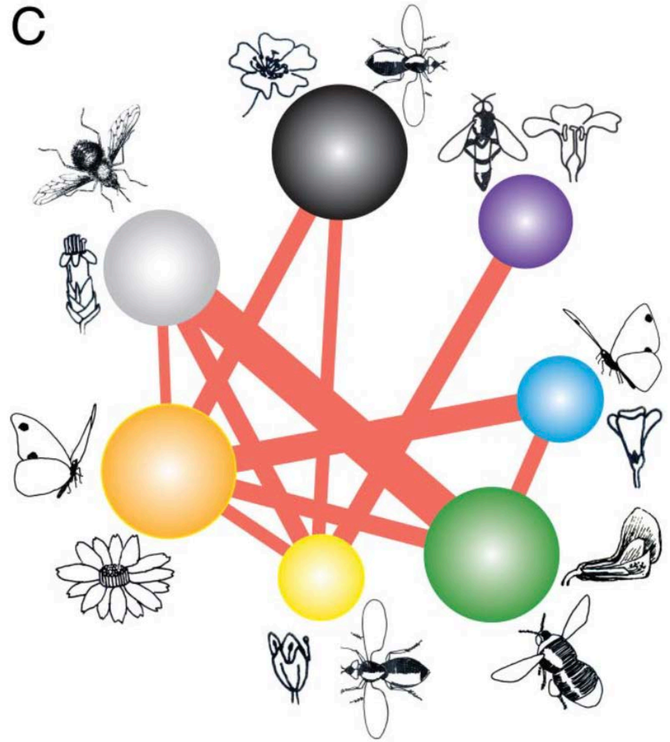
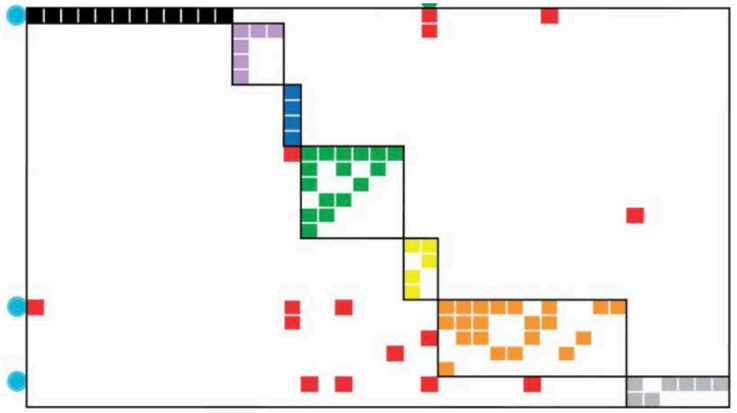
Community detection

Analysis of Biological-Ecological Networks 2026

Shai Pilosof





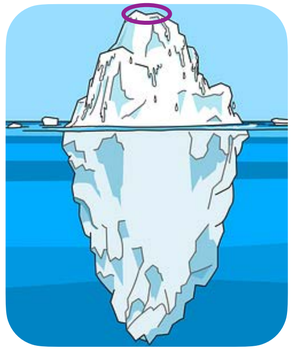
From Olesen et al 2007, PNAS

www.ecomplab.com
pilos@post.bgu.ac.il

Ben-Gurion University of the Negev

Class goals

1. Introduce the concept and analytical approaches for community detection*.
2. Demonstrate how community detection provides insights.

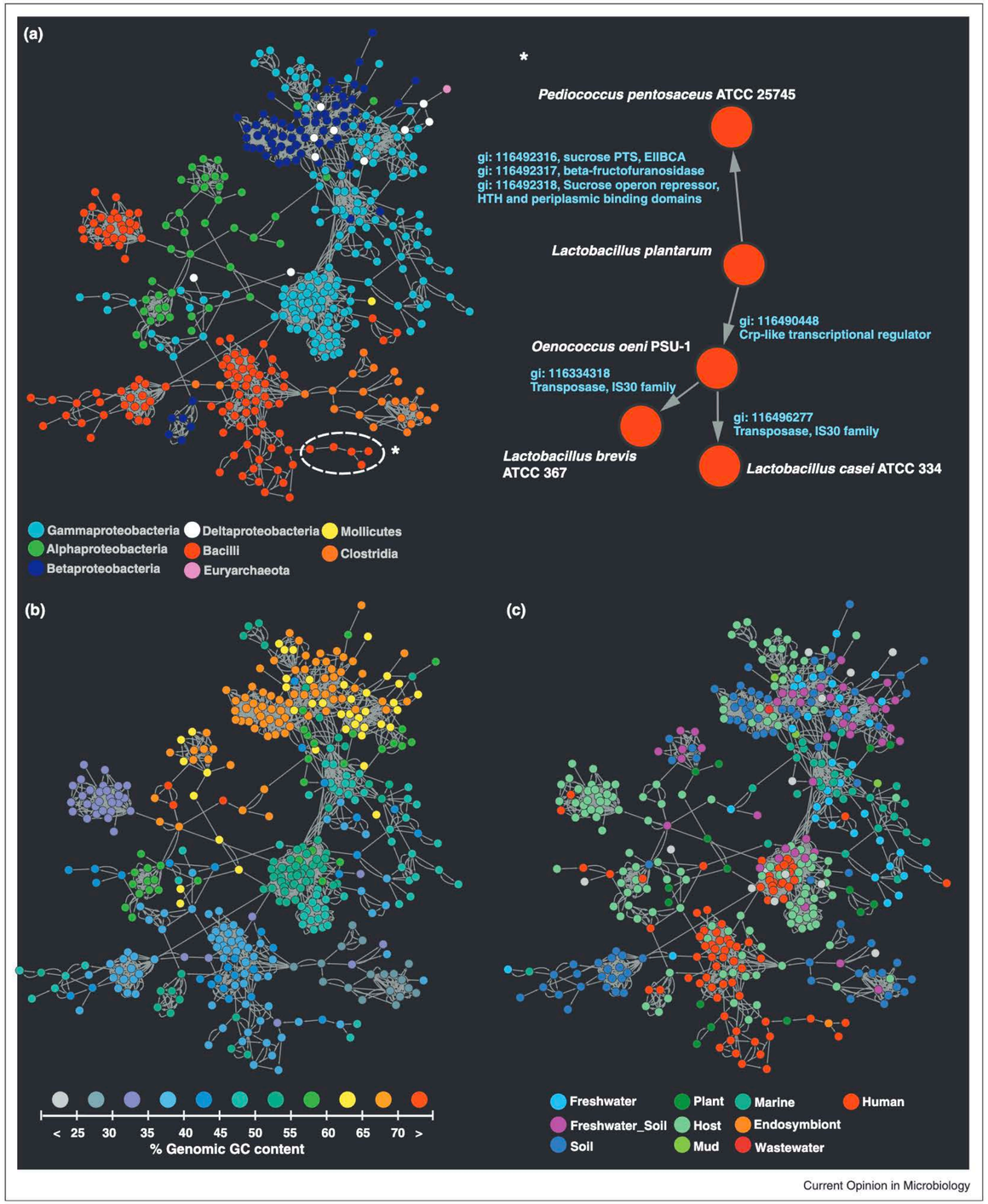


* This is not even the beginning of the tip of the iceberg. **You must consult papers and books.**

Outline

- What is community detection.
- Newman’s modularity.
- The map equation: flow-based community detection.
- Node “modularity roles”
- Characterizing and interpreting modules.

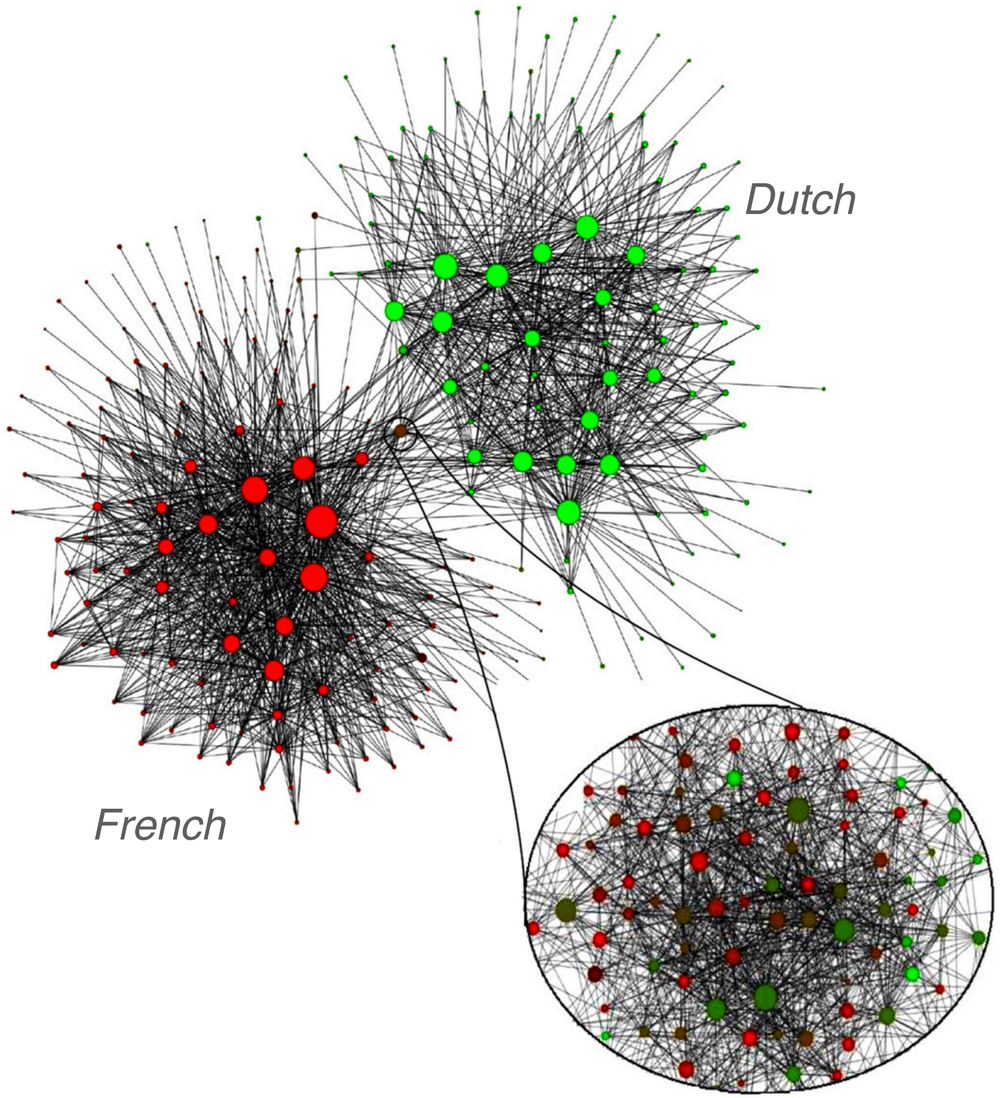
Phage HGT network



Popa et al 2011

What is community detection

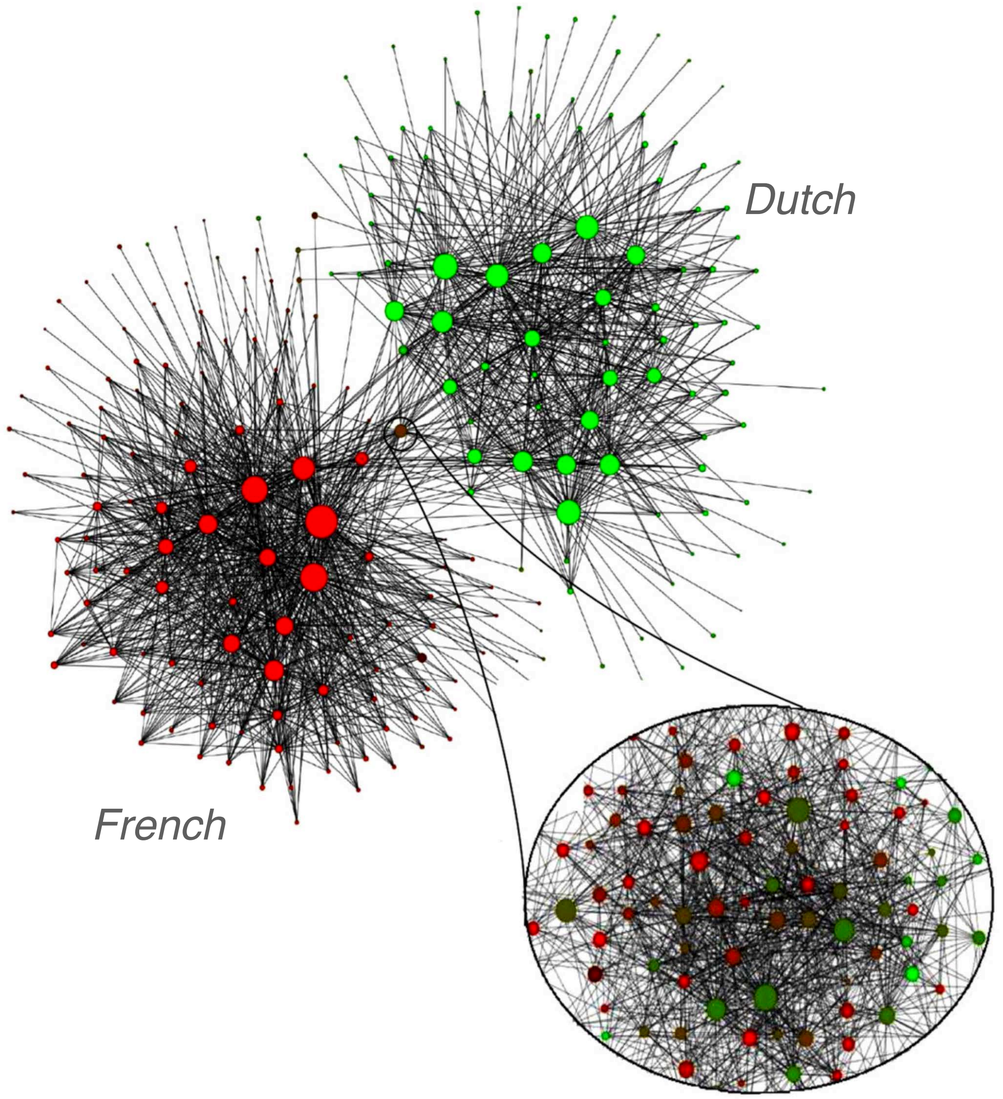
- **Objective**: find natural divisions of the network into groups of nodes.



Blondel et al. 2008, J Stat Mech

What is community detection

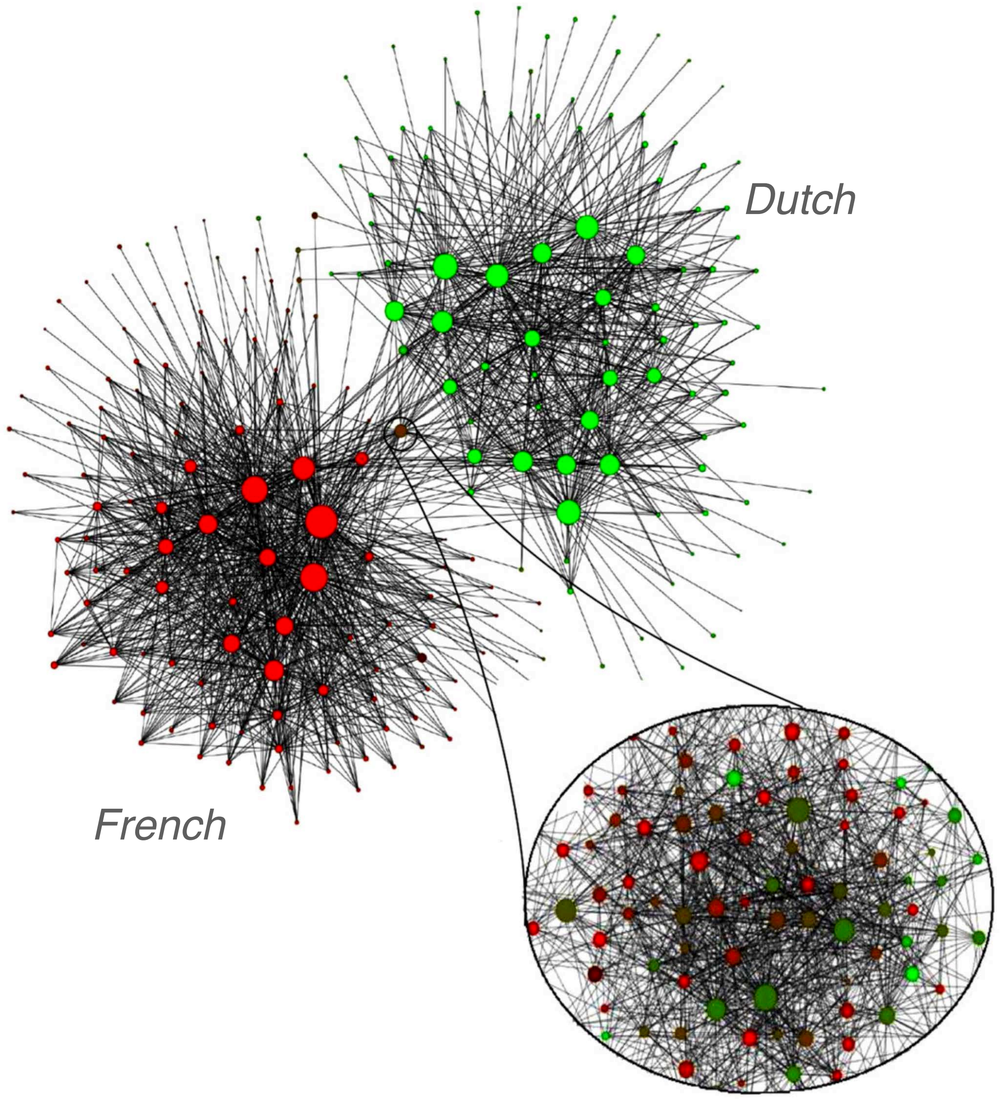
- **Objective**: find natural divisions of the network into groups of nodes.
- **Why**:
  - Reduce complexity (nodes in a group behave similarly).
  - Reveal signatures of generative processes.
  - Functional consequences.



Blondel et al. 2008, J Stat Mech

What is community detection

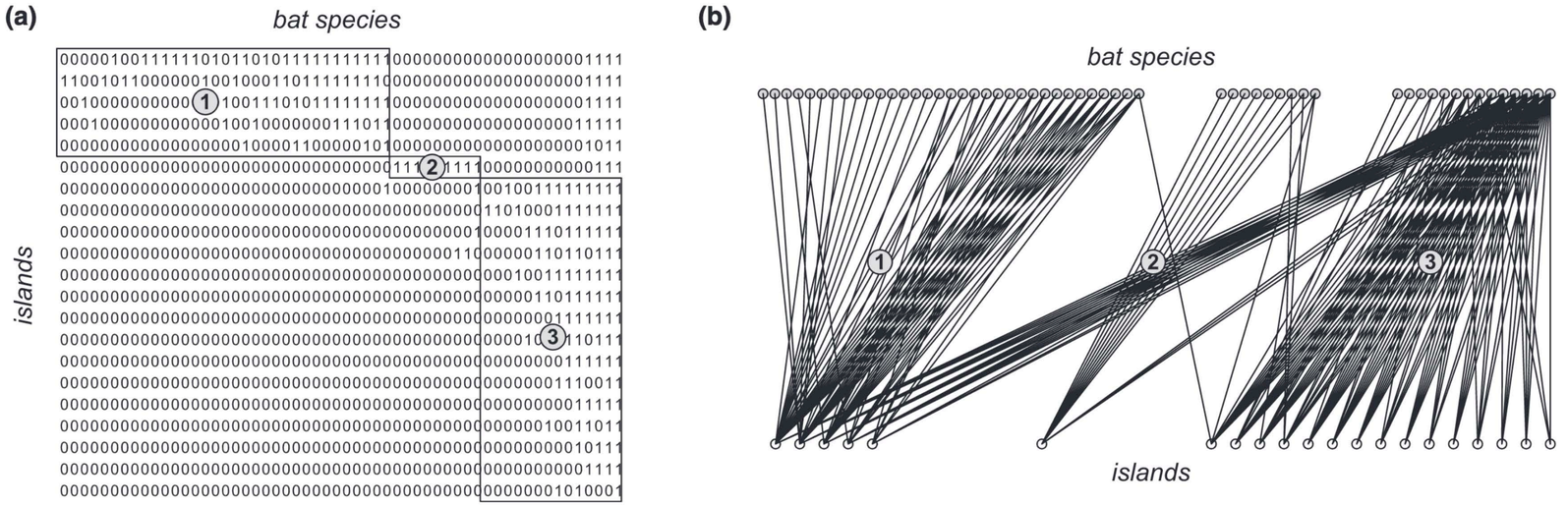
- **Objective**: find natural divisions of the network into groups of nodes.
- **Why**:
  - Reduce complexity (nodes in a group behave similarly).
  - Reveal signatures of generative processes.
  - Functional consequences.
- **But**: there is hardly ever an objectively true division.



Blondel et al. 2008, J Stat Mech

Terms

- Community detection
- Network partitioning
- Clustering
- Modularity
- Groups
- Block structure
- Compartmentalized structure

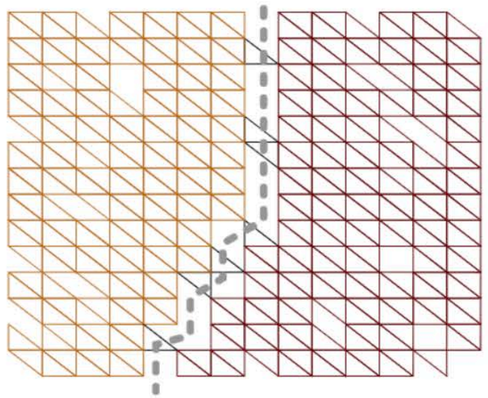


Thebault 2013

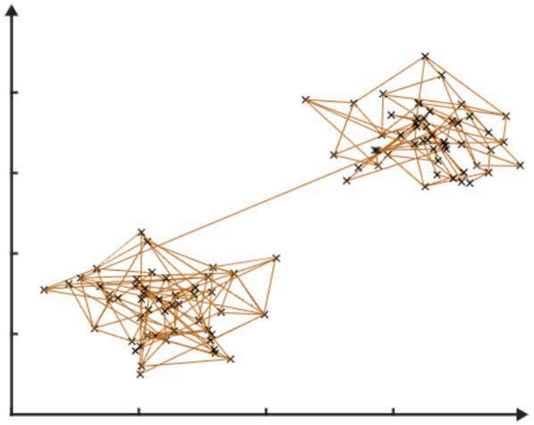
These terms carry different connotations and can refer to different concepts and algorithms of community detection.

Approaches for community detection

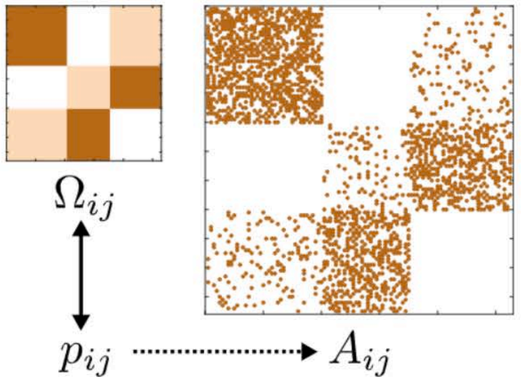
(i) Cut-based perspective



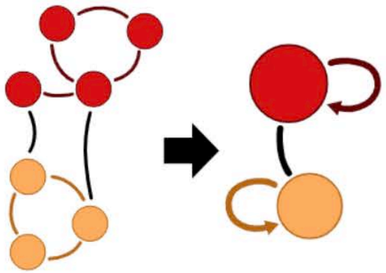
(ii) Clustering perspective



(iii) Stochastically equivalent nodes



(iv) Dynamical perspective



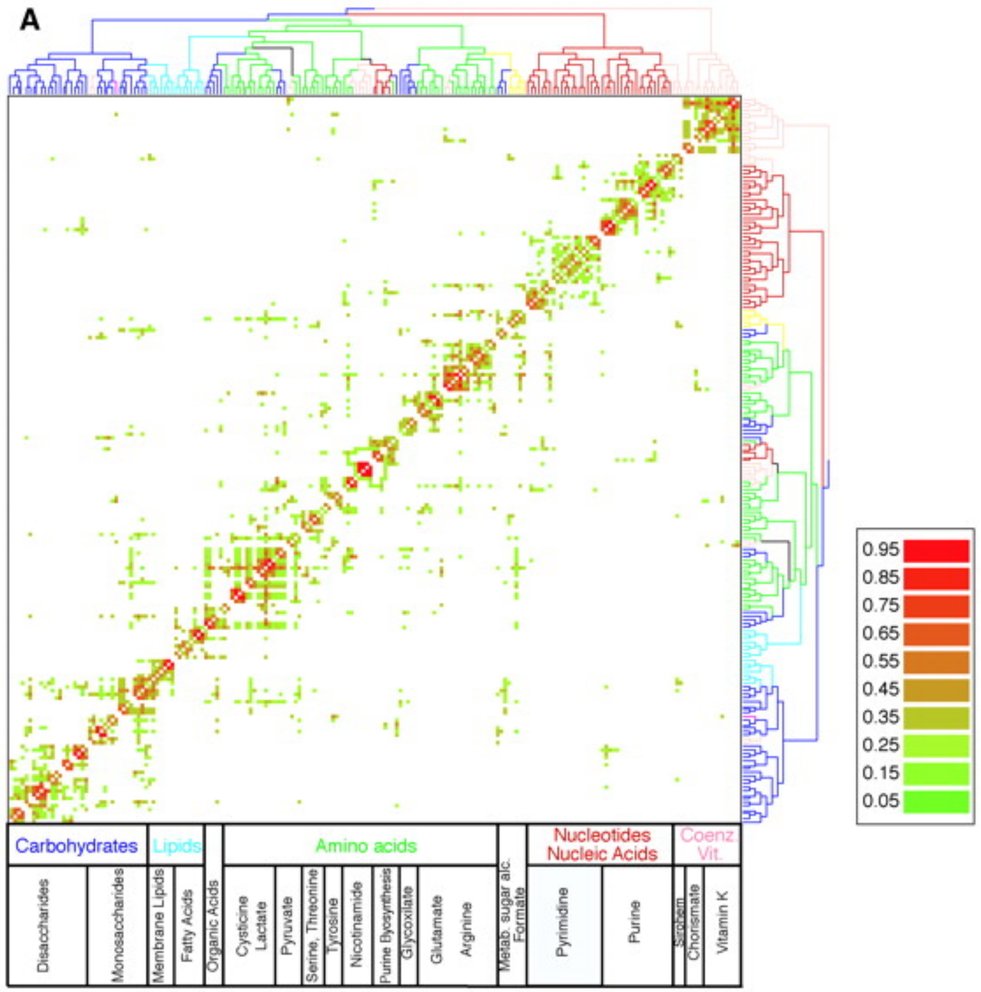
$\dot{x}=f(x) \approx \dot{y}=g(y)$

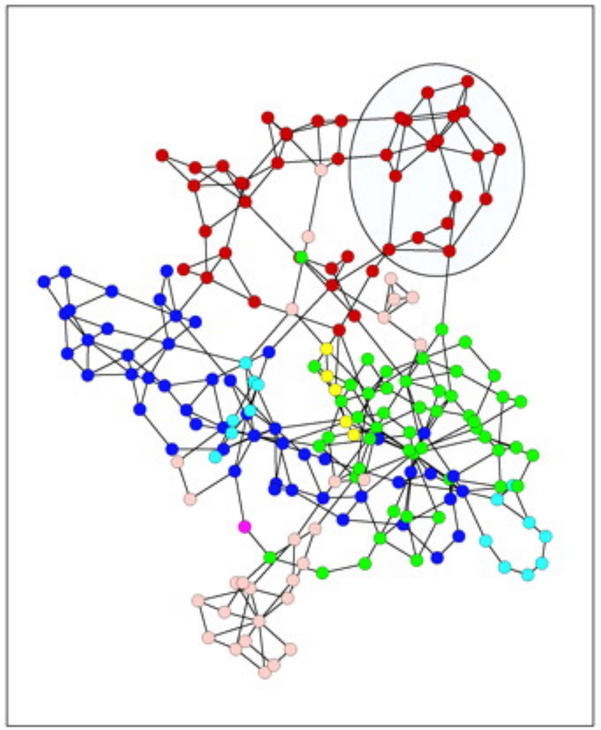
A guideline on the course website (adapted from Farage et al 2021):

https://ecological-complexity-lab.github.io/network_course/class_communities.html#A_guide_to_choosing_a_community_detection_approach

Rosvall et al 2020

Biological networks are (many times) clustered

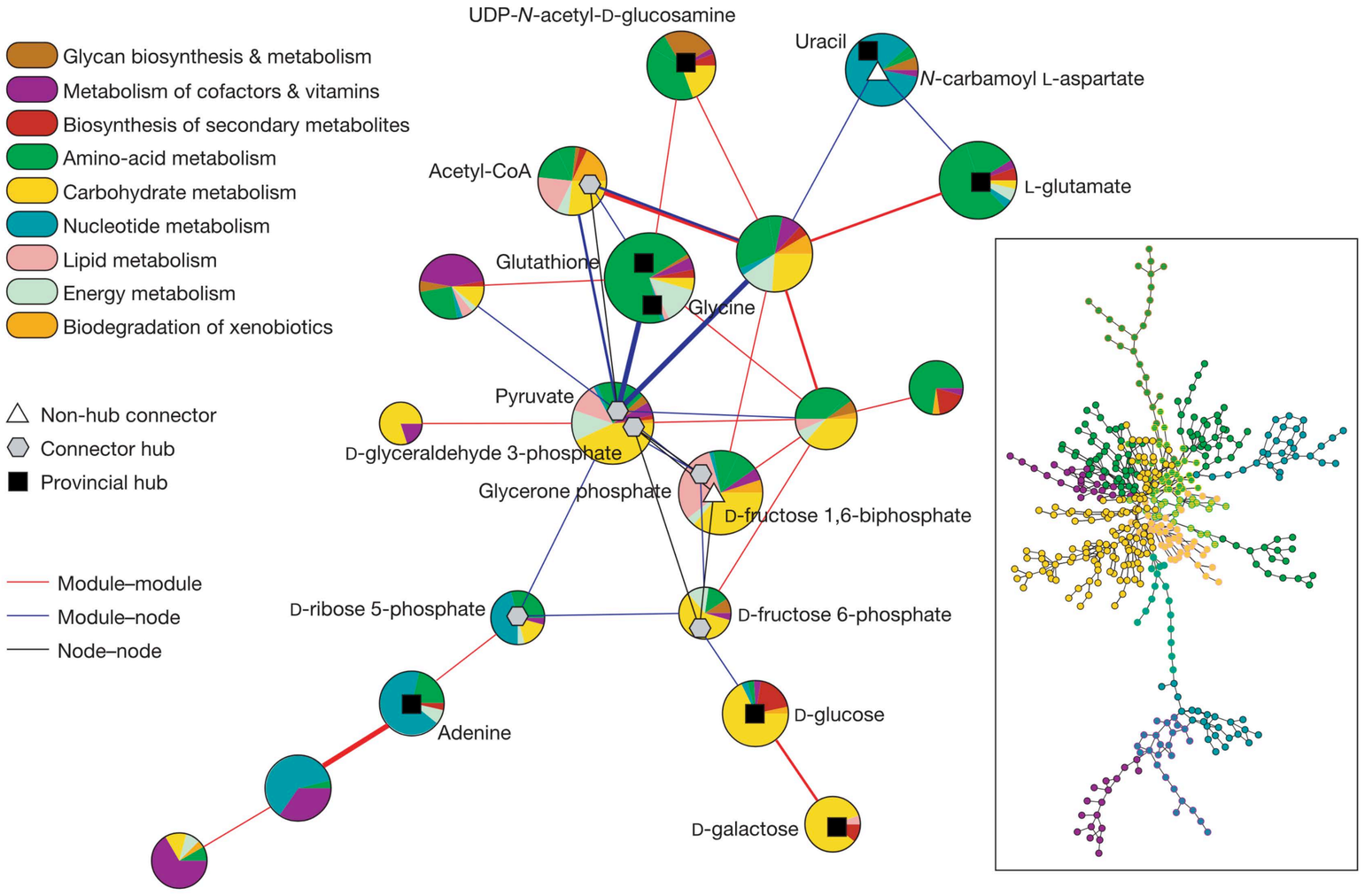




Metabolic network of *E. Coli*, partitioned using topological overlap (see paper for details)

Jeong et al 2000, Nature

Metabolic network of *E. Coli*, partitioned using modularity, with metabolite roles



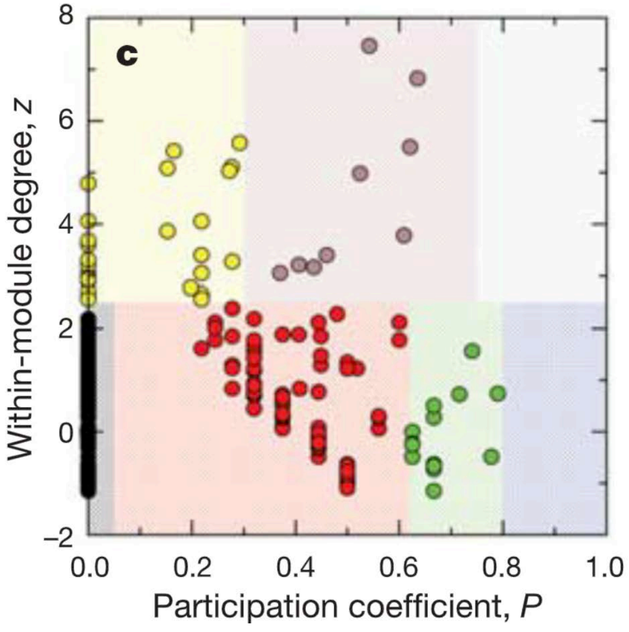
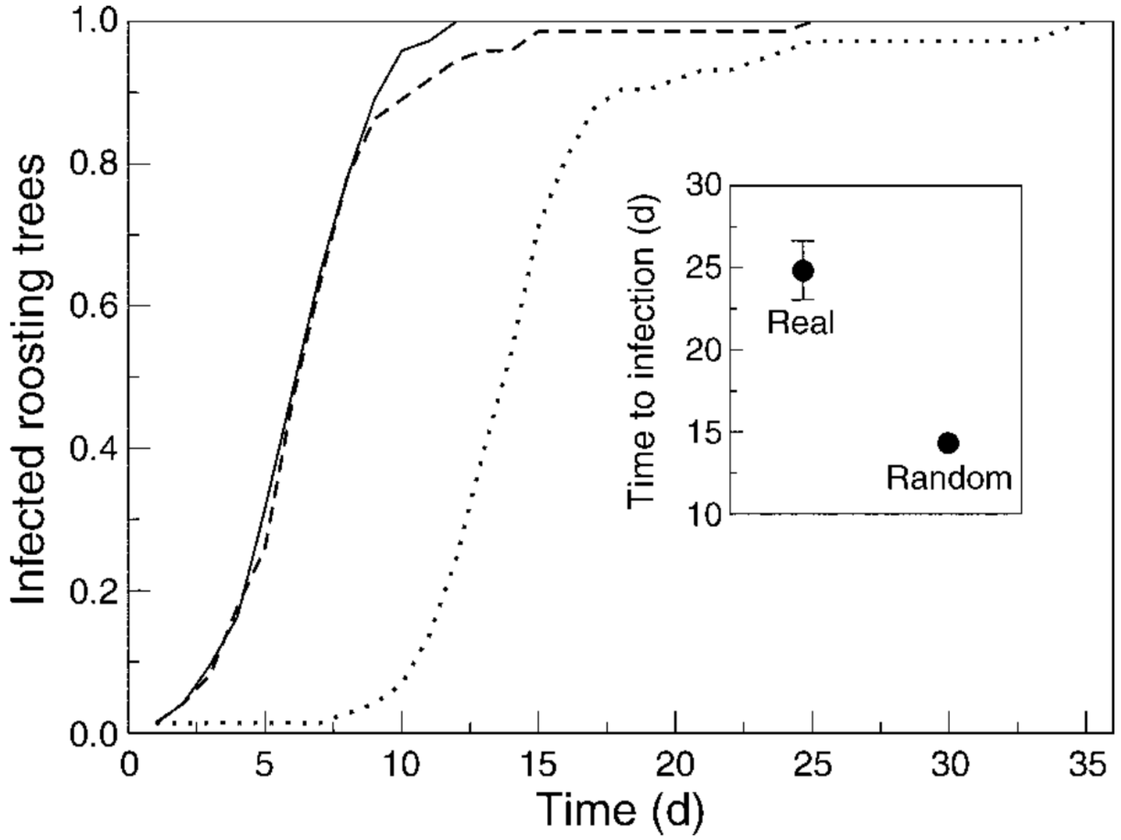
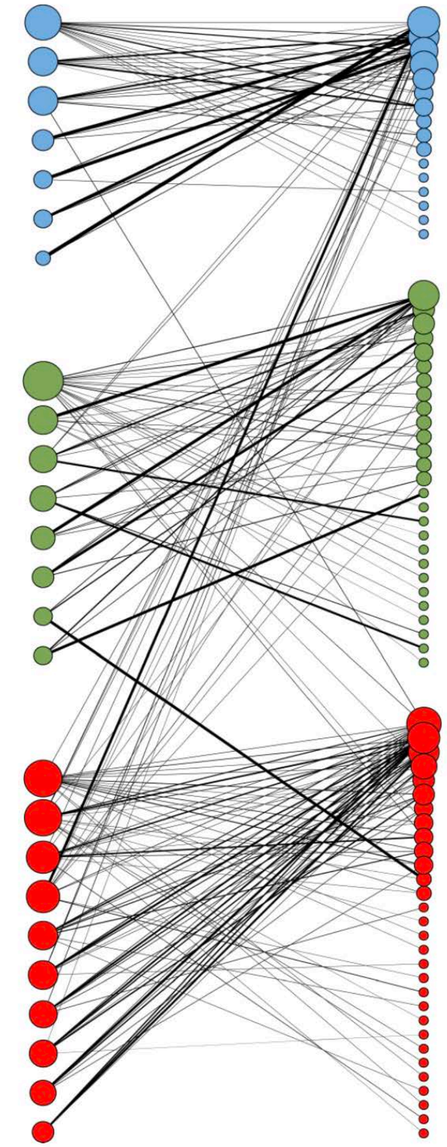


Figure 3 Cartographic representation of the metabolic network of *E. coli*. Each circle represents a module and is coloured according to the KEGG pathway classification of the metabolites it contains. Certain important nodes are depicted as triangles (non-hub connectors), hexagons (connector hubs) and squares (provincial hubs). Interactions between modules and nodes are depicted using lines, with thickness proportional to the number of actual links. Inset: metabolic network of *E. coli*, which contains 473 metabolites and 574 links. This representation was obtained using the program Pajek. Each node is coloured according to the 'main' colour of its module, as obtained from the cartographic representation.

Guimerà and Amaral 2005, Nature

Example: modularity of a roosting network

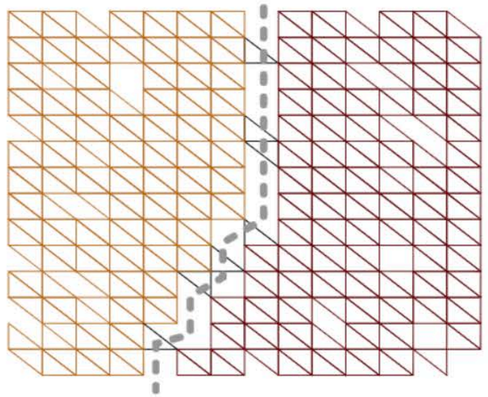




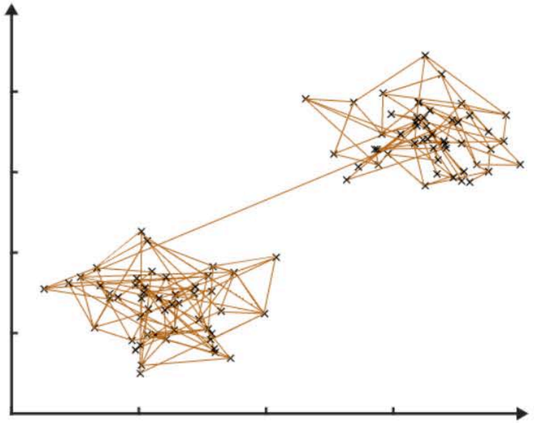
Fortuna el al 2009, Ecology

Approaches for community detection

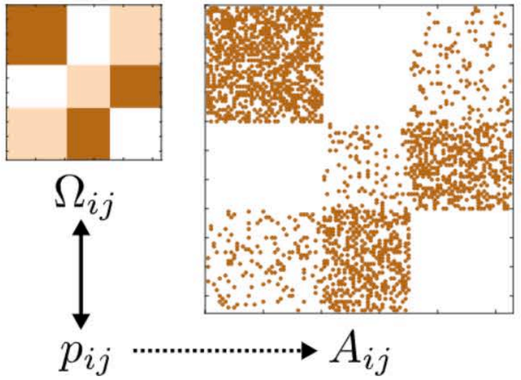
(i) Cut-based perspective



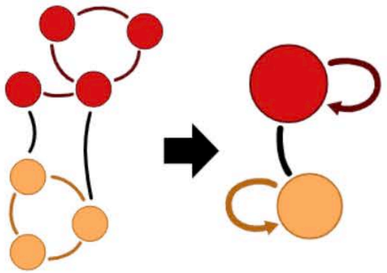
(ii) Clustering perspective



(iii) Stochastically equivalent nodes



(iv) Dynamical perspective



Rosvall et al 2020

Partitioning - fixed number of communities

Problem: partition a network into two non-overlapping subgraphs, such that the number of links between the nodes in the two groups, called the *cut size*, is minimized.

Brute force solution: inspecting all possible divisions into two groups and choosing the one with the smallest cut size. The number of possibilities is given by the Stirling number of the 2nd kind, with $\mathcal{N} = 2$.

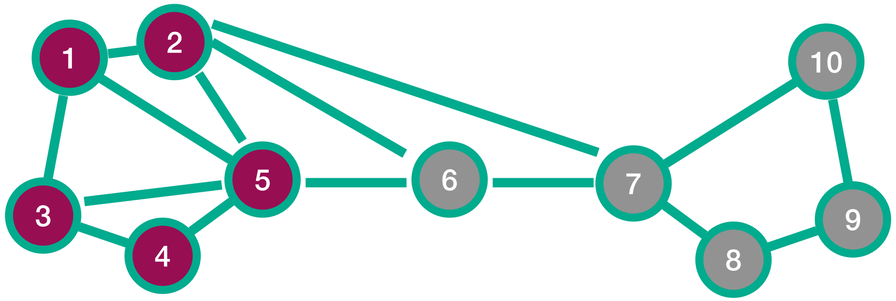
$$\mathcal{S}(S, \mathcal{N}) = \frac{1}{\mathcal{N}!} \sum_{j=0}^{\mathcal{N}} (-1)^{\mathcal{N}-j} \binom{\mathcal{N}}{j} j^S$$

For a graph with 10 nodes there are 511 options. For 100 nodes there are more than 10^29.

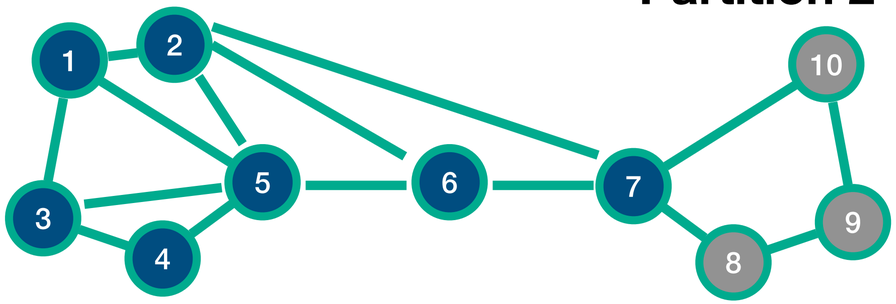
Challenge 1: Computational

Challenge 2: the number and size of groups is predefined.

Partition 1



Partition 2



N1=N2=5: 511 partitions; 1ms computation
N1=N2=50: 10^29 partitions=10^16 years computation

Age of the universe: 13.8*10^9 years

Allesina and Pascual 2009 Ecol Lett;
Barabasi 2021, Network Science Book

Community detection

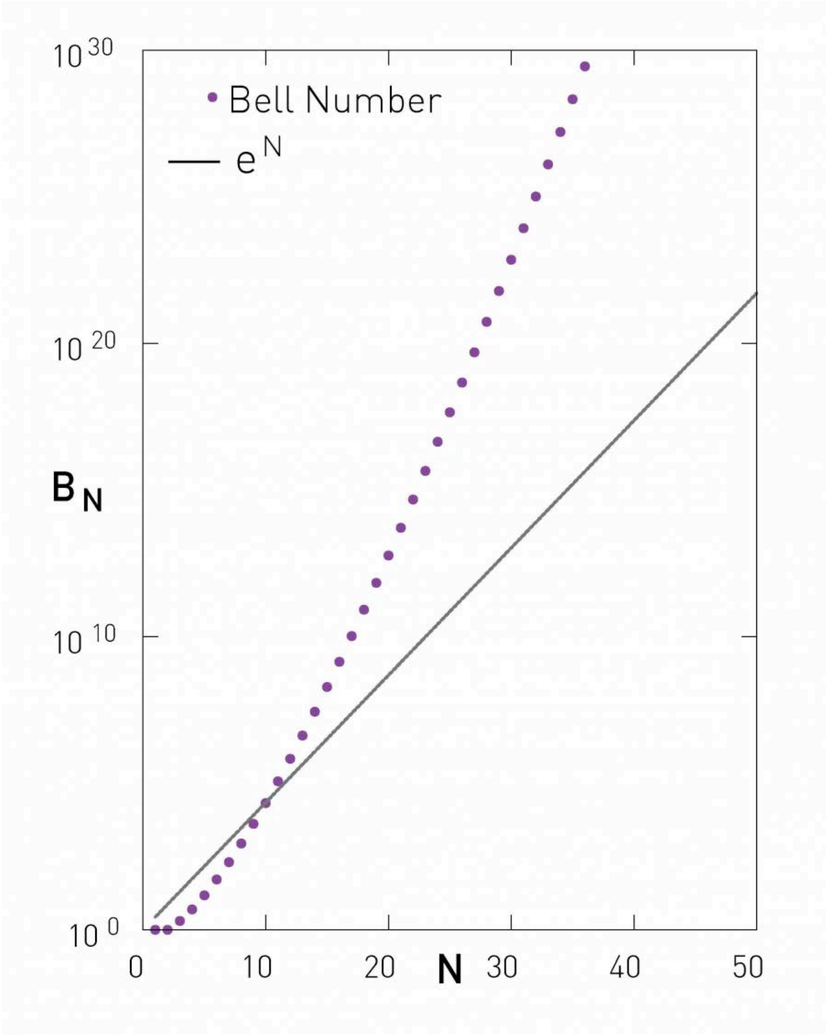
Define a ***partition*** as a division of a network into an arbitrary number of groups, such that each node belongs to one and only one group.

The number of possible partitions is given by the Bell number (sum of Stirlings numbers):

$$\mathscr{B}_{\mathcal{S}} = \sum_{\mathcal{N}=1}^{S} \mathcal{S}(S, \mathcal{N})$$

Fundamental challenge of community identification: The number of possible ways we can partition a network into communities grows exponentially or faster with the network size. Therefore it is impossible to inspect all partitions.

Solution: Need algorithms to identify communities without inspecting all partitions.



Allesina and Pascual 2009 Ecol Lett;
Barabasi 2021, Network Science Book

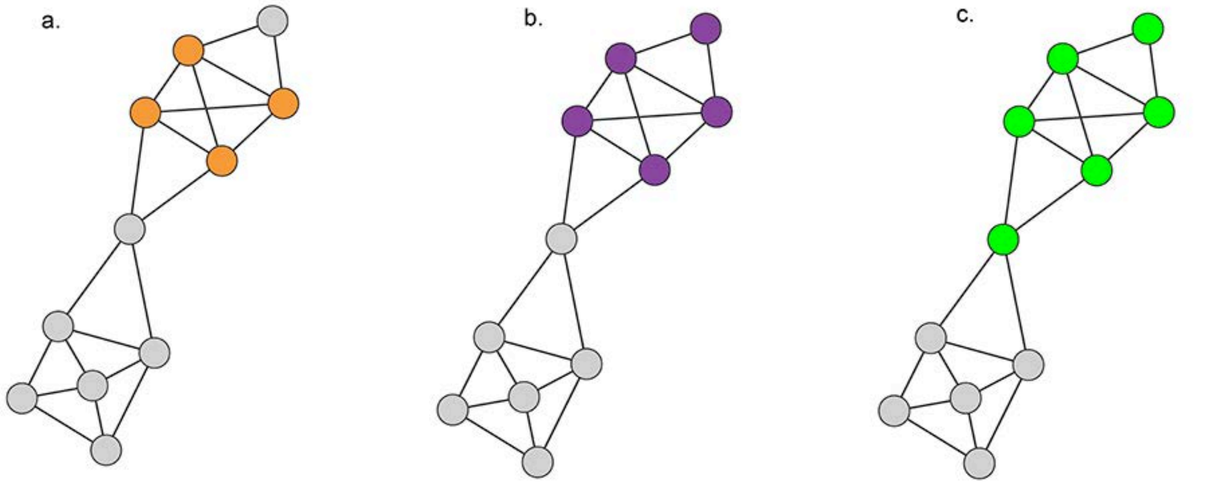
Weak and strong communities

- *Definition*: A community is a locally dense connected subgraph in a network.
- C is a *strong community* if each node within C has more links within the community than with the rest of the graph.
- A *weak community* allows some nodes to violate the strong community requirement. Hence, the inequality applies to the community as a whole rather than to each node individually.

4-node clique (complete subgraph)

Strong community

Weak community



$$k_i^{int}(C) > k_i^{ext}(C) \forall i \in C$$

$$\sum_{i \in C} k_i^{int}(C) > \sum_{i \in C} k_i^{ext}(C)$$

Barabasi 2021, Network Science

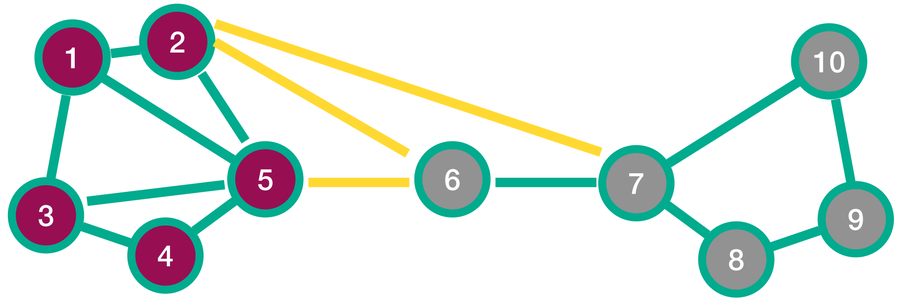
Community detection - general approach

Choose a definition for a community (e.g., weak/strong communities, modularity)

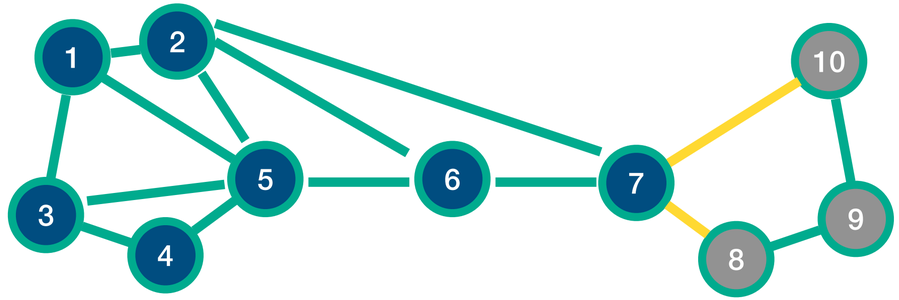
Identify communities by inspecting partitions for $1 \leq \mathcal{N} \leq S$ using an algorithm.

Select the partition that best satisfies the condition

Partition 1: cut size = 3



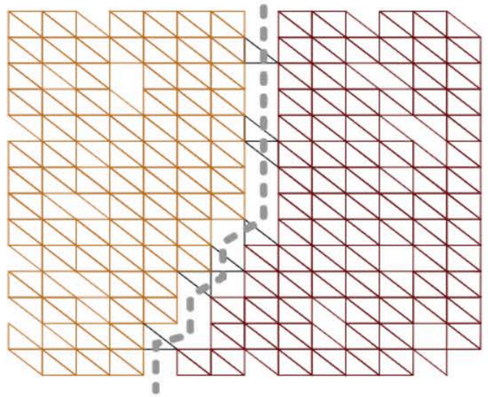
Partition 2: cut size = 2



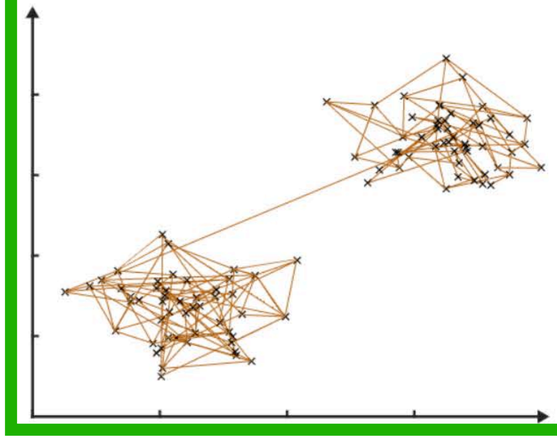
Partition 2 is better than 1 because it minimizes the **cut size** (number of links between nodes in the two groups).

Approaches for community detection

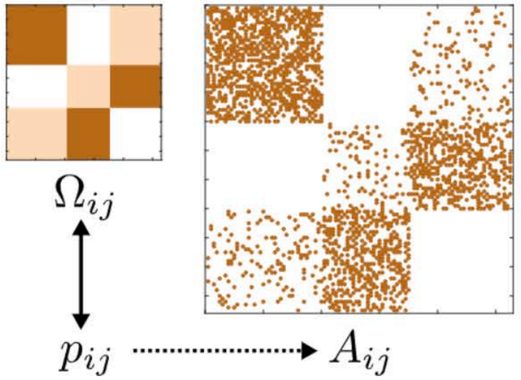
(i) Cut-based perspective



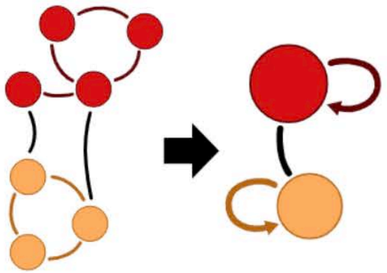
(ii) Clustering perspective



(iii) Stochastically equivalent nodes



(iv) Dynamical perspective

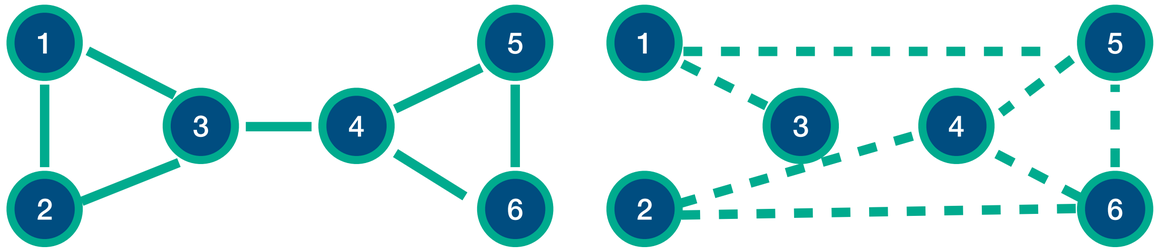


$\dot{x}=f(x) \approx \dot{y}=g(y)$

Rosvall et al 2020

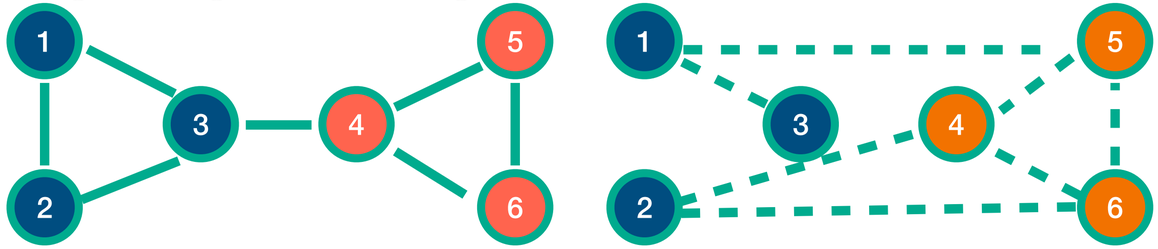
Modularity - for unweighted, undirected unipartite networks

Q = Observed internal connectivity–Expected internal connectivity.



In the 2-module solution:

- Observed internal edges are high.
- Expected internal edges (given degrees) are lower.



Newman and Girvan 2004, PRE; Barabasi 2021, Network Science Book

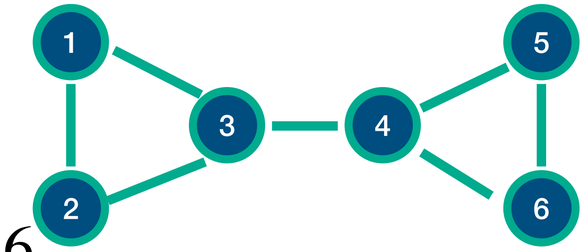
Modularity - for unweighted, undirected unipartite networks

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

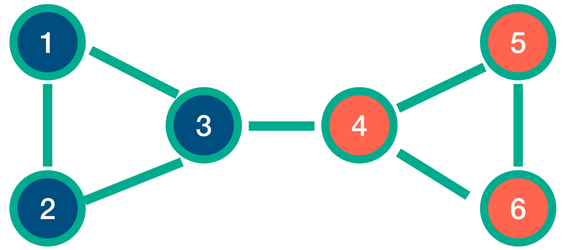
$$M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$$

where L_c is the total number of links within the community c and k_c is the total degree of the nodes in this community.

$N = 6$
$L = 7$
$n_c = 1$
$N_c : N_1 = 6$
$L_c : L_1 = 7$



$N = 6$
$L = 7$
$n_c = 2$
$N_c : N_1 = N_2 = 3$
$L_c : L_1 = L_2 = 3$
$k_c : k_1 = k_2 = 7$



- $Q \approx 0$: no more within-community edges than expected by chance.
- $Q > 0$ more within-community edges than expected.
- $Q < 0$: fewer within-community edges than expected

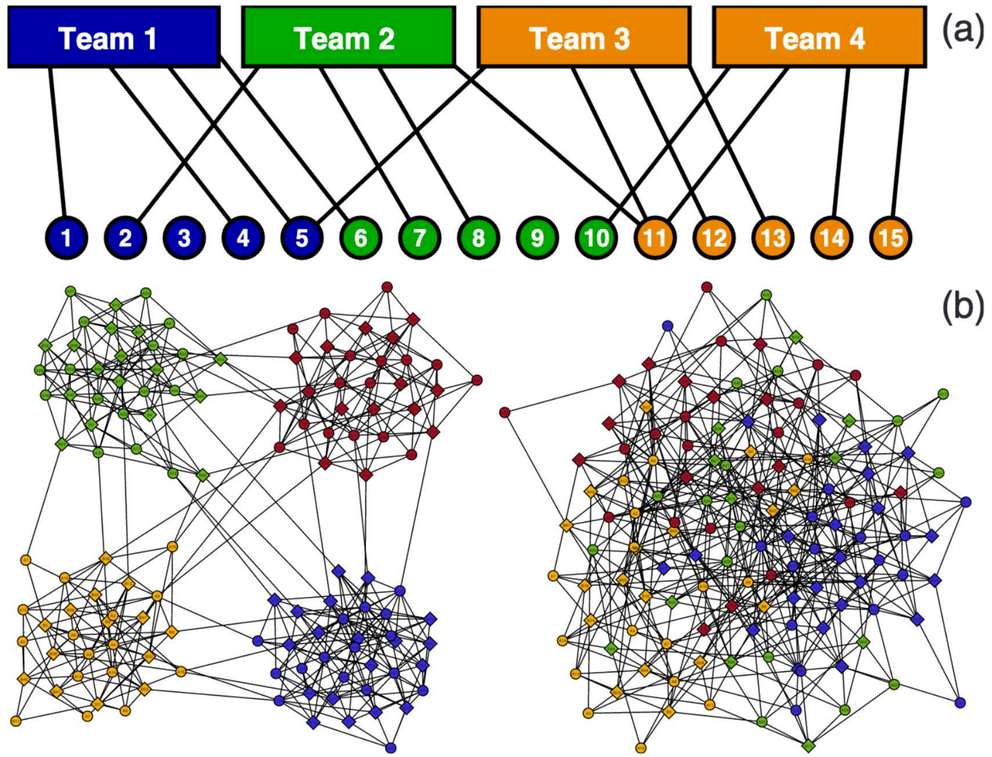
Newman and Girvan 2004, PRE; Barabasi 2021, Network Science Book

Modularity - for unweighted bipartite networks

Two solutions:

1. Detect modules of A-nodes that share a significantly high number of links with the same B-nodes.
2. Allows both parts of the network to be classified simultaneously (formula below).

See Thebault 2013 J Biogeog. for review.



$$M_B = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c^A \cdot k_c^B}{L^2} \right)^2 \right]$$

k_c^A and k_c^B: Sum of degrees of nodes within module c that belong to sets A or B

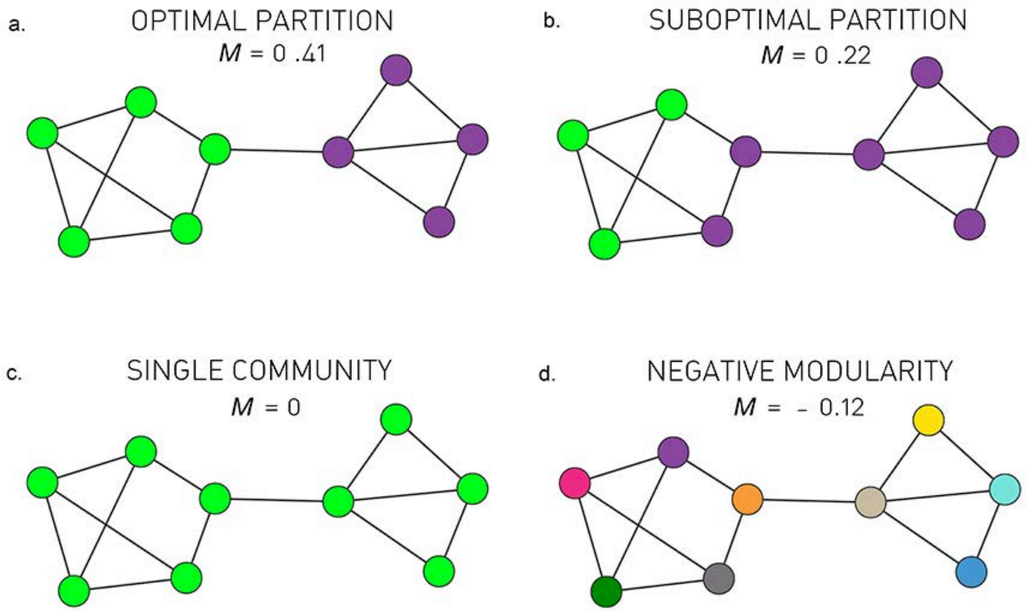
Guimerà et al 2007 PRE; Barber 2007 PRE

Define the objective function M

Identify communities by inspecting partitions for $1 \leq \mathcal{N} \leq S$ using an algorithm.

Select the partition that maximizes M

- Modularity ranges -1 to 1
- Higher modularity means better partition: our goal is to **maximize** M.



H: For a given network the partition with maximum modularity corresponds to the **optimal** community structure.

Barabasi 2021, Network Science Book

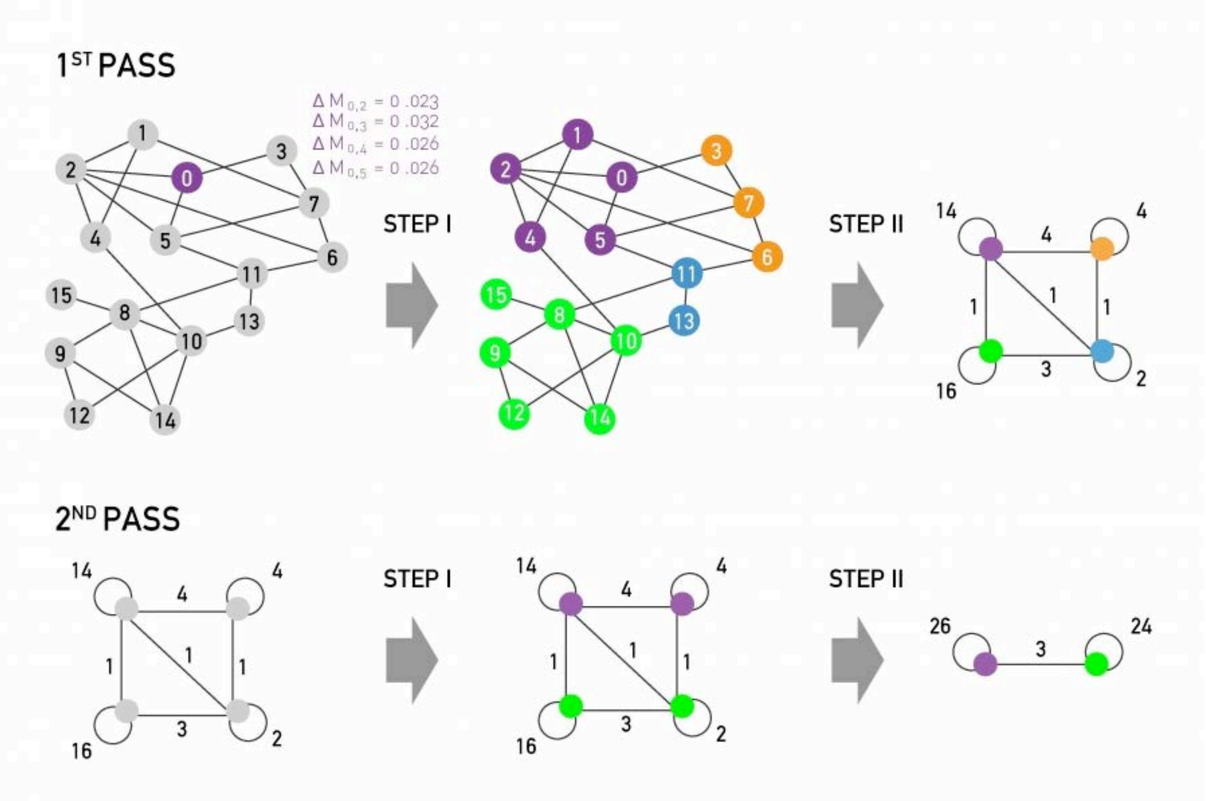
Louvain - a search algorithm for optimal solution

Concept: Modularity is optimized by local changes. Nodes are joined to modules if it improves the objective function.

Original paper: Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008;2008: P10008.

Simple explanation: https://www.mapequation.org/infomap/#Algorithm

Other popular algorithms: simulated annealing, and improvements to Louvain.



Define the objective function M

Identify communities by inspecting partitions for $1 \leq k \leq S$ using an algorithm.

Algorithm selects the partition that maximizes M

Because the algorithms are stochastic we need to run them many times. In other words, repeat the process on the left (typically 100 for ecological networks) and select the maximum M_{max}.

Modularity - more variations and developments

Heirarchical modularity

Weighted networks

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5**, 90–98 doi: 10.1111/2041-210X.12139

APPLICATION

A method for detecting modules in quantitative bipartite networks

Carsten F. Dormann[1*] and Rouven Strauss[2]

ROYAL SOCIETY OPEN SCIENCE

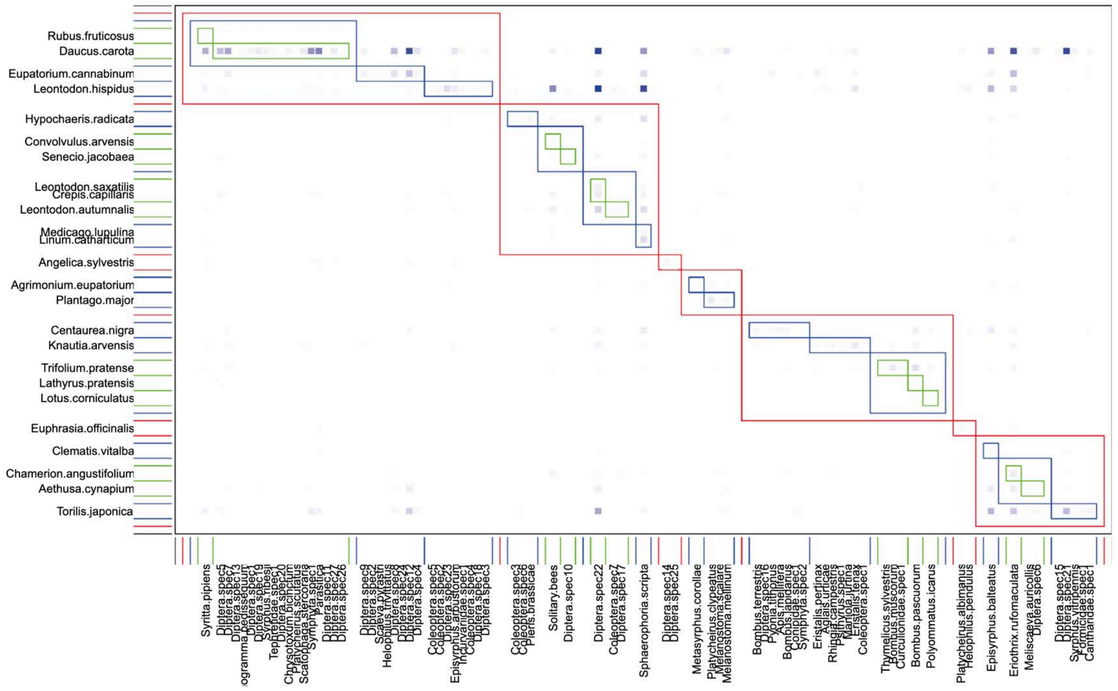
rsos.royalsocietypublishing.org

Research

CrossMark

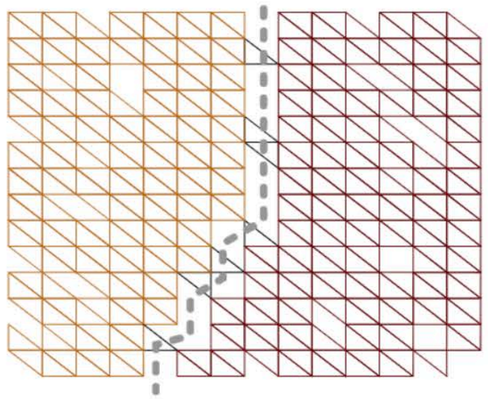
Improved community detection in weighted bipartite networks

Stephen J. Beckett[†]

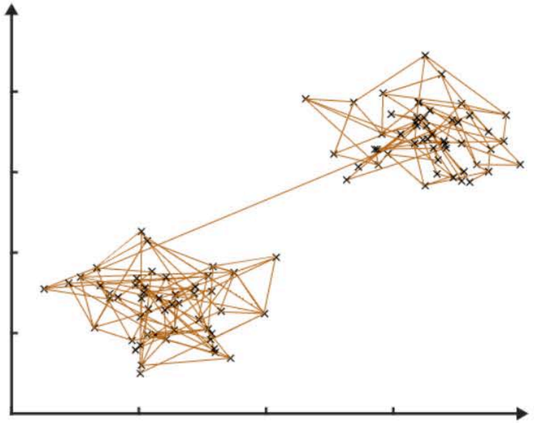


Approaches for community detection

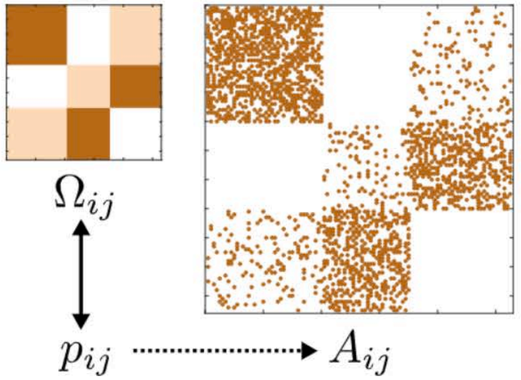
(i) Cut-based perspective



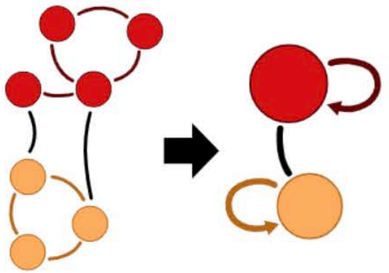
(ii) Clustering perspective



(iii) Stochastically equivalent nodes



(iv) Dynamical perspective

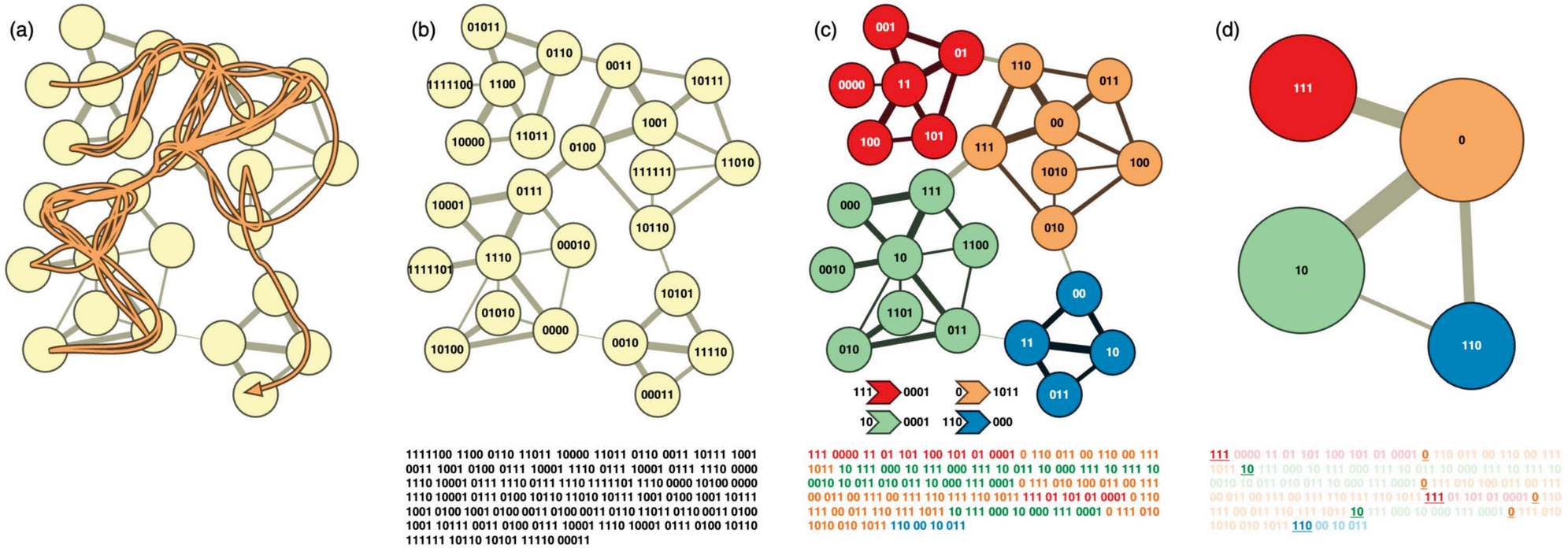


$\dot{x}=f(x) \approx \dot{y}=g(y)$

Rosvall et al 2020

From flow to information

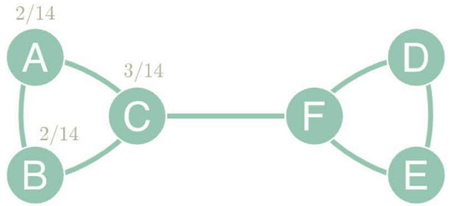
- Measure the amount of information required to describe a random walk within and between modules.
- For a given partition of the network, there is an associated information cost, measured in bits, for describing the movements of the random walker.



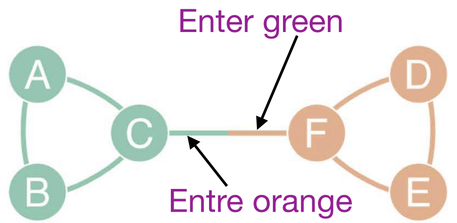
Rosvall and Bergstrom 2008., PNAS; Rosvall et al 2010, Eur Phys J

From flow to information

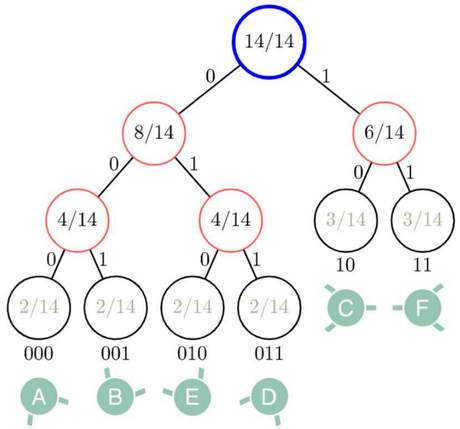
(a)



(b)

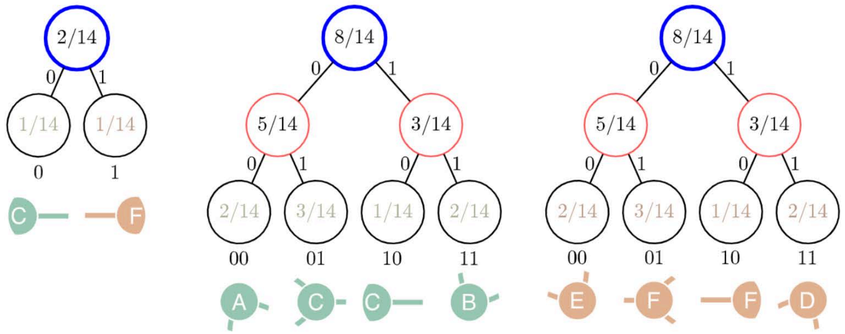


(c)



$$L = 14/14\, H(2/14, 2/14, 3/14, 2/14, 2/14, 3/14) \approx 2.56 \text{ bits}$$

(d)



$$L = 2/14\, H(1/14, 1/14) + 8/14\, H(2/14, 2/14, 3/14, 1/14) + 8/14\, H(2/14, 2/14, 3/14, 1/14) \approx 2.32 \text{ bits}$$

The walk C—>A—>B is encoded by: 10-000-001

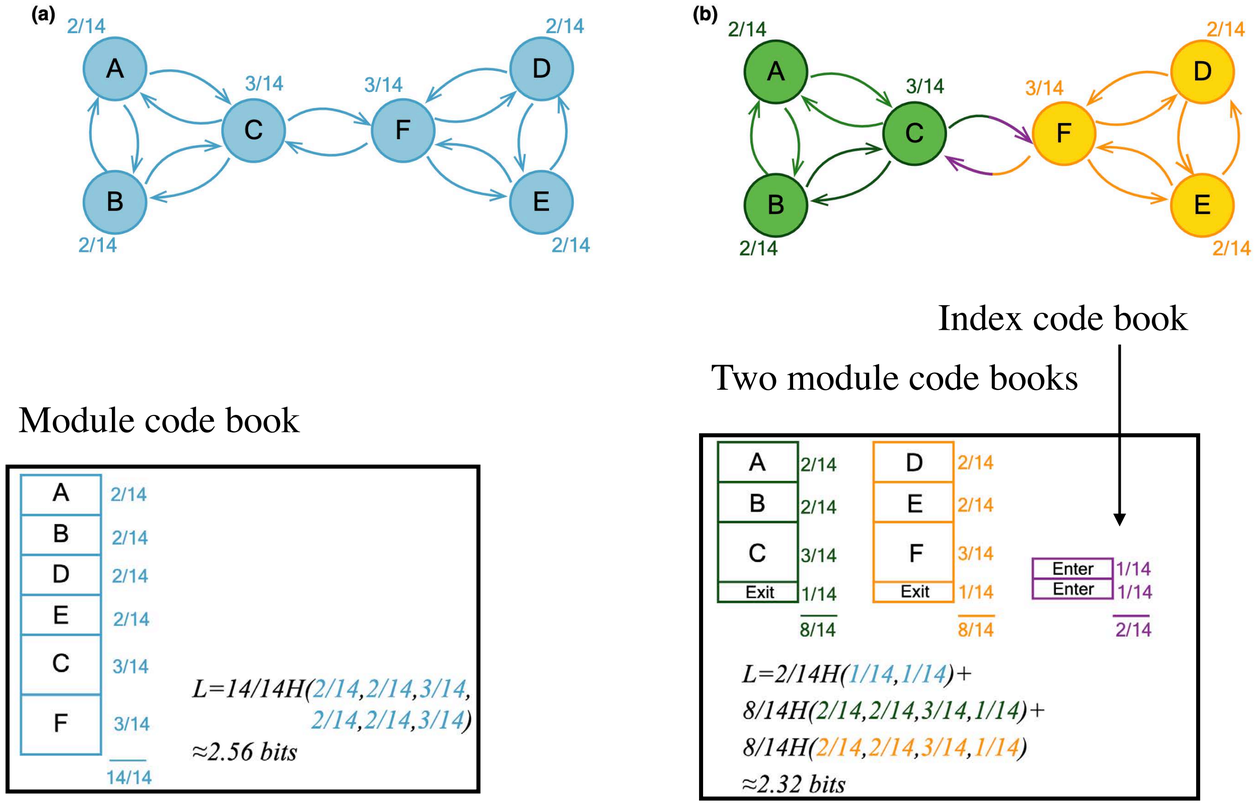
The walk C—>F—>D is encoded by: 1011011

The walk C—>A—>B is encoded by: 01-00-11

The walk C—>F—>D is encoded by: 01-10-1-01-11

Rosvall and Bergstrom 2008, PNAS; Farage et al 2021 Methods Ecol Evol

From flow to information



$$L(\mathsf{M}) = q_{\curvearrowleft} H(\mathcal{Q}) + \sum_{i=1}^{m} p^{i}_{\circlearrowright} H(\mathcal{P}^{i})$$

$$H = \sum_{j} p_j \cdot log_2(p_j)$$

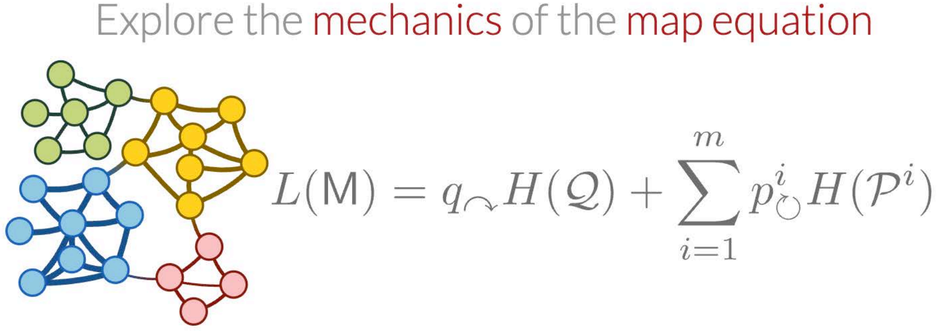
- Weighted average code length is the average amount of information necessary to describe a step:
$$L^{\text{Huffman}} = \sum_{j} p_j l_j$$
- In the two-module solution, we can use the same code words for different nodes in different modules, enabling shorter code words that save information.
- This requires “paying” information to describe exit and entry to modules.
- **But**: we are not interested in an actual description but only in **estimating** the description length.
- The map equation allows us to directly calculate the theoretical limit L of the description length using Shannon’s entropy H. **It measures the theoretical minimum description length L in bits given a network partition M.**
- The optimal partition is the one that best compresses a description of flows on the network. Therefore, **the goal is to minimize the map equation.**

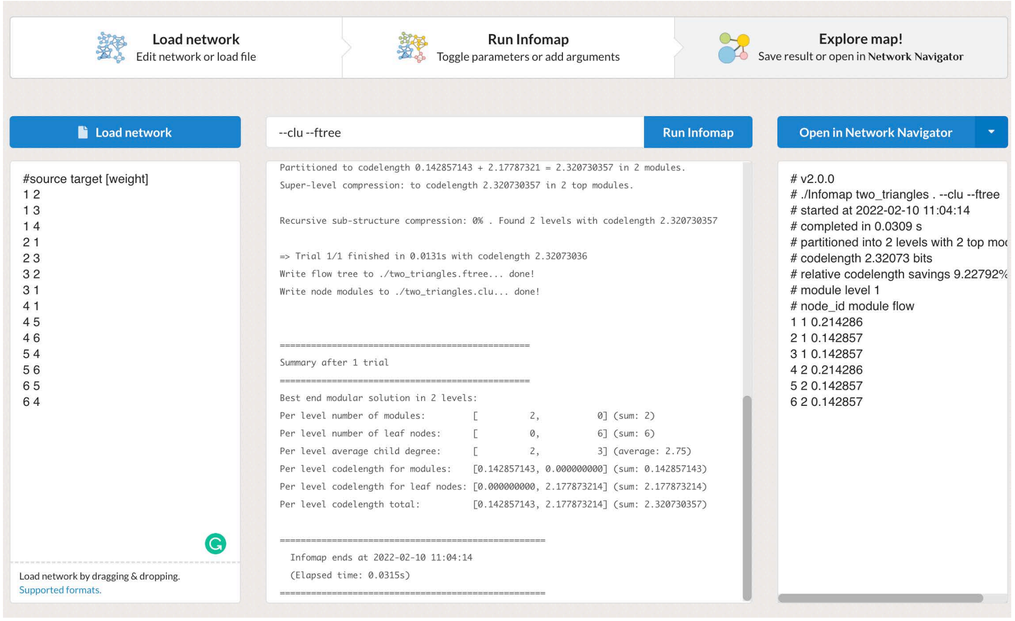
Rosvall and Bergstrom 2008, PNAS; Farage et al 2021 Methods Ecol Evol

Why use Infomap

- Efficient and fast algorithm.
- Handles almost all network types: unipartite, bipartite, directed, weighted, multilayer
- Adequate for describing flow and edge density.
- Friendly, maintained, constantly developed, super-documented
- One-stop shop: https://www.mapequation.org/. No need for multiple implementations
- R package (infomapecology) https://ecological-complexity-lab.github.io/infomap_ecology_package/

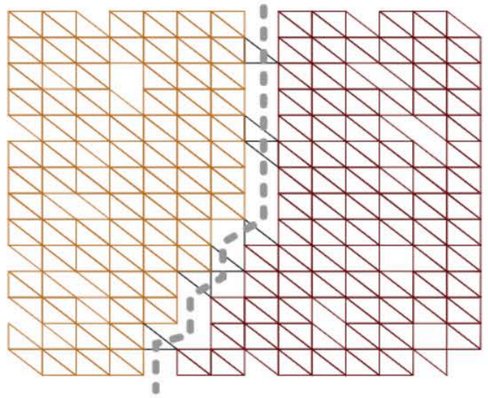
https://www.mapequation.org/demo/



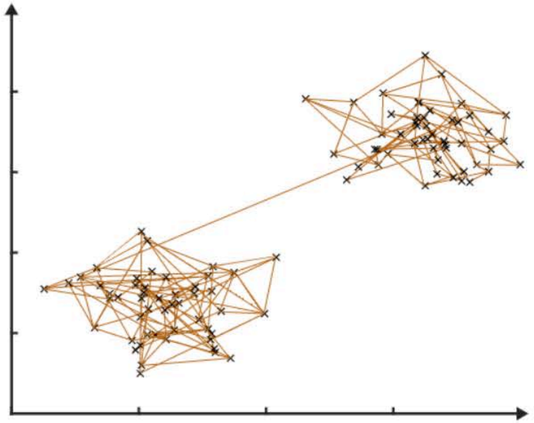


Approaches for community detection

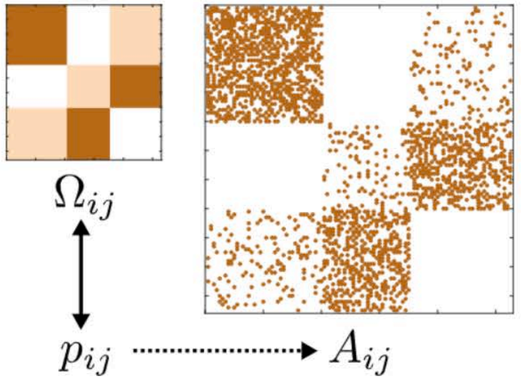
(i) Cut-based perspective



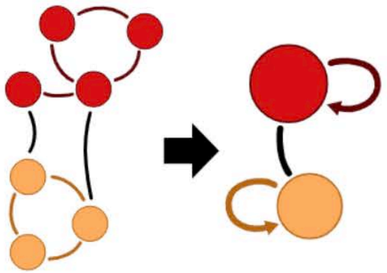
(ii) Clustering perspective



(iii) Stochastically equivalent nodes



(iv) Dynamical perspective



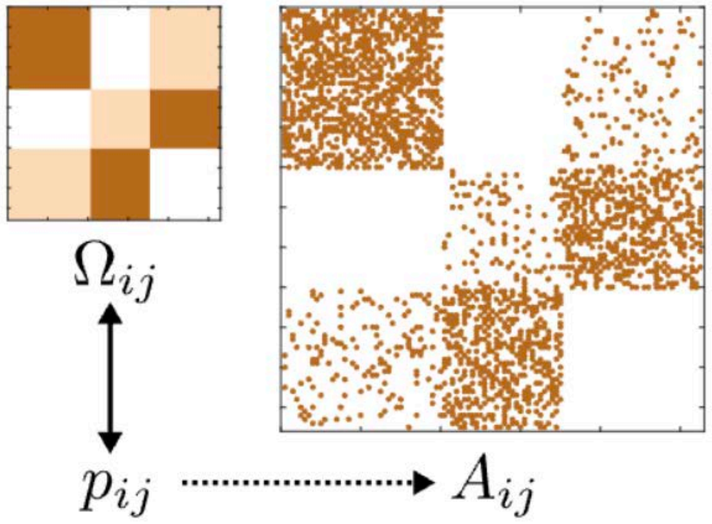
$\dot{x}=f(x) \approx \dot{y}=g(y)$

Rosvall et al 2020

Stochastic block models

$$P(A_{ij} = 1 \mid z_i, z_j) = \Omega_{z_i z_j}$$

Nodes within the same group share the same probability of connecting to nodes of another group.



Goal: Find the latent groups of nodes in a network.

How: Fit model parameters, use model selection to find optimal K

Likelihood: probability of observed data given the parameters.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \mathscr{L}(\theta \mid D)$$

Rosvall et al 2020, Peixoto 2021

Stochastic block models

The Goal: We want to find the best way to group S species into K communities and find the probability of links between those groups.

The "Likelihood" function: For any specific grouping we calculate the mathematical probability that the network would look exactly like the one we measured.

The "Maximum": The algorithm tries thousands of different groupings. The one that results in the highest probability of producing your actual observed network is the Maximum Likelihood Estimate.

SBM in ecology: The group model

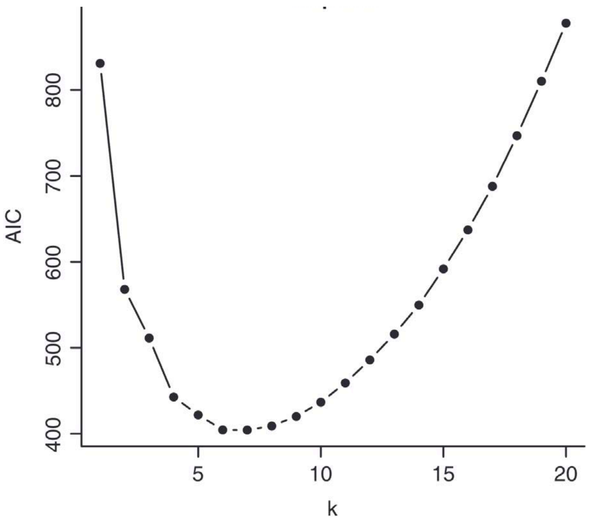
Simplest model: connect two nodes with probability p (= a single group).

Then what is the likelihood of obtaining the observed network?

$$\mathscr{P}(N(S,L)\,|\,p) = p^{L}(1-p)^{S^2-L} \Rightarrow \text{ML: } p = L/S^2$$

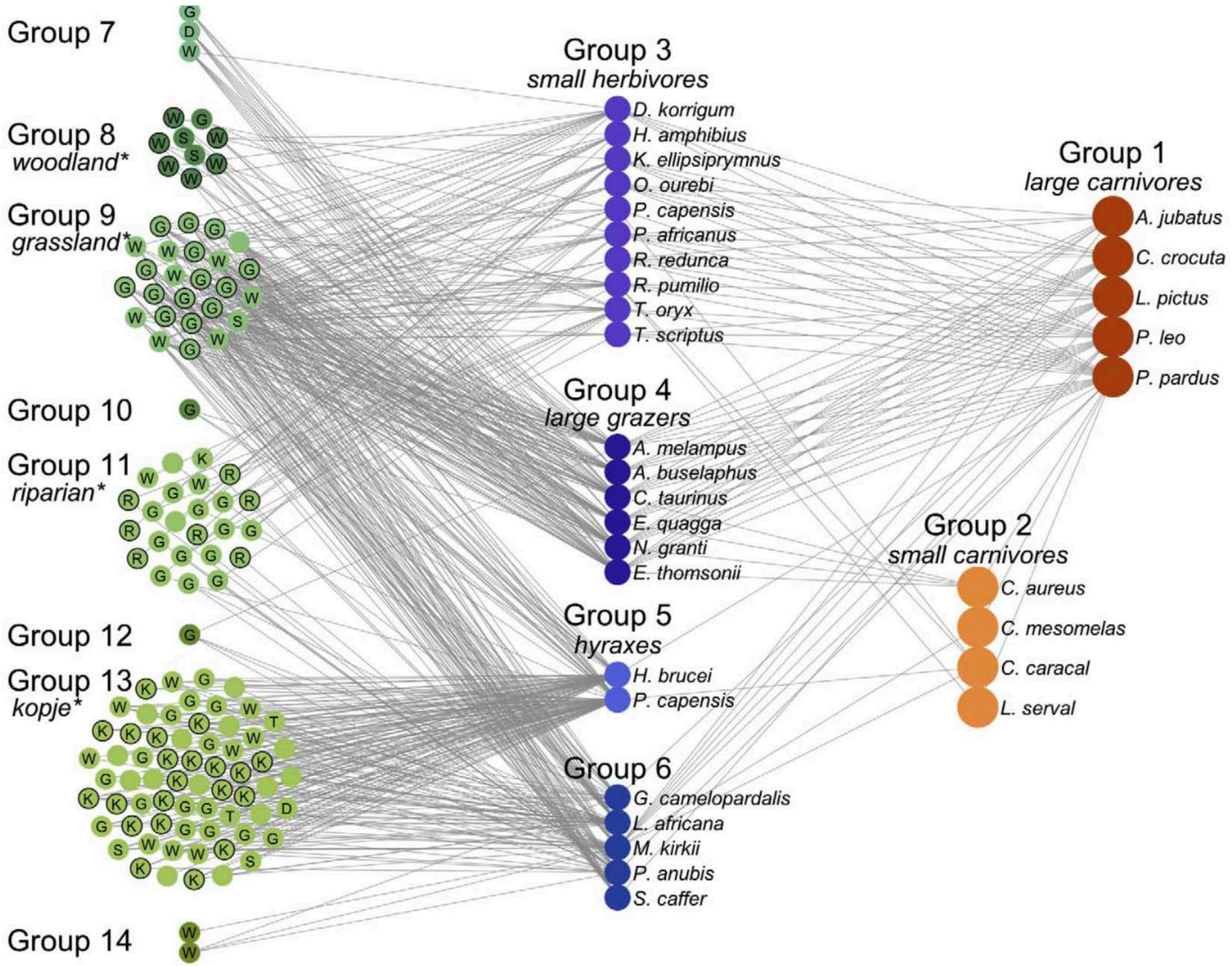
A generalization for k groups, with L_{ij} number of links connecting groups i and j

$$\mathscr{P}(N(S,L)|\,\vec{p}) = \prod_{i=1}^{k}\prod_{j=1}^{k} p_{ij}^{L_{ij}}(1-p_{ij})^{S_iS_j-L_{ij}}$$



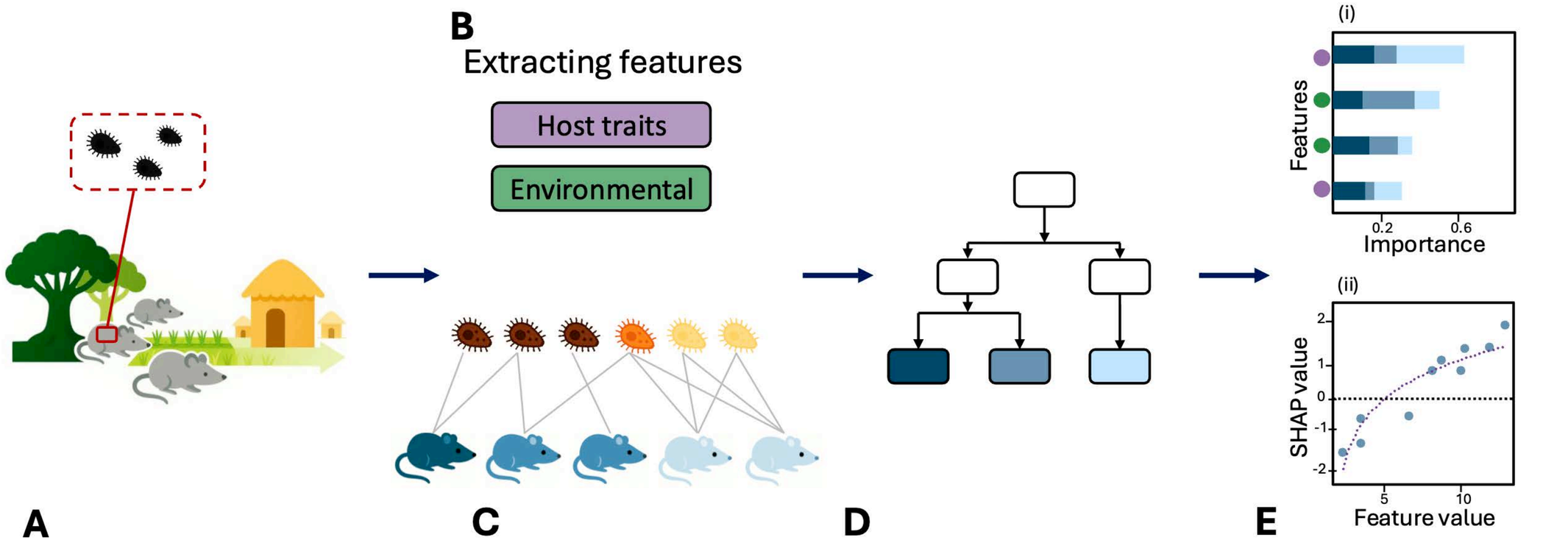
Allesina and Pascual 2009

SBM in ecology: The group model



Baskerville et al 2011

SBM identifies infection profiles



A
Host-parasite sampling along environmental gradients

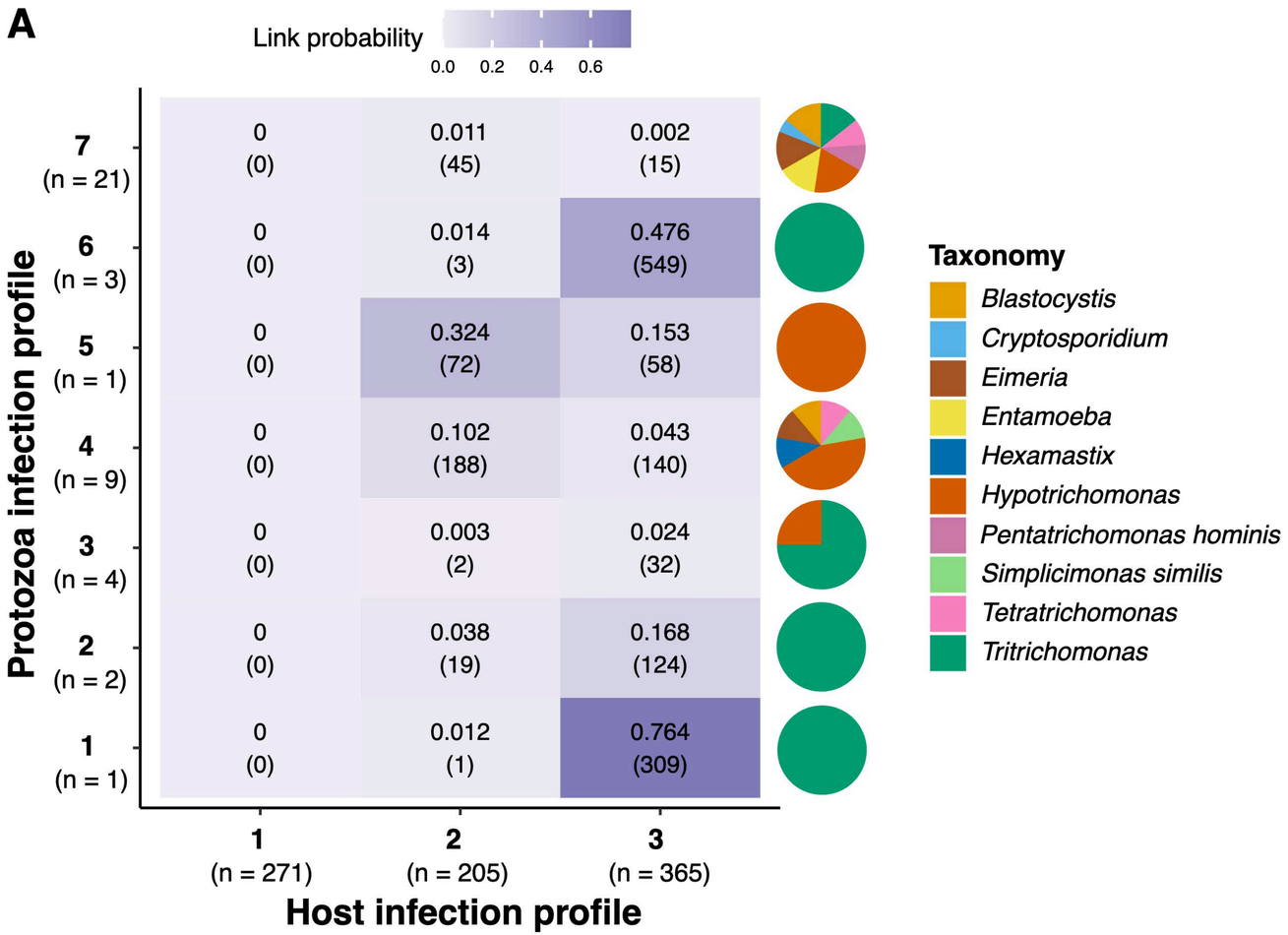
C
Detecting infection profiles (SBM)

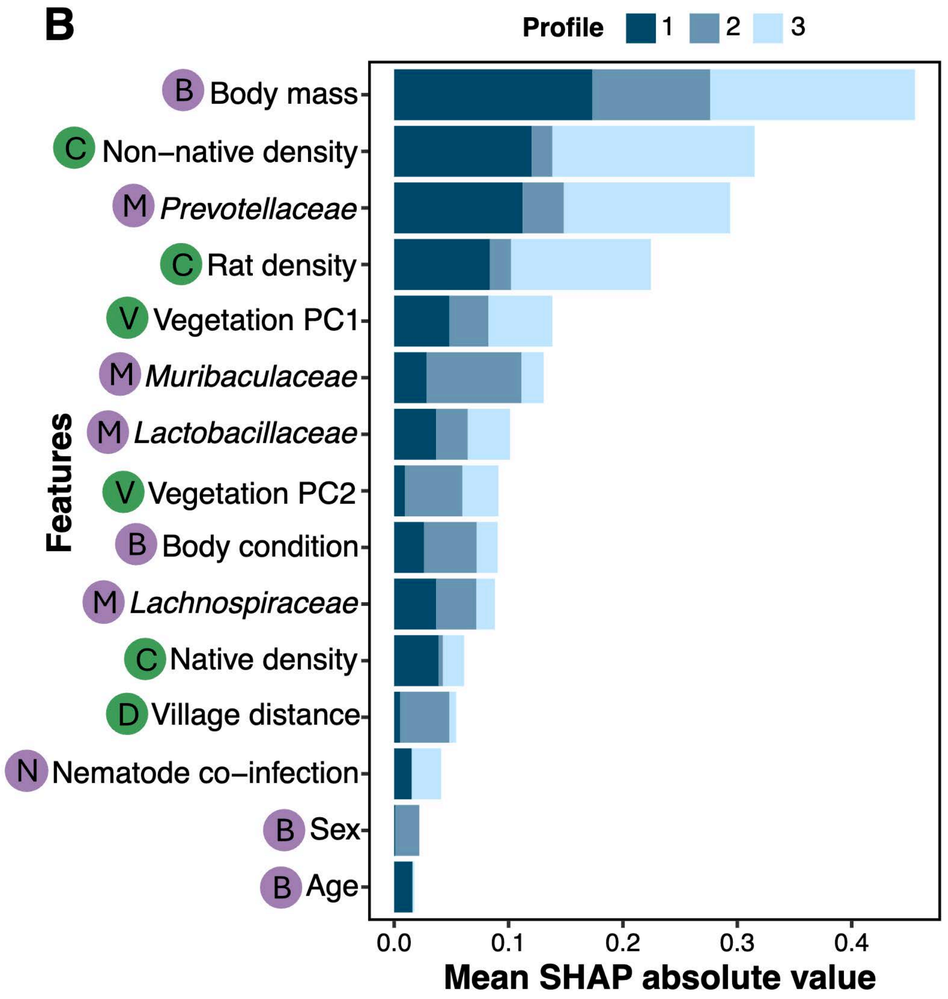
D
Training classification model to predict infection profiles (XGBoost)

E
Feature importance and dependence plots (SHAP)

Markfeld, Talpaz*, ISME Communications*

SBM identifies infection profiles





Markfeld, Talpaz*, ISME Communications*

What can we ask about the communities we find?

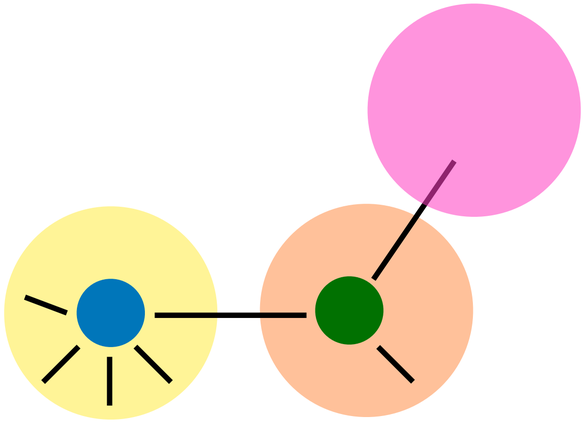
How can we characterize them?

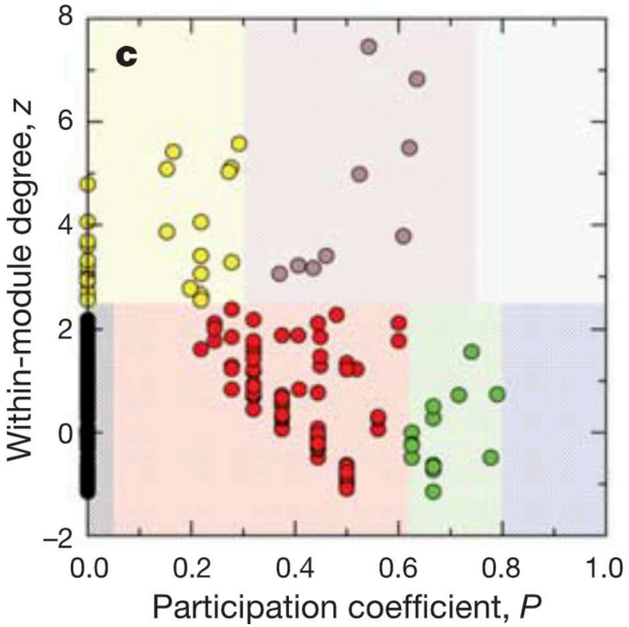
- Size distribution
- Node traits
- Phylogenetic (or other) signals
- Links within and between
- Dynamical consequences

Nodes' "module roles"

The role of a species *i* can be characterized by its **standardized within-module degree z:**

$$z = \frac{k_{is} - \bar{k}_s}{SD_{ks}},$$



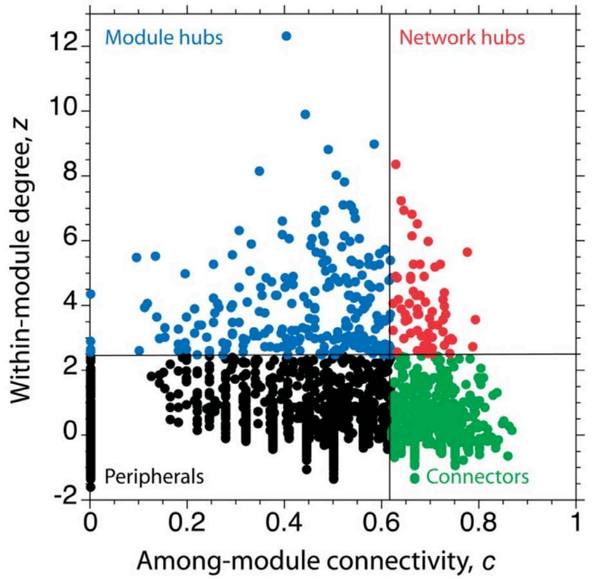


Guimerà and Amaral 2005, Nature

and its **among-module connectivity, c**

$$c = 1 - \sum_{t=1}^{N_M} \left(\frac{k_{it}}{k_i} \right)^2,$$

If i has all its links within its own module, c = 0; and if these are distributed evenly among modules, c → 1.



Olesen et al 2007, PNAS

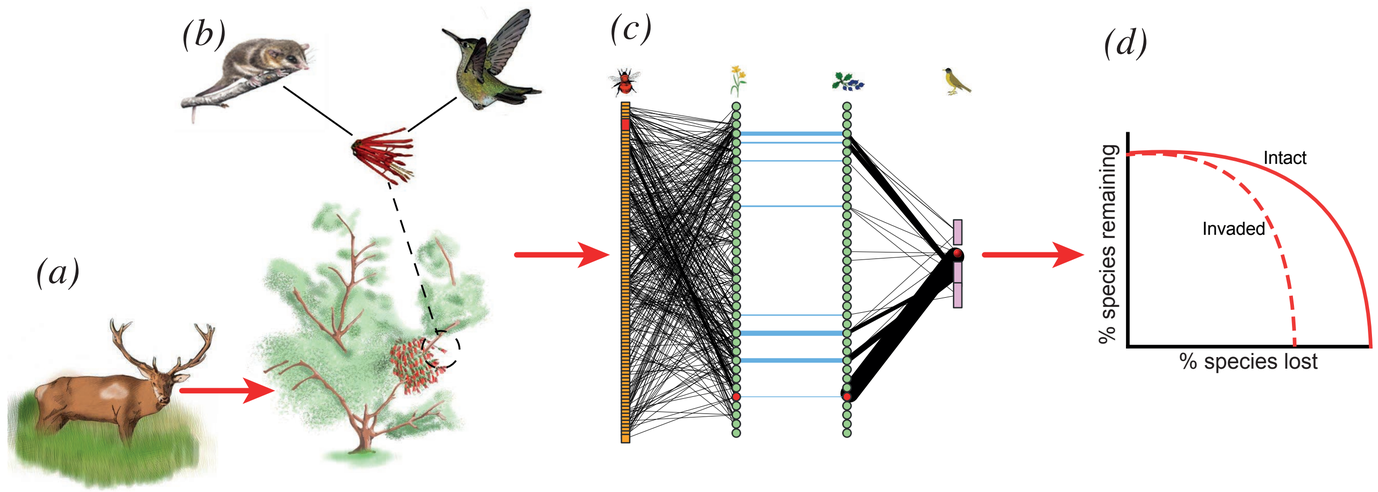
k_{is}: number of links of *i* to other nodes **in its own module *s***.

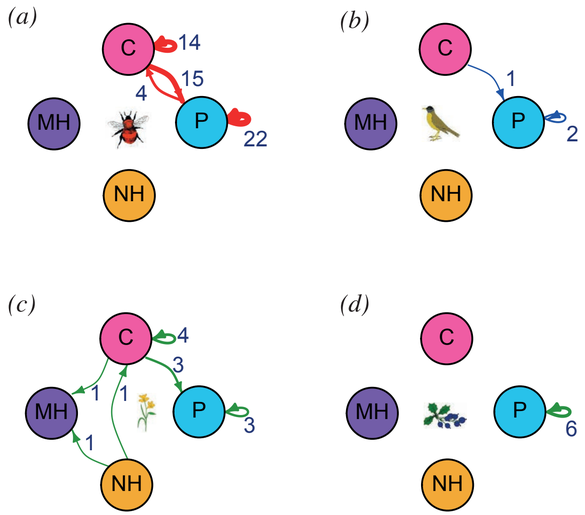
$\bar{k}_s$ and SD_{ks}: average and SD of the degree of all nodes in modeule *s*.

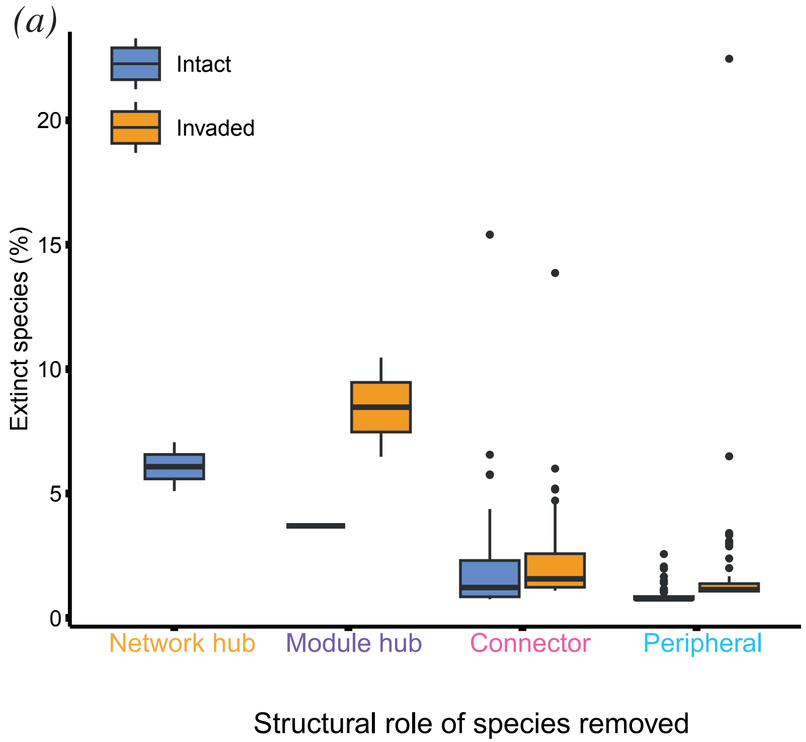
k_i: degree of node *i* (inter and intra module).

k_{it}: number of links from *i* to nodes in module *t* (including *i*'s own module).

Network robustness depends on species roles

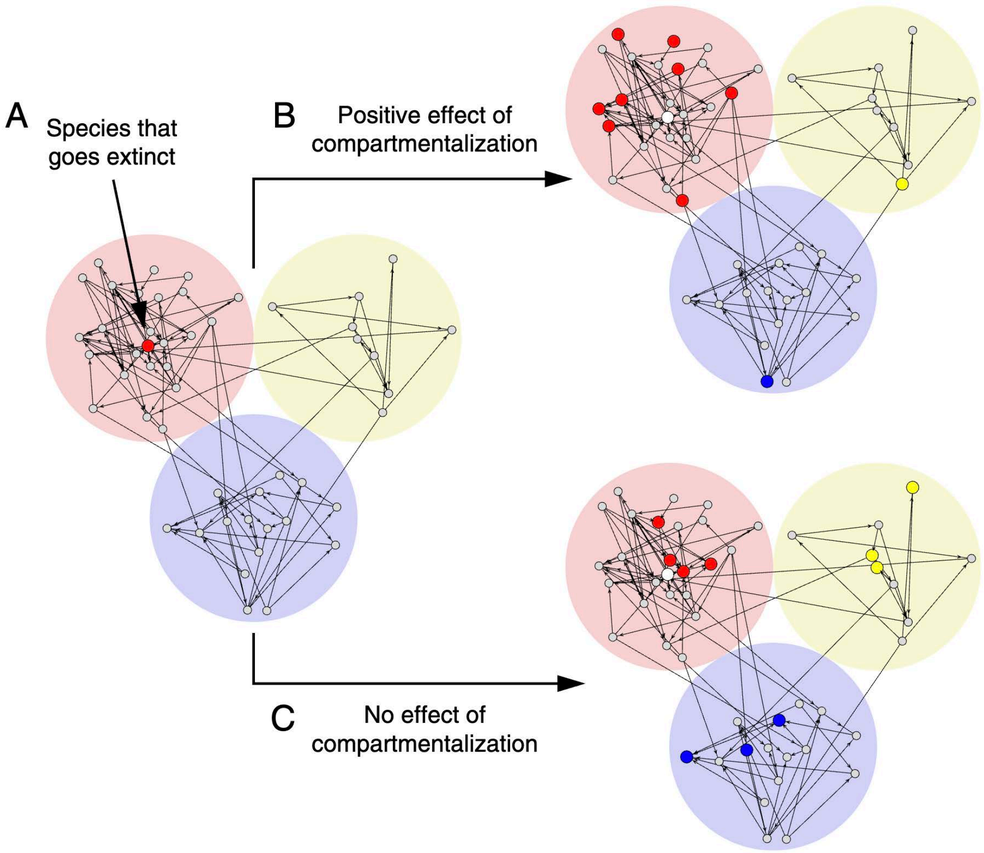






Vitali,…, Pilosof 2023

Modularity slows perturbations



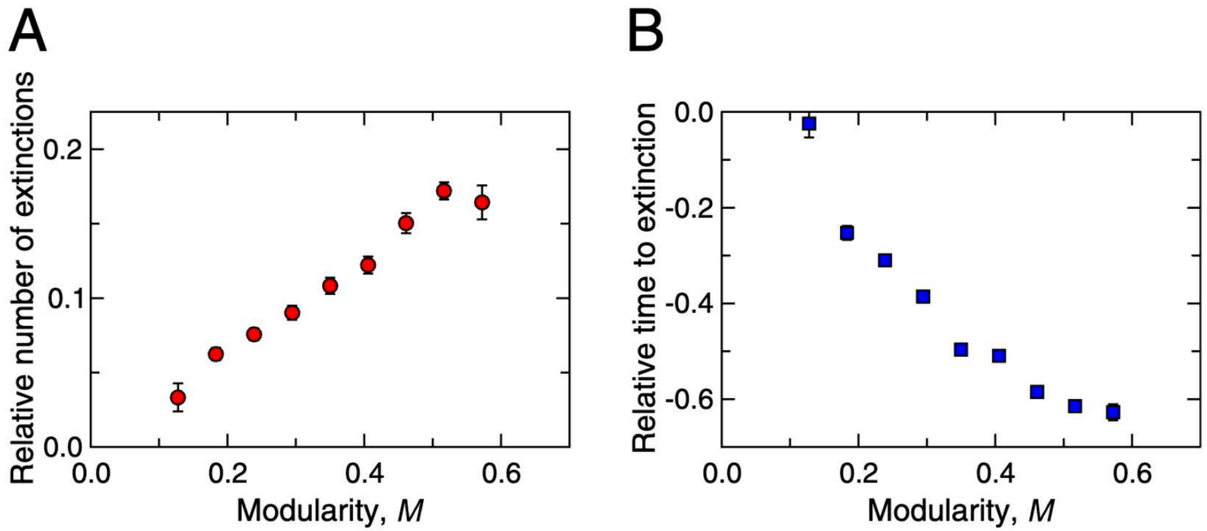
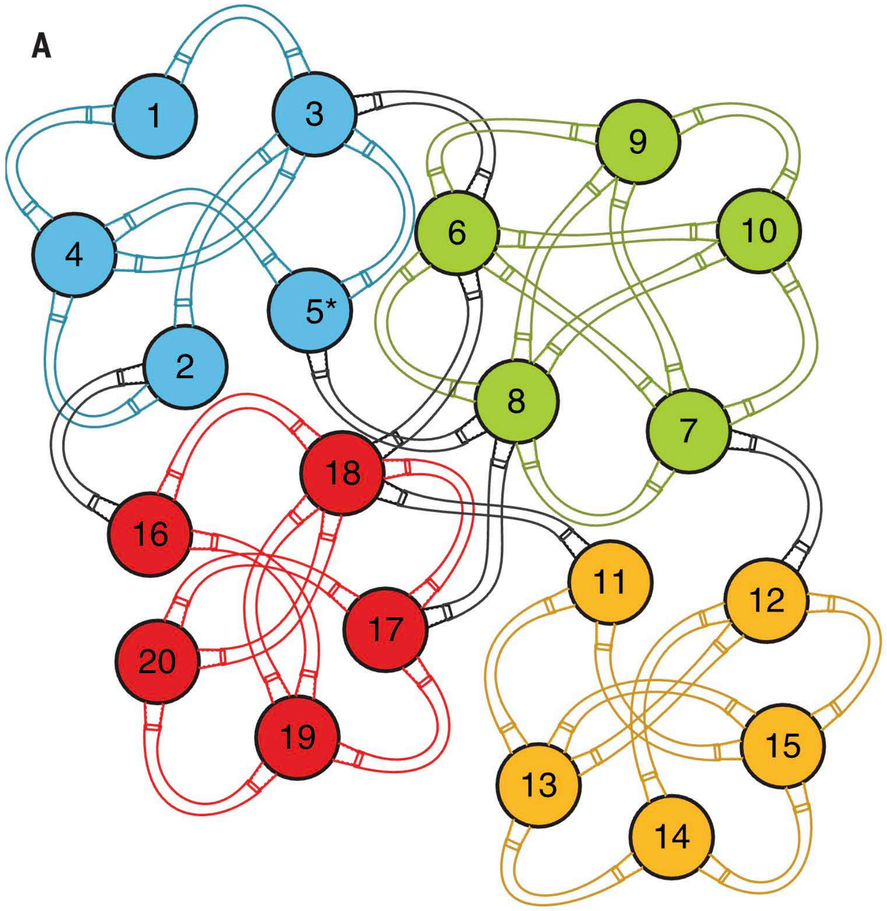
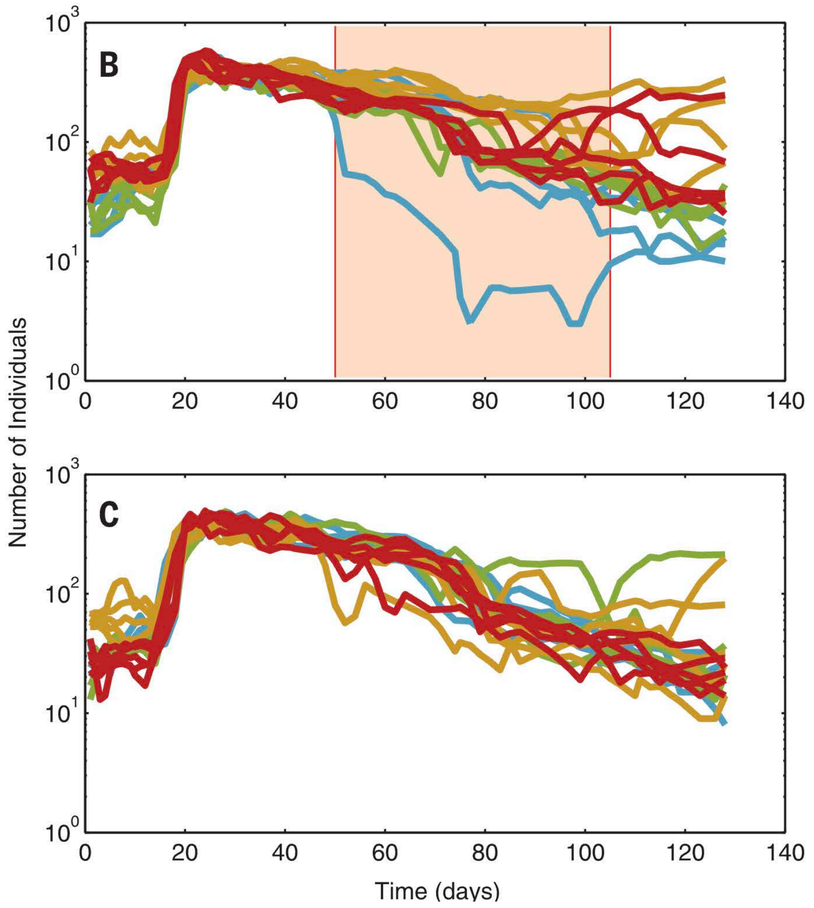


Fig. 3. Community response to manipulated species extinctions. (*A*) Mean relative number of extinctions that occur in the same compartment as an eliminated species, as a function of the web's modularity. Values greater than zero imply that the subsequent species that go extinct as a consequence of the original extinction have a higher probability of belonging to the same compartment. (*B*) Mean relative time to extinctions that occur in the same compartment as the eliminated species, as a function of the web's modularity. Values less than zero imply that these species tend to go extinct earlier, as a consequence of the original extinction. The SEs of the reported averages are shown as error bars.

Stouffer and Bascompte 2011, PNAS

Modularity slows perturbations

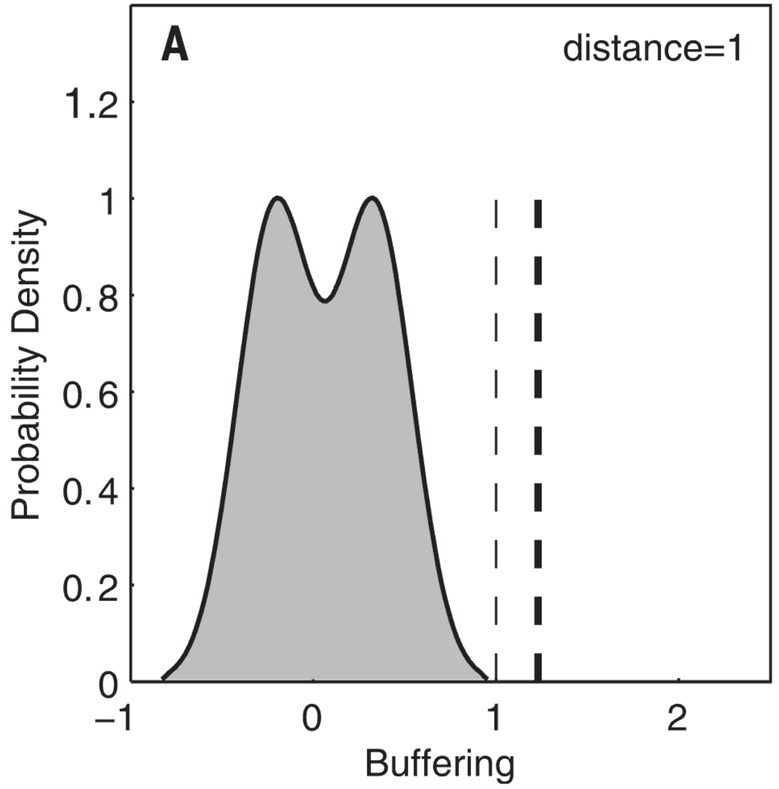




Gilarranz et al 2017, Science

Modularity slows perturbations

Buffering: the ratio between perturbation sizes inside
and outside the module to which the perturbed node belongs.

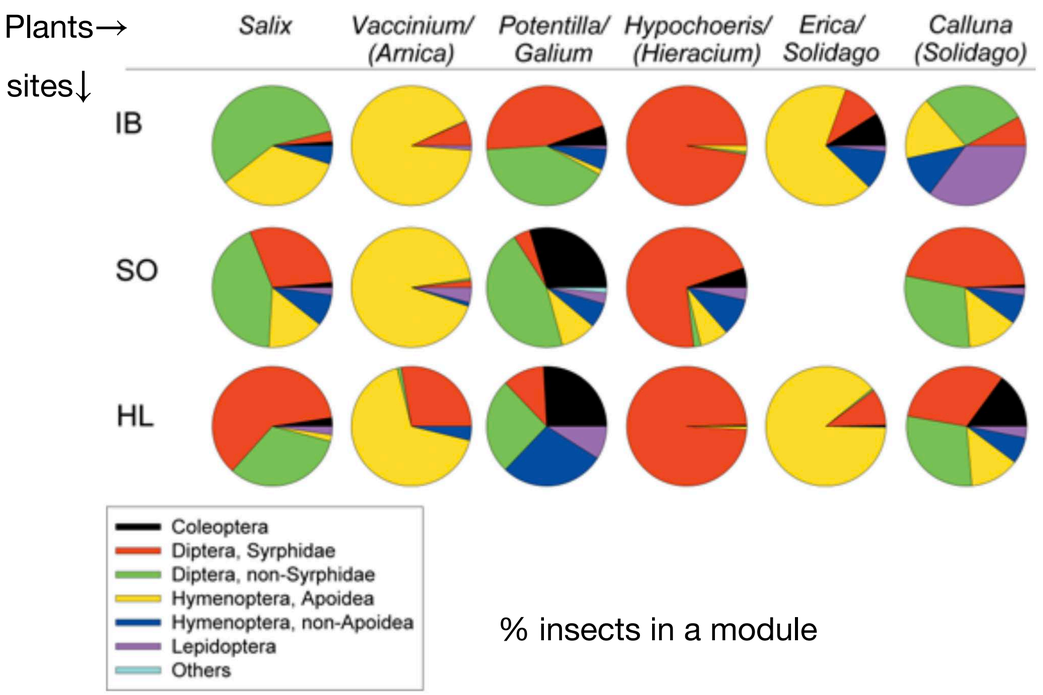


Gilarranz et al 2017, Science

Compartmentalization as a signature for group-generating processes

Remember! Any inference should be in light of the community detection approach

- Chance
- Habitat separation
- Temporal dynamics
- Competition
- Trait matching ⟶
- One-sided adaptation
- Coevolution



% insects in a module

Dormann 2017, Ann Rev Ecol Evol

Dupont and Olesen 2009

Summary

- Community detection is a cornerstone of network research.
- Community detection is a non-trivial problem.
- The methodology should match the question.

Choose a definition / objective function for a community

Identify communities by inspecting partitions using an algorithm.

Select the partition that best satisfies the condition