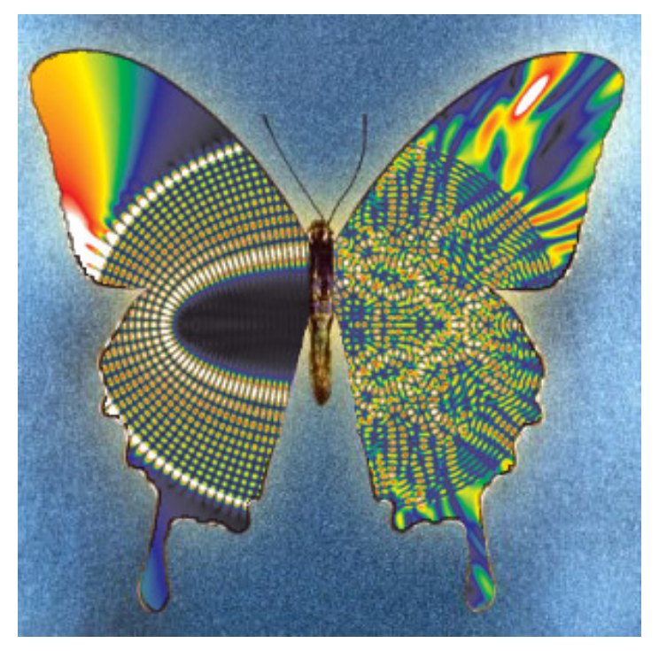
Local and global solutions to community detection - when resolution matters -

J.-G. Young, L. Hébert Dufresne, A. Allard and L.J. Dubé

Département de physique, de génie physique, et d'optique, Université Laval, Québec, Canada



Motivations

Community structure occurs at all scales, but an effective **limit of resolution** arises in commonly used community detection algorithms. This problem manifests itself in two distinct ways:

1. relevant structures are **merged** with neighboring communities.
2. small communities may be entirely **overlooked**.

New improved detections methods must address these limits for a fuller understanding of the **mesoscopic structure** of real systems.

Our goals are to solve these two problems by

1. combining global methods to a **local measure of quality**.
2. using a **global solution**, based on the sequential detection of communities.

Resolution problems

Resolution-limit-free methods yield **partitions that do not change when applied to any sub-partitions of the network** [1]. We illustrate two types of widespread problems that leads to a limit of resolution by studying two popular detection methods.

Modularity based algorithms: Merging

These algorithms seek to optimize the modularity quality function

$$Q = \sum\nolimits_{i,j}^{N} [A_{ij} - k_i k_j/m] \delta(\gamma_i, \gamma_j) \equiv \sum\nolimits_{i,j}^{N} B_{ij}\, \delta(s_i, s_j)$$

over the set of all possible nodes partitions. This objective function identifies groups of nodes that share more links than expected from the Configuration Model.

$$\begin{aligned} N &\equiv \text{number of nodes in the network} \\ A_{ij} &\equiv \text{adjacency matrix element} \\ m &\equiv \text{number of links in the network} \\ k_i &\equiv \text{degree of node } i \\ \delta(s_i, s_j) &\equiv 1 \text{ if both nodes are in the same group and 0 otherwise} \end{aligned}$$

Resolution problem: Modularity has an **intrinsic scale** related to the size of the network. There exists many cases where subgraphs will be further divided if fed back in the algorithm, i.e. algorithms based on modularity or similar measures **merge small communities**.

Link clustering algorithm: Shadowing

Ahn *et al.* link clustering algorithm (**LCA**) [2] aggregates links into communities based on the similarity

$$S\left(e_{ik}, e_{jk}\right) = \frac{|n_+(i) \bigcap n_+(j)|}{|n_+(i) \bigcup n_+(j)|}$$

of their respective neighborhoods, where e_{ij} denotes the link between nodes i and j, and where $n_+(i)$, $n_+(j)$ are their neighborhoods (central nodes i and j included). Communities are built by iteratively grouping adjacent links whose similarity exceeds a given threshold S_c, chosen to maximize the density

$$D = \frac{2}{M} \sum_c \frac{m_c - (n_c - 1)}{(n_c - 1)(n_c - 2)} \qquad (*)$$

Resolution problem: Links in the vicinity of dense communities exhibit **vanishing similarities**, and only $S_c \to 0$ will allow their detection. Algorithms that rely on a global parameter and that allow overlapping communities are all vulnerable to this **shadowing** effect.

Partitioning as a detection tool

The resolution limit of quality functions that feature an implicit scale can be circumvented by **forcing a partition** that contains more communities than the natural optimum. We take advantage of the fact that partitioning methods require the number of modules as an input. We illustrate our method using modularity optimization.

Spectral partitioning based on modularity

It is possible to rewrite the modularity of a partition in q modules as a matrix equation

$$Q \propto \mathrm{Tr}(\boldsymbol{S}^T \boldsymbol{B} \boldsymbol{S}),$$

where $\boldsymbol{S}$ is a $N \times (q-1)$ matrix of *simplex vector* $\boldsymbol{s}^T$ chosen from a set of q vectors. Each node is associated to a row of the matrix, and a choice of vectors corresponds to a choice of partition.

An approximate optimum of this expression can be found using a eigenvector based partition algorithm [3].

Meta-algorithm I:

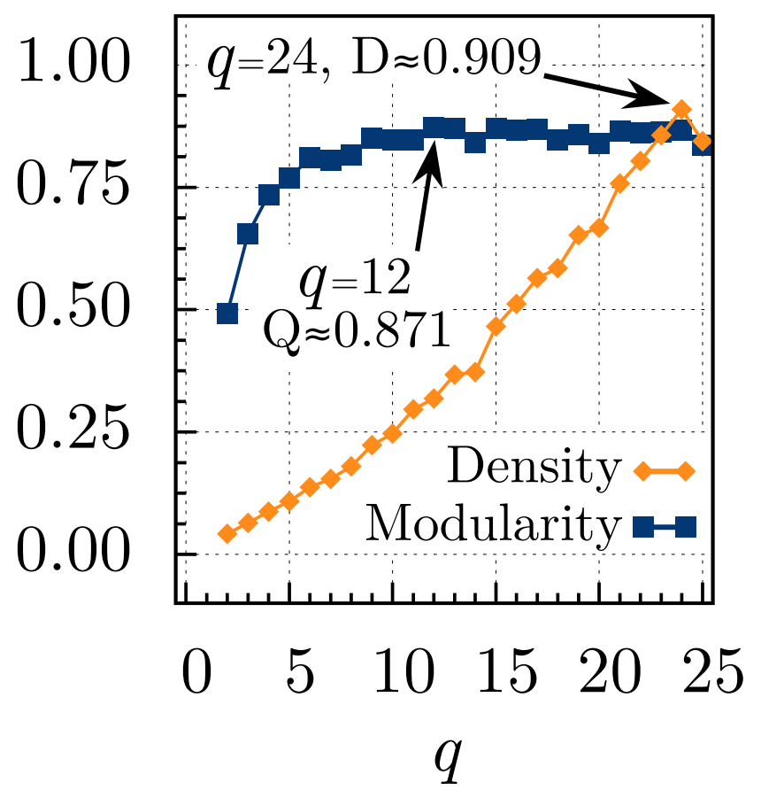
For a large range of number groups q:

1. obtain the optimal modularity partition.
2. evaluate a local measure of quality for this partition.

A peak in the local quality measure corresponds to the **resolution-limit free partition**. For sparse networks (most real networks qualify), the worst-case scenario has a $\mathcal{O}(q_{max}^2 N)$ running time.

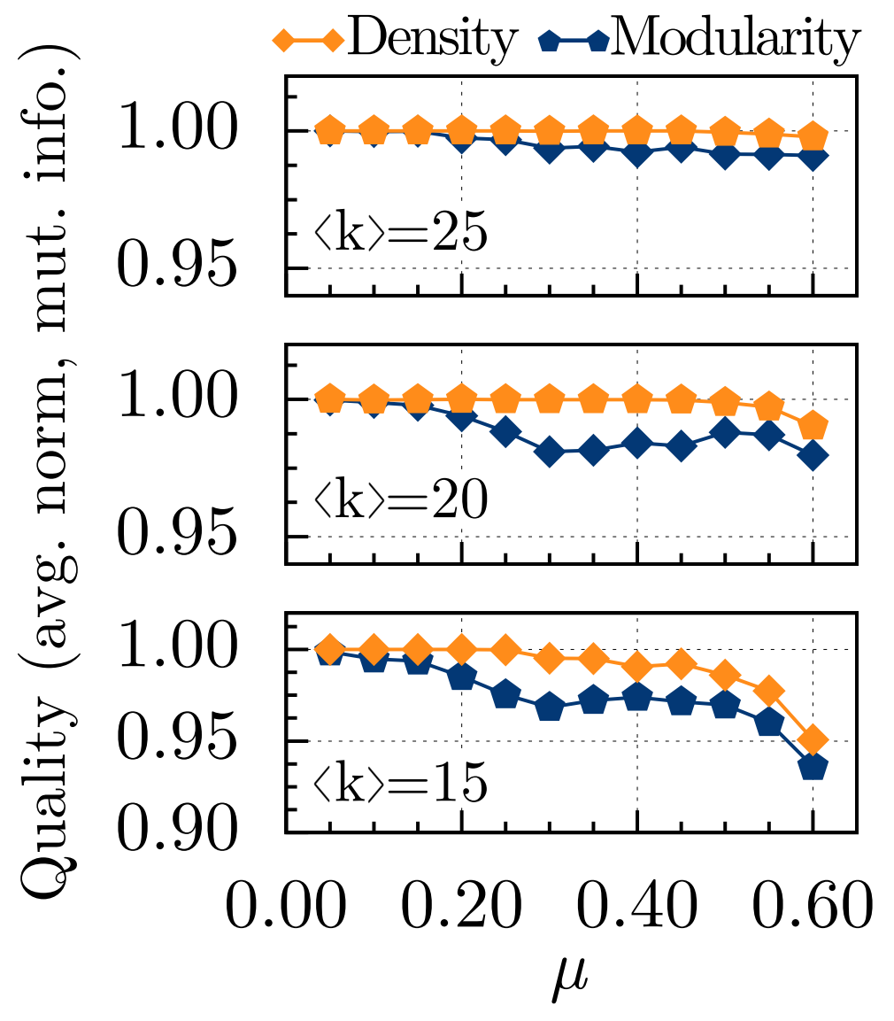
Selection of the optimal partition

A resolution limit free quality measure is necessary to select the optimal number of modules q. We propose the **average link density of the partition** (see eq. $*$), since it is *local* to each community and very sensitive to small changes in the partition. The 2 measures are pictured here for the standard **ring configuration** (24 fully connected cliques of 5 nodes connect by a minimal number of links). As expected, the maximum in density leads to the true partition ($q = 24$, $D \approx 0.909$, $Q \approx 0.867$), while modularity is maximal for pair-wise grouping of the cliques ($q = 12$, $D \approx 0.318$, $Q \approx 0.871$).



Benchmark results

We have applied our algorithm to non-overlapping Lancichinetti-Fortunato-Radicci benchmark networks of $N = 1000$ nodes with a heterogeneous community sizes distribution (power-law). We find that our algorithm **outperforms** the standard modularity based approaches. In other words, we identify the best partition **more consistently** and are able to do so **over a wider range of mixing parameter** μ (the ratio of external edges to the total number of edges per community).



Cascading detection

We have observed that the inability to detect small/sparse communities in the vicinity of larger/denser ones could be circumvented by **removing the troublesome structures** from the networks. We propose a cascading approach to community detection that address this *shadowing effect*.

Meta-algorithm II:

Using any given community detection algorithm:

1. identify large or dense communities.
2. remove the internal links of the communities identified in step 1.
3. repeat until no new significant communities are found.

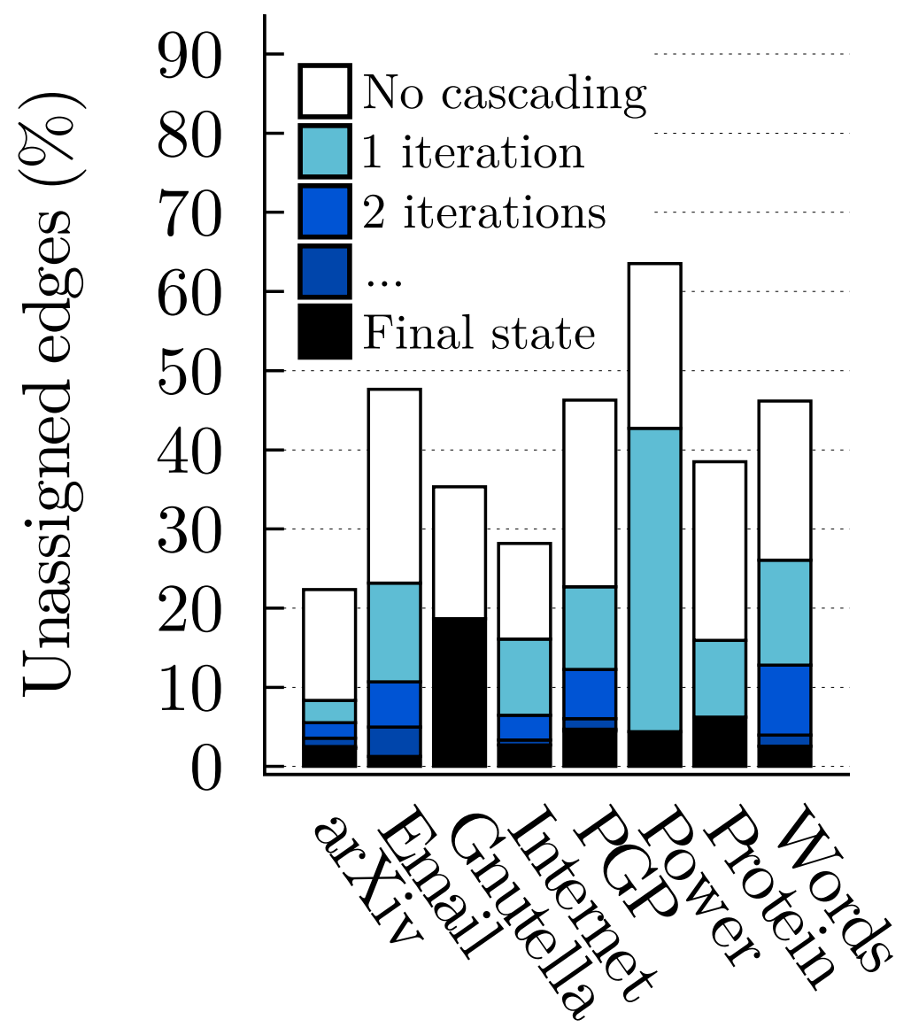
Internal links? Internal links are defined as links that join two nodes belonging to the same community.

The first iteration targets communities that are normally detected by the algorithm of choice. This ensures that the cascading approach **retains the main features of the community structure**. After removing links involved in the detected communities, a new iteration of the detection algorithm is then performed on a sparser network in which **previously hidden communities are now apparent**.

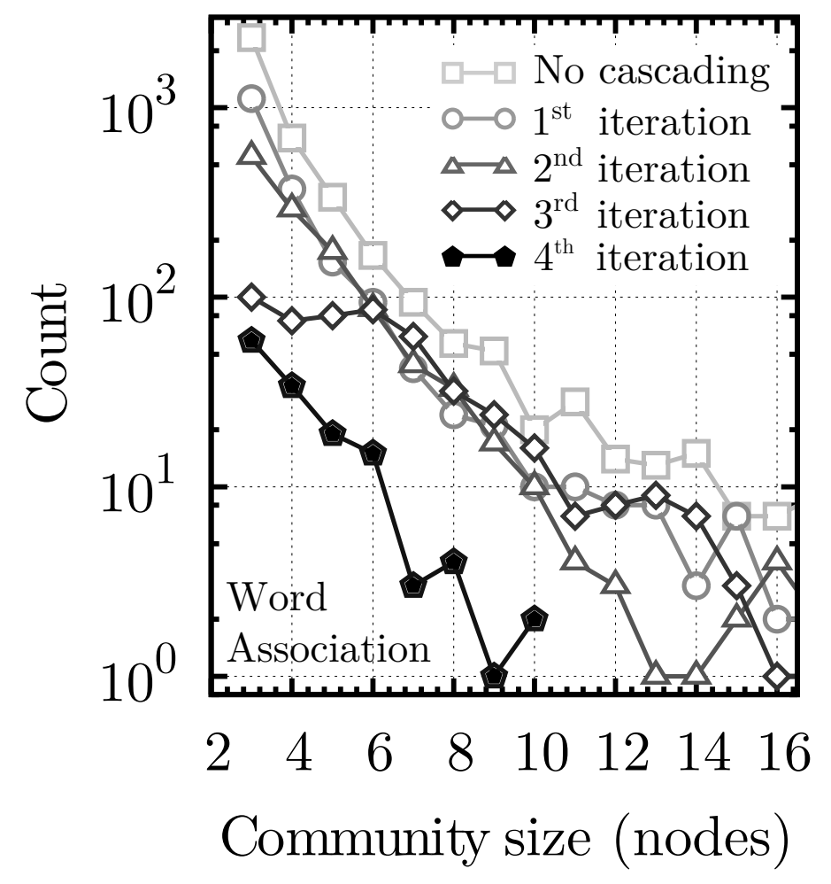
This repeated analysis does not increase the computational cost significantly, since very few iterations are usually necessary, and since the network becomes sparser at each step.

Results

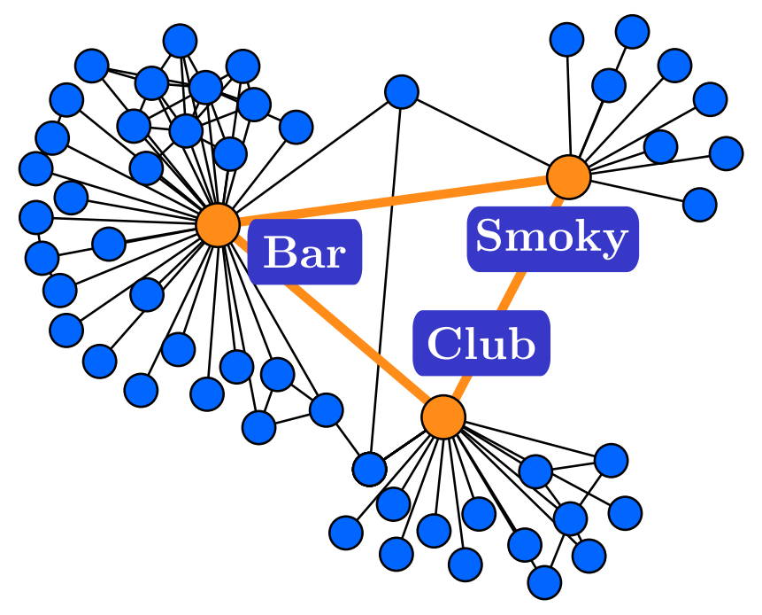
To investigate the efficiency and the behavior of cascading detection, we have applied our approach to **8 real networks** using the LCA. Our results show that cascading detection always improves the thoroughness of the community structure detection (the average percentage of unassigned links drops from 41.0% to 5.3%).



We find **meaningful communities at multiple layers of organizations**. Pictured here is the distribution of the community sizes for each iteration. It suggests that communities unveiled through cascading detection are similar to the ones detected at the first iteration (i.e. traditional use of the detection algorithm).



Visual inspection of the detected communities confirms the **quality of the hidden communities**, as well as our intuition of the shadowing effect. For example, this triangle was detected at the third iteration of the algorithm on the Words network. This structure was overlooked during the initial detection due to the high degree of its three nodes.



Conclusion

We have introduced two **meta-algorithms for community detection** that address two types of resolution problems.

Partitioning as a detection tool

- By analyzing multiple partitions of the same network using a local quality function, **we find resolution limit free community structures where modularity would have merged groups**.
- Since optimized partitioning algorithms scale nicely with the size of the system, our approach is applicable to large networks.

Cascading detection

- We recover multiple levels of meaningful organization, **discovering hidden communities along the way.**
- Our meta-algorithm is not significantly slower than traditional ones, permitting cascading detection on large networks.

Future directions

- Extension of the partitioning method to all **Potts models** based algorithm, and to line-graph partitioning.
- Improved cascading detection:
  - ◇ **Sequential detection** (i.e. detect communities one by one)
  - ◇ A less destructive approach to link removal (e.g. **avoid the removal of important links** that are shared by communities at lower layers of organization)

Acknowledgements

NSERC CRSNG

CIHR IRSC Canadian Institutes of Health Research Instituts de recherche en santé du Canada

Calcul Québec

[1] V. A. Traag *et al.*, "Narrow scope for resolution-limit-free community detection," Phys. Rev. E, **84**, 016114, (2011).

[2] Y.-Y. Ahn *et al.*, "Link communities reveal multiscale Complexity in Networks," Nature, **466**, 761, (2010).

[3] M. A. Riolo, & M. E. J. Newman, "First-principles multiway spectral partitioning of graphs," arXiv:**1209.5969**, (2012).

[4] J.-G. Young *et al.*, "Unveiling hidden communities through cascading detection on network structures," arXiv:**1211.1364**, (2012). *To appear in Springer's Lecture Notes in Computer Science.*

jean-gabriel.young.1@ulaval.ca