The unconstrained LFR benchmark

Bojan Evkoski[1,2], Igor Mozetič[1], and Petra Kralj Novak[3,1]

[1] Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia
bojan.evkoski@ijs.si
[2] Jozef Stefan Postgraduate School, Ljubljana, Slovenia
[3] Central European University, Vienna, Austria

The most common approach for evaluating and comparing community detection algorithms is to use networks with a priori known community structure. In the absence of real-world networks with known community structure, artificially generated networks are used. The Lancichinetti-Fortunato-Radicchi (LFR) benchmark [13] is the most widely accepted algorithm for generating artificial networks that resemble real-world networks. A common comparison setting, used for example in [18, 3, 14, 16, 17], is to vary only the LFR mixing parameter μ, which corresponds to partition difficulty (a higher μ means a higher percentage of edges going out of communities, making it harder to find the true communities). The community detection algorithms are then compared on different network sizes (LFR parameter n). In this setting, the diversity of LFR networks is limited. We argue that the performance of community detection algorithms may vary depending on other network properties, i.e., some algorithms perform better on one set of LFR parameters and other algorithms perform better on others. Consequently, conclusions based on only one set of LFR parameters can be misleading.

We propose the unconstrained LFR to perform a more comprehensive benchmarking of community detection algorithms while avoiding the shortcomings of the standard LFR benchmarking. The approach consists of two steps: **generating diverse LFR networks** and then **benchmarking by applying the Friedman test and the post-hoc Nemenyi test**. In this way, the full diversity of the LFR network space can be explored and the potential bias from a single set of LFR parameters is avoided.

Network creation. For the network creation part, we randomly generate values for the following parameters: n—number of nodes in the network (from n_{min} to n_{max} nodes); τ_1—power law exponent for the degree distribution of the network (from τ_{1min} to τ_{1max}); τ_2—power law exponent for the community size distribution in the network (from τ_{2min} to τ_{2max}); μ—fraction of inter-community edges of each node (from μ_{min} to μ_{max}); d_{max}—maximum degree allowed for a node (from $\sqrt{n}$ to $n/2$); d_{avg}—average degree of nodes (from $d_{avg,min}$ to $d_{avg,max}$); (c_{min}, c_{max})—minimum and maximum size of a community ($1 < c_{min}$ and $d_{max} < c_{max}$; $c_{min} \in [1, \sqrt{n}]$ and $c_{max} \in [d_{max}, n/2]$). If the combination of parameters fails to generate a valid network, the process is repeated until a valid combination is found.

Benchmarking. Once the network creation part is complete, we measure the performance and stability of the community detection algorithms, with three measures: Normalized Mutual Information (*NMI*) [11], Adjusted Rand Index (*ARI*) [10], and the BCubed F_1 score [1, 5]. We then compare the scores by the Friedman-Nemenyi test [7, 15, 4]. We use the Friedman-Nemenyi combination to compare several algorithms simultaneously on many different networks whose performances by *NMI*, *ARI*, and F_1 are not normally distributed. The result is visualized by critical difference diagrams.

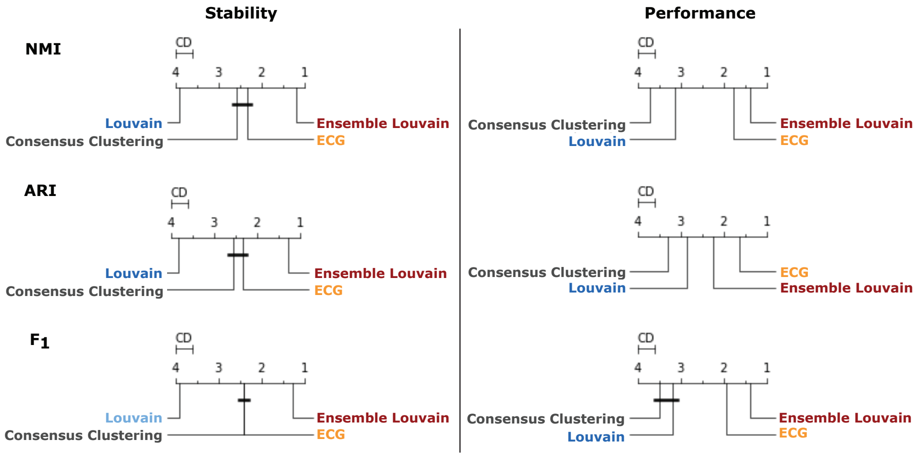

Fig. 1. Unconstrained LFR benchmark. We compare four algorithms on 500 unconstrained LFR benchmark networks, applying the Friedman-Nemenyi significance test. The left-hand side shows the stability results, and the right-hand side the matching to ground truth results. Each individual chart shows the ranks of the four algorithms as estimated by one of the evaluation measures (*NMI*, *ARI*, and F_1). CD denotes the critical difference, and the black bars connect ranks of procedures that are not significantly different at the 5% level.

Pilot study. As a pilot study, we apply the Unconstrained Friedman-Nemenyi LFR benchmark on four community detection algorithms: Louvain [2], Ensemble Louvain [6], Ensembles for Clustering Graphs (ECG) [17] and Consensus Clustering [12]. The last three algorithms use different ensemble techniques of combining multiple partitions to find a more stable and accurate partition than Louvain.

We generate 500 ($N = 500$) unconstrained LFR networks using the NetworkX [8] library with the following parameter settings:

- Number of nodes range, $n \in [100, 12500]$,
- Power law exponent for the degree distribution range, $\tau_1 \in [1.1, 3.0]$,
- Power law exponent for the community size distribution range, $\tau_2 \in [1.05, \tau_1]$,
- Fraction of inter-community edges of each node range, $\mu_{min} \in [0.05, 0.70]$,
- Maximum node degree range, $d_{max} \in [\sqrt{n}, n/2]$,
- Average node degree range, $d_{avg} \in [3, 25]$,
- Maximum community size, $c_{max} \in [d_{max} + 1, n/2]$,
- Minimum community size, $c_{min} \in [2, \sqrt{n}]$.

Note that this range of parameters is only a recommendation based on preliminary experiments. It is chosen so as to most likely yield a viable combination for a network to be generated, while preserving different network and community structures.

We apply the community detection algorithms ten times on the 500 LFR networks and calculate the *NMI*, *ARI* and F_1 measures on their partitions. For stability, we calculate the partition similarity between pairs of the multiple runs of the same algorithm. For performance, we compare the ten runs of each algorithm to the LFR ground truth. The scores are the input to the Friedman-Nemenyi combined test using the Autorank library in Python [9], where we generate rankings of the four algorithms, separately for stability and performance. The final results of this pilot study are presented in Figure 1. A clear distinction in stability and performance can be observed between the algorithms when their ranks differ more than the critical distance (CD). Hence, the suggested benchmark introduces a methodological upgrade and also improves interpretability of the results.

Acknowledgement. The authors acknowledge financial support from the Slovenian Research Agency (research core funding no. P2-103 and P6-0411).

References

1. A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proc. 17th Intl. Conf. on Computational Linguistics (COLING)*, pages 79–85, Stroudsburg, PA, USA, 1998.
2. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
3. T. Chakraborty, N. Park, A. Agarwal, and V. Subrahmanian. Ensemble detection and analysis of communities in complex networks. *ACM Transactions on Data Science*, 1(1):1–34, 2020.
4. J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30, 2006.
5. B. Evkoski, I. Mozetič, N. Ljubešić, and P. K. Novak. Community evolution in retweet networks. *PLoS ONE*, 16(9):e0256175, 2021. Also arXiv:2105.06214.
6. B. Evkoski, I. Mozetič, and P. K. Novak. Community evolution with Ensemble Louvain. In *Complex Networks 2021, Book of abstracts*, pages 58–60, 2021.
7. M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.
8. A. Hagberg, P. Swart, and D. S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
9. S. Herbold. Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173, 2020.
10. L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.
11. T. O. Kvålseth. On normalized mutual information: Measure derivations and properties. *Entropy*, 19(11):631, 2017.
12. A. Lancichinetti and S. Fortunato. Consensus clustering in complex networks. *Scientific Reports*, 2(1):2045–2322, 2012.
13. A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.
14. Z. Lu, J. Wahlström, and A. Nehorai. Community detection in complex networks via clique conductance. *Scientific reports*, 8(1):1–16, 2018.
15. P. B. Nemenyi. *Distribution-free Multiple Comparisons*. PhD thesis, Princeton University, USA, 1963.
16. G. K. Orman, V. Labatut, and H. Cherifi. Towards realistic artificial benchmark for community detection algorithms evaluation. *International Journal of Web Based Communities*, 9(3):349–370, 2013.
17. V. Poulin and F. Théberge. Ensemble clustering for graphs: comparisons and applications. *Applied Network Science*, 4(1):1–13, 2019.
18. Z. Yang, R. Algesheimer, and C. J. Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, 6(1):1–18, 2016.