INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTATIONSDAG 2015-06-04

EXAMENSARBETE Social Network Analysis of Open Source Projects
STUDENTER Nicklas Johansson, Christian Tenggren
HANDLEDARE Alma Orucevic-Alagic (LTH)
EXAMINATOR Martin Höst (LTH)

Social network analysis of open source projects

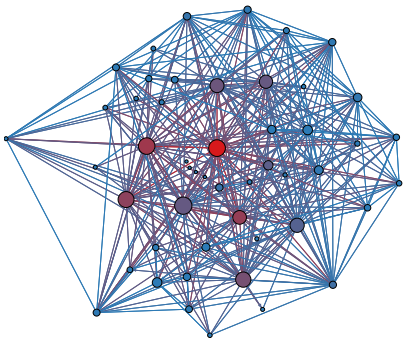
POPULÄRVETENSKAPLIG SAMMANFATTNING **Nicklas Johansson, Christian Tenggren**

By applying social network analysis on the collaboration of open-source developers for a wide variety of projects a few observations can be made that can give some valuable insight in the development process of open-source projects.

Motivation
In software development, open source projects are projects where the code is freely available for anyone to read and copy. These projects are developed through a decentralized structure, where anyone capable is able to submit changes and improvements to the codebase. This development process differs heavily from the way most companies do their development, yet it is still many times very successful and can produce products of considerable quality. As open source software is commonly used in many closed projects, they also get contributions from many companies. This makes them an interesting target for analysis of the development process as there's a large amount of data openly available, contrary to projects developed in closed environments.

This led us to perform a study on a large amount of open-source projects hosted by a well-known open-source community called Apache Software Foundation. The foundation hosts many projects with a wide variety of sizes, including several high profile projects such as OpenOffice.

Social network analysis
To analyze developer collaboration we have used the concept of social network analysis. This is done by studying networks representing developers and their collaboration. Performing a social network analysis means applying different metrics to the networks, where each metric looks at a specific property of the networks. Examples of metrics are the number of other developers a specific developer has collaborated with and centrality measures that shows the influence developers have over one another. The clustering coefficient is another metric that calculates if there exists subgroups in the developer networks. A subgroup is a set of vertices that are well connected to each other. Well connected means that each vertex in the set has an edge to a majority of all the other vertices in the set.



The developer network for Apache OpenOffice

Result
A few conclusions can be drawn from the data gathered from the various projects. The more developers a projects has, the higher the average centrality is. The average centrality follows a line very closely, suggesting that the networks have a similar structure. The majority of the developers have a small centrality index while only a few developers are very central in the projects. On the other hand, we found no real correlation between the clustering coefficient and any other data gathered throughout the study.