The Strength of Social Coding Collaboration on GitHub
Who?!

Gabriela Brant Alves
Graduate Student in Sistemas de Informação at UFMG

gabrielabrant@dcc.ufmg.br
Team

Michele A. Brandão
michelebrandao@dcc.ufmg.br

Diogo M. Santana
diogo.marques@dcc.ufmg.br

Ana Paula C. da Silva
ana.coutosilva@dcc.ufmg.br

Mirella M. Moro
mirella@dcc.ufmg.br
1. Introduction

What and why?
Social Coding is an approach of software development that enables cooperation among developers.
our goal is to use \textbf{correlation} analysis to identify the relationship between \textbf{semantic} and \textbf{topological} properties that measure the \textbf{strength of social coding collaboration}
What we want to validate?

- Higher correlated properties can be used in a **model** to measure the strength of collaborations

- Help to improve:
  - algorithms that **recommend** developers to work in a project
  - analysis of the **productivity** of a developer
2. Data Gathering
GHTorrent

September 15, 2015

1,987,760 Projects (32GB)
We prune the projects to consider only those developed using a JavaScript
529,405
Non-forked projects

90,363
Repositories of JavaScript

37,691
Developers in JavaScript repositories
3. Network Model
Nodes are developers
Developers work in a project
Edges are formed when they contribute to the same repository
Edges are formed when they contribute to the same repository.
Edges are formed when they contribute to the same repository
Edges are formed when they contribute to the same repository
3. Semantic Properties
Given any two developers A and B, the set of repositories they shared is given by R. The metric $\text{SR}(A,B)$ is the total number of repositories that they both worked at, and is given by the cardinality of R set (i.e., $|R|$).
Consider two repositories r1 and r2. The r1 repository is only shared by developers A and B. So, their jointly contribution to r1 ($JCSR(A,B,r1)$) is equal to 1. Instead, r2 is shared by developers C, D and F. So, the jointly contribution given by C and D to r2 ($JCSR(C,D,r2)$) is 0.66.
Jointly developers commits to shared repositories

Given NC(A, r_j) as the total number of commits by A into repository r_j, NC(B, r_j) as the total number of commits by B into repository r_j, and NC(r_j) the total number of commits by any developer into repository r_j.

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

\[
NC(r_j) = 18
\]

\[
\frac{100+40}{160} = 0.875
\]

\[
\frac{40+20}{160} = 0.375
\]
3. Topological Properties
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
3. Results
Observations

- In repositories with a large number of collaborators, all of them will be connected among themselves.
- The developers in the social network tend to form clusters and share a large number of neighbors.
Correlations

Pearson

Spearman
Correlations

SR alone does not capture the strength of a collaboration.
Correlations

They are directly related. Such properties should be considered together in a model to measure the collaborations strength.
Conclusion

- We proposed three semantic properties
- We investigated the correlation of semantic and topological properties
- The number of shared repositories is not a significant indicator of the collaborations strength
  - Should be considered together in a model
- The JCSR and JCOSR are very correlated
Future Work

- Collect number of lines added and deleted from each commit
- Run metrics in other programming languages and compare results
- Develop metrics considering temporal aspects of the network
THANKS!

Any questions?
You can find me at gabrielabrant@dcc.ufmg.br