Boa: a Virtual Collaboratory for Studying Software and its Evolution at a Large Scale

Hridesh Rajan*

Department of Computer Science, Iowa State University

A. Overview

Software is rapidly becoming one of the most fundamental building blocks of human interaction and activity. Ultra-large-scale software repositories, e.g. SourceForge (700k+ projects), GitHub (7M+ projects), and Google Code (300k+ projects) contain an enormous collection of software and information about software. These big-3 repositories amount to 1,000,000,000+ lines of code, 10,000,000+ revision logs, and 3,000,000+ issue reports. Scientists and engineers alike are interested in analyzing this wealth of information both for curiosity as well as for testing important hypotheses. For example, a social scientist may ask "how people perceive and consider the potential impacts of their own and others' edits as they write together? [1]"; a legal expert may wonder "what is the most widely used open source license? [8]"; a security expert may ask "how many projects continue to have the heartbleed vulnerability? [3]"; a foreign policy expert may ask "how many open source projects have a restricted export control policy? [7]"; a software engineering expert may ask "What is the average time to resolve a bug reported as critical? [10]".

However, the current barrier to entry is prohibitive and only a few with well-established infrastructure and deep expertise can attempt such ultra-large-scale analysis. Necessary expertise includes: programmatically accessing version control systems, data storage and retrieval, data mining, and parallelization.

The need to have expertise in these four different areas significantly increases the cost of scientific research that attempts to answer research questions involving the ultra-large-scale software repositories. As a result, experiments are often irreproducible, reusability of experimental infrastructure low, and data associated and produced by such experiments is often lost and becomes inaccessible and obsolete, because there is no systematic curation [6]. Last but not least, building analysis infrastructure to process ultra-large-scale data efficiently can be very hard [2]. There are efforts to provide repository data, as well as tools to mine that data; however, this community does not yet have an intuitive, easy to use, end-to-end solution.

A.1 Boa: A Virtual Collaboratory for Analyzing Ultra-large Software Repositories

To solve these problems, we have created *Boa* [4, 5, 9], a virtual collaboratory for research that analyzes software and its evolution at a large scale. *Boa* is a research infrastructure that consists of a domain-specific language, its compiler and data updating tools, terabytes (and growing) of raw data from open source repositories that contains 700,000+ open source projects as of this writing, a backend based on map-reduce to effectively analyze this dataset, a compute cluster, and a web-based frontend for writing analysis programs. Within *Boa*, research questions concerning human and technical aspects of open source software development can be answered by writing, often short, programs that are automatically parallelized by the infrastructure to process already curated dataset. This significantly decreases the barrier to entry for such research, improves scalability, and lowers complexity and size of analysis programs, which allows researchers to focus on their essential tasks [4, 5]. Since standardized datasets are available within *Boa*, collaboration and comparison of research results is facilitated. Reproducing an experiment conducted using *Boa* is just a matter of re-running *Boa* programs provided by previous researchers.

In May 2013, we also opened our preliminary infrastructure on a limited basis and Figure 1 presents a geo-coding of this initial set of users.

*Email: hridesh@iastate.edu, Address: 226 Atanasoff Hall, Ames, IA, 50010

Many other researchers have asked, and we would truly like to make our collaboratory broadly available while we continue to research *Boa*'s design, semantics and implementation. However, it would not be useful to make the infrastructure available for download as it would take months for that download to finish, whereas the live-data in open source repositories is updated every second. Ideally we would like to make our intuitive web-based interface available to the community, but there are three problems. First, our preliminary infrastructure does not currently implement key desired requirements of a shared data analysis infrastructure, e.g. scalable to large number of users, extensible to new needs, etc... Second, we do not currently have the necessary resources to make this web-based collaboratory available to the broader community and to support its maintenance until it becomes scalable, reliable, and community sustained. Finally, with increasing number of users we are seeing a significant rise in operational needs of the community infrastructure, e.g. maintaining the *Boa* cluster, responding to user queries, etc, which are somewhat orthogonal to the research goals.

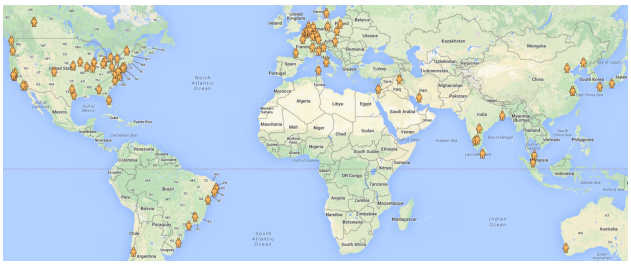

Figure 1: Map of currently registered *Boa* users: enhancing infrastructure for research and education, developing partnerships with international academic institutions and organizations, and building networks of U.S. colleges and universities.

Our goal in attending this NSF Cloud workshop is to investigate whether the NSF cloud project can help us address these research and operational needs.

References

[1] J. Birnholtz and S. Ibara. Tracking changes in collaborative writing: edits, visibility and group maintenance. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 809–818, New York, NY, USA, 2012. ACM.

[2] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. In *Proceedings of the 6th Symposium on Opearting Systems Design & Implementation - Volume 6*, 2004.

[3] Z. Durumeric, J. Kasten, D. Adrian, J. A. Halderman, M. Bailey, F. Li, N. Weaver, J. Amann, J. Beekman, M. Payer, et al. The matter of heartbleed. In *ACM Internet Measurement Conference (IMC)*, 2014.

[4] R. Dyer, H. A. Nguyen, H. Rajan, and T. N. Nguyen. Boa: A language and infrastructure for analyzing ultra-large-scale software repositories. In *Proceedings of the 35th International Conference on Software Engineering*, ICSE '13, pages 422–431. IEEE Press, 2013.

[5] R. Dyer, H. Rajan, and T. N. Nguyen. Declarative visitors to ease fine-grained source code mining with full history on billions of AST nodes. In *Proceedings of the 12th International Conference on Generative Programming: Concepts and Experiences*, GPCE'13, October 2013.

[6] 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.

[7] S. Goodman, P. Wolcott, and G. Burkhart. *Building on the Basics: An Examination of High-Performance Computing Export Control Policy in the 1990s*. Center for International Security & Cooperation, 1995.

[8] J. Lerner and J. Tirole. Some simple economics of open source. *The Journal of Industrial Economics*, 50:197–234, 2002.

[9] H. Rajan, T. Nguyen, R. Dyer, and H. Nguyen. Boa: A language and community research infrastructure for analyzing open source repositories. `http://boa.cs.iastate.edu/`, August 2012.

[10] C. Weiss, R. Premraj, T. Zimmermann, and A. Zeller. How long will it take to fix this bug? In *Proceedings of the Fourth International Workshop on Mining Software Repositories*, MSR '07, pages 1–, Washington, DC, USA, 2007. IEEE Computer Society.