Structural preferential attachment: scale-free benchmark graphs for overlapping community detection algorithms

Jean-Gabriel Young[§], Laurent Hébert-Dufresne[†], Edward Laurence[§] & Louis J. Dubé[§].

§*Département de physique, de génie physique et d'optique, Université Laval, Québec, QC, Canada.*

†*Santa Fe Institute, Santa Fe, NM 87501, USA*

We introduce a large class of scale-free benchmark graphs for overlapping community detection algorithms. Our benchmark relies on a realistic *and* efficient graph generator, namely the structural preferential attachment (SPA) model [1-2]. As a result, we are able to generate large, scale-free graphs in a timely manner (linear time in the number of edges). We use a bootstrap procedure for the internal structure of communities, and consequently reproduce the universal properties that are recovered by most detection algorithms (e.g. internal density that follows a power law of exponent $\gamma = 1$, skewed internal degree distribution).

In the SPA benchmark, structural properties (e.g. clustering coefficient) are functions of the model parameters, rather than imposed directly. Therefore, a limited exploration of the the low-dimensional parameter space (2 probabilities p, q, and a link creation ratio $r > 0$) allows one to visit multiple *qualitatively* different regions. By comparing the detected communities with the ground-truth at each point, one can then determine the type of communities that are easily uncovered by an algorithm. This organic approach to benchmarking opens the way to more thorough and insightful comparisons of overlapping community detection algorithms.

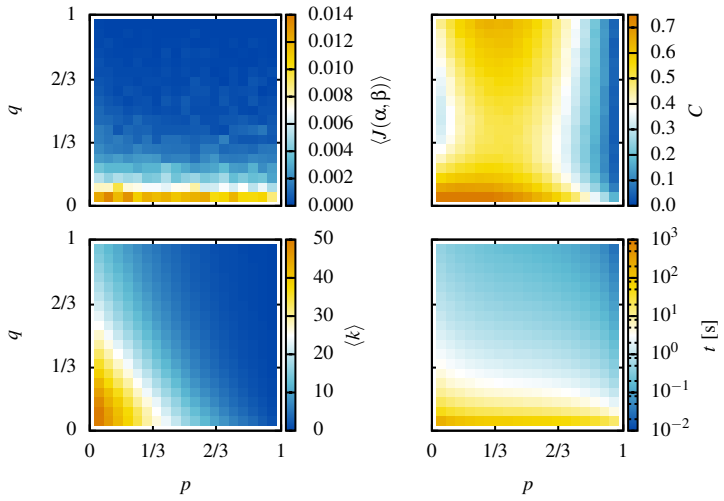
Submitted for a **poster contribution** to the International Workshop and Conference on Network Science (NetSci) held in Zaragoza from June 1st to June 5th 2015.

[1] Hébert-Dufresne, L., Allard, A., Marceau, V., Noël, P.-A., and Dubé, L.J., *Phys. Rev. Lett.*, **107**, 158702, 2011.

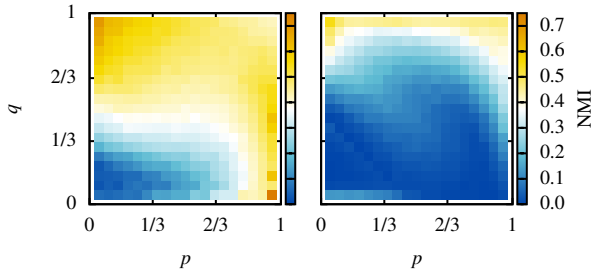
[2] Hébert-Dufresne, L., Allard, A., Marceau, V., Noël, P.-A., and Dubé, L.J., *Phys. Rev. E.*, **85**, 026108, 2012.

[3] Lancichinetti, A., Radicchi, F. Ramasco, J.J. and Fortunato, S., *PLoS ONE*, **6**, e18961, 2011.

[4] Yang, J. and Leskovec, J., *ACM International Conference on Web Search and Data Mining*, 2013.



(a) (*Preliminary results*) **Structural and meta information for benchmark graphs of $N = 5000$ nodes.** (*top-left*) Average overlap of randomly selected pairs α, β of communities. We define the overlap between communities α, β as the Jaccard index of the two node sets. (*top-right*) Average clustering coefficient. (*bottom-left*) Average degree. (*bottom-right*) Time complexity of the construction algorithm. Note how qualitative regions are easily discerned, e.g. low q, p yield dense, clustered and highly overlapping networks.



(b) (*Preliminary results*) **Case-study.** We applied algorithms based on drastically different definition of community structure to our benchmark graphs, and quantified the quality of the resulting cover with the normalized mutual information (NMI). (*left*) *OSLOM*, a local optimization algorithm based on statistical significance, defines communities as dense subgraphs of the complete network [3]. It performs well except in the dense, clustered and highly overlapping regime (low p, q). (*right*) *Bigclam* fits a modelled community structure to the network [4]. Within this framework, nodes that are shared by overlapping communities are assumed to be *more* densely connected than the rest of the network. Bigclam fails to uncover the ground-truth in most regimes, since its definition of community structure does not match the one implemented in our benchmark.