Survivability of Software Projects in Gnome
A Replication Study

Tom Mens, Mathieu Goeminne, Uzma Raja, Alexander Serebrenik
Context

• Study of ‘macro-level software evolution’
• Study the evolution of large collections of software projects/packages/distributions: GNOME, R, Debian, ...
• Coherent collections of systems: ecosystems
• Social/human aspects of these (eco)systems


Context

• Study of the GNOME ecosystem
• Taking into account the social aspects of software evolution


Goal

- Conceptual replication study of Open Source Software project survivability
  - Question: Can we build a model predicting project survivability?
  - Same research question, evaluated by using a different experimental procedure
  - Original paper: 136 SourceForge projects
  - Our paper: 183 GNOME projects
- Build a predictive model of project inactivity
  - Based on the official Git repositories and bug trackers
Viability dimensions
in the original study

• **Vigor**
  • ‘the ability of a project to grow over a period of time’
  • \# versions / \# years

• **Resilience**
  • ‘the ability of a project to recover from internal and external perturbations’
  • Mean time to react to an issue report

• **Organisation**
  • ‘the amount of structure exhibited by the contributors’ interaction’
  • ‘complexity’ of the relations among contributors

\[
O_{rg} = \sum_p \sum_q \frac{T_{pq}}{T} \log_2 \frac{T_{pq}T}{T_p T_q},
\]

\(O_{rg}\) = Organization of project \(0 \leq O_{rg} \leq 1\).
\(T_{pq}\) = Task originated by \(p\) and completed by \(q\).
\(T_p\) = All the tasks requested by \(p\).
\(T_q\) = All the tasks completed by \(q\).
\(T\) = Total tasks in the project.

• **Status**
  • Explicit on SF
Viability dimensions

our operationalization \((v, r, o) \rightarrow s\)

• Vigor

\[ V(p) = V_{int} + V_{ext} = \frac{\text{commits}(p)}{\text{age}(p)} + \text{authors}(p) \]

• Resilience

\[ R(p) = \frac{\text{conrrbs}(p)}{\text{MTTR}(p)} \]

where \( \text{MTTR}(p) = \sum_{i=1}^{\text{resolved}(p)} \frac{TTR_{(p,i)}}{\text{resolved}(p)} \)

• Organisation

\[ O(p) = \frac{\sum_{i=1}^{\text{resolved}(p)} \text{conrrbs}(p, i)(\text{conrrbs}(p, i) - 1)}{\text{conrrbs}(p)(\text{conrrbs}(p) - 1)} \]

• Status

Implicit: no activity during the last 360 days
## Project selection

<table>
<thead>
<tr>
<th>Constraint</th>
<th># projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing project</td>
<td>1,418</td>
</tr>
<tr>
<td>+ Existing bug tracker</td>
<td>197</td>
</tr>
<tr>
<td>+ Non empty data sets</td>
<td>187</td>
</tr>
<tr>
<td>+ Remove one-day projects</td>
<td>183</td>
</tr>
</tbody>
</table>
Viability Index

\[ VI(p) = \alpha + \beta_1 V(p) + \beta_2 R(p) + \beta_3 O(p) \]
Viability

Descriptive stats after log transformation

\[ VI(p) = \alpha + \beta_1 V(p) + \beta_2 R(p) + \beta_3 O(p) \]
Viability Index

Descriptive stats after log transformation

\[ VI(p) = \alpha + \beta_1 V(p) + \beta_2 R(p) + \beta_3 O(p) \]

Determined by Logistic Regression Analysis such that
- \( VI(p) \) is high (close to 1) if \( p \) is an active project
- \( VI(p) \) is low (close to 0) otherwise
We checked that

- Our model is globally meaningful: at least one of the predictor variables is meaningful.
  - Comparison between Full model (containing the predictors) and Reduced model (containing $\alpha$ only): $H_0$ stating that $\beta_1 = \beta_2 = \beta_3 = 0$ is rejected

- Each of the individual predictor variables is useful
  - $H_0^0, H_0^\gamma, H_0^r, H_0^o$ stating that $\alpha = 0, \beta_1 = 0, \beta_2 = 0$, and $\beta_3 = 0$ are rejected

- Our model fits the data well
  - Goodness of Fit test
Validation

• The 3 dimensions of Viability are significant to predict whether a project is active or inactive in the Gnome ecosystem.

• Good prediction for other projects?

• Stratified random sampling approach

• 20% (7 inactive, 30 active) Gnome projects used to create our model
## Validation

### Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Predicted as active</th>
<th>Predicted as inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>active</td>
<td>111</td>
<td>8</td>
</tr>
<tr>
<td>inactive</td>
<td>3</td>
<td>24</td>
</tr>
</tbody>
</table>

Values are well within statistically acceptable range

- Accuracy: 92%
- Precision: 82%
- Recall: 93%
Weaknesses

- SourceForge is not Gnome
  - active project?
  - versions?
  - are the different results due to the different projects? to the operationalization?
- Only a partial view of project’s history
  - official Git repositories and bug trackers
  - other data sources? mailing lists, StackExchange?
- general principles (e.g., developers involvement) operationalized with simple metrics using a single data source.
Conclusion

• Efficient model for predicting the project activity, based on meta-data and contributors’ involvement

• Different sets of projects → different operationalization

• Work in progress
  • dimensions closer to the original ones (+ extended model) to facilitate comparison
  • Predict the future?
  • Add other data sources (mailing lists, etc.)
References


https://bitbucket.org/mgoeminne/gnome-survivability/downloads/