Fast unfolding of communities in large networks

R. Lambiotte
Institute for Mathematical Sciences
Imperial College London

with V.D. Blondel, J.-L. Guillaume and E. Lefebvre

Imperial College London
100 years of living science
100

Contents

Paper available at: arXiv:0803.0476v1
C++ and matlab code available at http://findcommunities.googlepages.com

Networks

Mathematical tool to describe systems composed of elements and of the relations between them (directed or not)

Internet

Transport networks

Power grids

Protein interaction networks

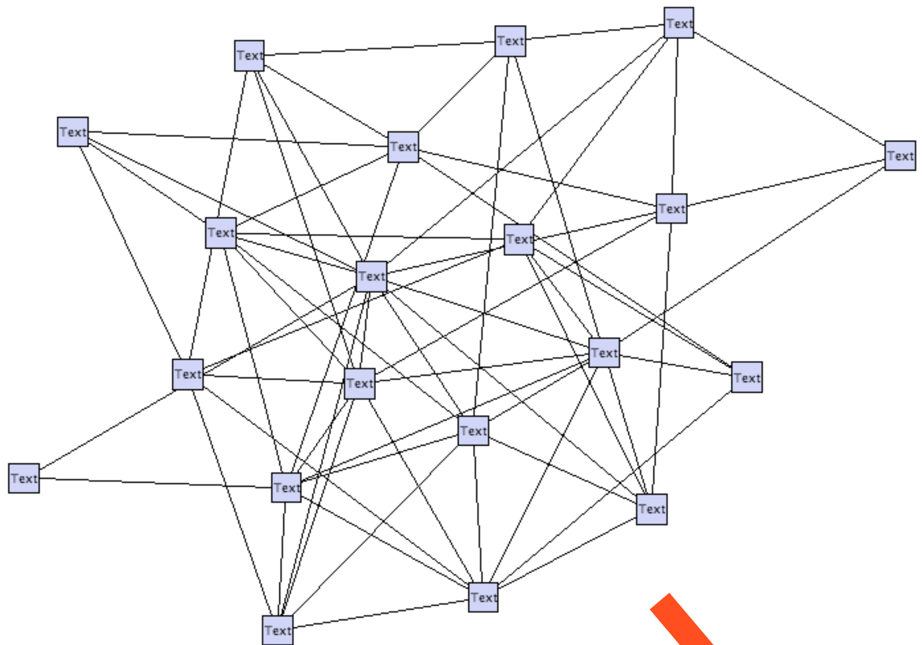
Food webs

Metabolic networks

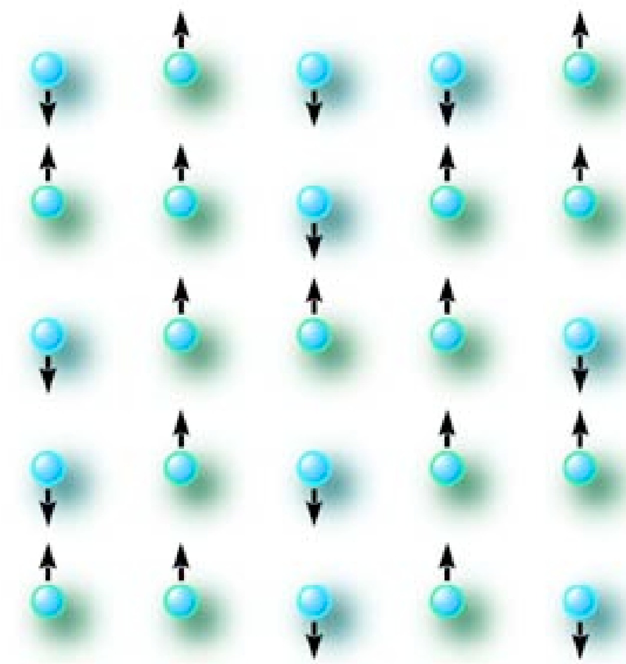
Social networks

Etc.

Complex networks



Random topology:
Mean Field



Regular lattice

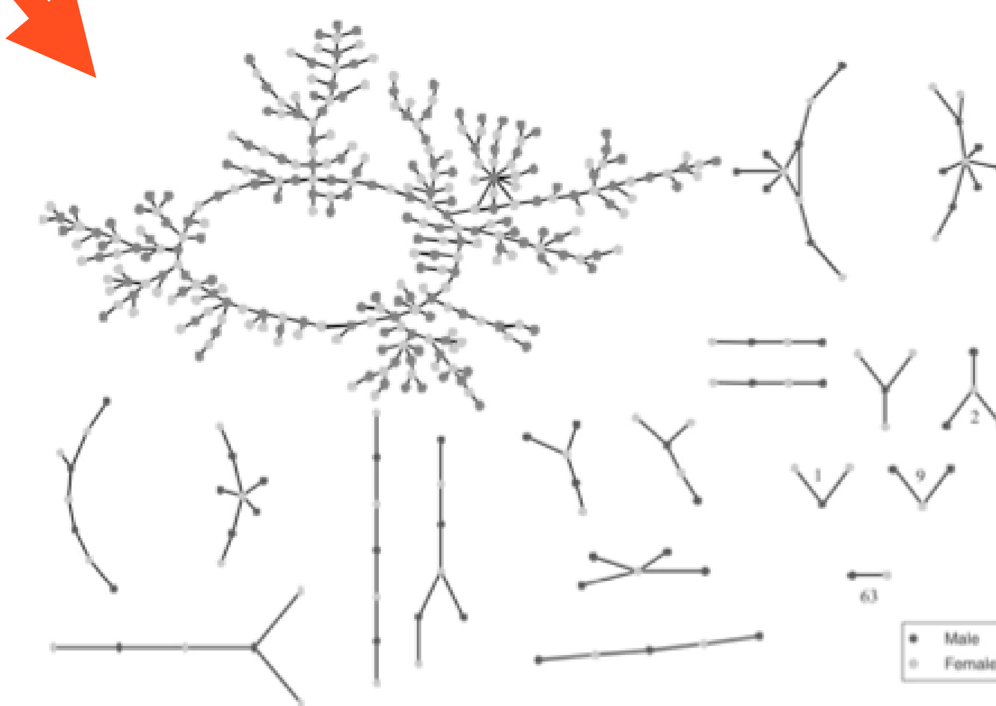


FIG. 2.—The direct relationship structure at Jefferson High

Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks:
Peter S. Bearman, James Moody and Katherine Stovel

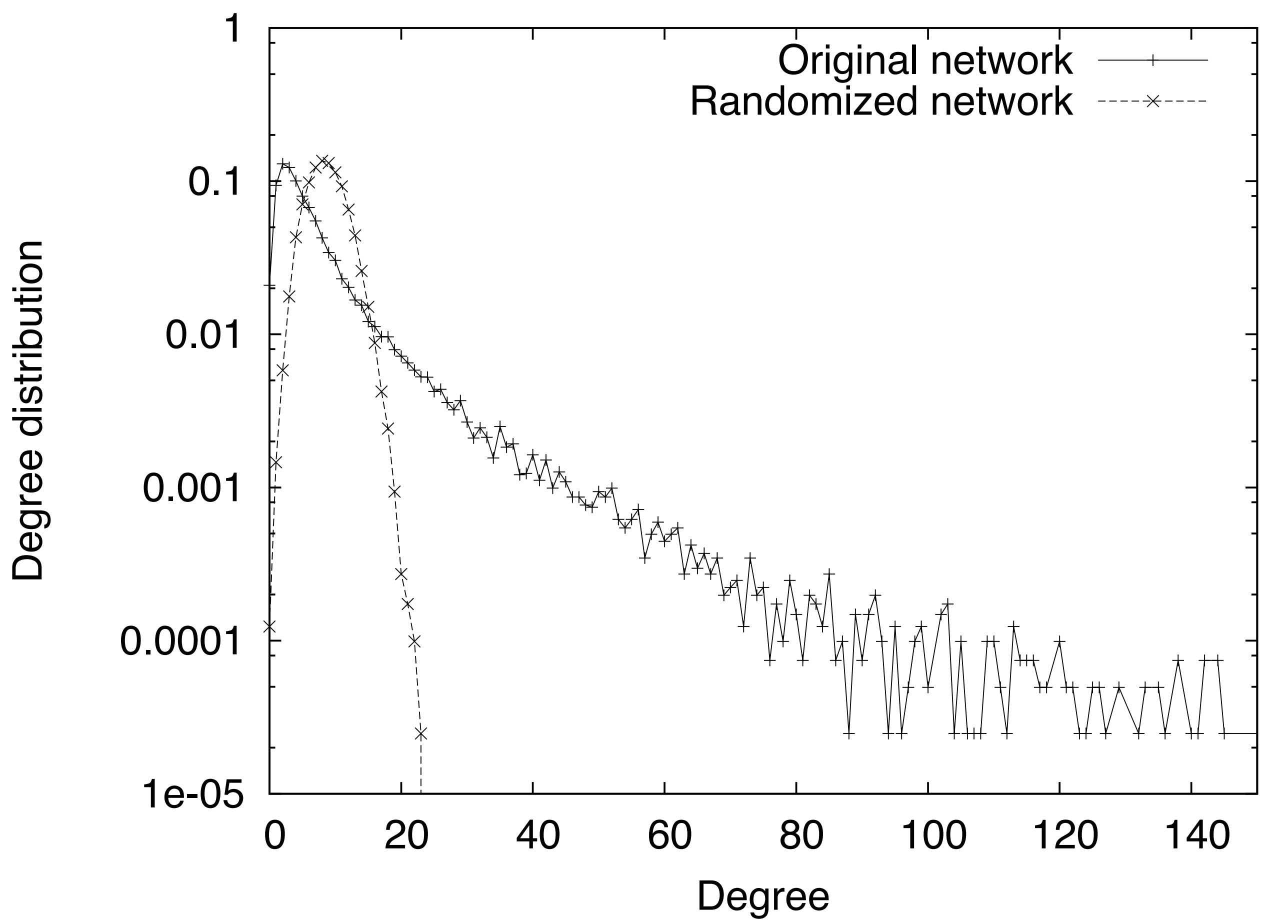
Complex networks

When composed by a large number of nodes and links

Statistical approach

- Degree distribution: overpopulated tail (scale-free or stretched exponential)
- Density of triangles (local structure; transitivity)
- Small "Diameter"
- Correlation of the degrees around one link

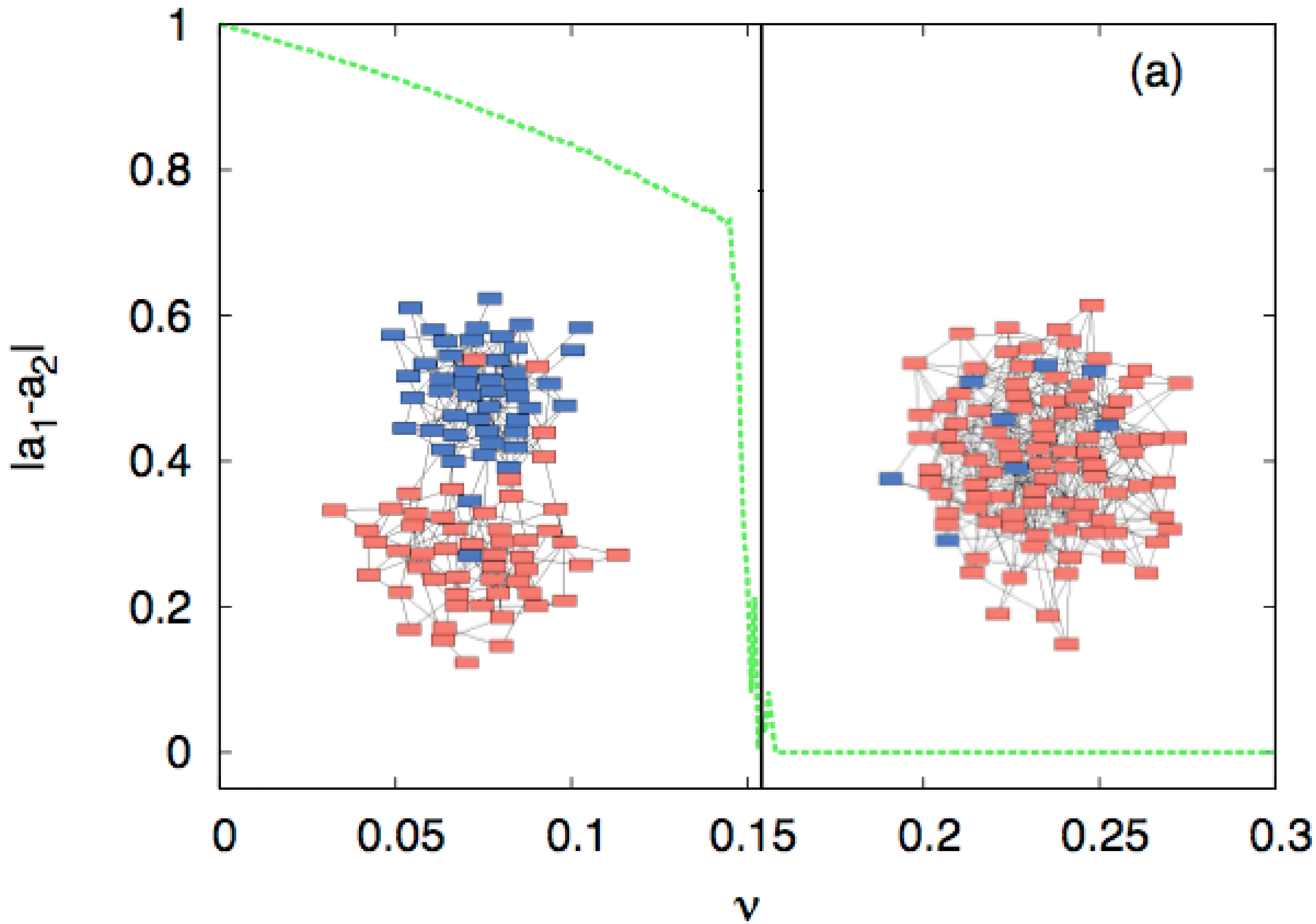
Etc.



Network of collaboration from the arXiv (1995-2005). Around 40000 nodes.

Complex networks and Dynamics

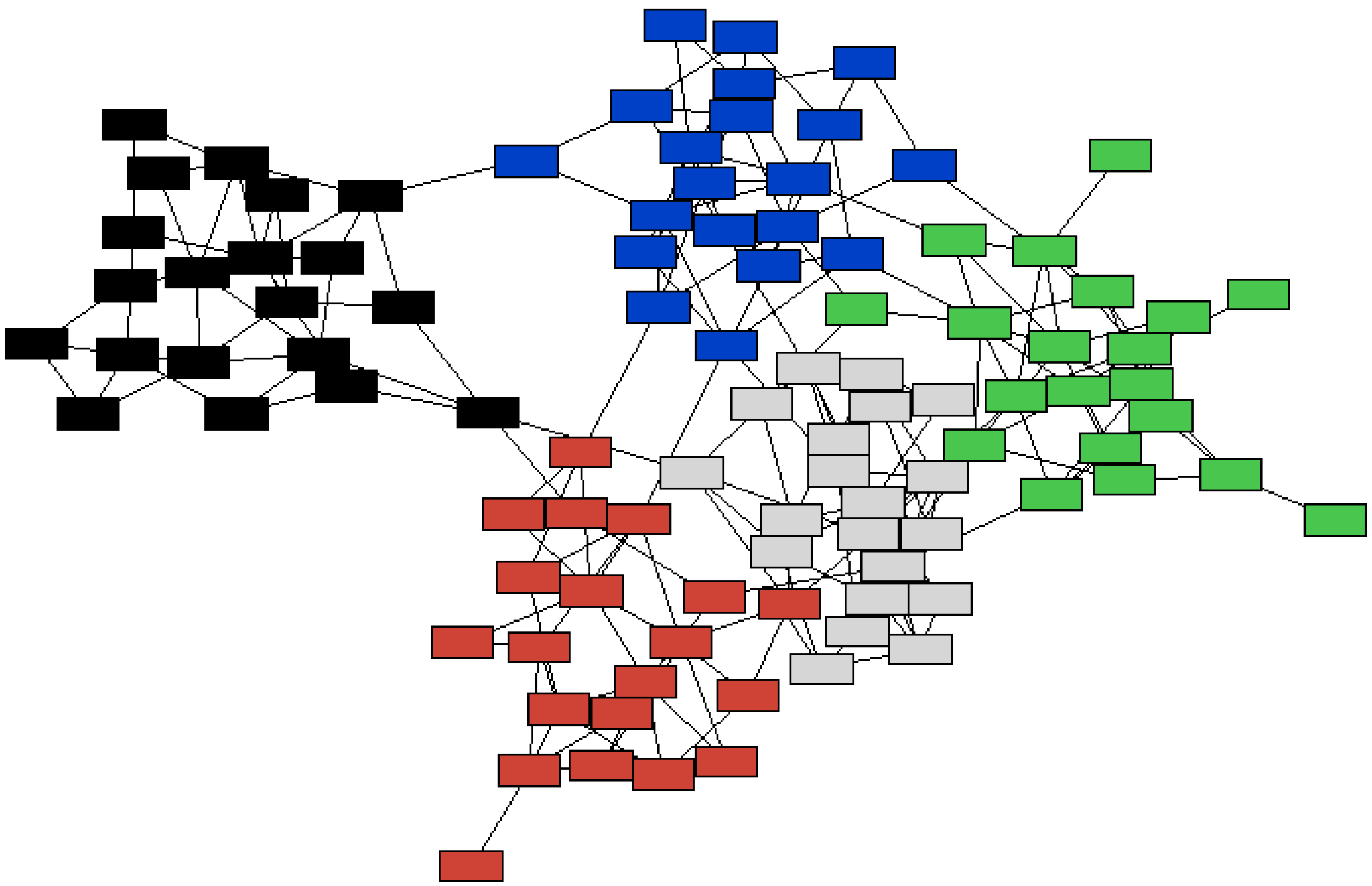
How does the topology affect the dynamics taking place on it (diffusion, synchronisation, etc.)



R. Lambiotte and M. Ausloos, J. Stat. Mech., (2007) P08026

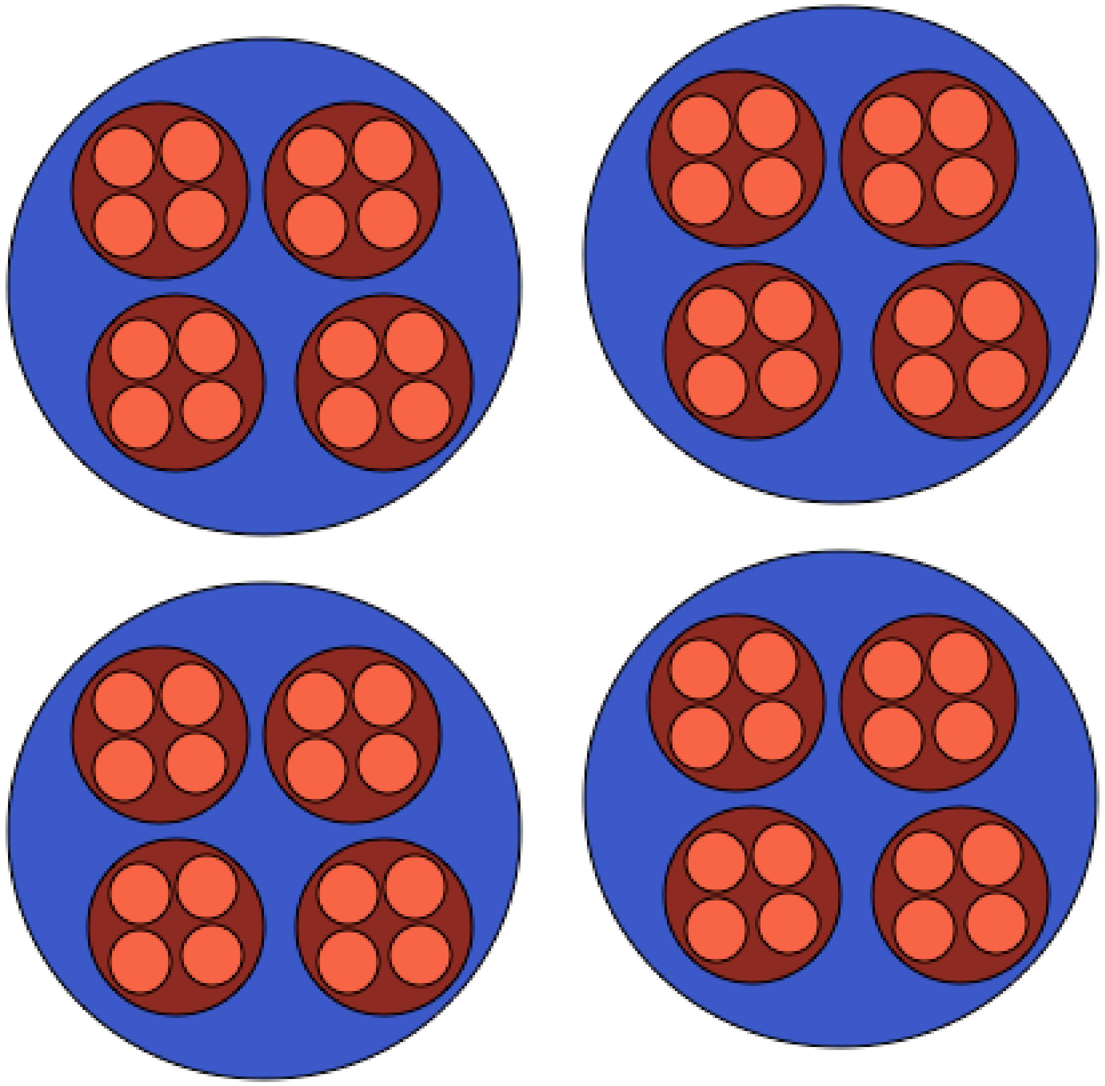
Modular Networks

Most networks are very inhomogeneous and are made of modules:
many links within modules and a few links between different modules



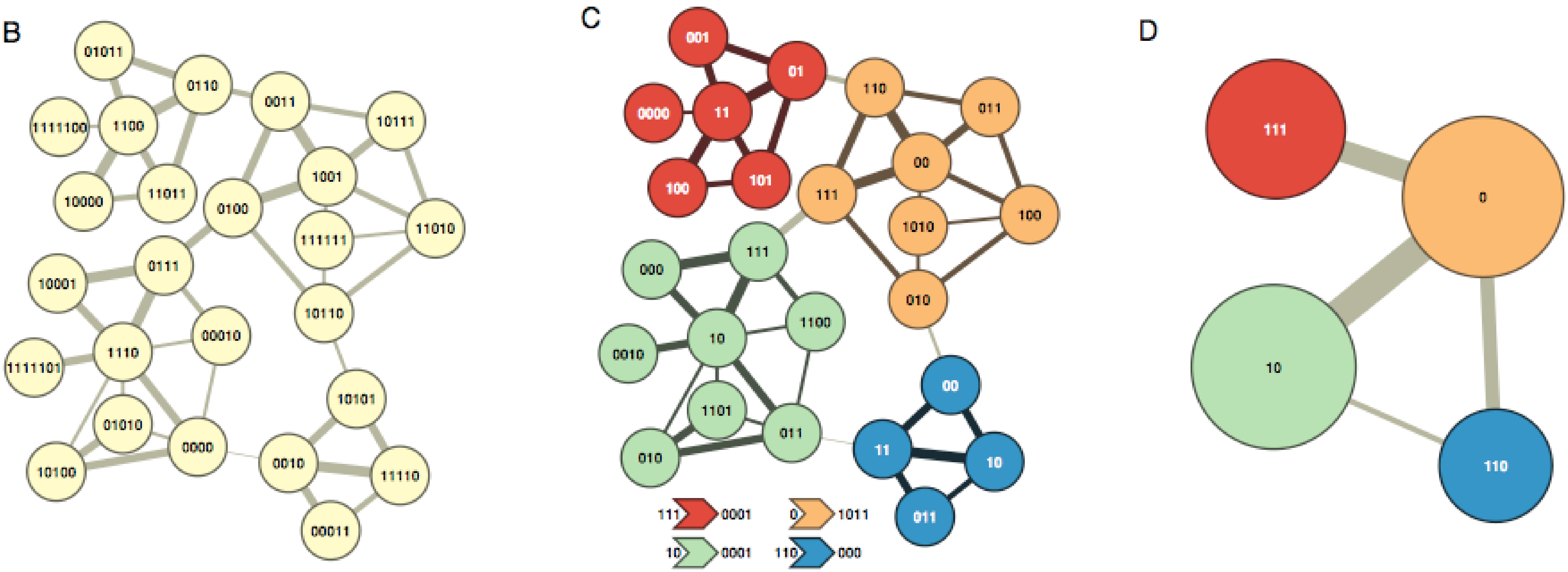
Modular Networks

Networks have a hierarchical structure: communities within communities



Modular Networks

Uncovering communities/modules helps to understand the structure of the network, to uncover similar nodes and to draw a readable map of the network (when N is large).



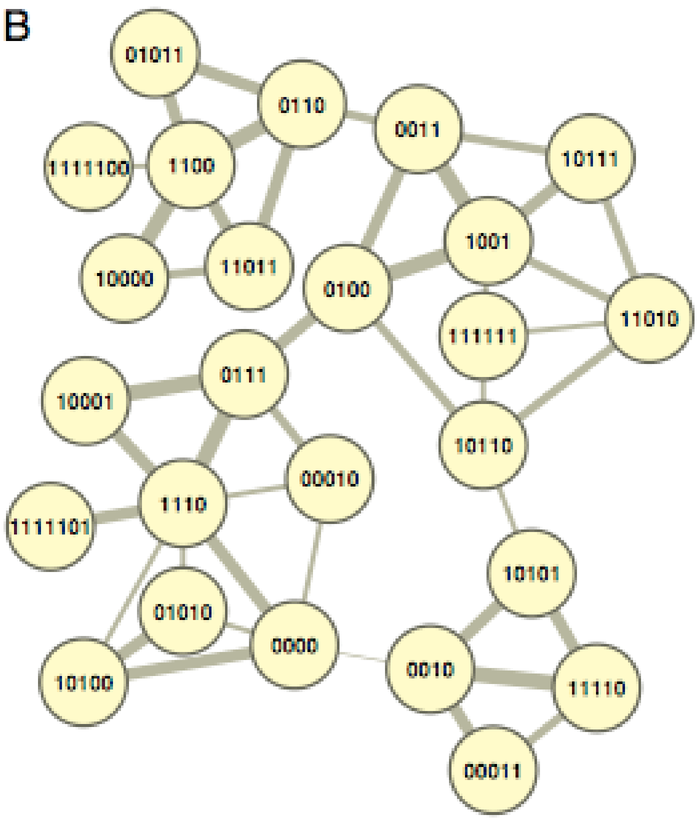
Find a partition of the network into communities

Coarse-grained description

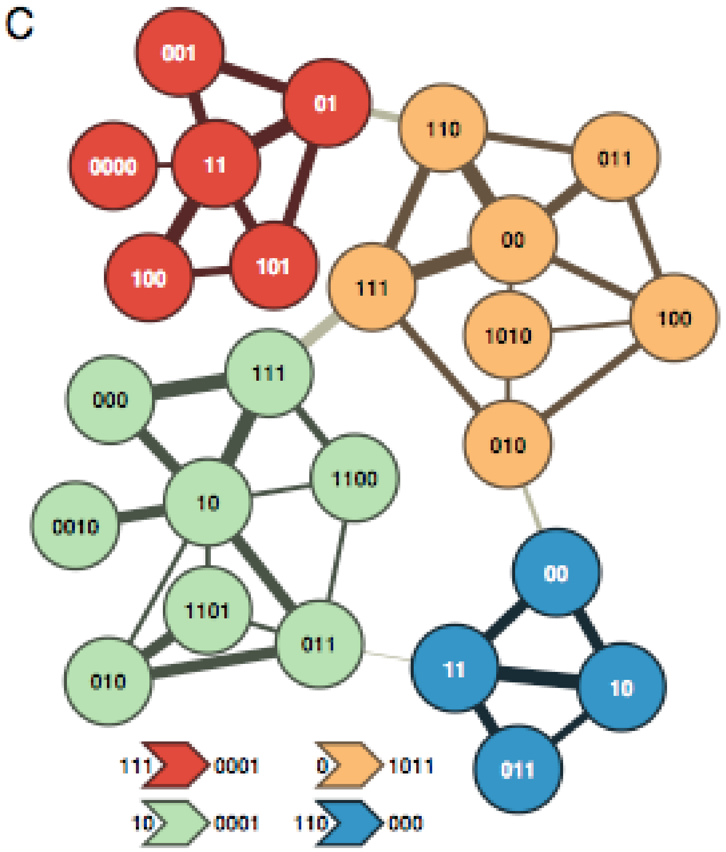
Martin Rosvall and Carl T. Bergstrom, PNAS 105, 1118 –1123 (2008)

Modular Networks

B



C



As rapid and accurate as possible

Unsupervised

Modularity

Modularity: function which measures the quality of a partition

Q=(number of edges within communities)-(expected number of such edges)

Larger values of Q=> stronger community structure

Comparison of the network with a null model (randomised model)

M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)

Modularity

Modularity: function which measures the quality of a partition

Let us focus on a weighted network and attribute to each node a community c_i

$$\boxed{Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)}$$

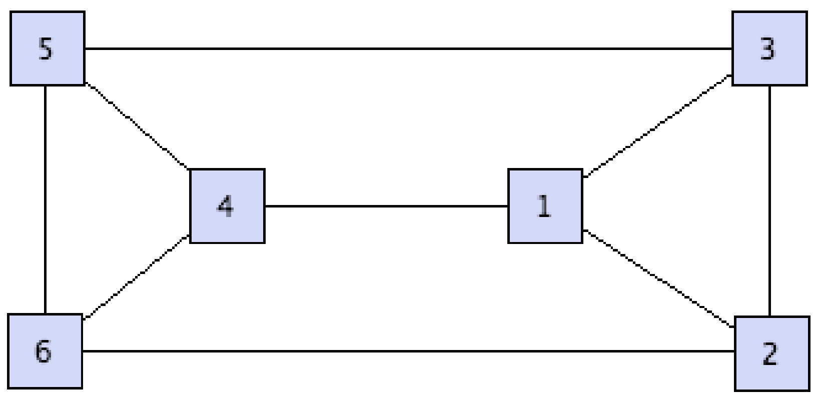
A_{ij} weight of the edge between i and j

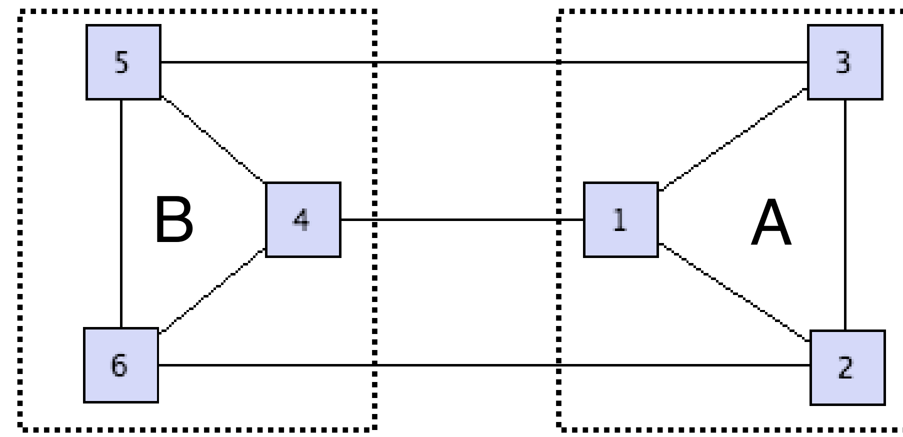
$k_i = \sum_j A_{ij}$ *degree* of i

$$Q \in [-1, 1]$$

$m = \frac{1}{2} \sum_{ij} A_{ij}$ total weight

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$





The degree is the same for each node: 3

The total number number of links is: 9

In community A, 6 pairs of nodes i,j, [(1,2),(2,1),(1,3), etc.] with a contribution: 1-9/18=1/2

In community A, 3 loops i,i missing, with a contribution -1/2

6 (1/2)-3(1/2)=3/2

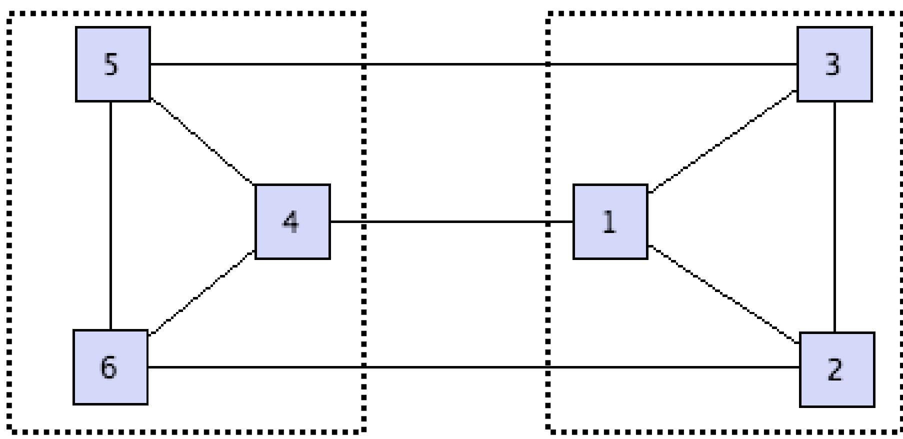
Same contribution from community B

Modularity is therefore:

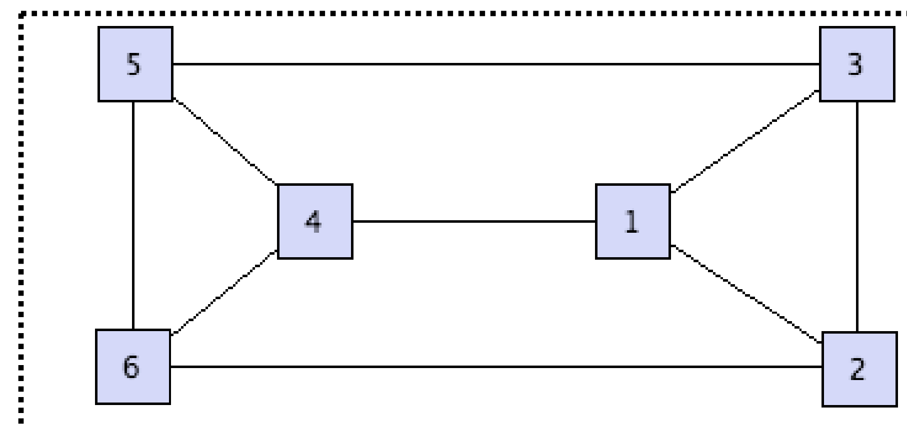
1/(2 * 9) * 2 * 3/2= 1/6

Modularity

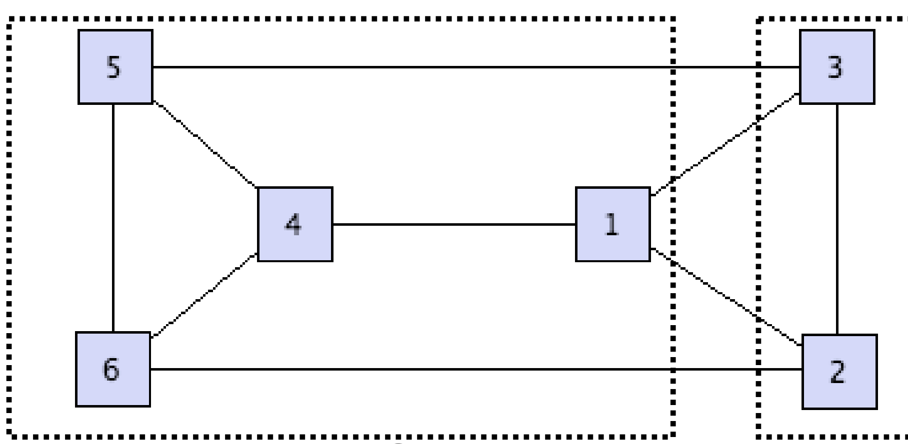
Modularity is used to validate the quality of a partition, but also as a quantity to optimise in order to find the best partition: Given a network, what is the partition with the highest value of Q



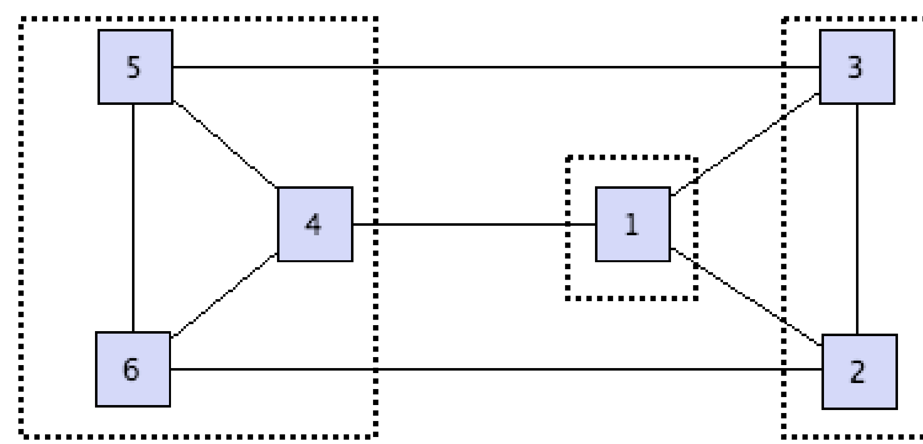
$$Q_{max} = 1/6$$



$$Q = 0$$



$$Q = 0$$



$$Q = 1/18$$

Modularity

Modularity is used to validate the quality of a partition, but also as a quantity to optimise in order to find the best partition: Given a network, what is the partition with the highest value of Q

Finding an exact answer is not possible in large (and less regular) networks

=> need for algorithms which find *good partitions* in short times

The largest network studied so far is a Japanese social networking systems of about 5.5 million (Wakita and Tsurumi) to be compared with:

Facebook has about 64 million active users

Vodaphone has about 200 million customers and Google indexes several billion web-pages.

Our algorithm

The algorithm is based on two steps that are repeated iteratively.

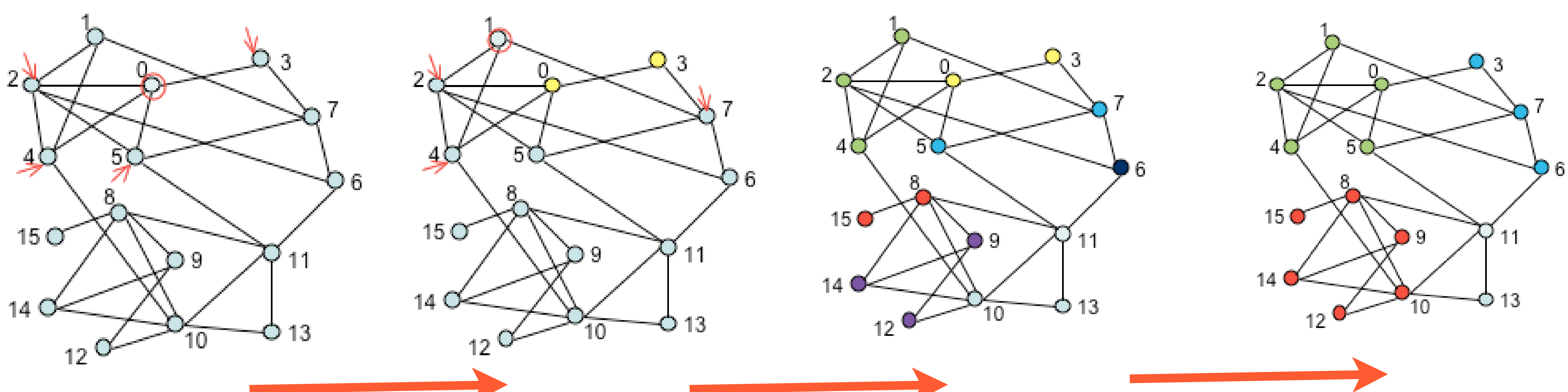
First phase: Find a local maximum

1) Give an order to the nodes (0,1,2,3,...., N-1)

2) Initially, each node belongs to its own community (N nodes and N communities)

3) One looks through all the nodes (from 0 to N-1) in an ordered way. The selected node looks among its neighbours and adopt the community of the neighbour for which the increase of modularity is maximum (and positive).

4)This step is performed iteratively until a local maximum of modularity is reached (each node may be considered several times).



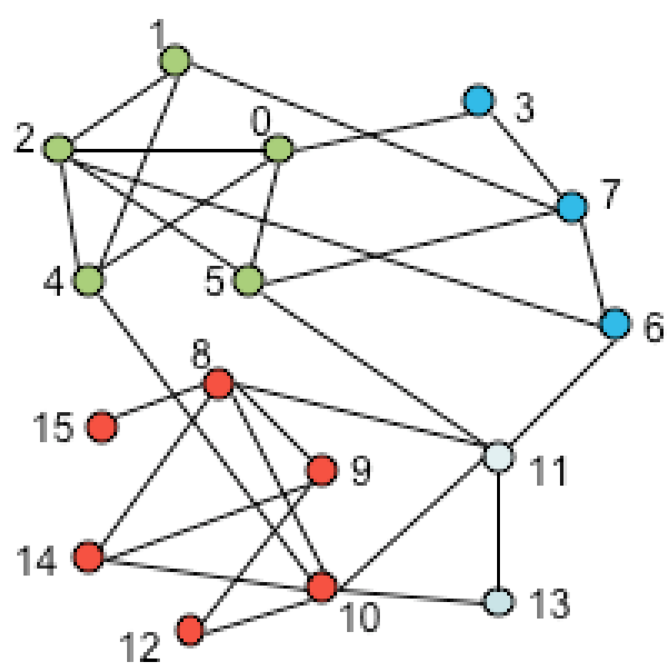
Node 0 moves to the community of Node 3

After N nodes have been considered

After each nodes has been considered 4 times

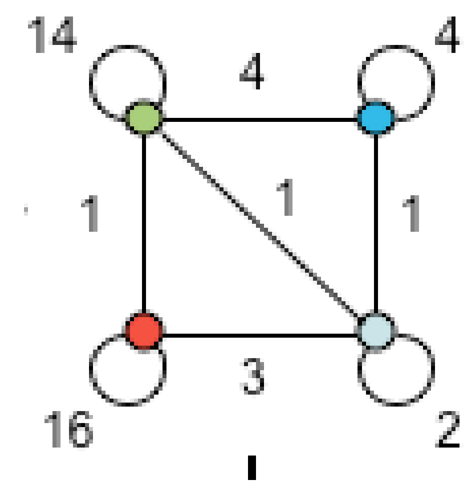
Our algorithm

Once a local maximum has been attained, we build a new network whose nodes are the communities. The weight of the links between communities is the total weight of the links between the nodes of these communities.



New network of 4 nodes!

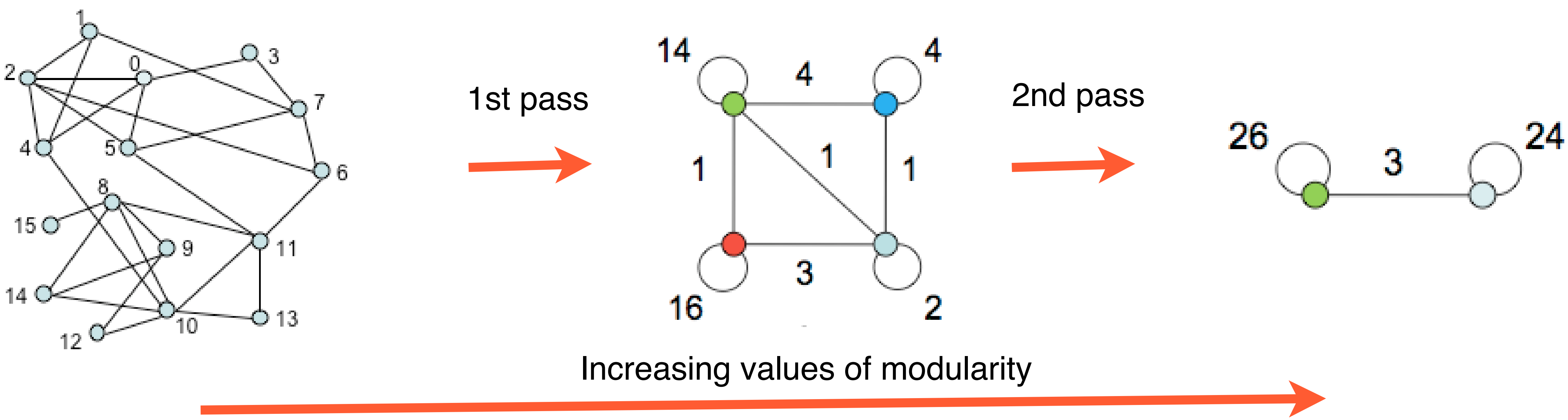
Note the self-loops



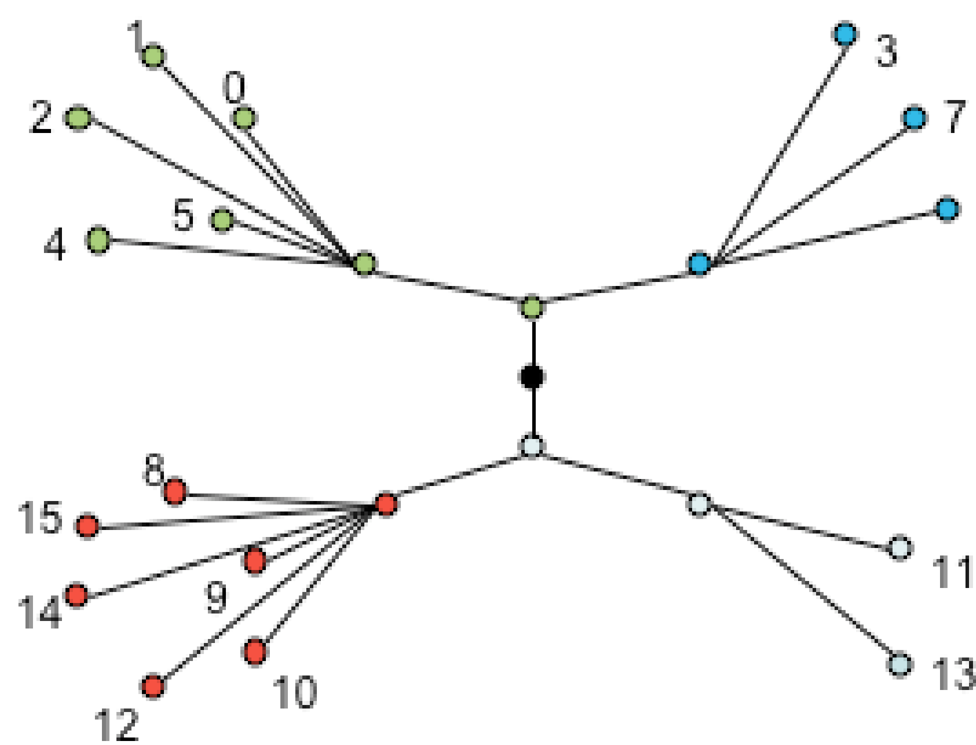
In typical realizations, the number of nodes diminishes drastically at this step.

Our algorithm

The two steps are repeated iteratively, thereby leading to a hierarchical decomposition of the network.



Hierarchical representation



Advantages

Incredibly simple (local greedy approach)=> easy to implement and to improve

Advantages

Incredibly simple (local greedy approach)=> easy to implement and to improve

Very fast!

Largest network: 5.5 million nodes

	Karate	Arxiv	Internet	Web nd.edu	Phone	Web uk-2005	Web WebBase 2001
Nodes/links	34/77	9k/24k	70k/351k	325k/1M	2.04M/5.4M	39M/783M	118M/1B
CNM	.38/0s	.772/3.6s	.692/799s	.927/5034s	-/-	-/-	-/-
PL	.42/0s	.757/3.3s	.729/575s	.895/6666s	-/-	-/-	-/-
WT	.42/0s	.761/0.7s	.667/62s	.898/248s	.553/367s	-/-	-/-
Our algorithm	.42/0s	.813/0s	.781/1s	.935/3s	.76/44s	.979/738s	.984/152mn

Our only limitation being the storage of the network in main memory

Computer simulations on large ad-hoc modular networks suggest that its complexity is linear on typical and sparse data

Clauset A, Newman M E J and Moore C, 2004 Phys. Rev. E 70 066111.
Wakita K and Tsurumi T, 2007 Proceedings of IADIS international conference on WWW/Internet 2007 153.
Pons P and Latapy M, 2006 Journal of Graph Algorithms and Applications 10 191.

Advantages

Incredibly simple (local greedy approach)=> easy to implement and to improve

Very fast!

Good accuracy

Advantages

Incredibly simple (local greedy approach)=> easy to implement and to improve

Very fast!

Good accuracy

	Karate	Arxiv	Internet	Web nd.edu	Phone	Web uk-2005	Web WebBase 2001
Nodes/links	34/77	9k/24k	70k/351k	325k/1M	2.04M/5.4M	39M/783M	118M/1B
CNM	.38/0s	.772/3.6s	.692/799s	.927/5034s	-/-	-/-	-/-
PL	.42/0s	.757/3.3s	.729/575s	.895/6666s	-/-	-/-	-/-
WT	.42/0s	.761/0.7s	.667/62s	.898/248s	.553/367s	-/-	-/-
Our algorithm	.42/0s	.813/0s	.781/1s	.935/3s	.76/44s	.979/738s	.984/152mn

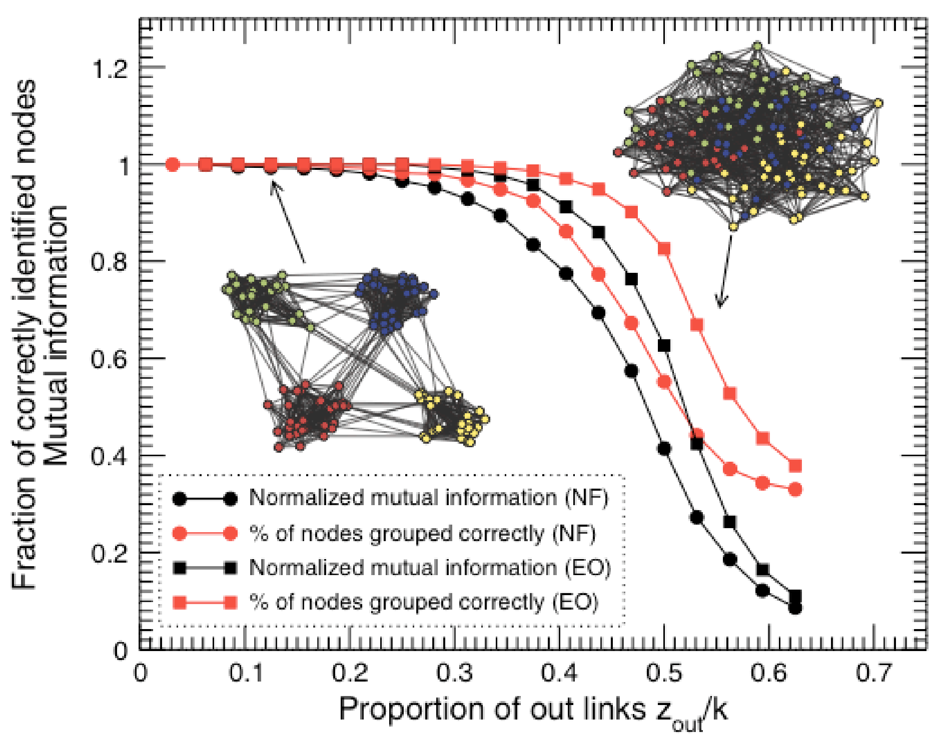
Advantages

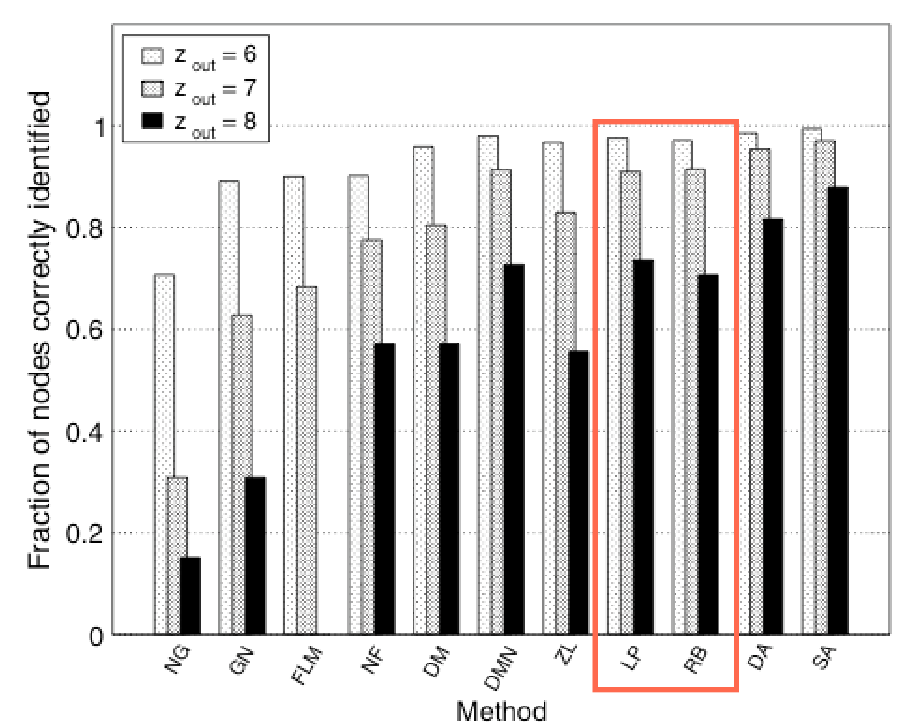
Incredibly simple (local greedy approach)=> easy to implement and to improve

Very fast!

Good accuracy

On ad hoc networks with a well-known community structure:





Comparing community structure identification , L Danon, A Diaz-Guilera, J Duch, A Arenas, *J. Stat. Mech.*, P09008 (2005)

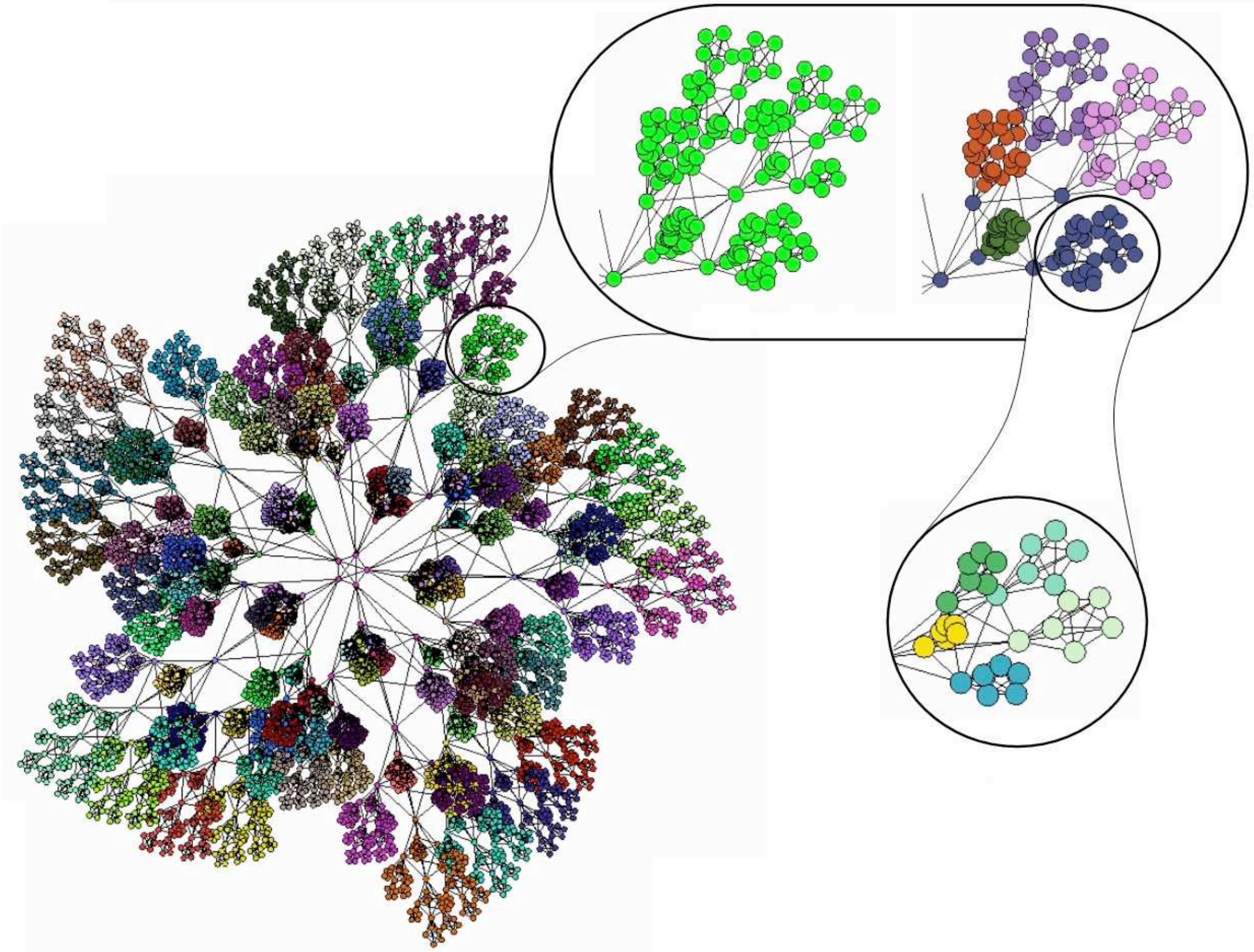
Advantages

Incredibly simple (local greedy approach)=> easy to implement and to improve

Very fast!

Good accuracy

Multiresolution



Conclusion

- Can deal with millions/billions nodes/links
- Achieves very good modularity
- Directly produces a hierarchy structure;
- Is strikingly simple;
- Can use other local quality functions (instead of modularity);
- C++ and matlab available on findcommunities.googlepages.com