Lean GHTorrent: GitHub data on demand

Georgios Gousios*, Bogdan Vasilescu†, Alexander Serebrenik†, Andy Zaidman*

*Delft University of Technology
Delft, The Netherlands
{g.gousios, a.e.zaidman}@tudelft.nl

†Eindhoven University of Technology
Eindhoven, The Netherlands
{b.n.vasilescu, a.serebrenik}@tue.nl

ABSTRACT

1. INTRODUCTION

During recent years, GITHUB (2008) has become the largest code host in the world, with more than 5M developers collaborating across 10M repositories. Due to its support for distributed version control (Git) and pull-based development [2], as well as its modern Web UI and focus on social coding [3], GITHUB has surpassed in size and popularity even much older forges such as Sourceforge (1999). As a result, numerous projects (especially open source) are migrating their code base to GITHUB (for instance, the Google query *migrate to github* returns more than 4M results), which now hosts popular projects such as Ruby on Rails, Homebrew, Bootstrap, Django or JQuery.

Researchers have quickly jumped on board and have started exploring the richness of GITHUB data. So far, studies focused on building language models of source code [1], understanding the effects of branching and pull-based software development [6, 10], uncovering associations between crowdsourced knowledge and software development [17], visualizing collaboration and influence [8], exploring the social network of developers [9, 14, 16], or investigating how the social nature of GITHUB impacts collaboration [3, 11] and could be used to improve development practices [12, 13].

To facilitate studies of GITHUB, we have created GHTorrent [5, 7], a scalable, queriable, offline mirror of the data offered through the GITHUB REST API. GHTorrent data has already been used in empirical studies (e.g., [6, 15, 17]), and a subset of it has been selected as the topic of the Mining Challenge at the 2014 edition of the Working Conference on Mining Software Repositories.

In this paper we present the extensions brought to GHTorrent since its official release [5], designed to offer customisable data dumps on demand. The new GHTorrent data-on-demand service offers users the possibility to request up-to-date GHTorrent data dumps for any subset of GITHUB projects, as indicated by completing a web form. This offers several advantages. First, while the GHTorrent project already offered data dumps of both its raw data (MongoDB, currently more than 2TB) and metadata (MySQL, currently more than 20GB), downloading and restoring these dumps can be very time consuming and might not be necessary if a particular analysis is restricted in scope to say a handful of "interesting" GITHUB projects (e.g., the Ruby on Rails project, for which separate data sets also started being collected [18]).

Second, while the idea of running queries with a restricted scope is not necessarily new with respect to the official release of GHTorrent [5], the data-on-demand service enhances replicability of results obtained using GHTorrent data. GHTorrent already offered an online query interface with access to an archived version of the relational database, which could be used to restrict the scope of a query. However, GITHUB is a very dynamic platform where developers, projects and wikis are created and deleted constantly. Therefore, online queries of GHTorrent data may return different results at different times if project data recorded by GHTorrent has been refreshed in the meantime. To enhance the replicability [4] of such results, it is therefore preferable to store the exact snapshot of the data set used in the analysis.

2. REFERENCES

[1] M. Allamanis and C. Sutton. Mining source code repositories at massive scale using language modeling. In *Proc. MSR*, pages 207–216. IEEE, 2013.

[2] E. T. Barr, C. Bird, P. C. Rigby, A. Hindle, D. M. German, and P. Devanbu. Cohesive and isolated development with branches. In *Proc. FASE*, pages 316–331. Springer, 2012.

[3] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb. Social coding in Github: transparency and collaboration in an open software repository. In *Proc. CSCW*, pages 1277–1286. ACM, 2012.

[4] J. M. González-Barahona and G. Robles. On the reproducibility of empirical software engineering studies based on data retrieved from development repositories. *Empirical Software Engineering*, 17(1-2):75–89, 2012.

[5] G. Gousios. The GHTorent dataset and tool suite. In *Proc. MSR*, pages 233–236. IEEE, 2013.

[6] G. Gousios, M. Pinzger, and A. van Deursen. An exploratory study of the pull-based software development model. In *Proc. ICSE*. ACM, 2014.

[7] G. Gousios and D. Spinellis. GHTorrent: Github's data from a firehose. In *Proc. MSR*, pages 12–21. IEEE, 2012.

[8] B. Heller, E. Marschner, E. Rosenfeld, and J. Heer. Visualizing collaboration and influence in the open-source software community. In *Proc. MSR*, pages 223–226. ACM, 2011.

[9] J. Jiang, L. Zhang, and L. Li. Understanding project dissemination on a social coding site. In *Proc. WCRE*, pages 132–141. IEEE, 2013.

[10] H. Lee, B.-K. Seo, and E. Seo. A git source repository analysis tool based on a novel branch-oriented approach. In *Proc. ICISA*, pages 1–4. IEEE, 2013.

[11] J. Marlow, L. Dabbish, and J. Herbsleb. Impression formation in online peer production: activity traces and personal profiles in Github. In *Proc. CSCW*, pages 117–128. ACM, 2013.

[12] R. Pham, L. Singer, O. Liskin, F. Figueira Filho, and K. Schneider. Creating a shared understanding of testing culture on a social coding site. In *Proc. ICSE*, pages 112–121. IEEE, 2013.

[13] R. Pham, L. Singer, and K. Schneider. Building test suites in social coding sites by leveraging drive-by commits. In *Proc. ICSE*, pages 1209–1212. IEEE, 2013.

[14] D. Schall. Who to follow recommendation in large-scale online development communities. *Information and Software Technology*, 2013.

[15] M. Squire. Forge++: The changing landscape of FLOSS development. In *Proc. HICSS47*. IEEE, 2014.

[16] F. Thung, T. F. Bissyandé, D. Lo, and L. Jiang. Network structure of social coding in GitHub. In *Proc. CSMR*, pages 323–326. IEEE, 2013.

[17] B. Vasilescu, V. Filkov, and A. Serebrenik. StackOverflow and GitHub: associations between software development and crowdsourced knowledge. In *Proc. SocialCom*, pages 188–195. IEEE, 2013.

[18] P. Wagstrom, C. Jergensen, and A. Sarma. A network of rails: a graph dataset of ruby on rails and associated projects. In *Proc. MSR*, pages 229–232. IEEE, 2013.