A 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

Summary

We introduce a large class of scale-free **benchmark graphs for overlapping community detection** algorithms.

The graphs and associated overlapping ground truth communities are produced by a realistic stochastic growth process that **generalizes preferential attachment**.

This organic approach to benchmarking allows us

- to generate a wide range of community structures;
- to identify qualitative structural regimes easily;
- to analyze the strengths and weaknesses of an algorithm **at a glance**.

The benchmark in a nutshell

We generate graphs using a modified version of the **Structural Preferential Attachment** (SPA) process [1-2]. This produces graphs with an overlapping community structures, and scale-free distributions of community sizes, node memberships, and degrees.

How to generate SPA graphs?

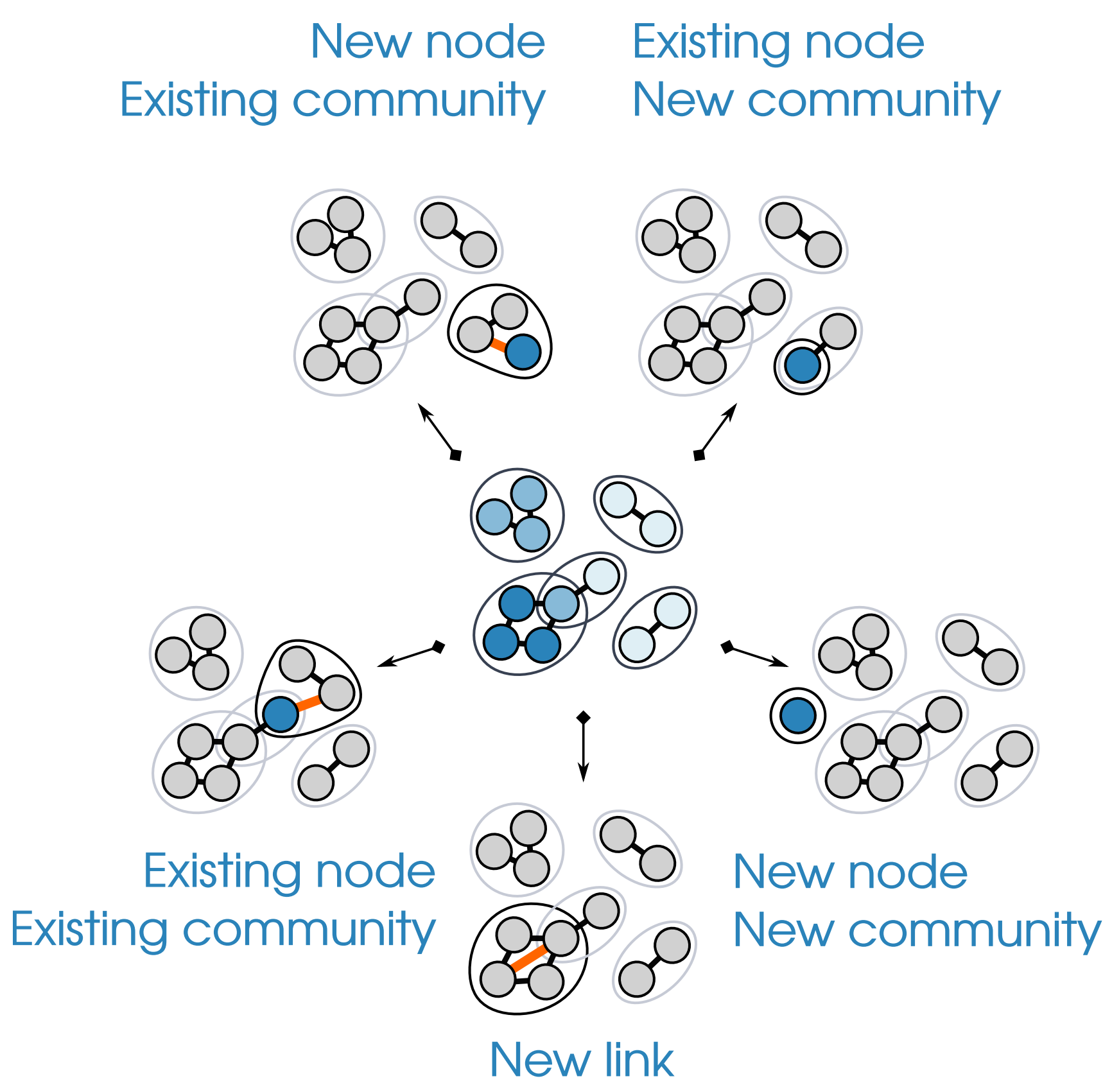
While the graph has fewer than N nodes,

1a. Introduce a new node with probability q, and a new community with probability p.

OR

1b. Increase the size (membership) of an existing community (node) with complementary probability. Select the community (node) *preferentially*.

2. Create a new *internal* link with probability $\propto r(1-p)$. Repeat.



Using SPA as a benchmark

SPA produces realistic networks with known community structures.

One can use overlapping community detection algorithm on these networks to try to identify the **ground-truth** communities.

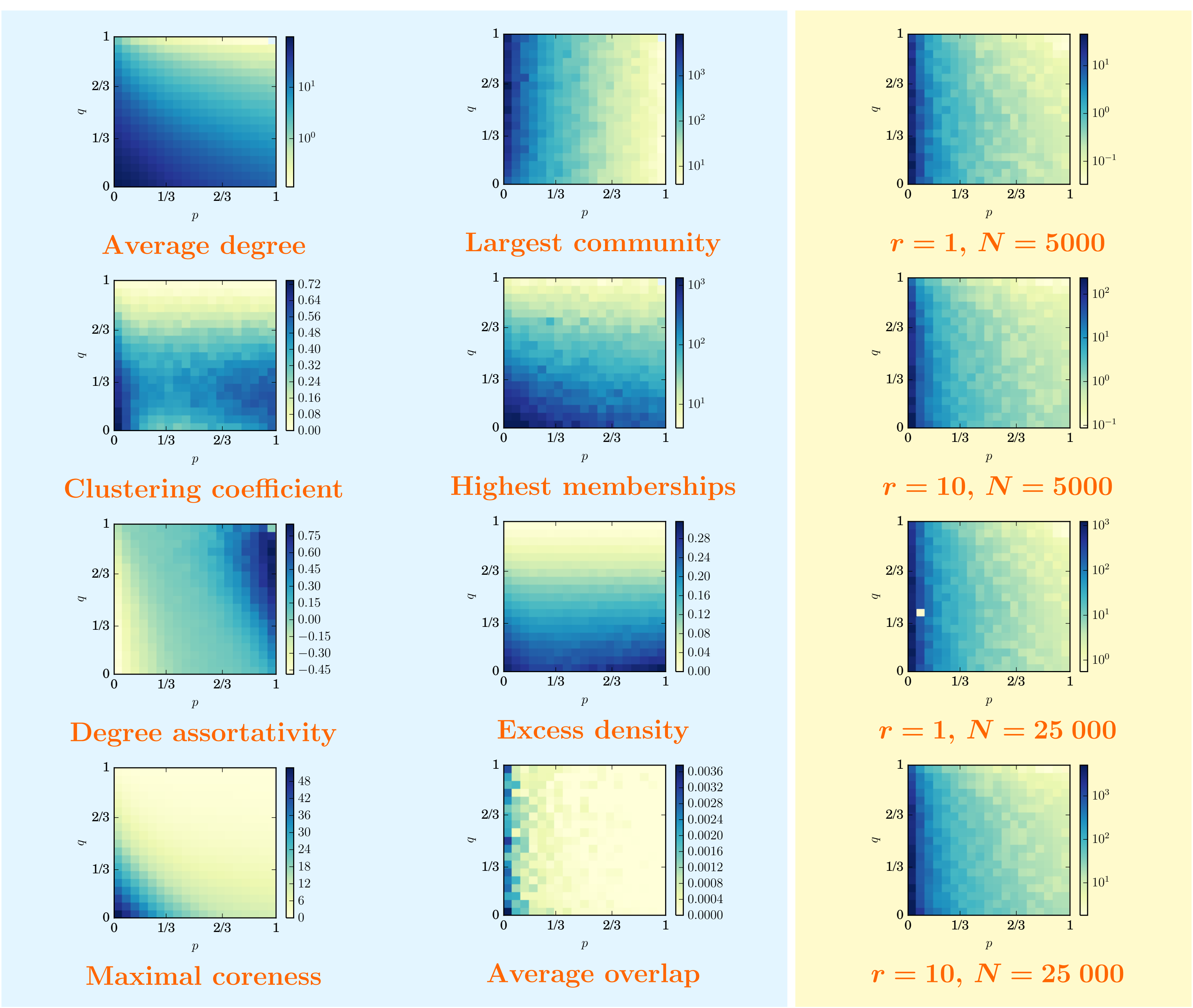
Using an information theoretic measure (NMI) to compare detected and ground-truth communities, it is then possible to quantify how successful the algorithm is in recovering the underlying structure.

Ground-truth communities?
The decomposition in ground-truth communities is considered as the 'true' community structure, i.e. the structure that must be recovered by a perfect detection algorithm.

Graph properties

The structural properties of the graph (e.g. clustering coefficient, degree) are functions of the input parameters (p, q, r, N), rather than imposed directly. These properties vary smoothly with the parameters.

Network structure | **Community structure** | **Running time [sec.]**

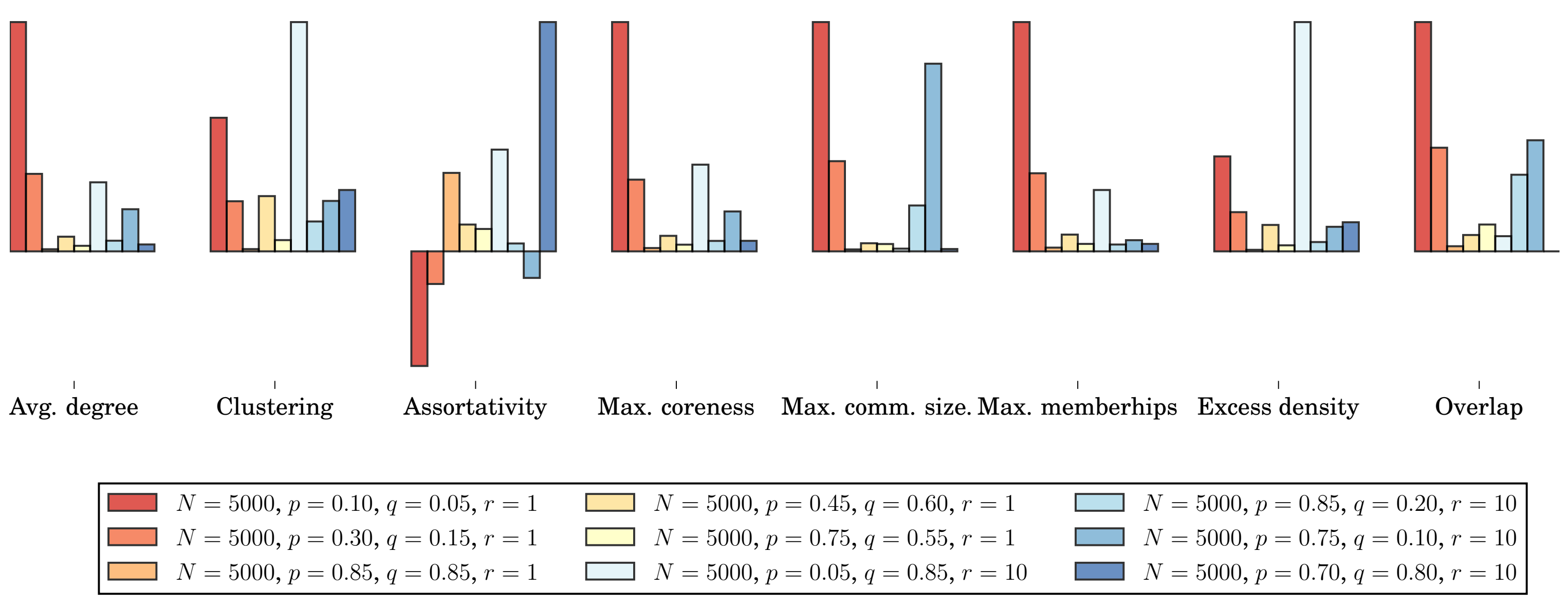


A subset of properties for fixed values of (N, r). Similar behaviors are obtained for all (N, r) pairs.

Relationship between (p, q, N, r) and the expected running time.

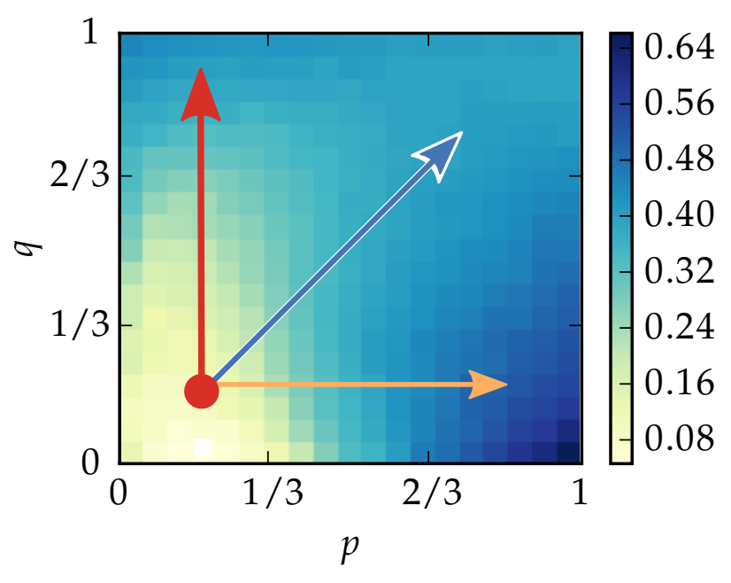
Structural classes

Mathematically, each point (p, q, r, N) can be embedded in a **property space**. Partitioning this space allows us to identify **qualitatively different** structural regimes.



N.B. The property space *is not* the (p, q, r, N) space; the coordinate of a point is given by 20 + loosely correlated properties (e.g. average degree, partition density). A non-euclidean metric defines the distance between each pair of points.

Case study: OSLOM

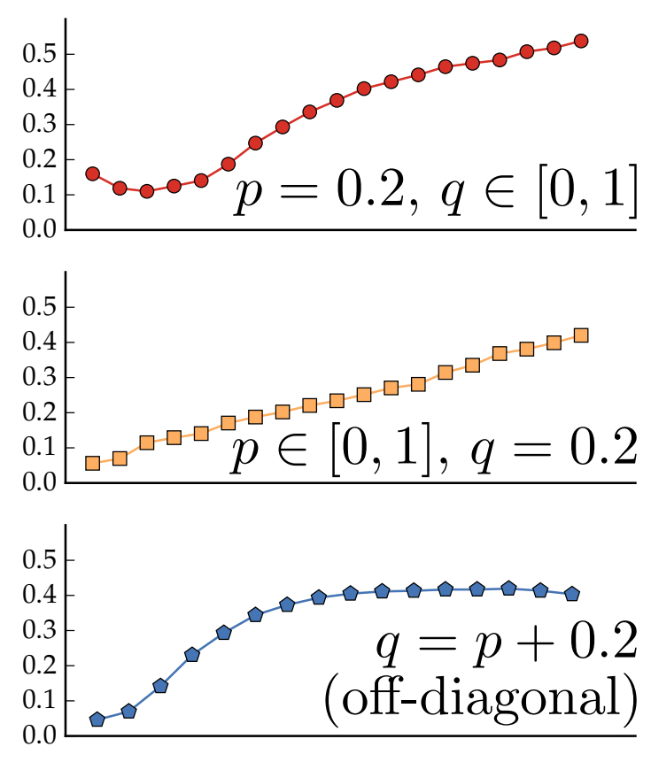


Accuracy (NMI)

We applied the OSLOM algorithm [3] to our benchmark for multiple (p, q) pairs (fixed N, r).

OSLOM performs poorly whenever p, q are small, i.e. for dense, clustered networks with large communities (left).

More importantly, we observe **transitions in detectability** along **multiple** trajectories in the configuration space (right).



Accuracy on trajectories

Case study: Algorithms at a glance

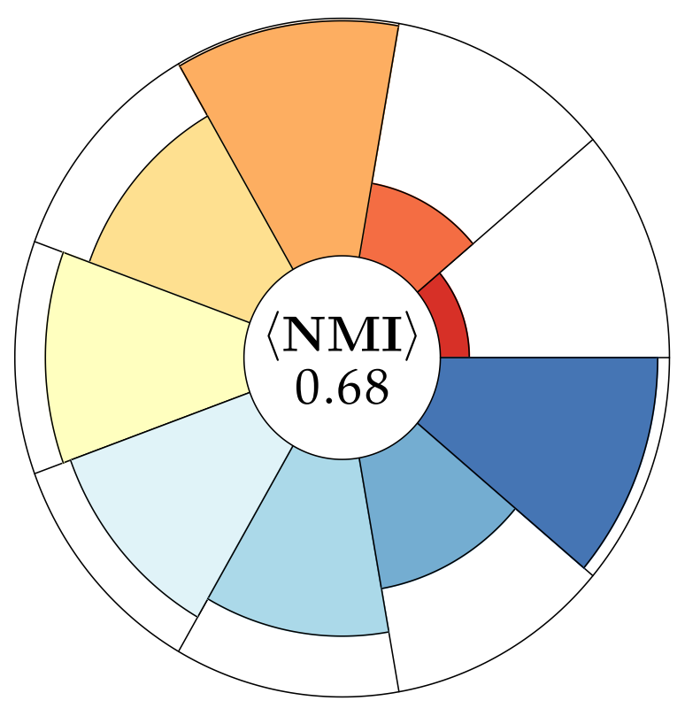
Testing an algorithm for every point of the configuration space is **time consuming**, because one must

- generate multiple graphs for each combination of parameters;
- apply the algorithm to these graphs.

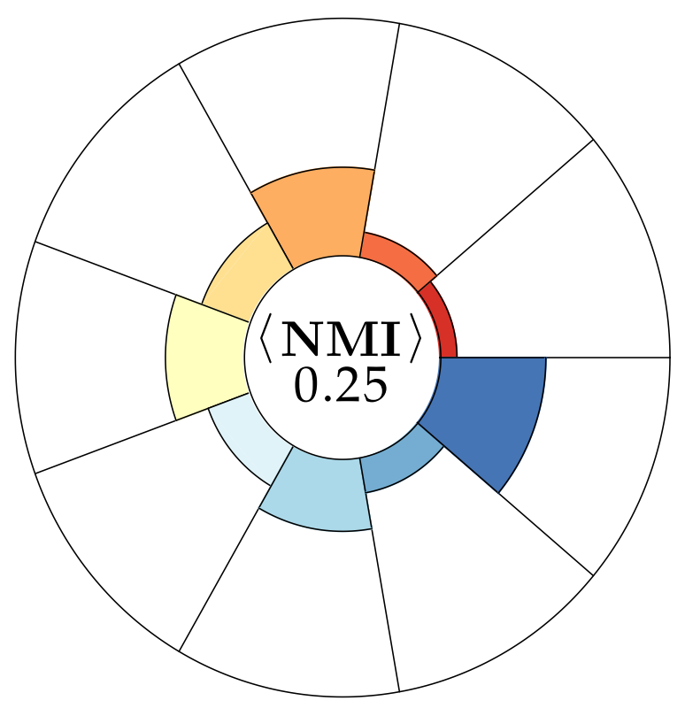
Fortunately, the strengths and weaknesses of an algorithm are easily captured by studying its behavior for a **small subset** of the possible configurations.

To the left and right, we show the average accuracy (NMI) of 4 algorithms, for representative networks of the 9 structural classes identified in the above box (longer leaf = better score). Their overall average score is shown in the center.

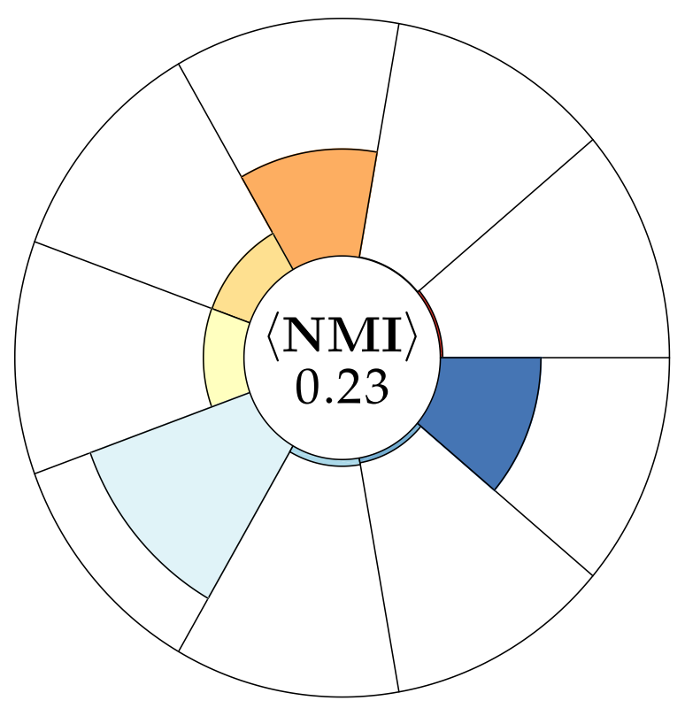
We see that **Infomap is the most versatile** algorithm (best overall score), but that **OSLOM is a reasonable alternative** for highly clustered networks, with few communities. BigCLAM and COPRA are outperformed by Infomap in all regimes.



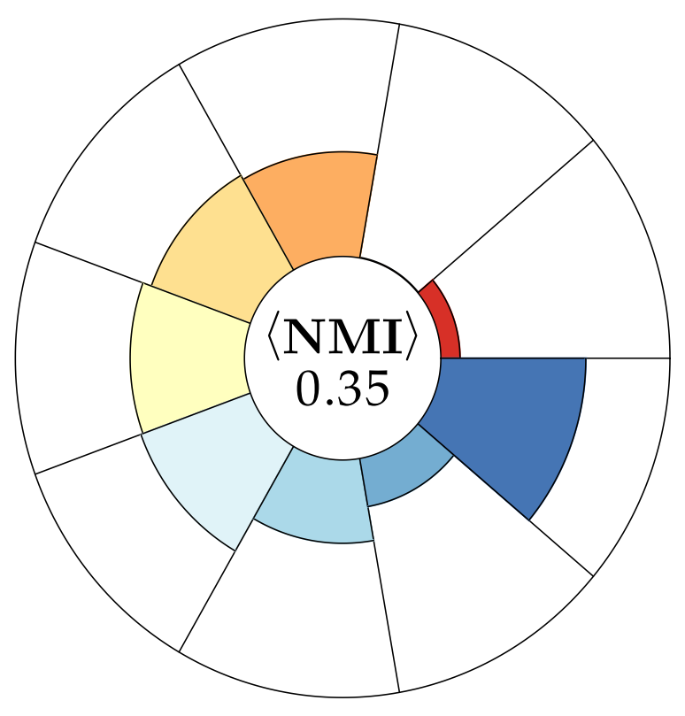
Infomap



COPRA



OSLOM



BigCLAM

Further information

Visit us at

www.spa-networks.org

Acknowledgements

Fonds de recherche sur la nature et les technologies Québec
NSERC CRSNG
Calcul Québec
James S. McDonnell Foundation
Université Laval
Santa Fe Institute

[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.