Beyond Space For Spatial Networks

R. Lambiotte
Institute for Mathematical Sciences
Imperial College London

with P. Expert, T.S. Evans and V.D. Blondel

Complexity & Networks

Imperial College London
100 years of living science
100

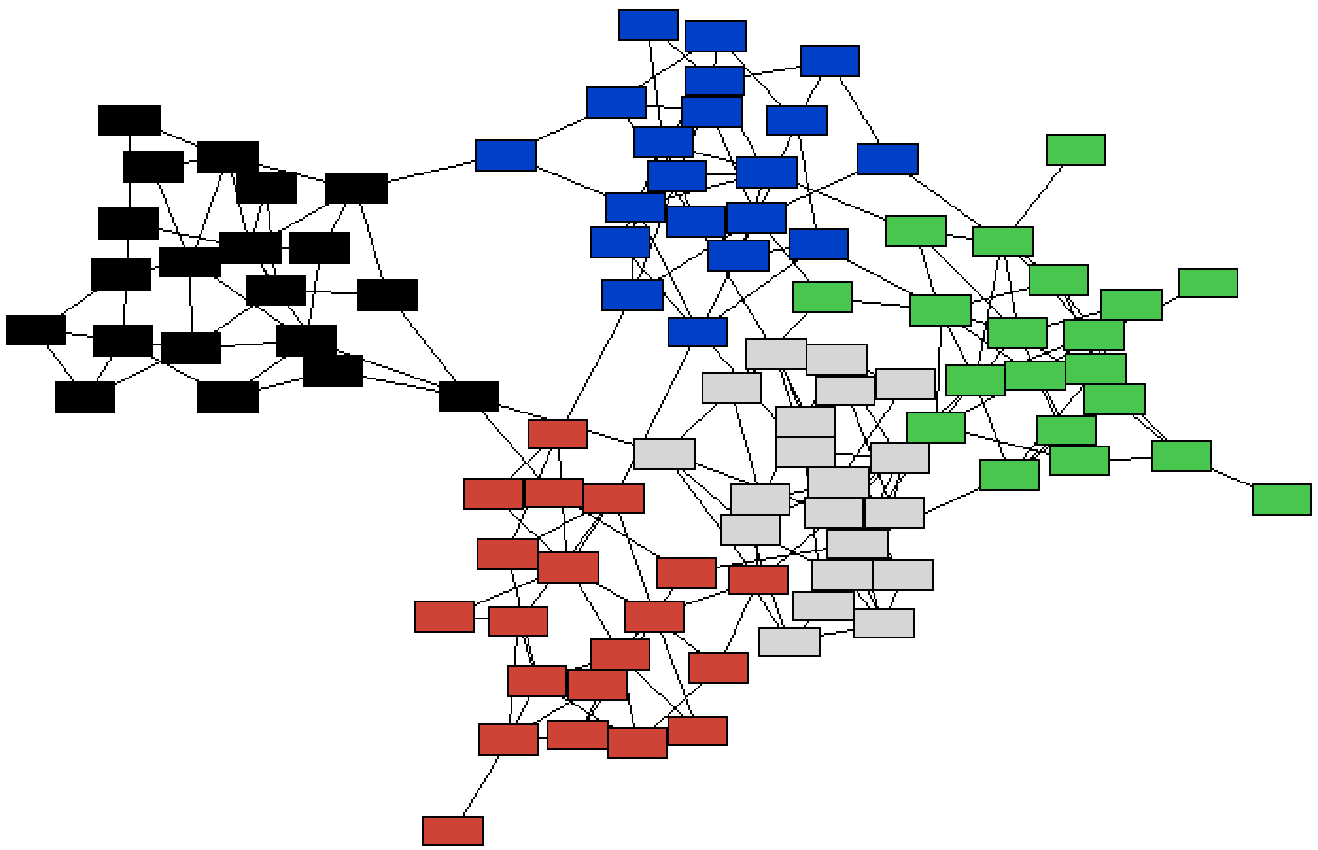
P. Expert, T.S. Evans, V.D. Blondel and R. Lambiotte, *arXiv:1012.3409*

1. Modules and Hierarchies
2. Modularity and chance
3. Spatial constraints on network topology
4. Empirical tests and benchmarks

Modular Networks

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

Modules=communities



Internet

Power grids

Food webs

Metabolic networks

Social networks

The brain

Etc.

Simon, H. (1962). The architecture of complexity. Proceedings of the American Philosophical Society, 106, 467–482.

Why communities, hierarchies?

Generic mechanism driving the emergence of modularity?

Why communities, hierarchies?

Generic mechanism**s** driving the emergence of modularity?

- Watchmaker metaphor: intermediate states/scales facilitate the emergence of complex organisation from elementary subsystems

- Separation of time scales: enhances diversity, locally synchronised states

- locally dense but globally sparse: advantages of dense structures while minimising the wiring cost

- in social systems, offer the right balance between dense networks (foster trust, facilitate diffusion of complex knowledge), and open networks (small diameter, ensures connectivity, facilitates diffusion of “simple” knowledge)

- naturally emerges from co-evolution and duplication processes

- Optimality of modular networks at performing tasks in a changing environment

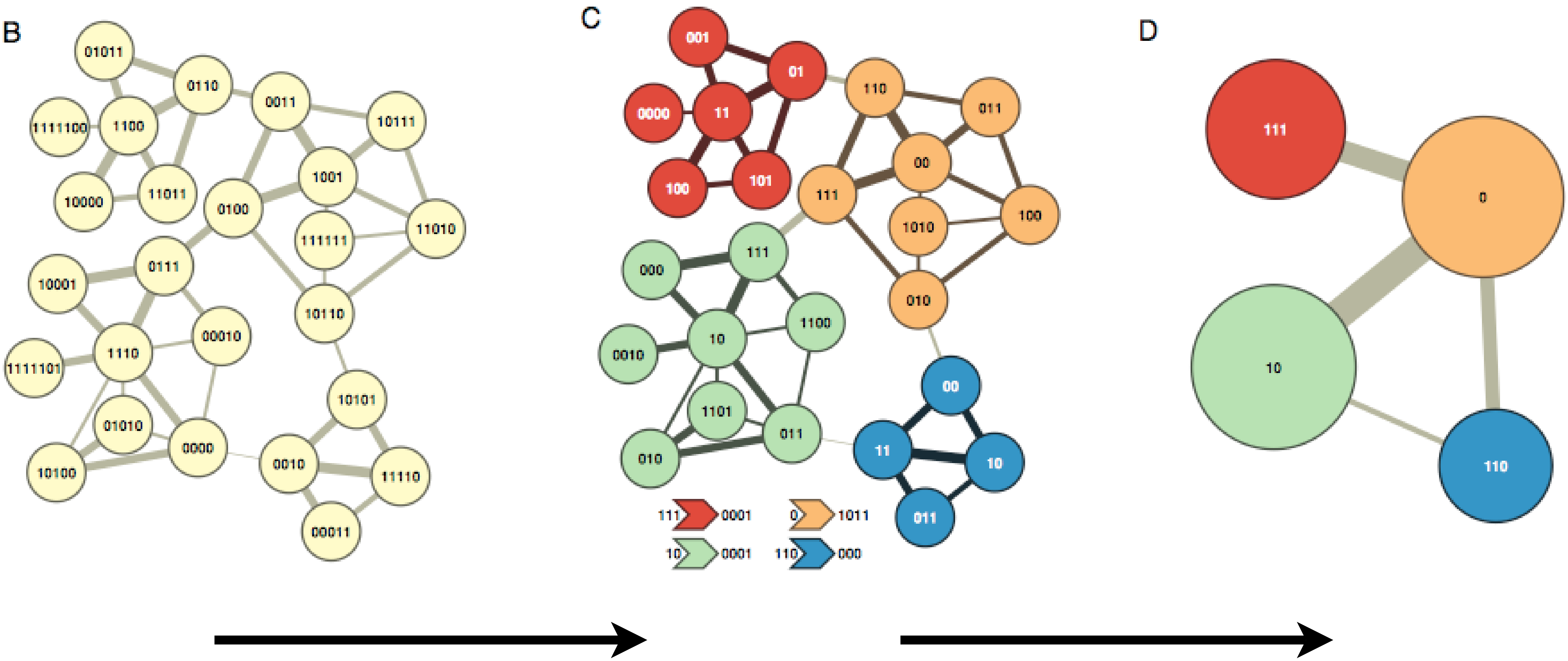
- enhanced adaptivity and dynamical complexity, e.g. transient “chimera” states

- delivers highly adaptive processing systems and to solve the dynamical demands imposed by global integration and functional segregation (brain organisation)

D. Meunier, R. Lambiotte and E.T. Bullmore, “Modular and hierarchical organisation in complex brain networks”, Frontiers in NeuroScience (2010)

Why community detection?

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).



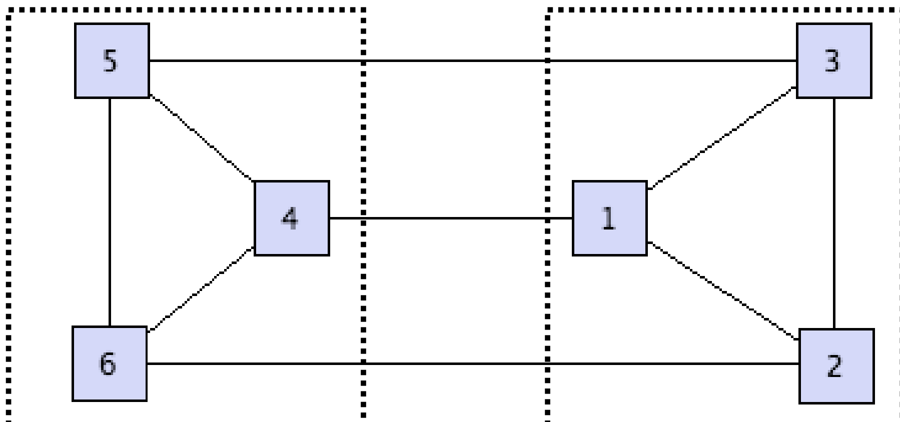
Find a partition of the network into communities

Coarse-grained description

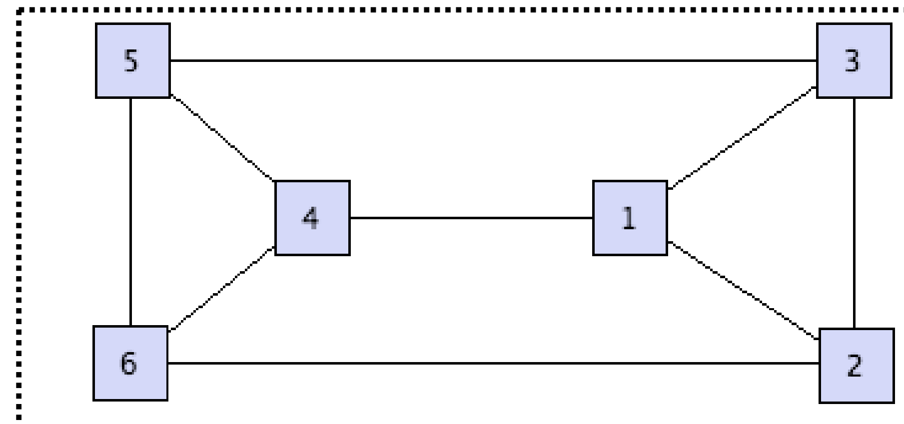
Martin Rosvall and Carl T. Bergstrom, PNAS 105, 1118 –1123 (2008)

Quality of a partition

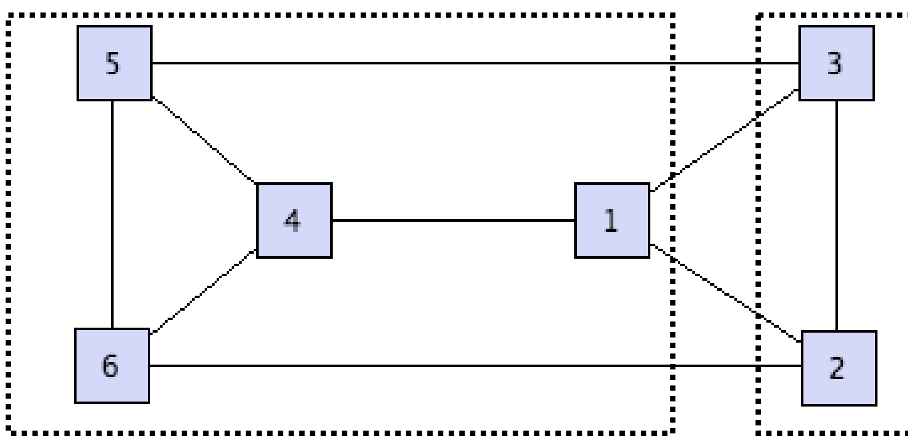
What is the best partition of a network into modules?



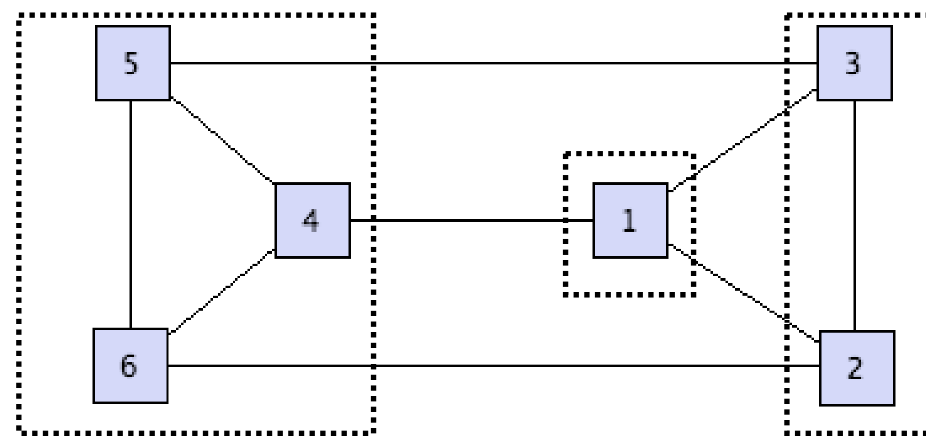
Q1



Q2



Q3



Q4

....

Modularity

Q = fraction of edges within communities - expected fraction of such edges

Let us attribute each node *i* to a community c_i

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

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

$$P_{ij} = \frac{k_i k_j}{2m}$$

expected number of links between i and j

$$\rightarrow \quad Q_C = \frac{1}{2m}\sum_{i,j}\Big[A_{ij} - k_i k_j/2m\Big]\delta(c_i, c_j)$$

Modularity optimisation

Different types of algorithm for different applications:

Small networks (<10^2): Simulated Annealing

Intermediate size (10^2 - 10^4): Spectral methods, PL, etc.

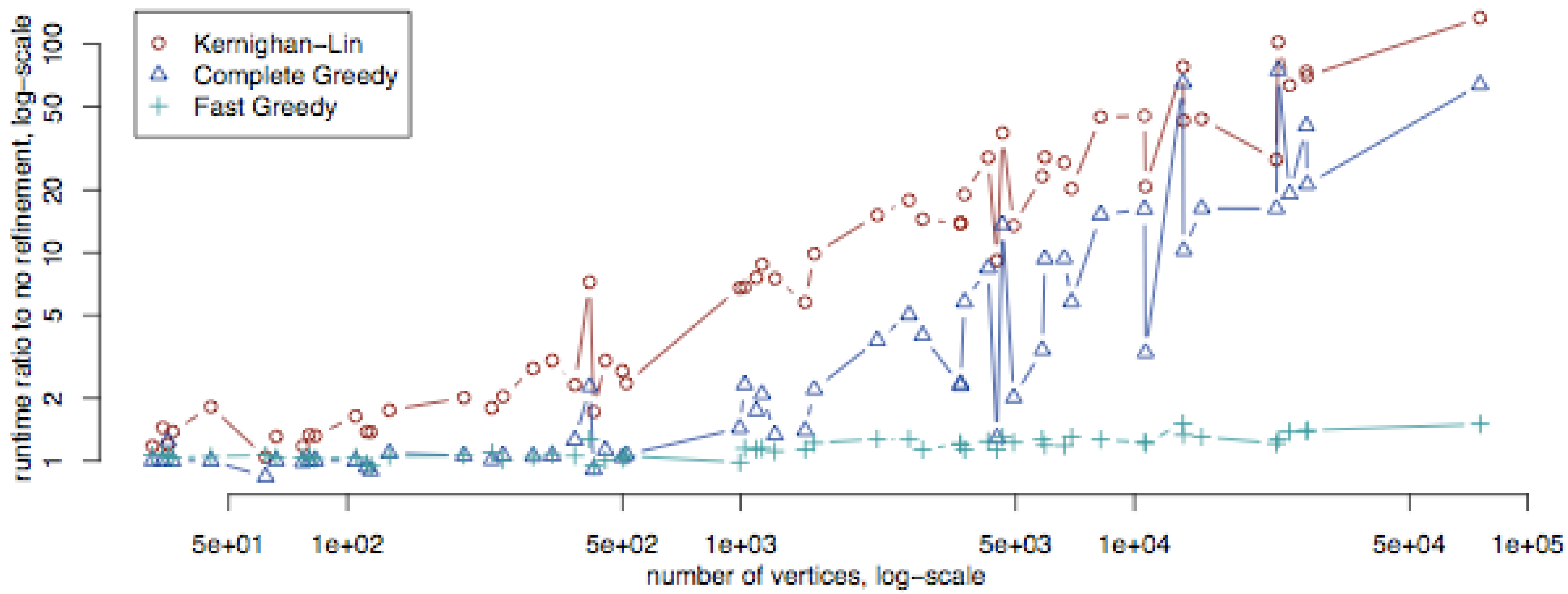
Large size (>10^4): greedy algorithms, e.g. multi-scale optimisation

	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

V.D. Blondel, J.-L. Guillaume, RL and E. Lefebvre, *J. Stat. Mech.*, P10008 (2008).

How to test the methods?

Test the heuristics: what is the value of Q obtained for different algorithms? Time complexity?



graph	size	subdivision		coarsening		local search		math prog		SS+ML
karate [42]	34	[29]	.419	[41]	.4198	[12]	.4188	[1]	.4197	.4197
dolphins [22]	62	[29]	*.4893*	[31]	*.5171*	[33]	*.5285*	[40]	.5285	.5276
polBooks [21]	105	[29]	*.3992*	[37]	*.5269*	[4]	*.5204*	[1]	.5272	.5269
afootball [14]	115	[39]	.602	[41]	.605	[4]	*.6045*	[1]	.6046	.6002
jazz [15]	198	[29]	.442	[9]	.4409	[12]	.4452	[1]	.445	.4446
celeg_metab [12]	453	[29]	.435	[36]	.450	[12]	.4342	[1]	.450	.4452
email [17]	1133	[29]	.572	[9]	.5569	[12]	.5738	[1]	.579	.5774
Erdos02 [16]	6927	[29]	*.5969*	[32]	.6817	[33]	*.7094*			.7162
PGP_main [5]	11k	[29]	.855	[9]	.7462	[12]	.8459			.8841
cmat03_main [25]	28k	[29]	.723	[41]	.761	[12]	.6790			.8146
ND_edu [2]	325k			[7]	.927	[4]	.935			.9509

How to test the methods?

Comparison with real-world data: do modules reveal nodes having similar meta-data?

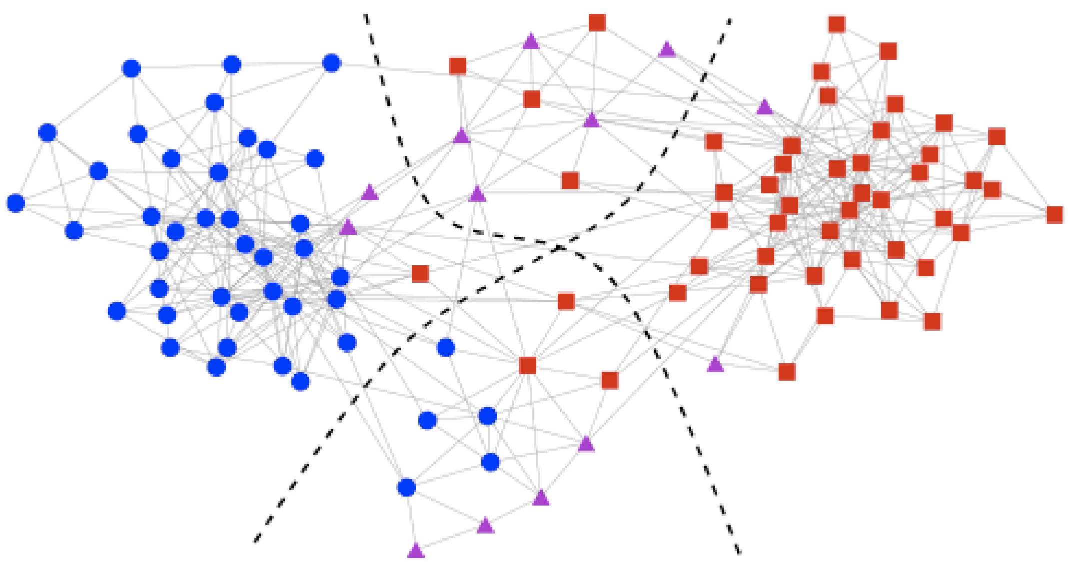
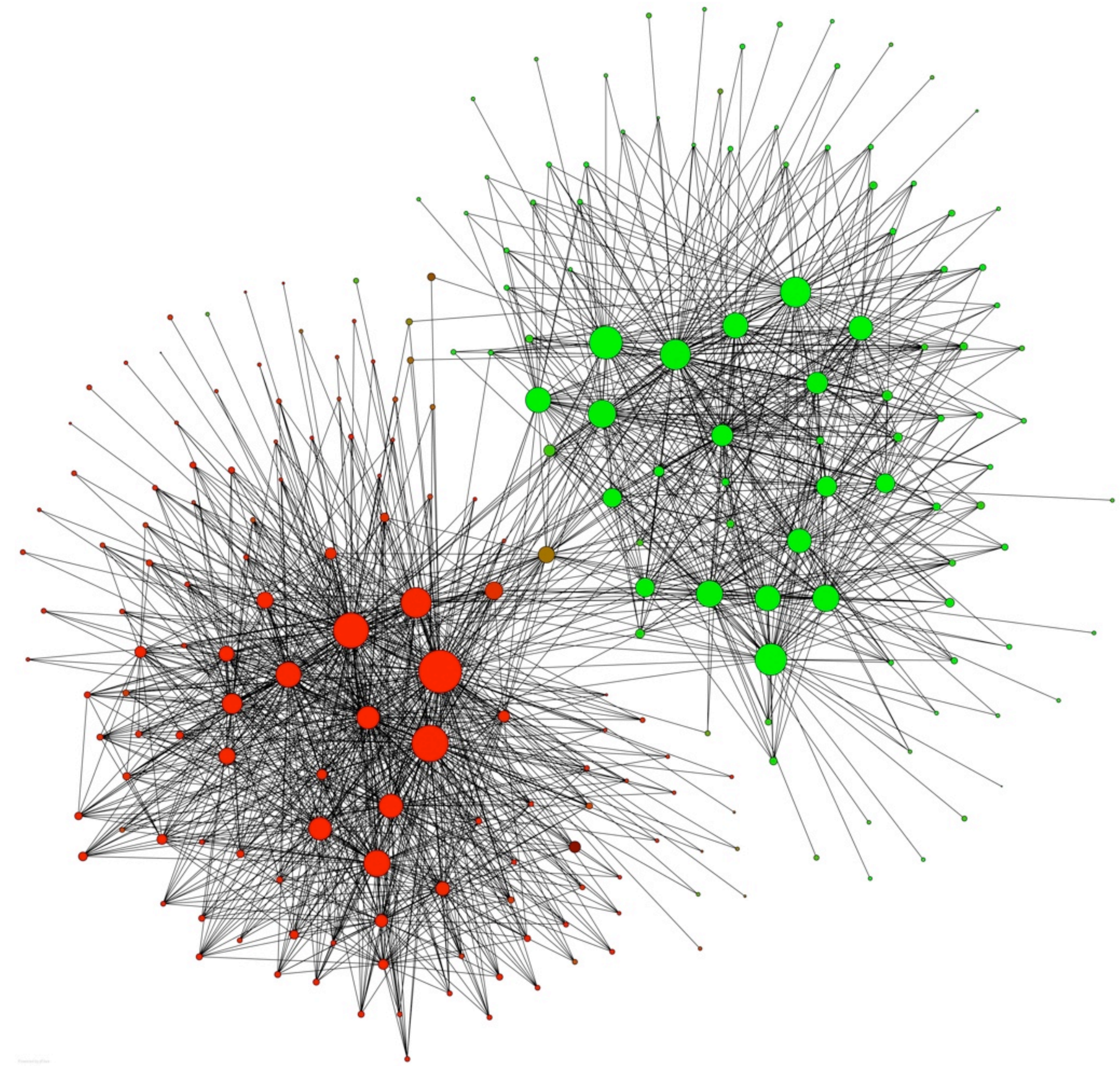


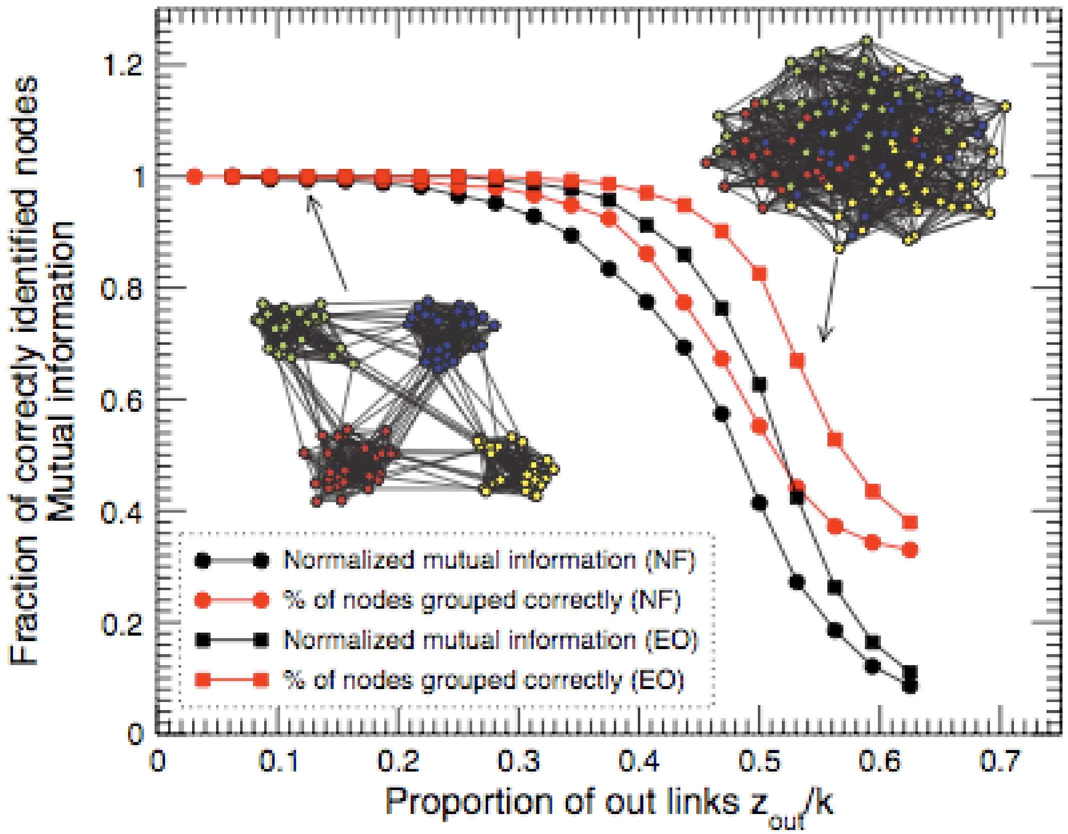
FIG. 3: Krebs' network of books on American politics. Vertices represent books and edges join books frequently purchased by the same readers. Dotted lines divide the four communities found by our algorithm and shapes represent the political alignment of the books: circles (blue) are liberal, squares (red) are conservative, triangles (purple) are centrist or unaligned.

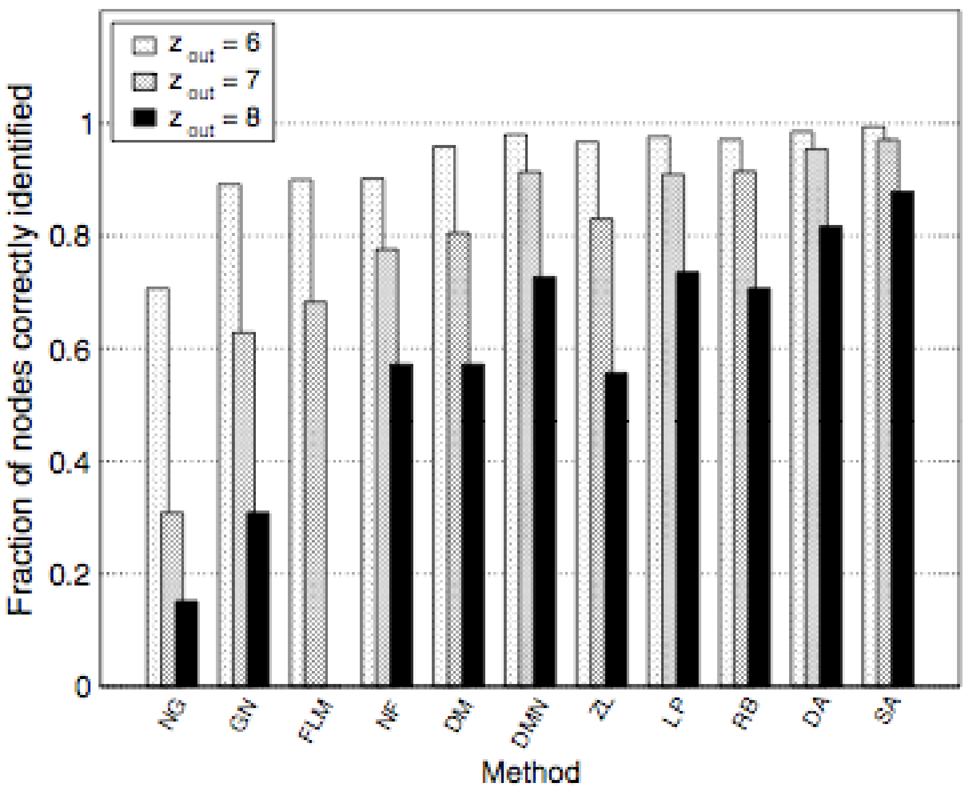


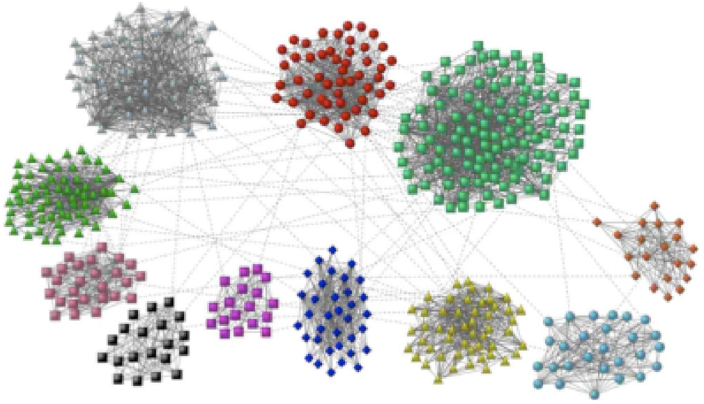
But: meta-data are often unknown. No insurance that modular organization coincides with semantic/cultural organisation

How to test the methods?

Benchmarks: artificial networks with known community structure.







But: random networks (their structure is quite different from real-world networks); in the way the benchmark is built, there is a (hidden) choice for what good partitions should be

Leon Danon, Jordi Duch, Albert Diaz-Guilera, Alex Arenas, J. Stat. Mech. (2005) P09008
Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi, Phys. Rev. E 78, 046110 (2008)

Modularity: what is chance?

Q = fraction of edges within communities - expected fraction of such edges

Let us attribute each node *i* to a community c_i

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

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

$$P_{ij} = \frac{k_i k_j}{2m}$$

expected number of links between i and j

Random network with constrained degrees

Is this constraint that important? What if one has extra information about the nodes?

Spatial networks

Networks where nodes are embedded in physical space (2d or 3d)

Many examples: Physical Internet infrastructure, road networks, flight connections, brain functional networks and social networks

Space induces constraints on the existence/strength of links between nodes, such that network architecture radically differs from that of random networks

E.g. planar graphs where links can not cross

More generally: cost (to create or maintain a link) associated to long-distance

Restricts the existence of hubs, i.e. high degree nodes, and thus the observation of fat-tailed degree distributions in spatial networks.

Configuration model: bad estimation of the way the network is organised

Gravity Model

Flow/number of interactions between regions separated by physical distance d_{ij} tends to be well modeled by

$$T_{ij} = N_i N_j f(d_{ij})$$

where N_i is the importance of node i (its population) and f is a deterrence function describing the influence of space.

In many socio-economic systems, f is well fitted by a power-law $d^{-\alpha}$ reminiscent of Newton's law of gravity, with population playing the role of a mass.

Used for systems as diverse as the International Trade Market, human migration, traffic flows or mobile communication between cities..

Which metrics for spatial nets?

When analyzing spatial networks, authors tend to use network metrics where the spatial arrangement of the nodes is ignored, thus disregarding that useful measures for non-spatial networks might yield irrelevant or trivial results for spatial ones.

Important examples are the clustering coefficient, as spatial networks are often spatially clustered by nature, and degree distribution, where high degree nodes are suppressed by long distance costs.

This observation underlines the need for appropriate metrics for the analysis and modeling of networks where spatial constraints play an important role

Modularity and Space

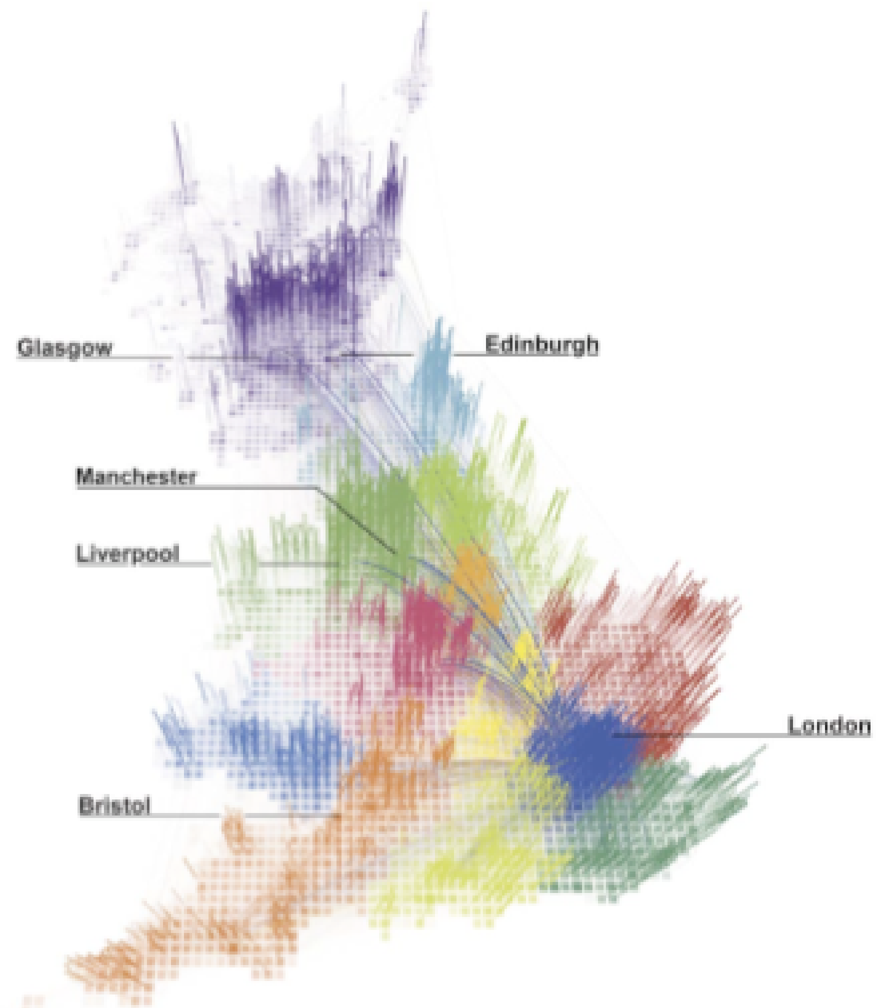
Applications include:

i) the design of efficient national, economical or administrative borders based on human mobility or economical interactions instead of historical or ad-hoc reasons

ii) the identification of functionally related brain regions and of principles leading to global integration and functional segregation

Modularity and Space

Standard modularity optimization



Interesting but modules are strongly determined by geographical factors and provide poor information about the underlying forces shaping the network.

In spatial networks, how can one detect patterns that are not due to space? In other words, are observed patterns only due to the effect of spatial distance, because of gravity-like forces, or do other forces come into play?

Redrawing the Map of Great Britain from a Network of Human Interactions, Ratti et al. (2010)

Space and Social Systems

Principles driving network organization,
link creation in social systems

Structural

Triadic closure (friend of my friend becomes my friend), reciprocity, etc.

Non-structural

Homophily (similar people tend to select each other)

Focus constraint (social relations depend on opportunities for social contact, including physical proximity)

When uncovering modules, one deals with an extremely intricate situation where structural and non-structural effects are mingled. Modules uncovered by modularity optimization are thus underpinned by an uncontrolled mixture of possibly antagonistic forces.

Our aim is the following: when the spatial positions of the nodes are known, as more and more often is the case, is it possible to take out the effect of space in order to identify more clearly homophilious effects and thus hidden structural or cultural similarities.

Spatial Null Model

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

The NG null-model only uses the basic structural information encoded in the adjacency matrix. Therefore, it is appropriate when no additional information on the nodes is available but not when additional constraints are known. In networks where distance strongly affects the probability for two nodes to be connected, a natural choice for the null model is

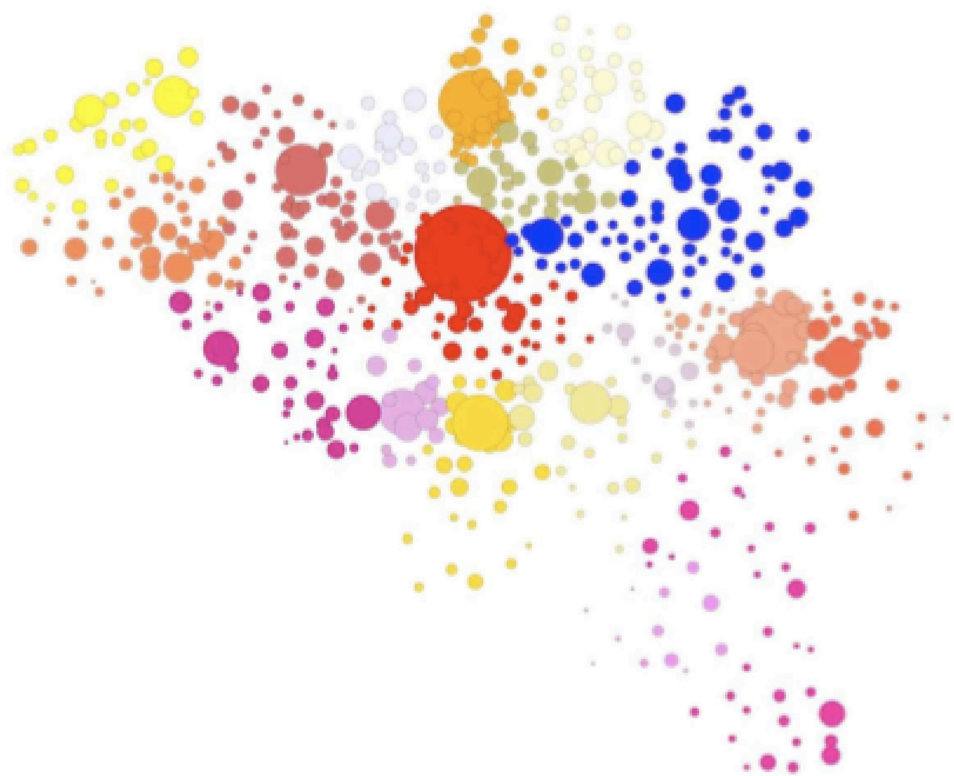
$$P_{ij}^{\mathrm{Spa}} = N_i N_j f(d_{ij}) \qquad f(d) = \frac{\sum_{i,j|d_{ij}=d} A_{ij}}{\sum_{i,j|d_{ij}=d} N_i N_j}$$

The probability for 2 nodes to be connected is thus directly measured from the data

Compared to QNG, QSpa tends to give larger contributions to distant nodes and its optimization is expected to uncover modules driven by non-spatial factors.

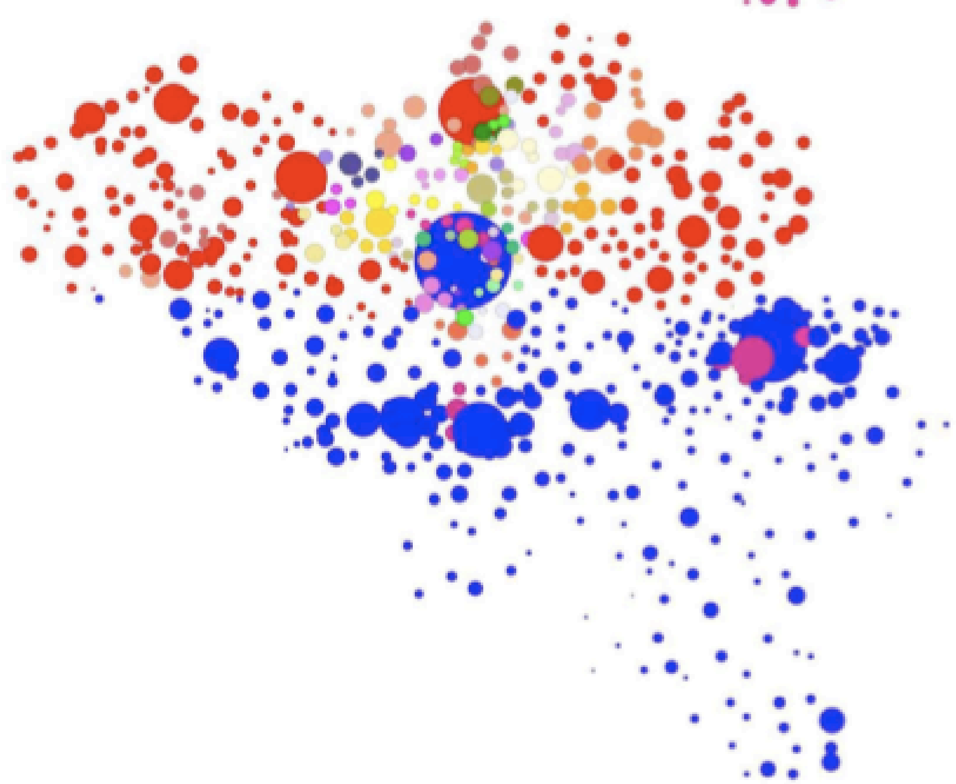
Numerical validation: empirical data

Mobile phone data: (undirected) weighted network between 571 zip codes in Belgium (=> linguistic frontier)



NG Modularity: 18 spatially compact modules mainly determined by short-range interactions between communes.

Although this partition coincides with the linguistic separation of the country, the unaware would not discover the existence of two linguistic communities.



Spa Modularity: almost perfect bipartition of the country where the two largest communities account for about 75% of all communes and nicely reproduce the linguistic separation of the country.

disentangle spatial and cultural effects

Numerical validation: benchmark

To test the validity of our method in a controlled setting, we focus on computer-generated benchmarks for spatial, modular networks.

The probability for two nodes to be connected depends on their distance and on the community to which they are assigned.

100 nodes placed at random in a two dimensional square of dimension 100 × 100 and randomly assigned into two communities of 50 nodes.

The probability that a link exists between nodes i and j has the form

$$p_{ij} = \frac{\lambda(c_i, c_j)}{Z d_{ij}} \qquad \sum_{i>j} p_{ij} = 1$$

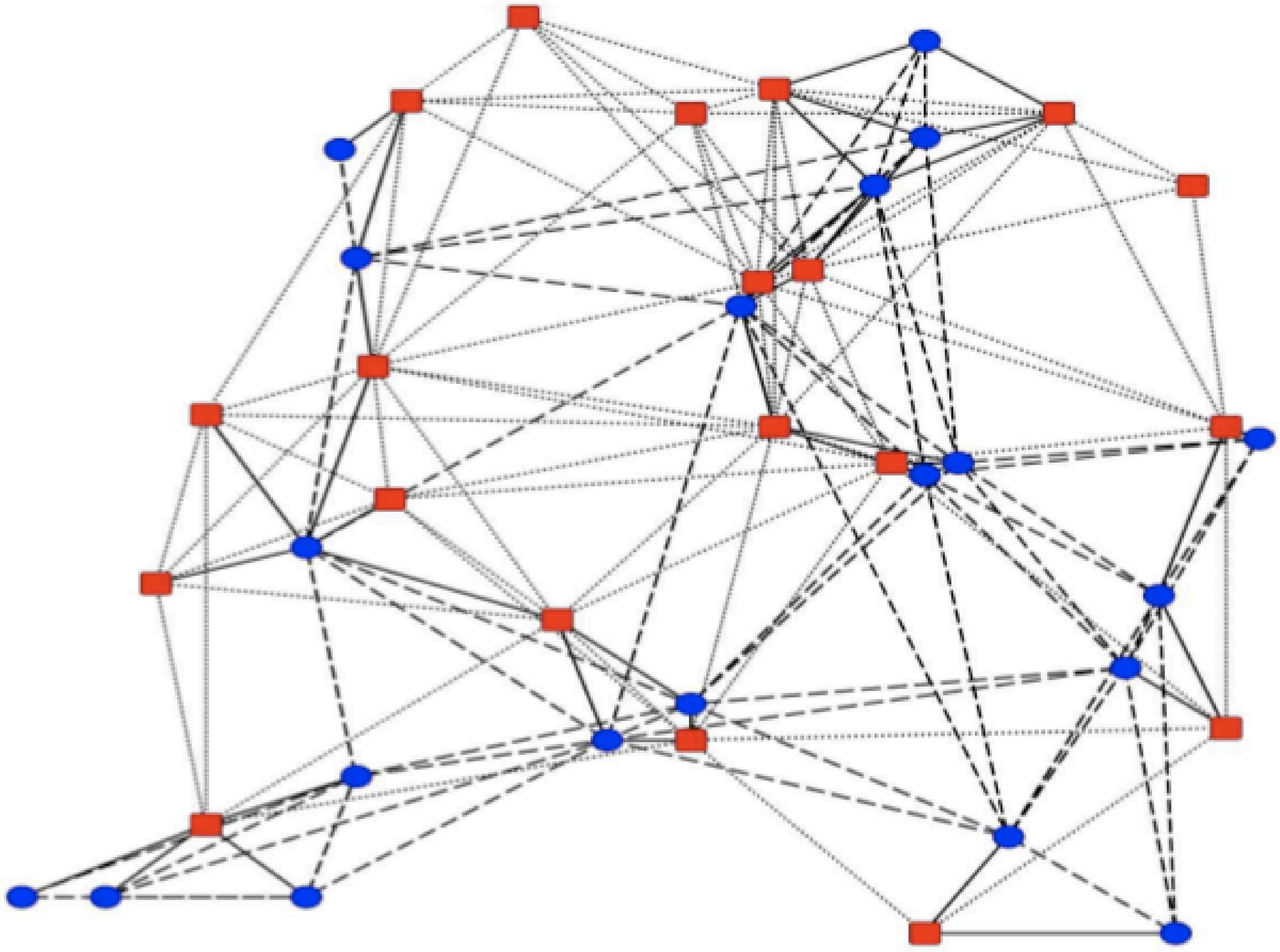
$$\lambda(c_i, c_j) \qquad 1 \text{ if } c_i = c_j \text{ and } \lambda_{\text{different}} \text{ otherwise}$$

$$L = \rho N(N-1)/2$$

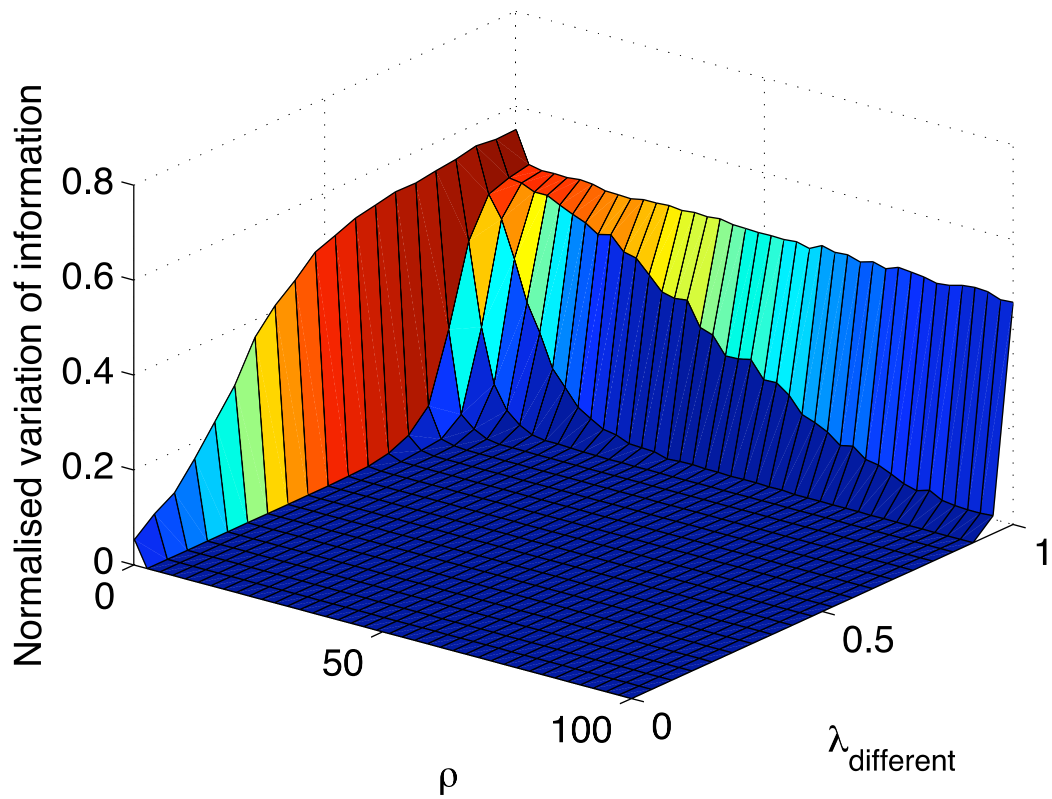
1) lambda: how discernible communities are

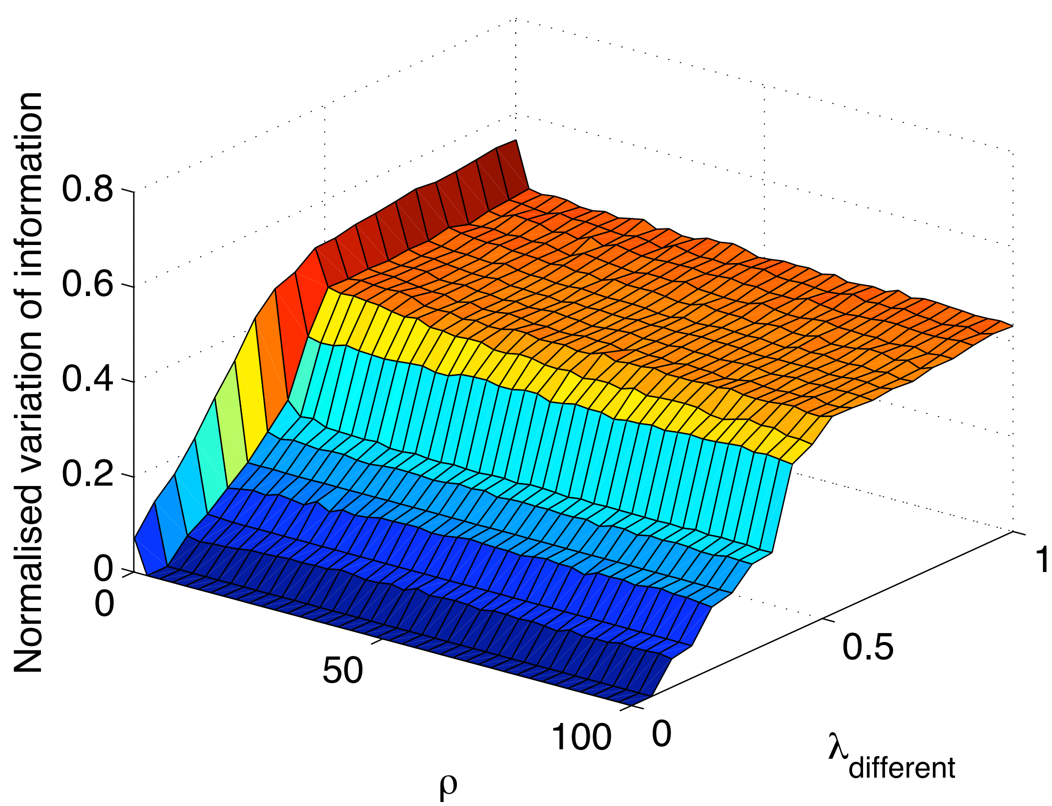
2) rho: density of links, finite-size fluctuations around the average

Numerical validation: benchmark



Numerical validation: benchmark





Conclusion

- Which metrics should we use to characterize and uncover useful information from spatial networks?

- Standard metrics => uncover the effect of space which is often trivial

- Carefully take into account spatial constraints affecting link creation

- Incorporate extra-information (non-structural) in the definition of modularity in order to portray more closely the network under scrutiny => exploit known attributes, e.g. spatial location, in order to uncover unknown ones, e.g. homophilious relations.

- P. Expert, T.S. Evans, V.D. Blondel and R. Lambiotte, *arXiv:1012.3409*
- R. Lambiotte, Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on, 546-553 (2010)
- V.D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of community hierarchies in large networks, J. Stat. Mech., (2008) P10008
- D. Meunier, R. Lambiotte and E.T. Bullmore, “Modular and hierarchical organisation in complex brain networks”, Frontiers in NeuroScience (2010)

Thanks to:

K. Christensen, M. Barahona, H.J. Jensen (Imperial College)
P. Panzarasa (Queen Mary)
D. Meunier and E. Bullmore (Cambridge)
J.C. Delvenne (Louvain)