Insights from Open Source Software Supply Chains (Keynote)

Audris Mockus
University of Tennessee
USA
audris@utk.edu

ABSTRACT

Open Source Software (OSS) forms an infrastructure on which numerous (often critical) software applications are based. Substantial research was done to investigate central projects such as Linux kernel but we have only a limited understanding of how the periphery of the larger OSS ecosystem is interconnected through technical dependencies, code sharing, and knowledge flows. We aim to close this gap by a) creating a nearly complete and rapidly updateable collection of version control data for FLOSS projects; b) by cleaning, correcting, and augmenting the data to measure several types of dependencies among code, developers, and projects; c) by creating models that rely on the resulting supply chains to investigate structural and dynamic properties of the entire OSS. The current implementation is capable of being updated each month, occupies over 300Tb of disk space with 1.5B commits and 12B git objects. Highly accurate algorithms to correct identity data and extract dependencies from the source code are used to characterize the current structure of OSS and the way it has evolved. In particular, models of technology spread demonstrate the implicit factors developers use when choosing software components. We expect the resulting research platform will both spur investigations on how the huge periphery in OSS both sustains and is sustained by the central OSS projects and, as a result, will increase resiliency and effectiveness of the OSS.

CCS CONCEPTS

• **General and reference** → **Measurement**; *Empirical studies*; • **Information systems** → **Data cleaning**; **Entity resolution**; *Deduplication*; • **Software and its engineering** → **Open source model**.

KEYWORDS

Empirical Software Engineering; Software Ecosystems; Software Measurement

ACM Reference Format:
Audris Mockus. 2019. Insights from Open Source Software Supply Chains (Keynote). In *Proceedings of the 27th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '19), August 26–30, 2019, Tallinn, Estonia.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3338906.3342813

BIOGRAPHY

Audris Mockus received the BS and MS degrees in applied mathematics from the Moscow Institute of Physics and Technology in 1988, the MS degree in 1991 and the PhD degree in statistics from Carnegie Mellon University in 1994. He is interested in recovering information and creating models of reality from big operational data. His latest interest concern models of the entire open source software ecosystem based on version control data and anthropological phenomena hidden in large image collections. He is the Ericsson-Harlan D. Mills Chair Professor in the Department of Electrical Engineering and Computer Science of the University of Tennessee. Previously he worked at Avaya Research and AT&T and Lucent Bell Labs.

ESEC/FSE '19, August 26–30, 2019, Tallinn, Estonia

ACM ISBN 978-1-4503-5572-8/19/08.
https://doi.org/10.1145/3338906.3342813