Network Analysis with NetworKit – Interactive *and* Fast

Henning Meyerhenke, Elisabetta Bergamini,
Moritz von Looz, and Christian L. Staudt

Faculty of Informatics, Karlsruhe Institute of Technology (KIT), Germany

Network science methodology is increasingly applied to study various real-world phenomena. Consequently, large network data sets comprising millions or billions of edges are more and more common. In order to process and analyze such massive graphs, we need algorithms whose running time is nearly linear in the number of edges. Many analysis methods have been pioneered on small networks, where speed was not the highest concern. Developing a scalable analysis tool suite thus often entails replacing them with suitable faster variants.

Here we present NetworKit (`http://networkit.iti.kit.edu`), an open-source software suite for analyzing the structure of large networks. We describe our methodology to develop scalable solutions to network analysis problems, including parallelization, fast heuristics for computationally expensive problems, efficient data structures, and modular software architecture. NetworKit is implemented as a hybrid: It combines performance-critical parts in C++ (using OpenMP for parallelism) with a Python user interface, enabling interactive workflows with a scripting language and integration into the Python ecosystem.

Our goal for the software is to put our algorithm engineering efforts into the hands of domain experts. The package provides a wide and growing range of functionality, including common and novel analysis algorithms and graph generators. Focus areas for novel analysis algorithms have been community detection, structure-preserving sparsification, and the ranking of vertices and/or edges based on their structural importance (so-called centralities). For scaling studies and benchmarking purposes, our fast generators can create graphs that exhibit typical complex network structure (e. g. random hyperbolic graphs) or scaled and obfuscated replicas of networks that could otherwise not be shared due to privacy concerns. Also, NetworKit can exploit the correspondence between graphs and matrices (GraphBLAS, `http://graphblas.org`) and has been used for supporting probabilistic range queries in large spatial data sets. In experiments with typical analysis and generation tasks on networks with millions or billions of edges, the ratio of graph size (number of edges) and running time (in seconds) is usually between 10^6 and 10^8 for NetworKit's nearly-linear time algorithms. Compared to the closely related software packages graph-tool and igraph, NetworKit shows consistently the highest speed. Our relevant publications can be found on the NetworKit website (`https://networkit.iti.kit.edu/publications.html`).