OPEN UNIVERSITY UK

MASTER DISSERTATION

Evaluating phylogenetic methods for quantifying risks and opportunities presented by forks in open source software

Author:
Alvaro ORTIZ TRONCOSO

Supervisor:
Dr. Doug LEITH

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Science in Computing (Software Engineering)

8. September 2017

OPEN UNIVERSITY UK

Abstract

Master of Science in Computing (Software Engineering)

Evaluating phylogenetic methods for quantifying risks and opportunities presented by forks in open source software

by Alvaro ORTIZ TRONCOSO

Software needs to evolve in order to deliver value to its stakeholders throughout its lifecycle. The most important drivers behind software evolution are changing user expectations, and a dynamic and innovative ecosystem. Open source software development is a software development paradigm that embraces change by blurring the distinction between users and developers. The cornerstone of open source development is a licensing scheme that grants anybody the right to examine, to copy and to modify the source code. Open source licenses have proven in many cases a viable alternative to strong intellectual property protection regimes. However, open source licenses expose software projects to a new kind of risk: forking. Forking happens when part of the team takes off in a new direction. Literature on the governance of open source projects disagrees on whether forking is a risk or an opportunity: the traditional view is that forking is the result of a failure of the project to keep its resources together however, as successful software products have lately emerged from forks, the traditional view is challenged. Notwithstanding, methods for quantifying the evolution of forks are currently scarce: the present research attempts to port methods from phylogenetics, a branch of evolutionary biology that attempts to unravel the mechanisms behind the evolution of living organisms, to the study of the evolution of forks and postulates that the progress of a fork can be modelled using these methods. Methods and concepts from evolutionary biology were validated by applying them to three cases of software forks. A statistical analysis shows that the history of a forked project can be reconstructed using phylogenetic trees, and finds evidence that the eventuality of a fork could be predicted. However, no evidence was found that the outcome of a fork can be foretold using these methods. The present research concludes by porting basic concepts from evolutionary biology into a software development context and elaborates how phylogenetic methods and concepts can be used by practitioners to increase their understanding of forking processes.

Acknowledgements

I am extremely grateful to the Open University staff, who have provided me with such professional and well informed support. Special mention must be made of my tutor, Dr. Doug Leith, whose guidance has been invaluable throughout. I also would like to thank my colleagues at the Technical University Berlin, in particular Erhard Zorn and Dr. Stefan Born, for balancing a busy workload. I am especially thankful for the encouragement shown by my family and in particular for the patience afforded me by my dear friend Ina Kemter.

Contents

Glossary

Branch
: A thread of development within a project or team; branches are common in open source development (Robles and González-Barahona, 2012).

Evolution
: Defined in paragraph 1.3.

Fork
: Defined in paragraph 1.3.

Merge
: A rejoining of separate development strands that had branched or forked previously, either by integrating source code or by dismissing parts of either project (Robles and González-Barahona, 2012).

Open source software
: Software development paradigm that blurs the difference between users and developers (Hippel and Krogh, 2003). Open source software licenses grant users the right to fork a project (Robles and González-Barahona, 2012).

Phylogenetic tree
: A pictorial representation of the degree of relationship between entities sharing a common ancestry (Baum and Offner, 2008).

Release
: A stage in the software lifecycle corresponding to a new generation of the system (Lehmann, 1980).