git2net: Mining Time-Stamped Co-Editing Networks from Large git Repositories

Presentation of work originally published in the Proc. of the 16th Intl. Conf. on Mining Software Repositories [GSS19]

Christoph Gote[1] Ingo Scholtes[2] Frank Schweitzer[3]

Keywords: repository mining; empirical software engineering; network science; data science; collaboration network; co-editing; social network analysis

Extended Abstract

Many software projects use version control systems like *git* to track changes in the developed source code. The analysis of collaboration networks constructed from co-editing relations stored in such *git* repositories can yield deep insights both into team-based software development processes as well as sociological theory. However, tools that allow to conveniently extract such rich, time-stamped collaboration networks for the large corpus of git repositories available are currently missing. Addressing this gap, the contributions of our work are as follows:

- We introduce `git2net`, an Open Source python tool that can be used to mine time-stamped and weighted co-editing relations between developers from the sequence of file modifications contained in *git* repositories. Here, each co-editing relationship $\big(A, B; t, w\big)$ represents developer A modifying w characters of code originally written by another developer B at time t. Utilising a parallel processing model the tool scales to massive software repositories with hundreds of thousands of commits and millions of lines of code.
- Analysing all file modifications contained in the *commit log*, `git2net` generates a database that captures fine-grained information on co-edited code either at the level of lines or contiguous code regions. It further analyses the overlap between co-edited code regions facilitating a character-based proxy estimating the effort behind code modifications. The approach is programming language agnostic.

[1] Chair of Systems Design, ETH Zürich, cgote@ethz.ch
[2] Data Analytics Group, Department of Informatics, University of Zürich, scholtes@ifi.uzh.ch
[3] Chair of Systems Design, ETH Zürich, fschweitzer@ethz.ch

doi:10.18420/inf2019_33

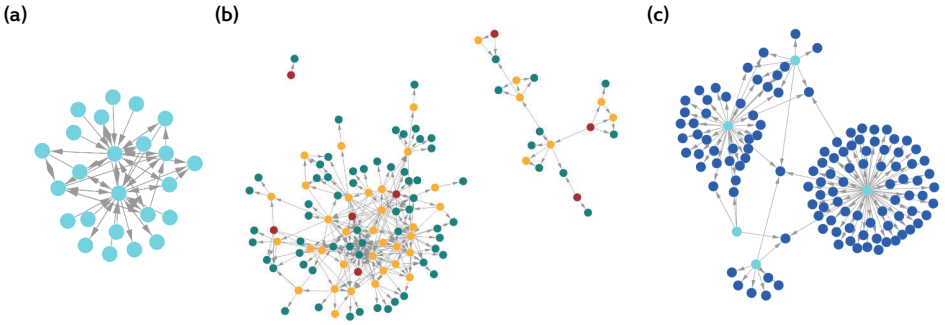

Fig. 1: Three time-aggregated collaboration networks generated by `git2net`. **(a)** Time-aggregated, static, directed network of co-editing relations. **(b)** Directed acyclic graph of edits of the a source code file. **(c)** Bipartite network linking developers (lightblue) to the files that they edited (blue).

- We develop methods to generate time-stamped collaboration networks based on multiple projections. Exemplary projections are shown in Figure 1.
- Applying `git2net` in a case study on two software projects, we show that the character-based analysis of file modifications yields considerably different network structures compared to previously used methods that have analysed code co-authorship at the level of files or modules. We further motivate how our tool can be used to address a set of research questions.

`git2net` facilitates the extraction of large-scale time-stamped network data that can be cross-referenced with project related information (e.g. project success, or organisational structures). We therefore expect the tool to be of considerable value for the network science community and researchers at the intersection of data science, computational science, and empirical software engineering.

Further details are available in the full paper [GSS19]. `git2net` is available as Open Source project on GitHub[4] and can be installed via `pip install git2net`. A tutorial reproducing the results from the full paper is available on `zenodo.org`[5].

References

[GSS19] Gote, Christoph; Scholtes, Ingo; Schweitzer, Frank: git2net: Mining Time-Stamped Co-Editing Networks from Large git Repositories. In: Proceedings of the 16th International Conference on Mining Software Repositories. MSR '19. IEEE Press, 2019.

[4] `https://github.com/gotec/git2net`
[5] 10.5281/zenodo.2587483