Geosci. Model Dev. Discuss., 8, C466–C473, 2015
www.geosci-model-dev-discuss.net/8/C466/2015/

Geoscientific
Model Development
Discussions
Open Access

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

H. Ihshaish et al.

hisham.ihshaish@uwe.ac.uk

Received and published: 15 April 2015

Dear C. Staudt (Referee):

Thank you for your comments and insightful review. On behalf of the co-authors, I address hereby the questions and concerns you have raised:

- 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").

With regards to the first argument, I only can agree with you. The paper is not only focusing on the performance of network analysis algorithms. Otherwise it would have been deemed unsuitable to the general GMD community. That being said, the work under review has been presented within the context of complex climate networks, particularly to introduce software tools that facilitate the reconstruction and analysis of large-scale climate networks from climatological observations and model datasets.

Of course this has been driven by the usefulness of complex networks' approach in climate research, and the interesting physical interpretations that could be obtained from a variety of network features/measures. References to original contributions based on the physical interpretation, and therefore usefulness, of common network properties and measures have been provided in the first paragraph of the "Introduction" section [p. 320 - 321]. Amongst these contributions, for example, the physical interpretation of measures like betweenness centrality, degree centrality and clustering coefficient (between others) can be found in [1, 2], and community detection in [3].

I recognise, however, that not much of discussion on the interpretation of the provided experimental results has been made. Indeed in this work we particularly focus on presenting primary results of the reconstructed and analysed large-scale climate networks as well as on introducing efficient software tools suitable for its computation. Further in-depth experiments and research to deliver new findings in the climate research field are therefore needed, and it is hoped that Par@Graph will accordingly be useful.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

- 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.

I take it you refer to the example in the “Abstract” – the only place where this network has been reported. Indeed the analysed networks in the experimental part of the paper are significantly larger, mainly in terms of edge density as in the case of POP 0.4 dataset. However, we believe a network of $\sim 3\times 10^6$ edges is known large in the context of climate networks. Especially if one considers that no work has been done to date to reconstruct a climate network from 3 $\times$ 10^5 time series. Moreover, the computational time for the reconstruction and analysis of the described network is, certainly, very short. And this leads to your following remark providing performance comparison with NetworKit:

-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

Whilst wiki-Talk's network is meant to be taken as a similar network to the one we provide in the paper in order to provide meaningful comparison of running times between the algorithms in Par@Graph and NetworKit, I nonetheless find very few similarity if at all exists between the SSH network as in the paper, and the wiki-Talk one. Firstly on a different matter for the sake of precision, the provided Wikipedia Talk (wiki-Talk) network in SNAP collection has around 5.02 million edges rather than 4.6. Secondly and thirdly, contrary to the provided SSH networks, it is directed and unweighted respectively. And last but most importantly, it is extremely sparse compared to the already sparse SSH network provided in our "Abstract".

So given that the wiki-Talk network has around 2.4 million nodes and 5.02 million edges, this means that there is on average 2 neighbours for each node (edge density of wiki-Talk is around 8.75 $\times 10^{-7}$), and that is why the computation of the provided measures is notably fast. Now in Par@Graph, we have these runtimes for wiki-Talk network on 16 cores (same hardware as in the paper):

- Degree centrality: 13 ms
- Eigenvector centrality: 56s
- Connected components: 247 ms
- Clustering coefficient: 9.3 s

Of course more extensive experiments are required to evaluate the performance of the different algorithms. In the publications discussing the performance for NetworKit [4], only very sparse networks have been tested. In interaction climate networks, however, such sparse networks (where only significantly correlated time series are considered as linked) are known likely to discard much detail of the underlying climate physical system or its interdependencies. In this paper we reconstruct (and analyse) large-scale climate networks with edge densities up to (and even higher than) 1 $\times 10^{-2}$.

- 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 recognise 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 2logn) on weighted graphs) on a network of m = 10^{12} edges seems impractical, and so does a O(n3) 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.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

As previously addressed, here as well as in the paper, an extensive evaluation of the performance of Par@Graph will be presented in a future work. Accordingly we will be comparing its performance (speedup and runtime) and scalability with existing software, including NetworKit. However, to report execution times corresponding to the studied networks with Par@Graph, we provide part of the log file (see Supplement) corresponding to the execution of Network # 3 as in Table 1 in the paper.

Please note that we also provide parallel implementation for fast approximation algorithms - see "Estimate Clustering Coef." in the provided log file. Indeed, igraph provides approximation algorithms, and so does Par@Graph. For instance, following the provided log file, for the "Estimated Clustering Coefficient", we were able to obtain excellent approximation of the local clustering coefficient.
With regards to the network of 10^{12} edges (Network # 4 in Table 1), neither the betweenness centrality nor the clustering coefficient were calculated, which will be mentioned in revised version of the manuscript.

This case (the network of 10^{12} edges) has been presented as the upper-limit of the size of a network that has been possible to reconstruct on the given hardware, providing as well some analysis - e.g., we were able to calculate various other metrics including the degree, eigenvector, entropy, etc. Of course it is possible to calculate the betweenness centrality for this given network, and we aim at showing this in our coming work, although then there might be a relevant discussion on whether it is practical or not to calculate the exact values of, for example, betweenness centrality, rather than on the possibility to do so.

Another important issue which has been out of discussion here is the reconstruction of such large-scale networks from time series. And this addresses the remark on the single-machine parallel code, where an argument to be made might be whether it is possible to calculate a correlation matrix of, say, 1 million time series or not – mainly by considering the memory required for such calculation, distributed parallelism seems an obvious approach to tackle such problem, which we provide in Par@Graph.

- 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?

Firstly regarding the inclusion of Par@Graph's source code in the original igraph project, we have parallel igraph patches for the last three versions of igraph (0.7.1, 0.7.0 and 0.6.5) and we contacted the developing team. Gábor Csárdi raised no additional concerns regarding any addition to igraph's project apart from the the original copyright under GPL. However, we have taken a step further to provide an independent package rather than a parallel igraph version and therefore, for reproducibility concerns, a copy of the source code of Par@Graph with a manual, as well as 3 SSH networks (used in the paper) in Pajek format will be sent to the editor, David Ham. In that way, he can send it to the referees, or any additional referees, without compromising their anonymity. That being said, since the development of Par@Graph has been carried out through a collaborative project in which a private company has been a partner, there will be a decision to make shortly about the type of licence to provide the software for the interested communities. For now, however, interested researchers are recommended to contact the authors for the source code, and a remark will be included accordingly in a "Code availability" section in the revised version of our manuscript.

-Please check the notation: V is used both for the set and the number of nodes. Writing "O(103) nodes” when "about 103 nodes" is meant is an unnecessary abuse of notation.

We will provide a correction to the “V” and the notation in the final manuscript following the referee’s recommendation.

References

[1] Donges, J. F., Zou, Y., Marwan, N., and Kurths, J.: Complex networks in climate dynamics, The European Physical Journal Special Topics, 174, 157–179, 2009a.

[2] Donges, J. F., Zou, Y., Marwan, N., and Kurths, J.: The backbone of the climate network, EPL (Europhysics Letters), 87,48 007, 2009b.

[3] Tantet, A. and Dijkstra, H. A.: An interaction network perspective on the relation between patterns of sea surface temperature variability and global mean surface temperature, Earth System Dynamics, 5, 1–14, 2014.

[4] Staudt, C. L. Sazonovs, A., and Meyerhenke, H.: NetworKit: An Interactive Tool Suite for High-Performance Network Analysis, arXiv preprint arXiv:1403.3005, 2014.

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

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper