Quantifying the Diffusion of Software between Countries

Géza Sümegi

Abstract

The importance and relevance of Open Source Software (OSS) today is unarguable. Such software promotes collaboration and can be created or developed all around the world. Consequently, geolocating OSS activity is not a straightforward task. This challenge is compounded by the fact that, although most OSS projects have detailed documentation, the source country is almost always missing and can hardly be inferred. However, it would be useful to gain information about where successful libraries come from because it can be a good indicator of the economic potentials of origin countries.

In this thesis, we introduce a few methodologies with which we try to infer the origin countries of open source Python packages. Not only do we apply intuitive approaches on our temporal download data, but also we use three main statistical measures, namely the Cross Correlation Function, Granger Causality and Transfer Entropy. With these methods, we infer significant causality links between countries, and visualize the results using heatmaps and networks. Our findings indicate that in general, it is challenging to extract true causal links from empirical data, but it is even harder in case of packages coming from the United States (US). All of our methods struggle with these packages and produce false results. Nonetheless, we find Transfer Entropy the most suitable algorithm for our purpose, which is able to find the true origin countries in a few cases, and it is well-complemented with the network representation of the influencing relationships between countries.

Budapest, Hungary
2024