On the variation and specialisation of workload
The Gnome case

B. Vasilescu, A. Serebrenik, M. Goeminne, T. Mens
Gnome as an ecosystem

- Ecosystem: set of interconnected projects
- ~1400 projects
- ~3000 contributors
- 15 years of activity
How does workload vary across contributors?

• Who are they?
• What do they do?
• How do they do it?

A partial answer by analysing the git repositories.
Who are the contributors?
Identity matching

- Contributors have an account per project repository…
- … and sometimes more than one.
- No explicit links between the accounts, need to guess them.
- Based on names and e-mails found in the git repositories.
Identity matching (cont.)

- (semi) automatic classification techniques.
- Must take into account variations, abbreviations, permutations, misspelling, nicknames, etc.
- No perfect process: even a manually post-checked result can contain false positives and false negatives.
- Since Gnome has no strict identification regulation on the whole, some matches are not detectable without an extra context information. Fictitious example:
  - Robbie Williams <robbiew@gnome.org>
  - Euphegenia Doubtfire <euphegenia@gmail.com>
What do the contributors do?
13 activity types

- Identified by the path, name and extension of the touched files.
- Coding : *.c, *.java, etc.
- Translation : *.po, etc.
- Testing : */test/*, etc.
- ...
How do the contributors contribute?
Metrics

• $\text{APTW}(p,c,t)$ : Number of files touched by the contributor $c$ performing an activity of type $t$ in a project $p$.

•Derived metrics, by aggregation: max, sum, etc.
Workload

- 50% contributors made < 14 changes.
- 1 contributor made 185,874 changes.
The more things you do, the more things you can!

- Correlations
  - Between the number of activity types and the workload.
  - Between the number of projects and the workload.
Favorite activities of contributors having $\geq 14$ changes

- Most frequent contributors specialise in coding and development documentation.
- The other activities are not subject to specialisation.
Favorite activities of contributors having < 14 changes

- Most occasional contributors specialise in translation and coding.
- The other activities are not subject to specialisation.
How strongly do the contributor’s focus?

- Basic measure: $\text{RATW}(c,t)$
- % of the total workload of $c$ dedicated to $t$.

Use of Gini as inequality index:

- Value in $[0, 1[$
  - 0 if the workload is equally distributed.
  - Close to 1 if the workload is concentrated in few activity types.
Contributor’s focus (cont.)

• **Occasional** contributors typically participate in a **single** activity type.

• **Frequent** contributors typically participate in **few** activity types.
To summarise
What did we learn?

• Most contributors are occasional and are involved in only one activity type; few are very active; frequent contributors are involved in few activity types.

• The more things you do, the more things you can.

• Occasional contributors are translators, involved in many projects. Frequent contributors are coders and are involved in few projects.

• And more again in our paper.
How did we do it?

• **Contributor matching**: semi-automatic and automatic methods.

• **Activity identification** based on file path/name/extension rules.

• **Advanced statistical analysis** (among others for the partial ordering of activity types).

• **Specialisation**: aggregation with inequality indices.
In the future

- Add a temporal aspect: How does the contributors’ behaviour change over time?
- Consider subsets of Gnome: subecosystems composed by projects sharing stronger properties than all projects on average: archived, by theme, etc.
- Combine both by studying migration trends.
- …
Thank you

On the variation and specialisation of workload – A case study of the Gnome ecosystem community
B. Vasilescu, A. Serebrenik, M. Goeminne, T. Mens
Empirical Software Engineering
Waiting for being accepted