern says about Stability of Software Project?

Ayushi Rastogi and Ashish Sureka

Indraprastha Institute of Information Technology, Delhi (IIITD), India

ayushir@iiitd.ac.in, ashish@iiitd.ac.in
Presentation Outline

1. Research Motivation and Aim
2. Stability Characterization Metrics
3. Research Questions
4. Estimating Future Participation
5. Conclusion
6. References
Presentation Outline

1. Research Motivation and Aim
2. Stability Characterization Metrics
3. Research Questions
4. Estimating Future Participation
5. Conclusion
6. References
Research Motivation and Aim

1. Majority of Free/Libre Open Source Software (FLOSS) projects fails due to lack of sustained contributors [1][2][3]

2. Lack of sustained contributors in FLOSS projects causes schedule overrun and team regeneration [3]

3. Adversely affects the existence and quality of the projects [2]

4. Contributors join and leave projects (contributor churn) at will [3][4].
Research Motivation and Aim - II

The voluntary participation in FLOSS projects makes it hard to understand contributor churn [1] thereby escalating the challenge of estimating its effects. 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.
Presentation Outline

1. Research Motivation and Aim
2. Stability Characterization Metrics
3. Research Questions
4. Estimating Future Participation
5. Conclusion
6. References
Experimental Dataset

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>
<thead>
<tr>
<th>Start Date</th>
<th>End Date</th>
<th>Bug Reports</th>
<th>Contributor Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009-01-01</td>
<td>2009-12-31</td>
<td>025841</td>
<td>13006</td>
</tr>
<tr>
<td>2010-01-01</td>
<td>2010-12-31</td>
<td>044041</td>
<td>26654</td>
</tr>
<tr>
<td>2011-01-01</td>
<td>2011-12-31</td>
<td>051683</td>
<td>28170</td>
</tr>
<tr>
<td>2012-01-01</td>
<td>2012-12-31</td>
<td>057970</td>
<td>30510</td>
</tr>
</tbody>
</table>
1. Each measure of a metric uses data from two consecutive time intervals as input.

2. We generate time-series data for 15 time intervals (out of 16 data points for each metric measured quarterly) for analysis.
Attrition Rate (AR)

1. 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 $T_{t-1}$.

2. 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.

\[ |AR_{T_t}| = \frac{|Con_{T_{t-1}} \setminus Con_{T_t}|}{|Con_{T_{t-1}}|} \times 100 \quad \text{where } t > 1 \quad (1) \]

---

1. 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 (RgR)

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.

2. 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.

\[
|RgR_{T_t}| = \frac{|Con_{T_t} \setminus Con_{T_{t-1}}|}{|Con_{T_t}|} \times 100 \quad \text{where } t > 1
\]
Retention Rate (RtR)

1. Retention Rate (RtR) for time interval $T_t$ measures percentage of contributors retained out of all participants in time interval $T_{t-1}$.

2. 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 $T_{t-1}$.

$$|RtR_{T_t}| = \frac{|Con_{T_t} \cap Con_{T_{t-1}}|}{|Con_{T_{t-1}}|} \times 100 \quad \text{where } t > 1$$ (3)
We normalize the metrics on the union of contributor count in two consecutive time intervals $T_t$ and $T_{t-1}$. Together the three metrics sum to 100.

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

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

\[
|RtR_{T_t}| = \frac{|Con_{T_t} \cap Con_{T_{t-1}}|}{|Con_{T_t} \cup Con_{T_{t-1}}|} \times 100 \quad \text{where } t > 1
\]
Presentation Outline

1. Research Motivation and Aim
2. Stability Characterization Metrics
3. Research Questions
4. Estimating Future Participation
5. Conclusion
6. References
Do We Observe High Attrition Rate in Google Chromium Project?

![Figure 1: Attrition Rate of contributors across Four Years](image)

Figure 1: Attrition Rate of contributors across Four Years
Do We Observe High Attrition Rate in Google Chromium Project?

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.

2. Every three months one-third of the contributors discontinue to participate in ITS.

3. High attrition rate in GC-ITS raises concerns regarding the stability of the project.
Do We Observe Comparable Attrition Rates for All Roles? If No, How Does it Vary with Role Relevance?

1. We observe a marked difference in Attrition Rates for four roles where role relevance follows the order: owner $\geq$ cc’ed $\geq$ commenter $\geq$ reporter

2. We observe that Attrition Rate increases with decreasing relevance of the role

3. We see minimum Attrition Rate for owner (shown in blue) and maximum for reporter (shown in red)

4. Attrition Rate for Reporter varies from 28.9% to 44.2% with mean 37.7% and standard deviation 4.3%.
Regeneration Rate of Contributors across Four Years

Figure 2: Regeneration Rate of Contributors across Four Years
Is High Attrition Rate accompanied by High Regeneration Rate?

1. 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.
2. This indicates that the resource lost is regenerated.
3. We observe a marked difference in regeneration rates for four roles.
Retention Rate of Contributors across Four Years

Figure 3: Retention Rate of Contributors across Four Years
Do We Observe Increasing or Decreasing Trend in Retention Rate Over Time?

1. Contributor Retention Rate ranges from 23% to 32% with wide variations across four activities

2. Retention Rate is high for owner and cc’ed compared to retention rate for reporter and commenter

3. Commenters stay in projects longer than the contributors who report bugs
Changes in Contributor Participation Pattern

Figure 4: Changes in Contributor Participation Pattern
Changes in Contributor Participation Pattern

1. We observe that Regeneration Rate is decreasing, Attrition Rate is increasing, and Retention Rate is decreasing.

2. Over time contributors leaving the project are not regenerated and the retention of contributors is decreasing over time.
Presentation Outline

1. Research Motivation and Aim
2. Stability Characterization Metrics
3. Research Questions
4. Estimating Future Participation
5. Conclusion
6. References
Prediction Model

Figure 5: Model I: Simple Statistical Model to Analyze Trends in Time-Series Data of Attrition Rate
Prediction Model

Table 1: 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>RtR</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>
Linear Regression Model

Figure 6: Model II: Linear Regression Model to Analyze Trends in Time-Series Data of Attrition Rate
Exponential Smoothing Forecasting

Figure 7: Model III: Exponential Smoothing Forecasting to analyze trend in time series data of Attrition Rate
### Prediction Model Accuracy

**Table 2: Accuracy of Prediction of Exponential Smoothing**

<table>
<thead>
<tr>
<th>Metrics</th>
<th>ME</th>
<th>RMSE</th>
<th>MAE</th>
<th>MPE</th>
<th>MAPE</th>
<th>MASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attrition Rate</td>
<td>5.04</td>
<td>6.99</td>
<td>5.04</td>
<td>12.21</td>
<td>12.21</td>
<td>1.19</td>
</tr>
<tr>
<td>Regeneration Rate</td>
<td>-4.68</td>
<td>6.93</td>
<td>5.16</td>
<td>-15.14</td>
<td>16.31</td>
<td>1.15</td>
</tr>
<tr>
<td>Retention Rate</td>
<td>-2.86</td>
<td>3.67</td>
<td>2.86</td>
<td>-11.58</td>
<td>11.58</td>
<td>1.15</td>
</tr>
</tbody>
</table>
## Accuracy of Linear Regression Model

Table 3: Accuracy of Linear Regression Model

<table>
<thead>
<tr>
<th>Metrics</th>
<th>ME</th>
<th>RMSE</th>
<th>MAE</th>
<th>MPE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attrition Rate</td>
<td>-2.20</td>
<td>4.51</td>
<td>3.99</td>
<td>-7.15</td>
<td>11.30</td>
</tr>
<tr>
<td>Regeneration Rate</td>
<td>2.37</td>
<td>5.01</td>
<td>4.40</td>
<td>5.02</td>
<td>11.75</td>
</tr>
<tr>
<td>Retention Rate</td>
<td>1.48</td>
<td>2.20</td>
<td>1.79</td>
<td>5.24</td>
<td>6.55</td>
</tr>
</tbody>
</table>
Presentation Outline

1. Research Motivation and Aim
2. Stability Characterization Metrics
3. Research Questions
4. Estimating Future Participation
5. Conclusion
6. References
Conclusion

1. Predictive accuracy of linear regression model is less than the predictive accuracy of exponential smoothing model.
2. We presented a generalized framework to characterize the stability of software maintenance projects on community participation patterns by mining Issue Tracking System.
3. We presented a generalized framework to characterize the stability of software maintenance projects on community participation patterns by mining Issue Tracking System.
4. We modeled participation patterns, and predicted future participation.
5. We proposed three statistical prediction models, compared the models for prediction accuracy, and stated the inferences.
Presentation Outline

1. Research Motivation and Aim
2. Stability Characterization Metrics
3. Research Questions
4. Estimating Future Participation
5. Conclusion
6. References
Daniel Izquierdo-Cortazar, Gregorio Robles, Felipe Ortega, and Jesus M Gonzalez-Barahona.
Using software archaeology to measure knowledge loss in software projects due to developer turnover.

Andreas Schilling, Sven Laumer, and Tim Weitzel.
Who will remain? an evaluation of actual person-job and person-team fit to predict developer retention in floss projects.

Pratyush N Sharma, John Hulland, and Sherae Daniel.
Examining turnover in open source software projects using logistic hierarchical linear modeling approach.