Collaboration Strength Metrics and Analyses on GitHub
Authors

Natércia A. Batista
Michele A. Brandão
Gabriela B. Alves
Ana Paula C. Silva

**Mirella M. Moro**

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais, Brazil
Introduction

Social Coding is a software development’s approach that allows the collaboration among developers.
Relevance

- Evaluating the strength of the developer's relationship can help to improve:
  - Recommendation algorithms of developers to work in a project
  - Productivity analysis of a development team
How to measure the strength of collaboration between developers on GitHub?
Related Work

▷ Casalnuovo et al. [2015] analyze the productivity of developers in projects
▷ Bartusiak et al. [2016] predict developers collaboration
▷ Tsay et al. [2014] investigate the acceptance of pull requests
Related Work

▷ Such studies measure the strength of interactions differently. However,
▷ none evaluates the best way to measure such strength and
▷ none investigates the correlation among such metrics
Goals

Analyze the strength of social collaboration measured by distinct metrics through different programming languages: insights on relationships’ patterns

1. GitSED - GitHub Socially Enhanced Dataset: curated (filtered on two programming languages), augmented (data not available on GHTorrent), and enriched (social network information).
2. Three new metrics for the strength of social coding collaboration: commits’ number of lines, potential of contribution, prior social interaction.
3. Evaluate all metrics over two social networks for JavaScript and Ruby.
Database

GHTorrent: database from September, 2015

- **Not** including: forked and deleted projects
- Two programming languages
  - JavaScript
  - Ruby
Database

Number of repositories and contributors

- JavaScript: 90,393 repositories, 88,586 contributors
- Ruby: 59,225 repositories, 51,475 contributors
Database

Number of connections

- JavaScript: 3,196,846
- Ruby: 4,620,128

Number of connections
Database

**GitSED**: GitHub Socially Enhanced Dataset

![Database Diagram]

- **PROGRAMMING LANGUAGE**
  - `programming_language_id`
  - `name`
  - `acronym`

- **REPOSITORY**
  - `repository_id`
  - `name`
  - `description`
  - `programming_language_id`
  - `url`
  - `create_date`
  - `end_date`
  - `duration_days`
  - `number_add_lines`
  - `number_del_lines`
  - `number_commits`
  - `number_committers`

- **DEVELOPERS SOCIAL NETWORK**
  - `repository_id`
  - `developer_id_1`
  - `developer_id_2`
  - `begin_contribution_date`
  - `end_contribution_date`
  - `contribution_days`
  - `number_add_lines`
  - `number_del_lines`
  - `number_commits`

- **USER**
  - `user_id`
  - `name`
  - `login`
  - `company`
  - `location`
  - `email`
  - `type`
  - `fake`
  - `deleted`

- **SOCIAL NETWORK METRIC**
  - `programming_language_id`
  - `developer_id_1`
  - `developer_id_2`
  - `NO`
  - `AA`
  - `PA`
  - `SR *`
  - `JCSR *`
  - `JCODSR *`
  - `JWCOSR *`
  - `PC *`
  - `GPC *`
Collaboration Network

▷ Links
  Two developers contribute to same repository

▷ Weights
  By topological and semantic metrics
Collaboration Network
Collaboration Network
Collaboration Network

Weight attributed from proposed metrics.
Topological Properties

▷ Clustering Coefficient  is the tendency of the nodes to cluster
▷ Neighborhood Overlap  computes the strength of the links
▷ Adamic-Adar  more weight to low-degree common neighbors
▷ Preferential Attachment  the rich get richer
▷ Resource Allocation  how a node indirectly influences its pair’s neighborhoods
▷ Tieness  combination of Neighborhood Overlap and weight
SR - Number of Shared Repositories

- Number of shared repositories between a pair of developers

5 repositories
JCSR - Jointly developers contribution to shared repositories

Contribution of a pair of developers relative to the others in a same repository

\[ \frac{2}{2} = 1 \]

\[ \frac{2}{3} = 0.66 \]
JCOSR - Jointly developers commits to shared repositories

▷ Number of commits of a pair of developers in shared repositories

\[ \frac{15 + 3}{18} = 1 \]

\[ \frac{20 + 40}{160} = 0.375 \]
JWCOSR - Jointly developers weighted commit to shared repositories

- Number of lines on commits of a pair of developers in shared repositories

\[
\frac{(|200 - 12| + |300 - 0|)}{|500 - 12|} = 1 \\
\frac{(|500 - 200| + |1.000 - 0|)}{|2.500 - 500|} = 0.65
\]
PC - Previous Collaboration

Collaborations in past repositories relative to the number of developers

\[
\frac{\frac{1}{3} + \frac{1}{2}}{2} = 0.416
\]
GPC - Global Potential Contributions

▷ Potential time of collaboration between a pair of developers in the network

(2 + 3 + 1 + 1 + 6) / 20* = 0.65

5 repositories:
R1: 2 months
R2: 3 months
R3: 1 month
R4: 1 month
R5: 6 months

* longer network collaboration time
Analysis and Results

▷ The average number of connections between developers varies according to the programming language.

▷ Few pairs of developers have interactions in more than one repository.
To define a computational model to measure the strength of collaboration, we must analyze which properties best classify such a strength.

- Combine semantic + topological metrics
  - More importance to strong relationships
  - \{Tieness, Resource Allocation\} + topological
Analysis and Results

- Just one metric should be considered between T_SR, T_JCOSR, T_JWCOSR, T_PC and T_GPC because they are strongly correlated.
Analysis and Results

▷ Just one metric should be considered between T_SR, T_JCOSR, T_JWCOSR, T_PC and T_GPC because they are strongly correlated

▷ Individually, T_JCSR should be considered
Analysis and Results

▷ Just one metric should be considered between JCOSR, JCSR and PC because they are strongly correlated
Analysis and Results

▶ Just one metric should be considered between AA and PA because they are strongly correlated.
Analysis and Results

- Just one metric should be considered between AA and PA because they are strongly correlated
- All metrics SR, GPC e JWCOSR should be considered
Example: Collaboration Ranking

Table 1: Top-10 pairs of developers ranked by T_JWCOSR followed by the other properties values.

<table>
<thead>
<tr>
<th>#</th>
<th>D_1</th>
<th>D_2</th>
<th>T_JWCOSR</th>
<th>T_PC</th>
<th>T_GPC</th>
<th>NO</th>
<th>AA</th>
<th>D_1</th>
<th>D_2</th>
<th>T_JWCOSR</th>
<th>T_PC</th>
<th>T_GPC</th>
<th>NO</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>001</td>
<td>002</td>
<td>0.500</td>
<td>0.500</td>
<td>0.258</td>
<td>0.333</td>
<td>1.443</td>
<td>020</td>
<td>025</td>
<td>0.500</td>
<td>0.333</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>2</td>
<td>002</td>
<td>006</td>
<td>0.485</td>
<td>0.375</td>
<td>0.250</td>
<td>0.333</td>
<td>1.443</td>
<td>021</td>
<td>027</td>
<td>0.484</td>
<td>0.300</td>
<td>0.063</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>3</td>
<td>003</td>
<td>007</td>
<td>0.281</td>
<td>0.417</td>
<td>0.250</td>
<td>0.333</td>
<td>1.443</td>
<td>021</td>
<td>024</td>
<td>0.442</td>
<td>0.375</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>4</td>
<td>004</td>
<td>003</td>
<td>0.273</td>
<td>0.375</td>
<td>0.250</td>
<td>0.333</td>
<td>1.443</td>
<td>022</td>
<td>025</td>
<td>0.395</td>
<td>0.312</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>5</td>
<td>005</td>
<td>008</td>
<td>0.268</td>
<td>0.333</td>
<td>0.258</td>
<td>0.833</td>
<td>4.085</td>
<td>022</td>
<td>028</td>
<td>0.394</td>
<td>0.312</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>6</td>
<td>005</td>
<td>009</td>
<td>0.267</td>
<td>0.278</td>
<td>0.254</td>
<td>0.833</td>
<td>4.085</td>
<td>020</td>
<td>022</td>
<td>0.391</td>
<td>0.312</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>7</td>
<td>001</td>
<td>006</td>
<td>0.266</td>
<td>0.375</td>
<td>0.250</td>
<td>0.333</td>
<td>1.443</td>
<td>023</td>
<td>029</td>
<td>0.369</td>
<td>0.375</td>
<td>0.063</td>
<td>0.333</td>
<td>1.443</td>
</tr>
<tr>
<td>8</td>
<td>005</td>
<td>010</td>
<td>0.266</td>
<td>0.273</td>
<td>0.254</td>
<td>0.833</td>
<td>4.085</td>
<td>024</td>
<td>027</td>
<td>0.353</td>
<td>0.300</td>
<td>0.063</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>9</td>
<td>005</td>
<td>011</td>
<td>0.266</td>
<td>0.292</td>
<td>0.258</td>
<td>0.833</td>
<td>4.085</td>
<td>025</td>
<td>024</td>
<td>0.346</td>
<td>0.333</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>10</td>
<td>005</td>
<td>012</td>
<td>0.266</td>
<td>0.275</td>
<td>0.254</td>
<td>0.833</td>
<td>4.085</td>
<td>026</td>
<td>030</td>
<td>0.345</td>
<td>0.500</td>
<td>0.251</td>
<td>0.333</td>
<td>1.443</td>
</tr>
</tbody>
</table>

Tieness +Jointly developers Weighted Commit to Shared Repositories
Example: Collaboration Ranking

Table 1: Top-10 pairs of developers ranked by T_JWCOSR followed by the other properties values.

<table>
<thead>
<tr>
<th>#</th>
<th>D.1</th>
<th>D.2</th>
<th>JavaScript</th>
<th>T_JWCOSR</th>
<th>T_PC</th>
<th>T_GPC</th>
<th>NO</th>
<th>AA</th>
<th>Ruby</th>
<th>T_JWCOSR</th>
<th>T_PC</th>
<th>T_GPC</th>
<th>NO</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>001</td>
<td>002</td>
<td>0.500</td>
<td>0.500</td>
<td>0.258</td>
<td></td>
<td>0.333</td>
<td>1.443</td>
<td></td>
<td>0.500</td>
<td>0.333</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>2</td>
<td>002</td>
<td>006</td>
<td>0.485</td>
<td>0.375</td>
<td>0.250</td>
<td></td>
<td>0.333</td>
<td>1.443</td>
<td></td>
<td>0.484</td>
<td>0.300</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>3</td>
<td>003</td>
<td>007</td>
<td>0.281</td>
<td>0.417</td>
<td>0.250</td>
<td></td>
<td>0.333</td>
<td>1.443</td>
<td></td>
<td>0.442</td>
<td>0.375</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>4</td>
<td>004</td>
<td>003</td>
<td>0.273</td>
<td>0.375</td>
<td>0.250</td>
<td></td>
<td>0.333</td>
<td>1.443</td>
<td></td>
<td>0.395</td>
<td>0.312</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>5</td>
<td>005</td>
<td>008</td>
<td>0.268</td>
<td>0.333</td>
<td>0.258</td>
<td></td>
<td>0.833</td>
<td>4.085</td>
<td></td>
<td>0.394</td>
<td>0.312</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>6</td>
<td>005</td>
<td>009</td>
<td>0.267</td>
<td>0.278</td>
<td>0.254</td>
<td></td>
<td>0.833</td>
<td>4.085</td>
<td></td>
<td>0.391</td>
<td>0.312</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>7</td>
<td>001</td>
<td>006</td>
<td>0.266</td>
<td>0.375</td>
<td>0.250</td>
<td></td>
<td>0.333</td>
<td>1.443</td>
<td></td>
<td>0.369</td>
<td>0.375</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>8</td>
<td>005</td>
<td>010</td>
<td>0.266</td>
<td>0.273</td>
<td>0.254</td>
<td></td>
<td>0.833</td>
<td>4.085</td>
<td></td>
<td>0.353</td>
<td>0.300</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>9</td>
<td>005</td>
<td>011</td>
<td>0.266</td>
<td>0.292</td>
<td>0.258</td>
<td></td>
<td>0.833</td>
<td>4.085</td>
<td></td>
<td>0.346</td>
<td>0.333</td>
<td>0.250</td>
<td>0.667</td>
<td>2.485</td>
</tr>
<tr>
<td>10</td>
<td>005</td>
<td>012</td>
<td>0.266</td>
<td>0.275</td>
<td>0.254</td>
<td></td>
<td>0.833</td>
<td>4.085</td>
<td></td>
<td>0.345</td>
<td>0.500</td>
<td>0.251</td>
<td>0.333</td>
<td>1.443</td>
</tr>
</tbody>
</table>
In general, a computational model to measure the strength of collaboration should consider:
- Metrics T_JCSR, SR, GPC and JWCOSR
- Just one metric between T_SR, T_JCOSR, T_JWCOSR, T_PC and T_GPC
- Just one metric between AA and PA
- Just one metric between JCOSR, JCSR and PC
Future Work

- More programming languages + forks
- Build full computational model to measure collaboration strength
Thank you!

Mirella M. Moro
mirella@dcc.ufmg.br