Predictive Health Modelling of Evolving Software Packaging Ecosystems

University of Mons – Action de Recherche Concertée – 2021-2025

Tom Mens (Software Engineering Lab) & Souhaib Ben Taieb (Big Data and Machine Learning Lab)

Keywords

software ecosystem, software health, empirical analysis, data analytics, predictive modeling, machine learning, multi-dimensional point process, neural networks, dynamic network modeling

Summary

Open Source Software (OSS) is indispensable in today's software-driven society and industry. OSS communities manage and evolve ecosystems containing millions of interconnected software packages released and maintained by thousands of geographically distributed contributors. Packaging ecosystems face a wide range of health issues induced by bugs, security vulnerabilities, incompatible component updates, and unmaintained, deprecated or outdated package releases. Because of the highly connected and inherently collaborative nature of the socio-technical networks of packaging ecosystems, these issues frequently impact (transitively) related packages, resulting in a combination of fine-grained (package-level) and coarse-grained (network-level) health problems.

> This raises the need for efficient software health prediction models and techniques addressing OSS packaging ecosystem health, at the level of individual packages as well as at the socio-technical network level.

To address this need, we will extract and combine fine-grained events related to the development of individual software packages (e.g., new package releases, new source code commits, code reviews, reported bugs and their associated fixes, message exchanges between developers), and coarse-grained events related to the evolving socio-technical network (e.g. package versions, dependency constraints, new or abandoning contributors). This event data will be gathered from various sources: software package managers, version control systems, bug and issue trackers, and online communication channels. Modelling such data is particularly challenging notably due to the complex temporal dynamics, as well as the heterogeneity, quality, size and complexity of the date. We will develop and apply machine learning models for prediction and causal discovery of software health problems based on temporal point processes and dynamic network modelling techniques to analyse large-scale, multi-granular and evolving software ecosystem data.

> Based on recent developments in machine learning, we will develop prediction, simulation and discovery models to analyse and predict the health of OSS packaging ecosystems and their constituent packages.

Main objectives

Historical events of software package development activity will be modelled using *multi-dimensional point processes*. These processes allow to model the inherent property of software development data where past development events can have an important influence on future events affecting (in a negative or positive way) the health of a software package. We will consider state-of-the-art point process models based on *deep neural networks* to capture more diverse and more complex influences of past events on future events.

In addition to modelling the intrinsic temporal structure at the level of individual packages, we will use *dynamic network modelling* to capture the temporally evolving socio-technical network of the ecosystem. *Causal graph learning* techniques will be used to infer the causal effects of events and associated health metrics across ecosystem packages. Multi-level models will be conceived to combine package-level and network-level health prediction models by capturing the complex dynamic interplay between different levels of granularity and different time scales. To do so, *dynamic graph representation learning* techniques will be combined with flexible *neural models for temporal point processes*. Efficient learning techniques will be designed to estimate the parameters of the neural models based on data extracted from selected OSS packaging ecosystems.

The resulting machine learning algorithms and models will be used to provide practical predictive and discovery models for health analysis of OSS packaging ecosystems, taking into account the diversity in activities, granularities and temporal scales, as well as the socio-technical aspects. They will be used to analyse and predict health issues in upcoming package releases, and to assess the network-level health impact by predicting how events with a (positive or negative) effect on health affect other packages in the ecosystem.