Geosci. Model Dev. Discuss., 8, C112–C114, 2015
www.geosci-model-dev-discuss.net/8/C112/2015/

Geoscientific Model Development Discussions

Interactive comment on "Par@Graph – a parallel toolbox for the construction and analysis of large complex climate networks" *by* H. Ihshaish et al.

C. Staudt (Referee)

christian.staudt@kit.edu

Received and published: 4 March 2015

In the following I comment on the paper "Par@Graph – a parallel toolbox for the construction and analysis of large complex climate networks" from my perspective as a researcher on graph algorithms, as well as the maintainer of the NetworKit tool suite for high-performance network analysis (which the authors briefly mention as related work).

The framework targets clusters of multi-core machines and consists of two stages: a) the extraction of a network from correlations of time series of meteorological data b) an optimized, parallelized version of the igraph network analysis package

Climate research is a promising application for complex network analysis, and the sub-

ject is therefore very interesting and relevant. However, I would like to point out several areas where I think the paper falls short of making its points:

- The framework provides a set of basic network analysis methods, including degree centrality, eigenvector centrality and betweenness centrality. Granted that it is not the focus of the paper, it is nonetheless a bit disappointing that nothing is written about the interpretation or usefulness of these measures in the context of climate networks (except that they have "interesting physical interpretations").

- The paper claims that Par@graph enables the analysis of graphs with at least 10^{12} edges. These are impressive performance claims that raise interest in the toolbox. Unfortunately, the argumentation and evidence that follows is not what I would expect. The example that follows is strangely underwhelming: 3 million edges do not constitute a particularly large-scale network, distributed parallelism is not needed for such a small network, and the analysis is not exceptionally fast. For comparison, these are the running times I get for NetworKit on a network with 4.6 million edges (wiki-Talk, available from the SNAP collection), using a single machine with 16 physical cores:

- centrality.DegreeCentrality: 10 ms
- centrality.PageRank (comparable to Eigenvector-Centrality): 1min 22s
- properties.ConnectedComponents: 309 ms
- properties.ClusteringCoefficients.exactLocal: 6.9 s

Those 5 1/2 minutes for the network of 3M edges is one of the rare occasions when absolute running times for the network analysis stage are reported. Running times for the truly large network of 10^{12} edges are notably absent. If one zooms in very much into the plots (e.g. Figure 5), one can recognize that they show speedup factors, but not running times. I suggest that the authors substantiate their performance claim with more extensive running time experiments, besides making the plots more readable.

- Comparative experiments with preexisting software (besides igraph, on which the implementations are based) are also markedly absent. Some likely candidates for comparison are mentioned in related work. Such experiments are required to show that existing single-machine parallel codes are not scalable enough.

- Several of the proposed algorithms seem impractical for the scenario of very large networks: Completing a run of Brandes' betweenness algorithm (actually $O(nm + n^2 log n)$ on weighted graphs) on a network of $m = 10^{12}$ edges seems impractical, and so does a $O(n^3)$ algorithm for clustering coefficients. Figure 6 says that they actually calculated these values on a 10^{12} edge graph, which is amazing. How long did that take? Unfortunately no running times are reported. Clearly the scalability issue here is not single-machine versus distributed software, nor sequential versus parallel implementations, but algorithm complexity. Scaling to massive networks calls for different algorithmic approaches, such as fast approximation algorithms for these standard measures. As we have demonstrated within NetworKit, such algorithms can yield qualitatively comparable results in a tiny fraction of the time required for the exact result.

- When reproducibility in science is concerned, computational scientists really have one of the easiest jobs. Therefore I would strongly encourage the authors to make their program source code openly available. Also, are the modifications to the open-source software igraph being considered for inclusion in the main project?

- Please check the notation: V is used both for the set and the number of nodes. Writing "$O(10^3)$ nodes" when "about 10^3 nodes" is meant is an unnecessary abuse of notation.

Interactive comment on Geosci. Model Dev. Discuss., 8, 319, 2015.