UMONS
Université de Mons

Software Engineering Lab

THE UNIVERSITY OF ALABAMA

Dept. Information Systems,
Statistics & Management Science

TU/e Technische Universiteit Eindhoven University of Technology
Where innovation starts

Dept. Math. & Computer Science

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

T. Mens, M. Goeminne, *'Analysing Ecosystems for Open Source Software Developer Communities'* in *'Software Ecosystems: Analyzing and Managing Business Networks in the Software Industry'*, Slinger et al, 2013

J.M. Gonzalez-Barahona et al, *'Macro-level software evolution: a case study of a large software compilation'*. Empirical Software Engineering 14(3): 262-285, 2009

M. Lungu et al, *'The Small Project Observatory: Visualizing software ecosystems'*. Sci. Comput. Program. 75(4): 264-275 (2010)

Context

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

M. Goeminne, "*Understanding the Evolution of Socio-technical Aspects in Open Source Ecosystems: An Empirical Analysis of GNOME*". PhD thesis, UMONS, July 2013.

B. Vasilescu et al. "On the variation and specialisation of workload — a case study of the gnome ecosystem community". Empirical Software Engineering 19: 955-1008. 2014.

Goal

- Conceptual replication study of Open Source Software project survivability
  - U. Raja and M. J. Tretter, *"Defining and evaluating a measure of open source project survivability,"* IEEE Trans. Softw. Eng., vol. 38, pp. 163–174, Jan. 2012
  - 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

$(v,r,o) \longrightarrow s$

- **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) ⟶ s

- **Vigor**

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

- **Resilience**

$$R(p) = \frac{contribs(p)}{MTTR(p)}$$

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

- **Organisation**

$$O(p) = \frac{\sum_{i=1}^{resolved(p)} contribs(p,i)(contribs(p,i)-1)}{contribs(p)(contribs(p)-1)}$$

- **Status**

  Implicit: no activity during the last 360 days

Project selection

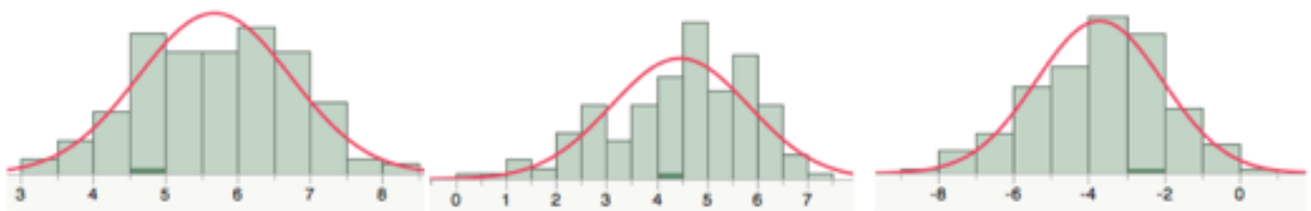
Constraint	# projects
Existing project	1,418
+ Existing bug tracker	197
+ Non empty data sets	187
+ Remove one-day projects	183

Viability Index

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

Viability

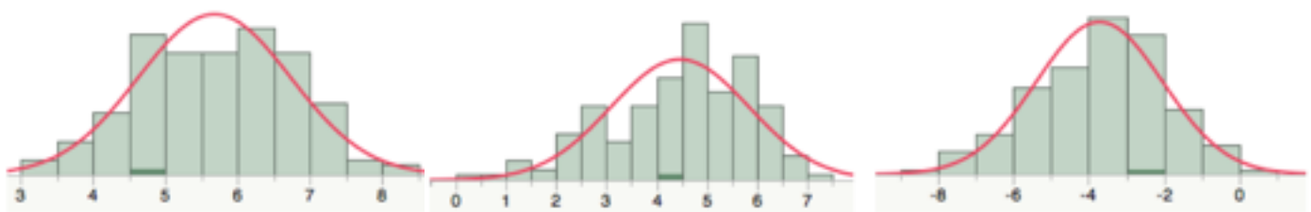
Descriptive stats after log transformation



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

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 α 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^v, H_0^r, H_0° 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

	Predicted as active	Predicted as inactive
active	111	8
inactive	3	24

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

Values are well within statistically acceptable range

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

U. Raja and M. J. Tretter, "Defining and evaluating a measure of open source project survivability," IEEE Trans. Softw. Eng., vol. 38, pp. 163–174, Jan. 2012

M. Goeminne, "*Understanding the Evolution of Socio-technical Aspects in Open Source Ecosystems: An Empirical Analysis of GNOME*". PhD thesis, UMONS, July 2013.

B. Vasilescu et al. "On the variation and specialisation of workload — a case study of the gnome ecosystem community". Empirical Software Engineering 19: 955-1008. 2014.

M. Goeminne et al, "*A historical dataset for the gnome ecosystem*," in MSR (T. Zimmermann, M. D. Penta, and S. Kim, eds.), pp. 225–228, IEEE / ACM, 2013.

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