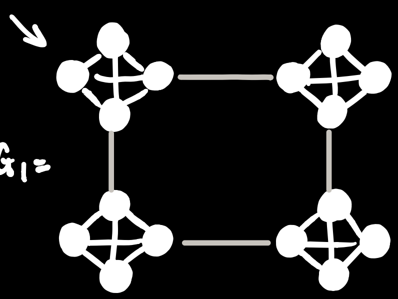
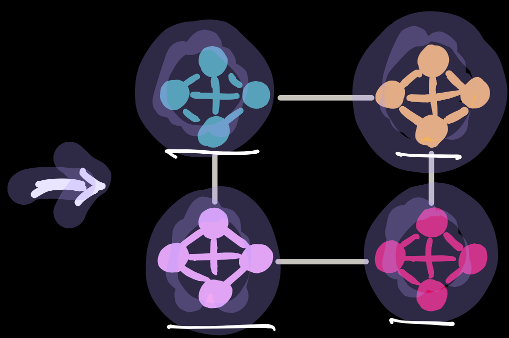
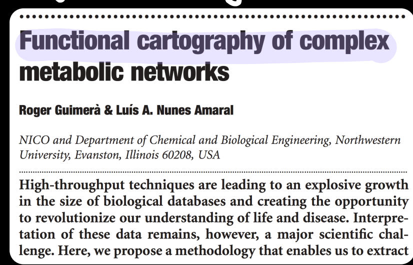
GraphT

Graph modularity
1. Module / community definition
2. Agglomerative and divisive clustering
3. Cartographic classification of nodes

$G_1 =$






Functional cartography of complex metabolic networks

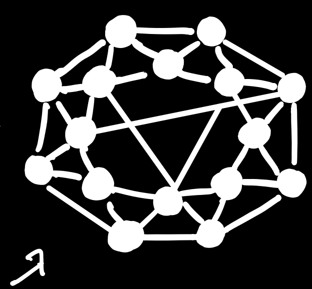
Roger Guimerà & Luís A. Nunes Amaral

NICO and Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA

High-throughput techniques are leading to an explosive growth in the size of biological databases and creating the opportunity to revolutionize our understanding of life and disease. Interpretation of these data remains, however, a major scientific challenge. Here, we propose a methodology that enables us to extract

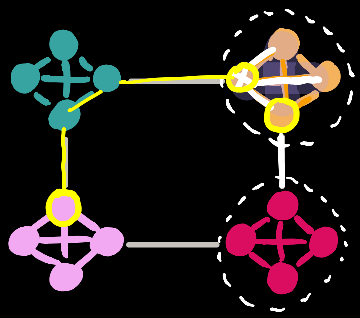
Guimerà et al., Nature 2005

$G_2 =$



there are no noticeable modules

1. Modularity definition



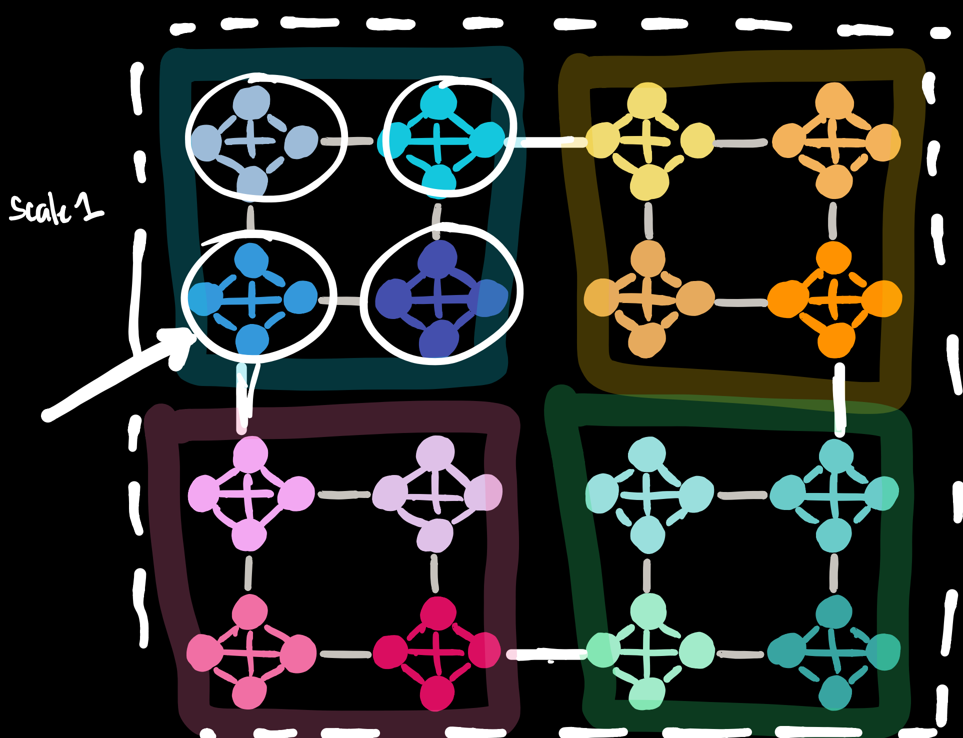
* Nodes in real-world graphs aggregate into densely connected subgroups called modules or communities.

* Nodes within these modules are strongly connected with each other than other parts of the network.

Example :

Social networks → friendships ties tend to be dense within certain group of people (sharing similar social interests)

* **Hierarchical modularity :**
Real-world graphs are organized hierarchically in their modules such that they contain modules within modules → over several topological scales of resolution.



Example : age → gender → social interest.

* Modular graphs have strong within-module connectivity results and a small number of inter-modular links to maintain a low characteristic path length of the network.

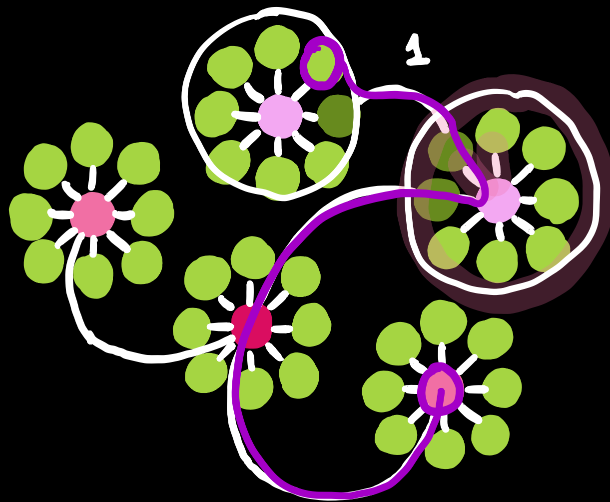
Reminder GT6

$$L = \frac{1}{N}\sum_{i} \boxed{l_i} = \underbrace{\frac{1}{N(N-1)}}_{\text{number of node pairs excluding self-pairing}} \sum_{i \neq j} \boxed{l_{ij}}$$

average shortest path from node i to all other nodes

shortest path length from node j to node i.

Intuition: A short characteristic path length L means that information can, on average, be routed between pairs of nodes using only a few connections.



Modular graph comprising five communities.

* Benefits of modularity in graph topology:

Nobel Laureate Herbert A. Simon (American economist and cognitive psychologist) 1916 – 2001

→ argued that modularity and hierarchical organization are essential ingredients for

evolvability flexibility adaptability complexity

→ To illustrate this point he compared two hypothetical watch makers

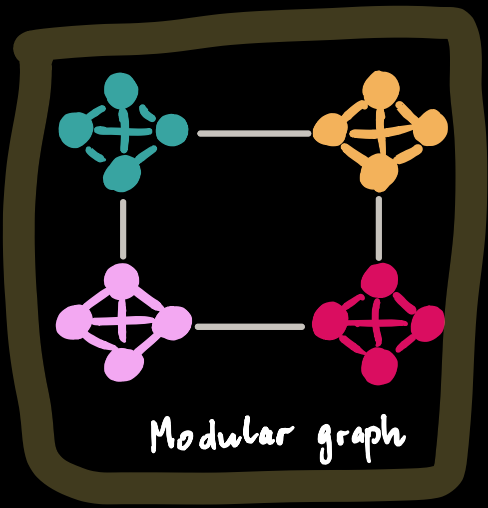
Hora

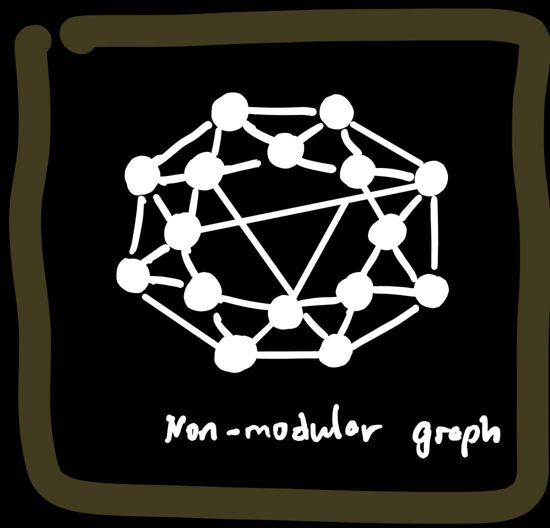
assembles his watch using a piecemeal approach, in which he constructs subassemblies consisting of 10 parts each. Last he combines them all.

if they get interrupted during the construction what is most likely to happen?

Tempus

assembles his watch so that each part is dependent on another.

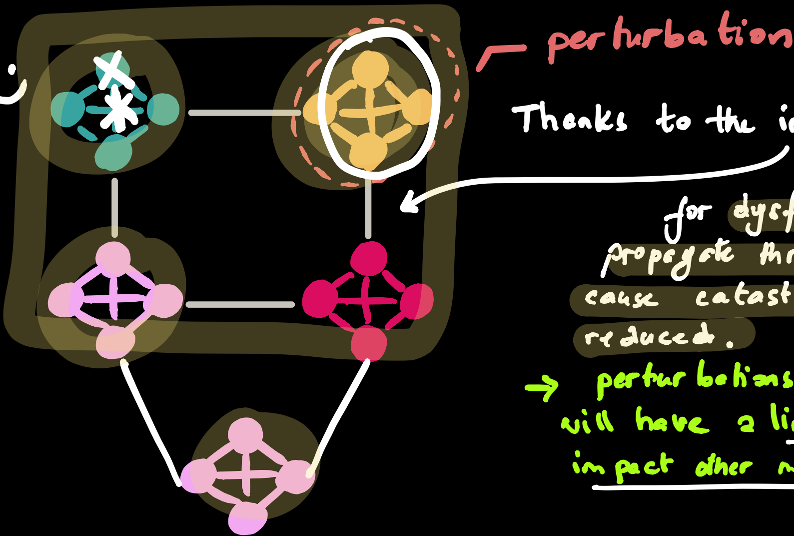




What are the advantages of modular hierarchical architectures over flat non-modular systems?

(Think about evolving systems, dynamically changing graphs, perturbations)

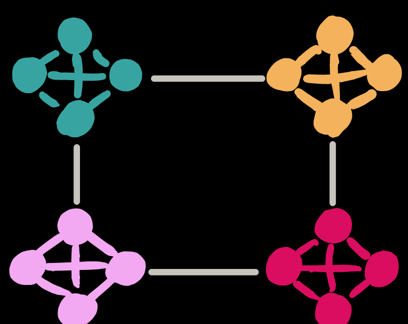
- Hierarchical modular architectures can adapt and evolve to changing environmental circumstances.
- New modules can be added to the system without drastically altering the existing elements (adaptability).
- Modeling studies have shown that modular architectures naturally arise in systems that must optimize performance under changing environmental circumstances (Kashton and Alon, 2005).



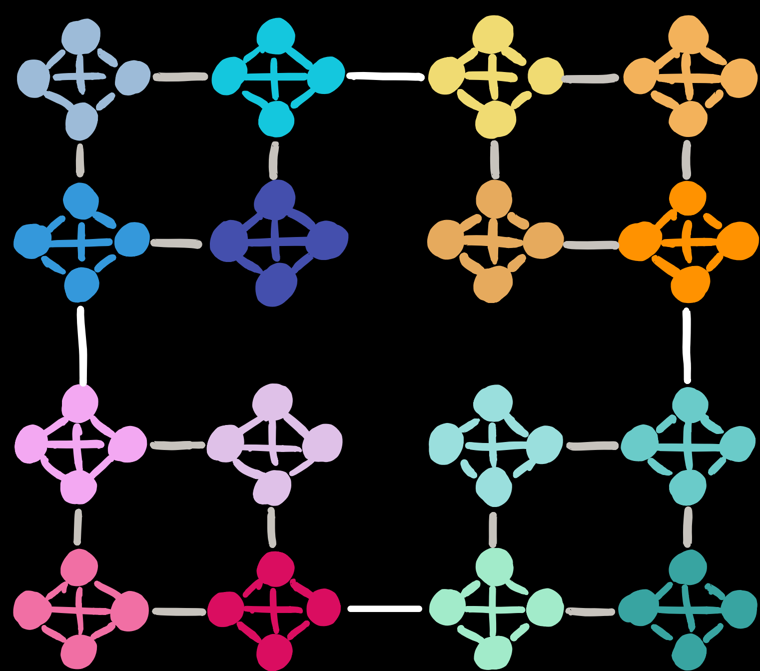
Thanks to the independence between subsystems, the potential for dysfunction or damage to propagate throughout the system and cause catastrophic failures is reduced.

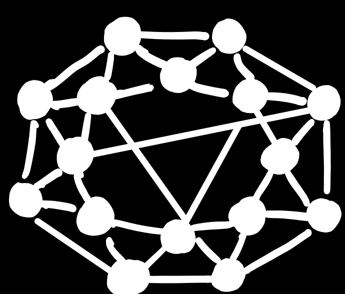
→ perturbations to a modular system will have a limited capacity to impact other modules.

modular



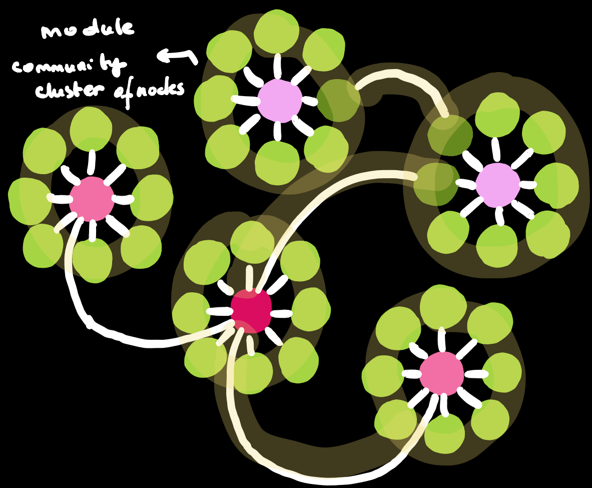
hierarchically modular





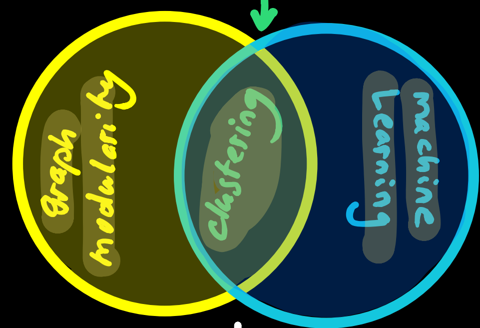
non-modular

2. How to discover the modules of a given graph?

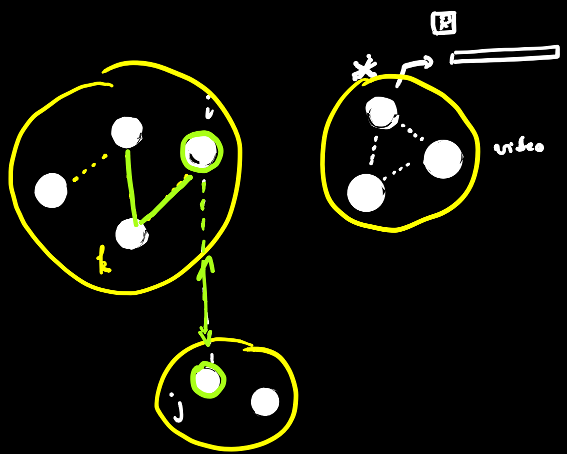


The accurate decomposition of a graph G into modules with high intrinsic connectivity and weaker extrinsic connectivity belongs to a more general class of problems in statistics and machine learning:

CLUSTERING



Basic idea: group similar nodes (observations) together in a fully data-driven manner.



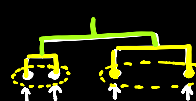
AGGLOMERATIVE methods

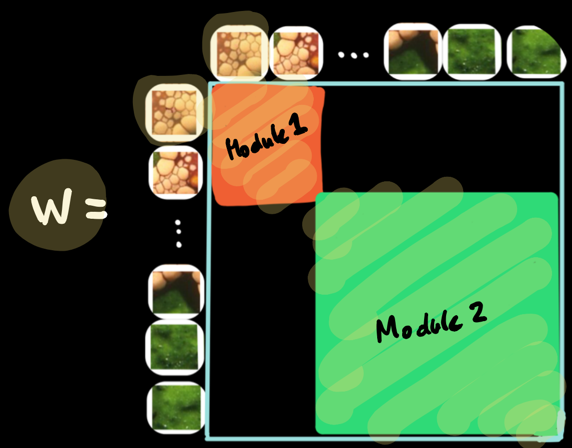
(large → small)

* Start with all points in a single cluster and attempt to find divisions that delineate cohesive subsets of observations.

(small → large)
(hierarchical clustering)

* Start with each point as a cluster and aggregate points into larger sets of similar observations.





GT 7.1 on self-diffusion

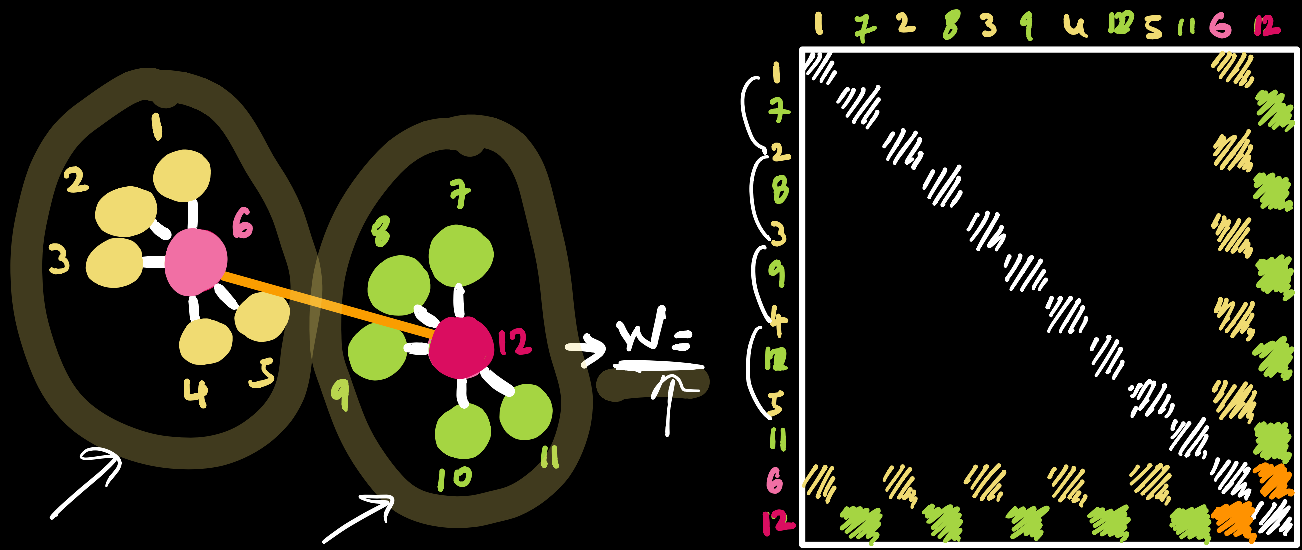
* To cluster graph nodes appropriately, we need to use a 'good' similarity measure between samples (nodes).

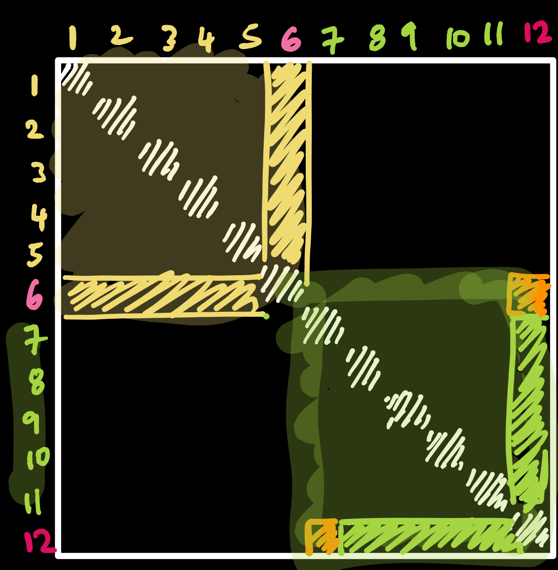
* One can use the direct edge weight between graph nodes.

* Choose an appropriate similarity metric between nodes i and j → generate N x N sample similarity matrix W.

* The values of the similarity matrix W are used to group nodes into clusters, such that:

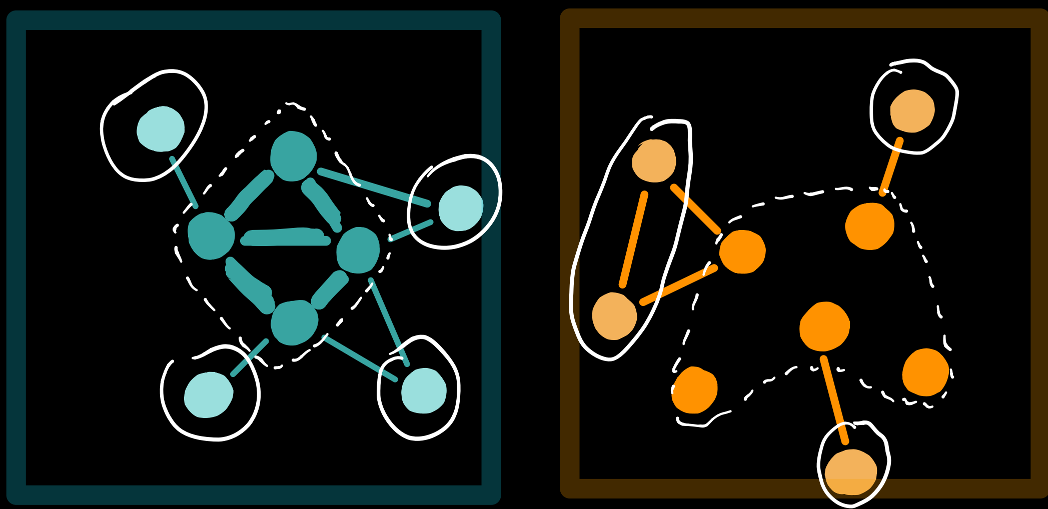
  1) the similarity within clusters is high
  2) the similarity between clusters is low





Observation

This graph can be split into two modules : blue and orange.



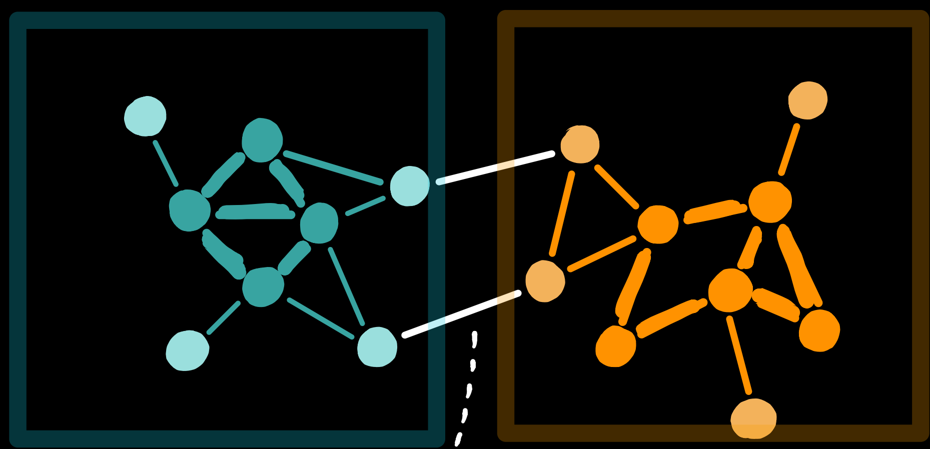
☺ Agglomerative methods cluster well core nodes of modules with strong pairwise similarity (bold node and edge color).

☹ However , they fail to group together peripheral community members with only weak similarity to the rest of the group (pale node and edge colors).

Think about a new clustering approach that detects peripheral submodules .

Girvan and Newman (2002)

This graph can be split into two modules : blue and orange.



Idea : remove edges with high betweeness centrality to identify the natural modules of a network.

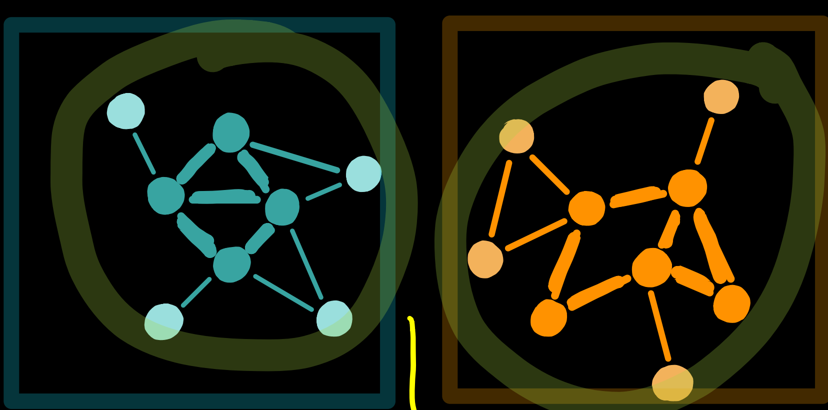
Why this works ?

GT 2.

- Betweeness centrality is a popular measure proposed by Freeman (1977) and Anthonisse (1971).

$C_B(v_i)$ measures the proportion of shortest paths between all node pairs in G that pass through node v_i.

↳ Nodes lying on a large fraction of these shortest paths represent putative bottlenecks of traffic flow under the assumption that information travels along the shortest paths.



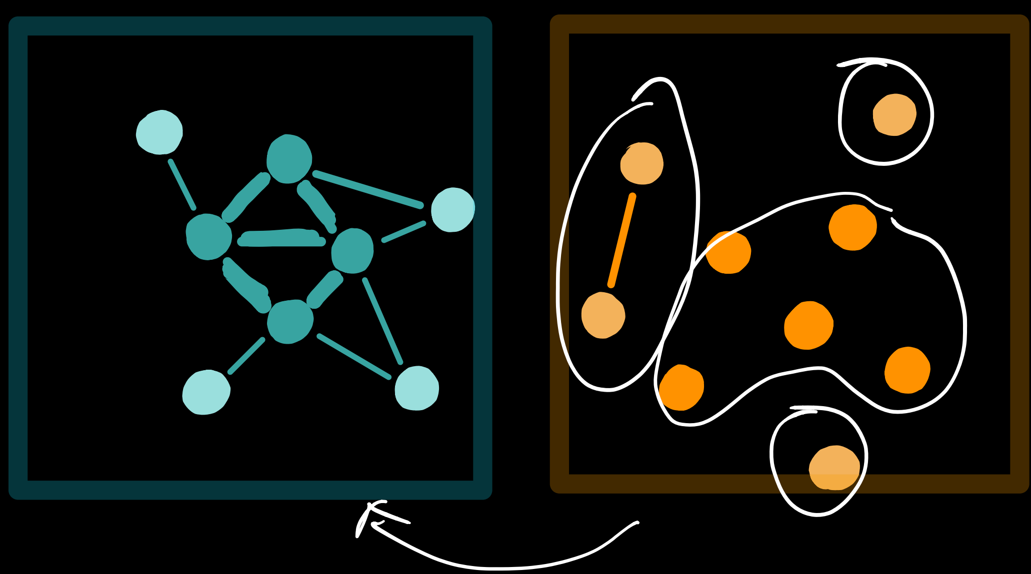
any message traveling from one module to another must pass through one of a sparse set of intermodular links (which means those have high betweeness centrality.

* Girvan and Newman (2002) proposed a divisive method for decomposing a network into hierarchical modules, as an alternative to agglomerative clustering.

* They define modules based on the ordered removal of edges according to their betweenness centrality:

  1. Calculate the betweenness centrality of each edge in G.
  2. Remove the edge with the highest betweenness centrality. If two edges or more have the same score, remove one randomly.
  3. Recalculate the betweenness centrality of all remaining edges.
  4. Repeat steps 3 and 4 until no edges remain.

→ as the algorithm progresses, more edges are removed and smaller and smaller modules appear.



3. Cartographic classification of nodes

Functional cartography of complex metabolic networks

Roger Guimerà & Luís A. Nunes Amaral

NICO and Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA

High-throughput techniques are leading to an explosive growth in the size of biological databases and creating the opportunity to revolutionize our understanding of life and disease. Interpretation of these data remains, however, a major scientific challenge. Here, we propose a methodology that enables us to extract