Community structure

Introduction to Network Science

Instructor: Michele Starnini — https://github.com/chatox/networks-science-course

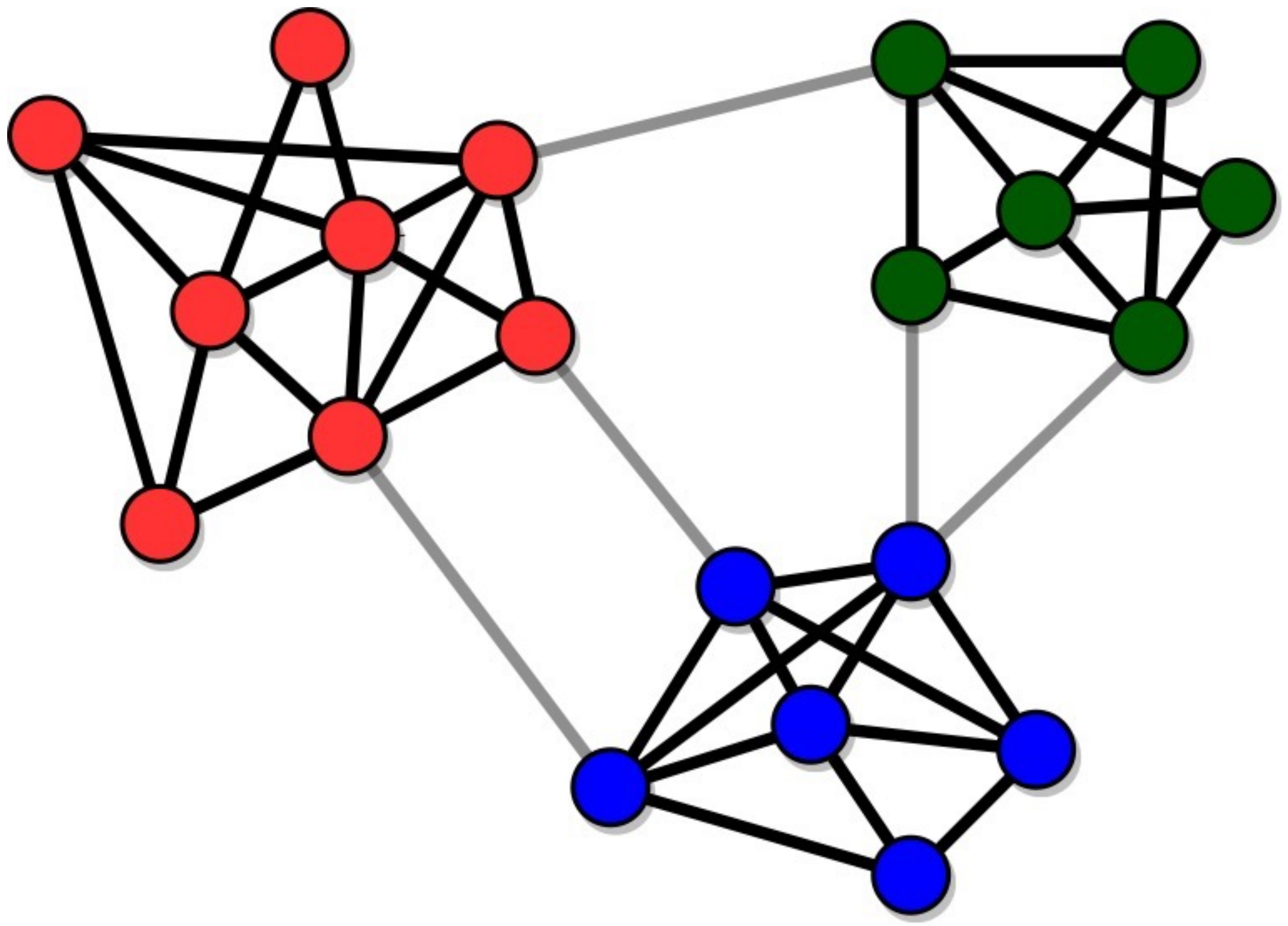
Universitat Pompeu Fabra Barcelona

Content

- Different definitions of a community structure
- Examples: two groups, multiple groups, hierarchical
- K-cores decomposition
- Network partitioning
- Hierarchical clustering

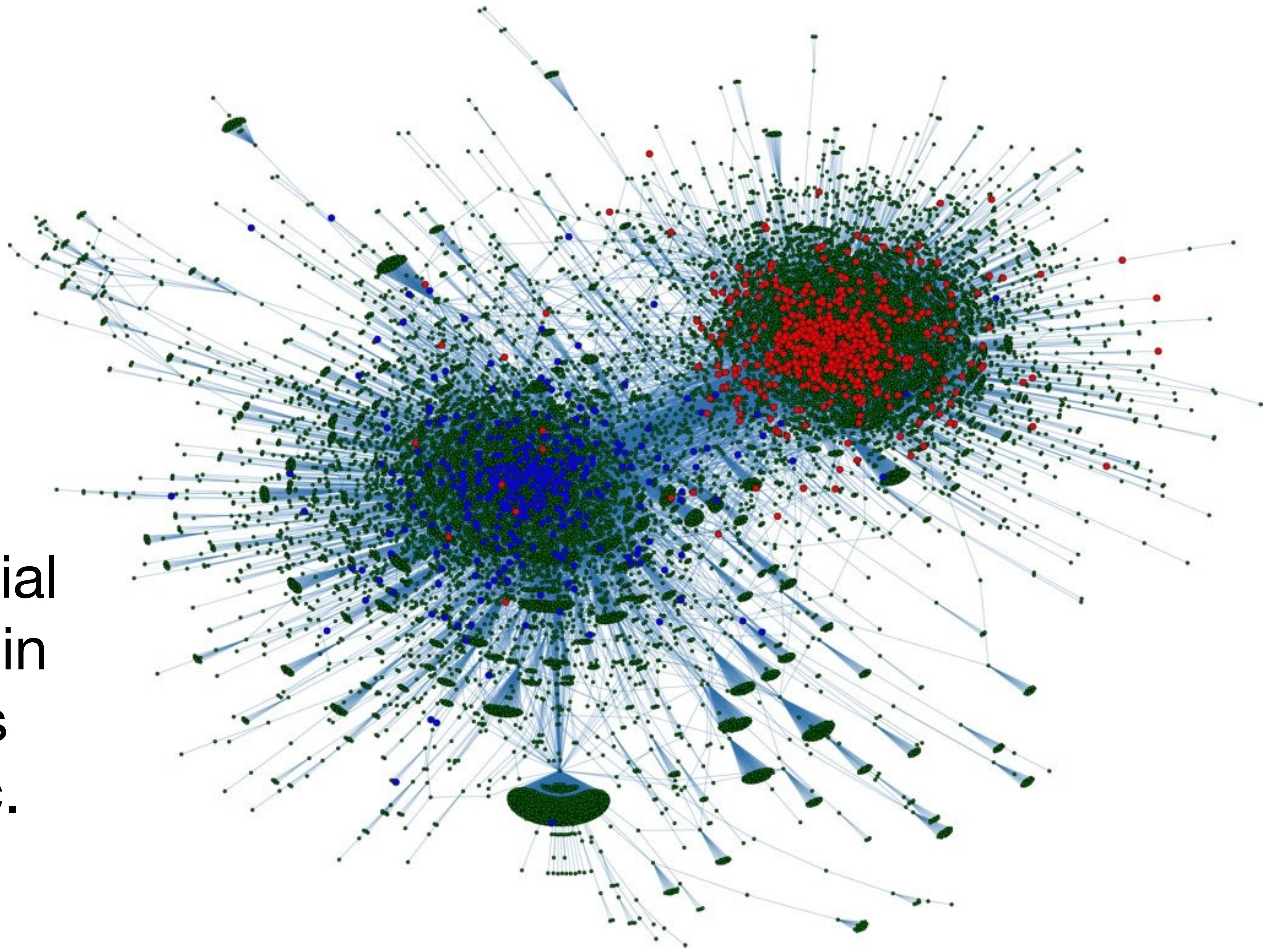
Community structure

Communities (or clusters): sets of tightly connected nodes



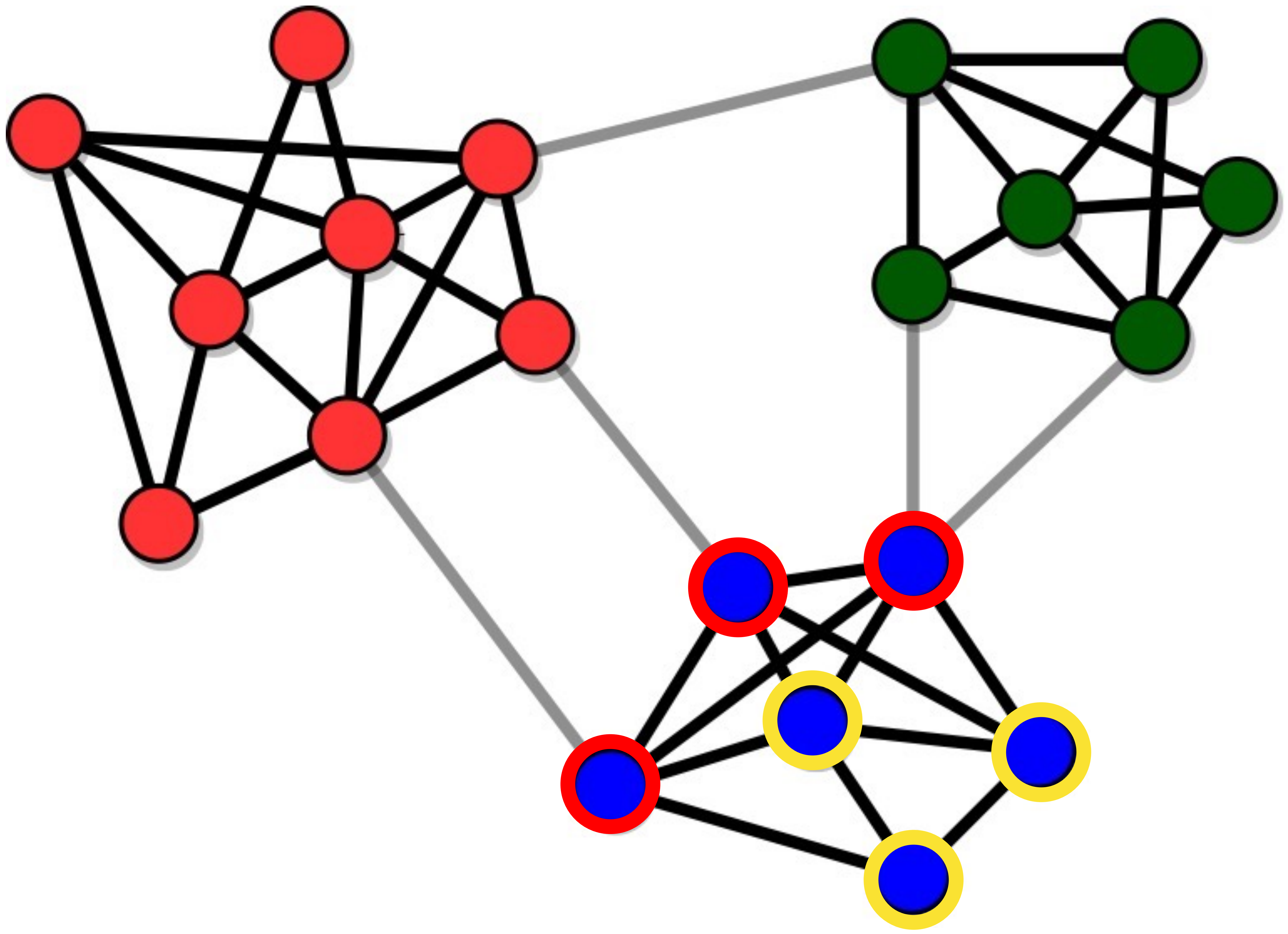
Community structure

- **Example:** Twitter users with strong political preferences tend to follow those aligned with them and not to follow users with different political orientation
- **Other examples:** social circles in social networks, functional modules in protein interaction networks, groups of pages about the same topic on the Web, etc.

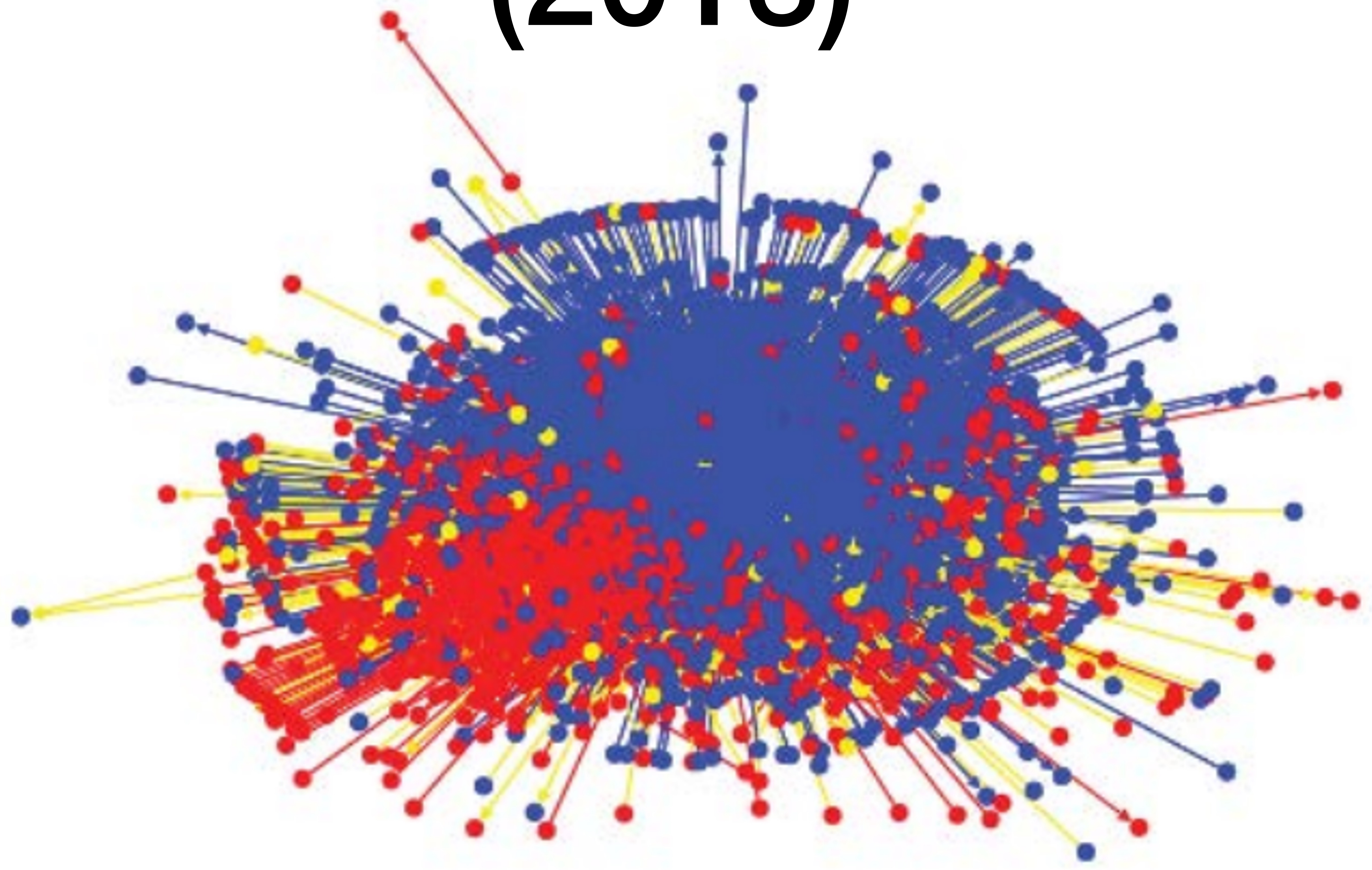


Why studying communities?

- Uncover the organization of the network
- Identify features of the nodes
- Classify the nodes based on their position in the clusters
- Find missing links



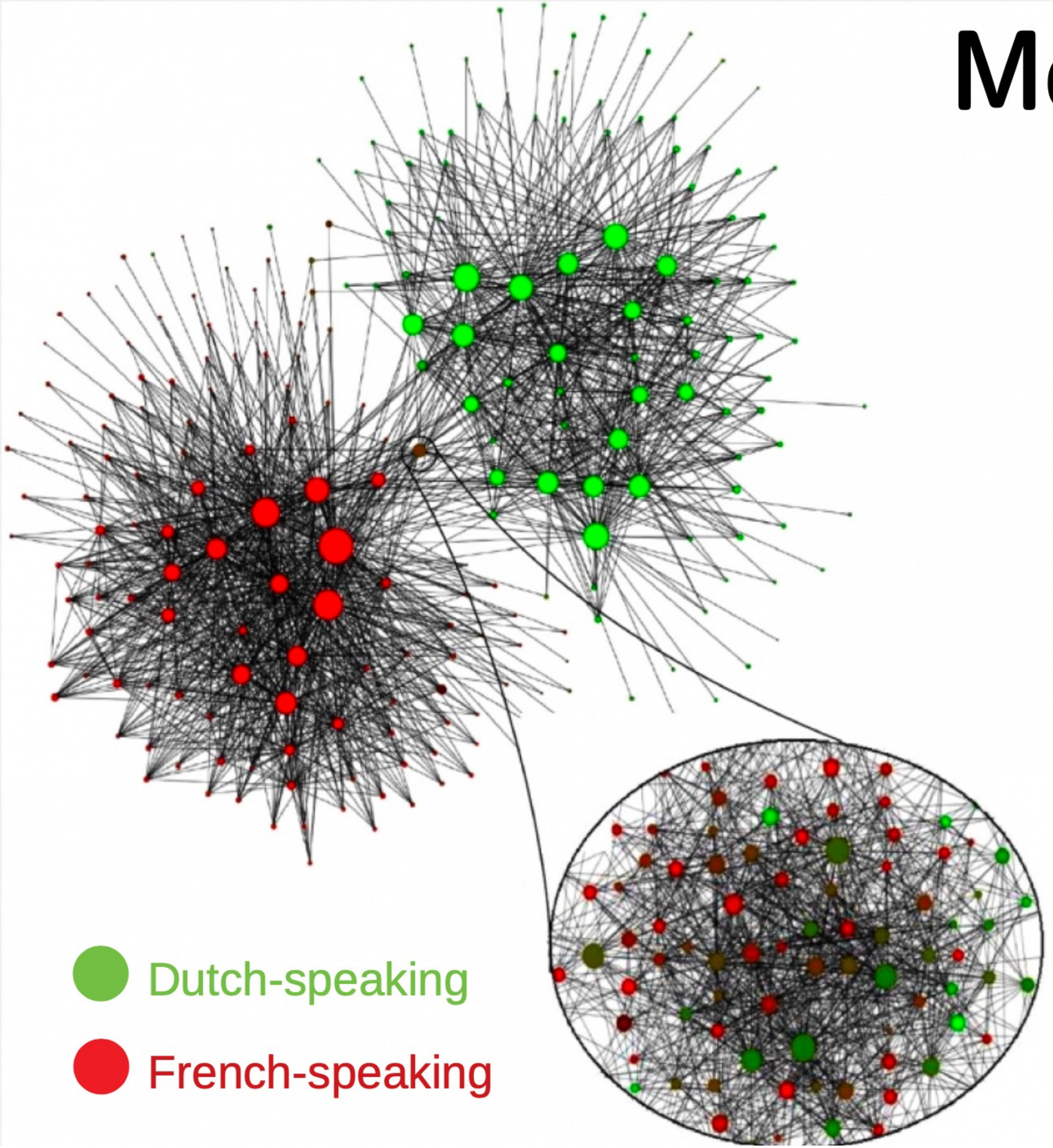
Egyptian Twitter Users (2013)



Islamist

Secularist

Mobile phone users in Belgium (2008)



Each node is a community of 100 mobile users or more that tend to call each other

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. J. Stat. Mech., 2008.

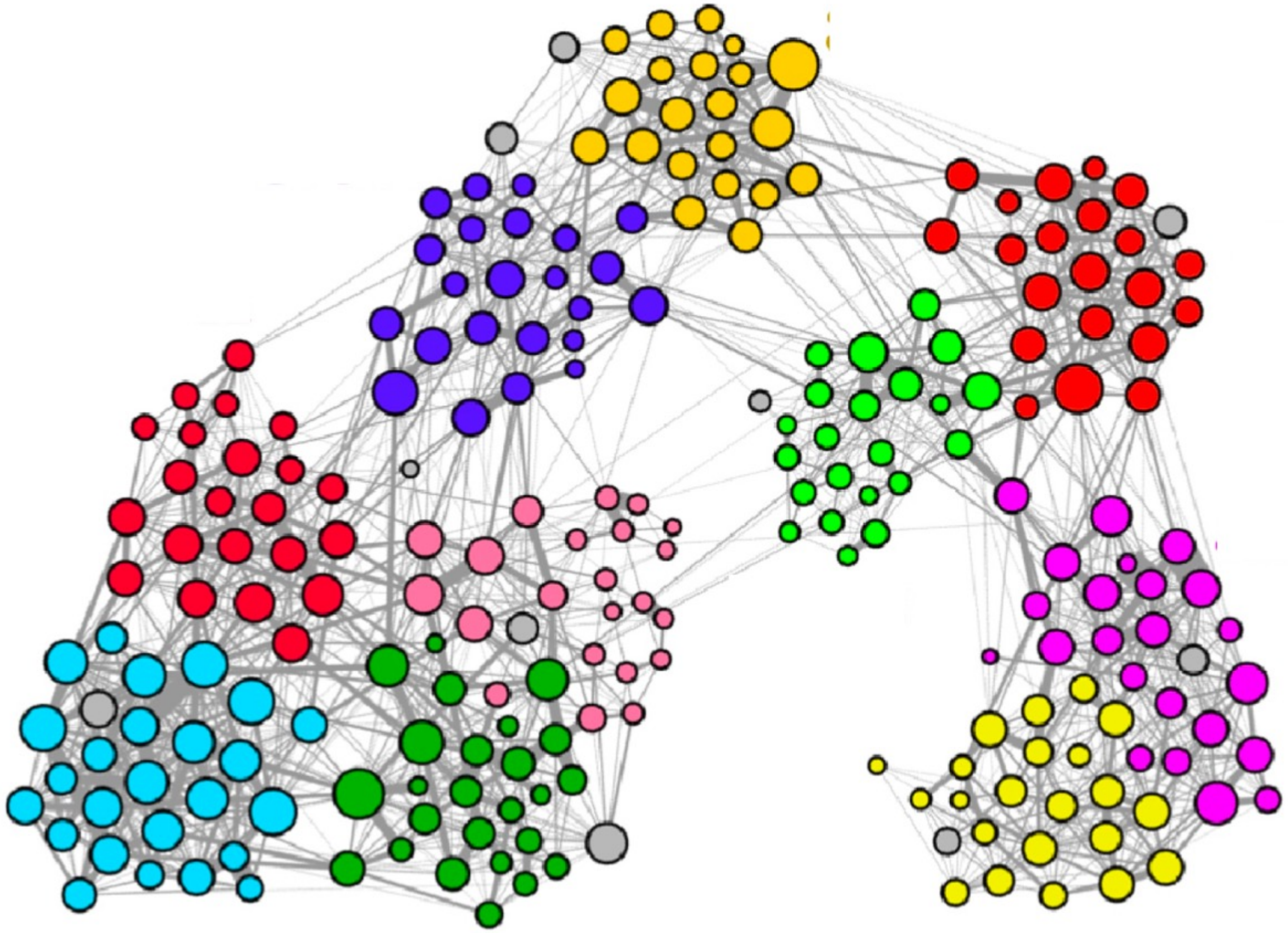
Political Books

Source:
Valdis Kreb
& The Economist

Primary school contacts

Links connect students who spent more than two minutes face to face

Students wore RF-ID badges hanging on their chest, which have a range of about 1.0-1.5 meters

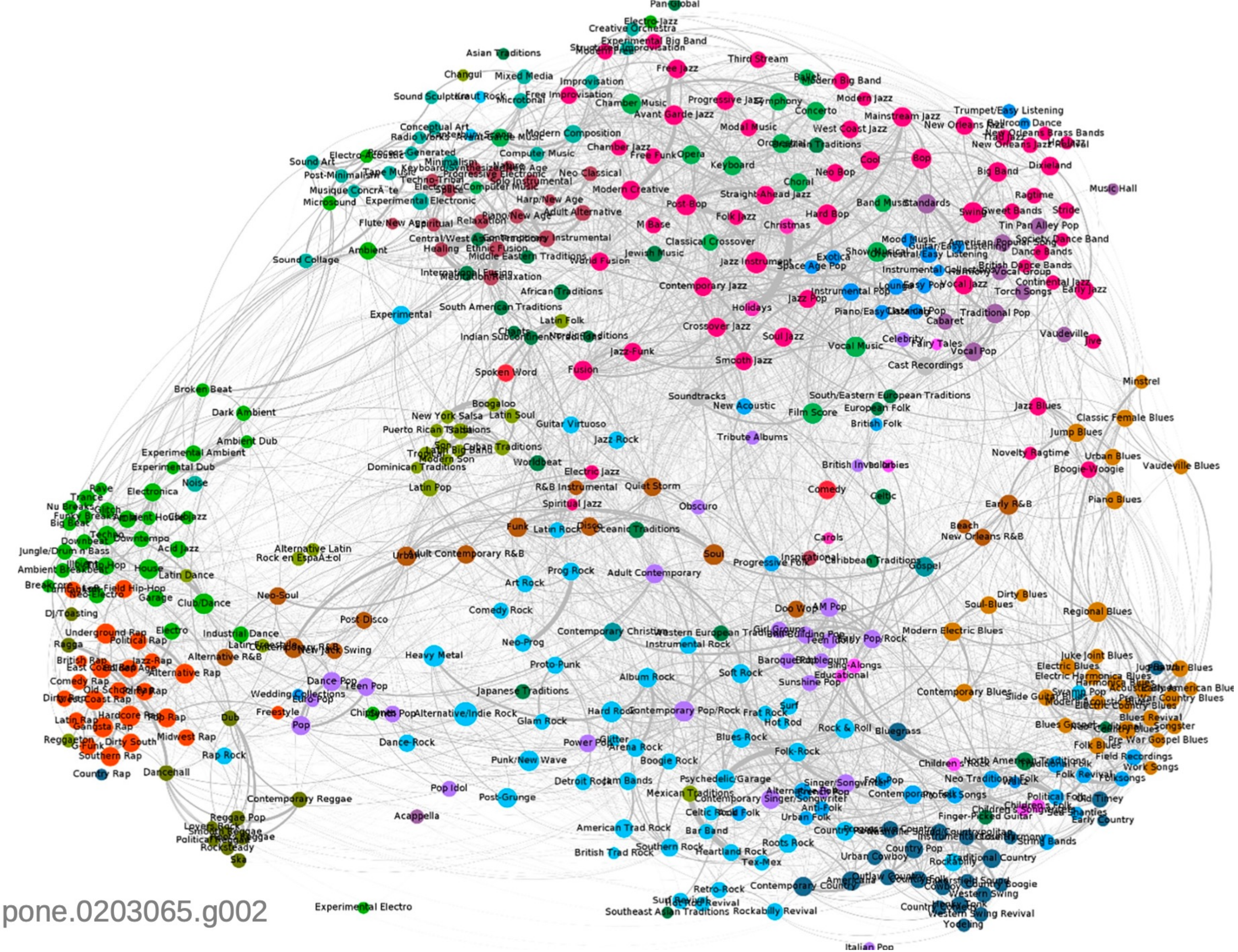


What do you think the colors represent in this visualization?

Stehlé, J., et al. (2011). High-resolution measurements of face-to-face contact patterns in a primary school. PloS one, 6(8), e23176.

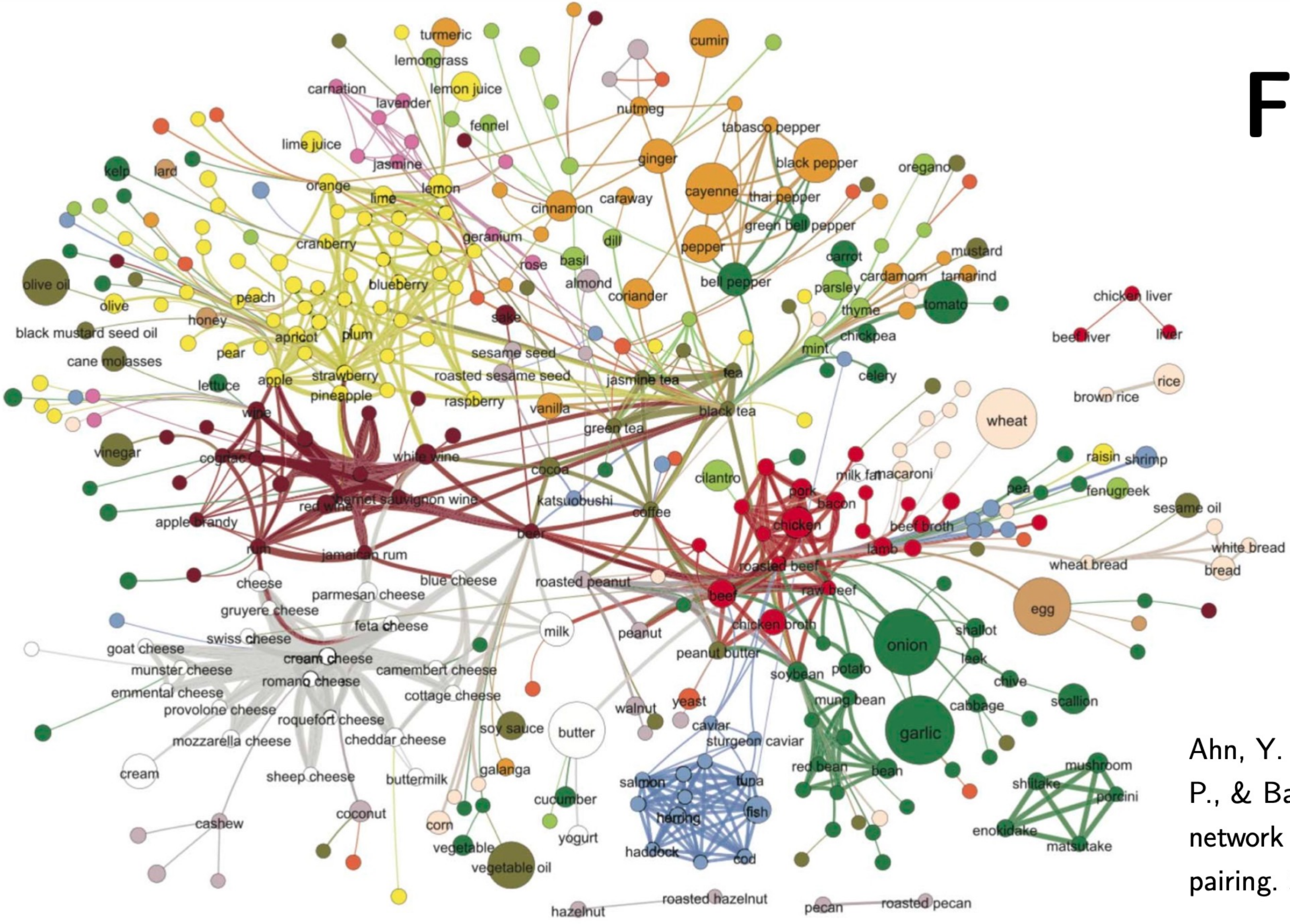
Music

Two Genres, G1, G2, are connected if there is a musician producing tracks in both genres; width of link is number of musicians



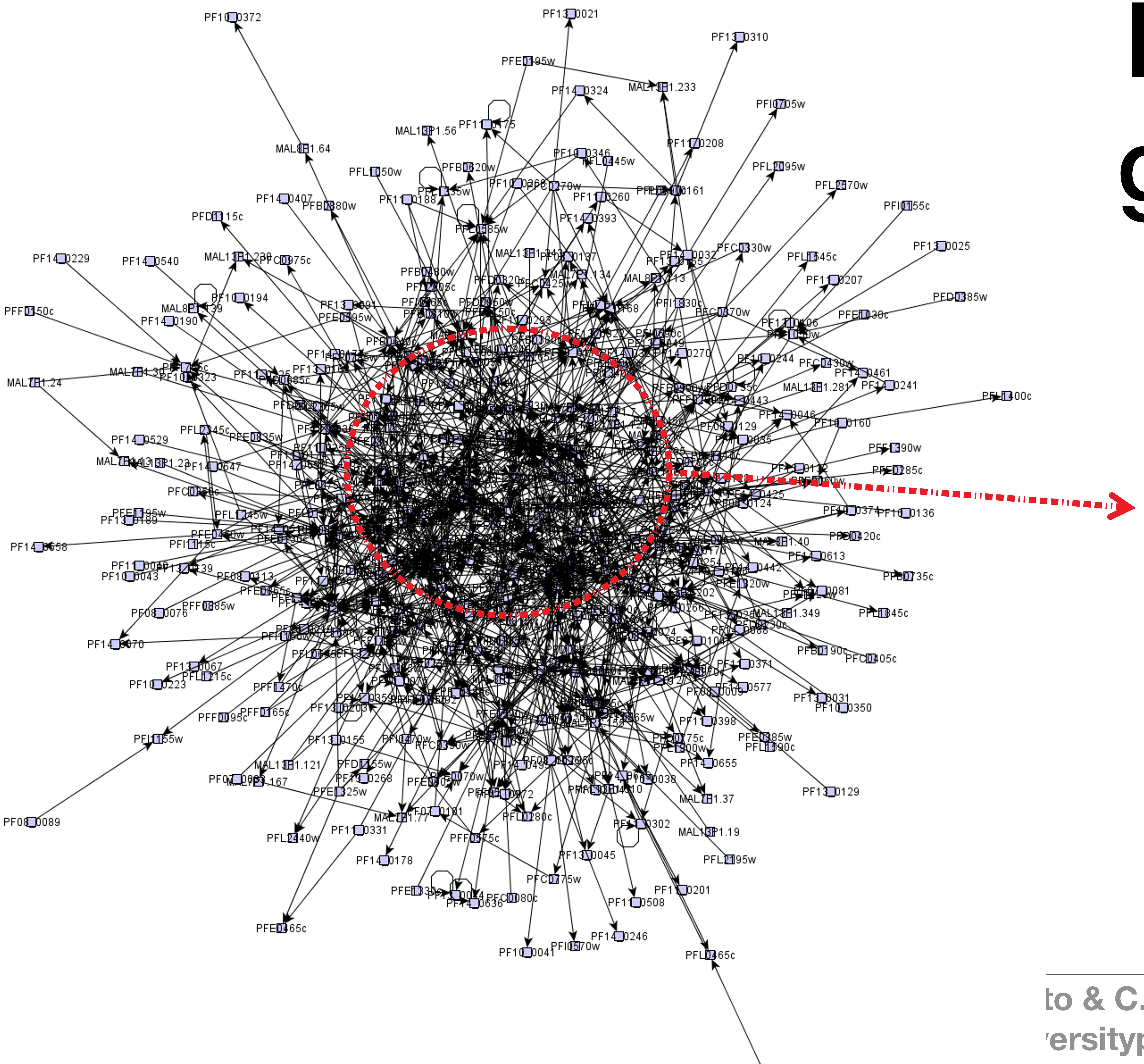
https://doi.org/10.1371/journal.pone.0203065.g002

Flavors



Ahn, Y. Y., Ahnert, S. E., Bagrow, J. P., & Barabási, A. L. (2011). Flavor network and the principles of food pairing. Scientific reports, 1, 196.

However, many graphs look like “hairballs”



Sometimes, at the center these graphs may have an interesting dense sub-graph

Dense subgraphs

- They may represent communities
- They can sustain the spreading of an epidemic
- They can represent coordinated inauthentic behavior in social networks
- They can be money-launders in financial networks
- They can represent functional modules in protein interaction networks

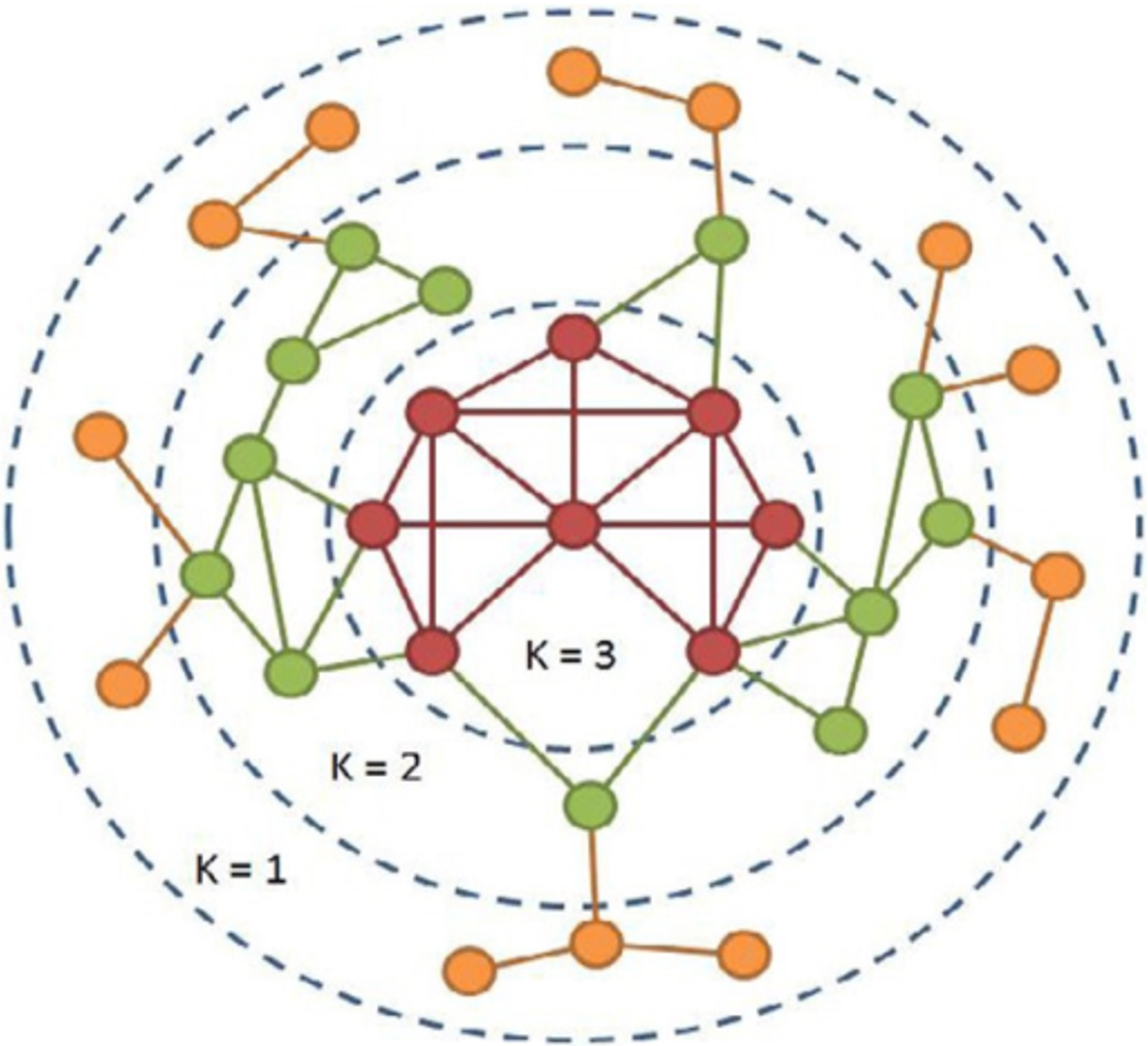
k-core decomposition
is a method to decompose a graph into *layers*

k-core decomposition

- Remove all nodes having degree ≤ 1 until there are no such nodes
  - Those are in the 1-core

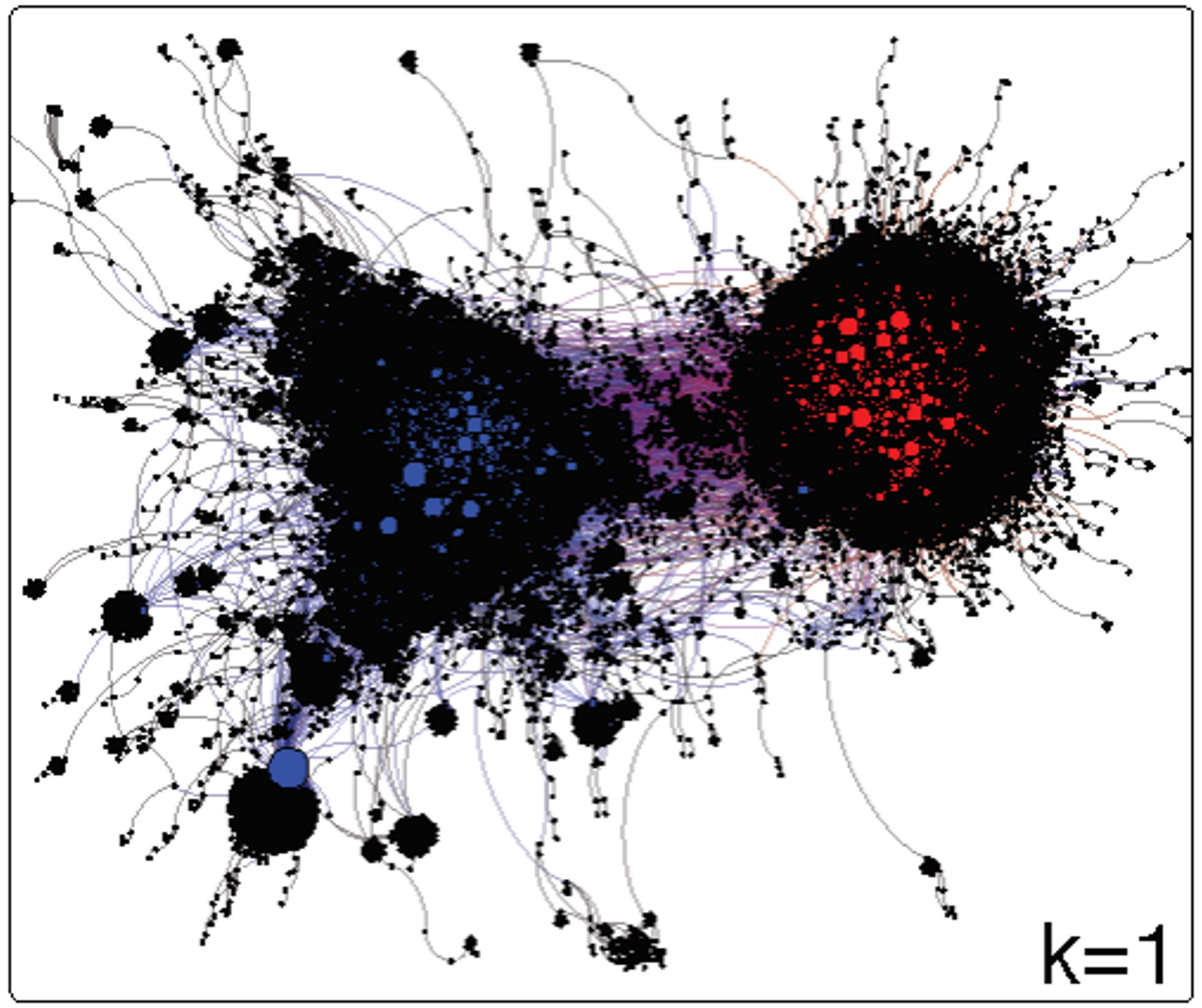
- Remove all nodes having degree ≤ 2 until there are no such nodes
  - Those nodes are in the 2-core
- Remove all nodes having degree ≤ 3 until there are no such nodes
  - Those nodes are in the 3-core
- Etc.

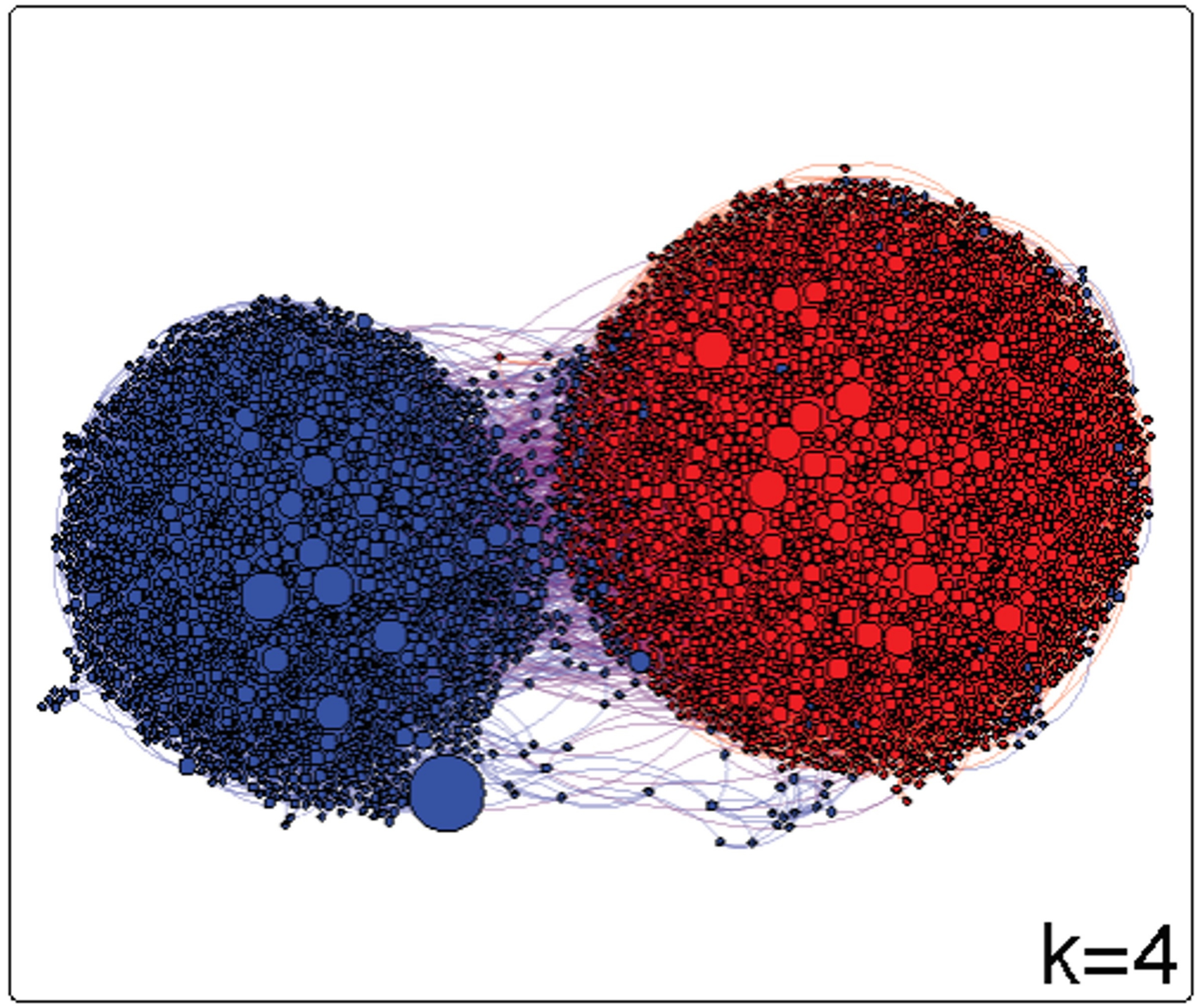
Example 1

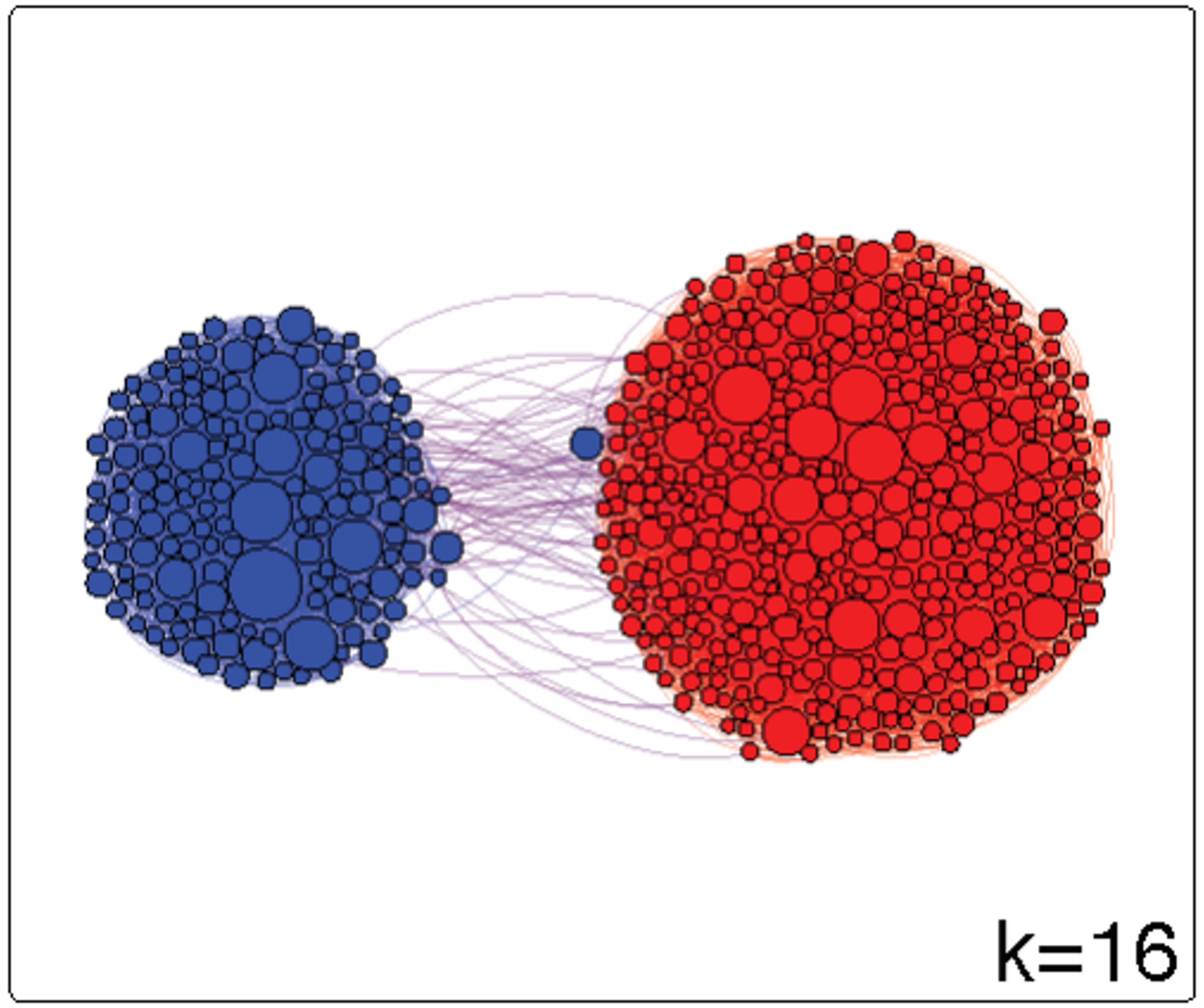


https://openi.nlm.nih.gov/detailedresult.php?img=3368241_fnagi-04-00013-g0001&req=4

Example 2

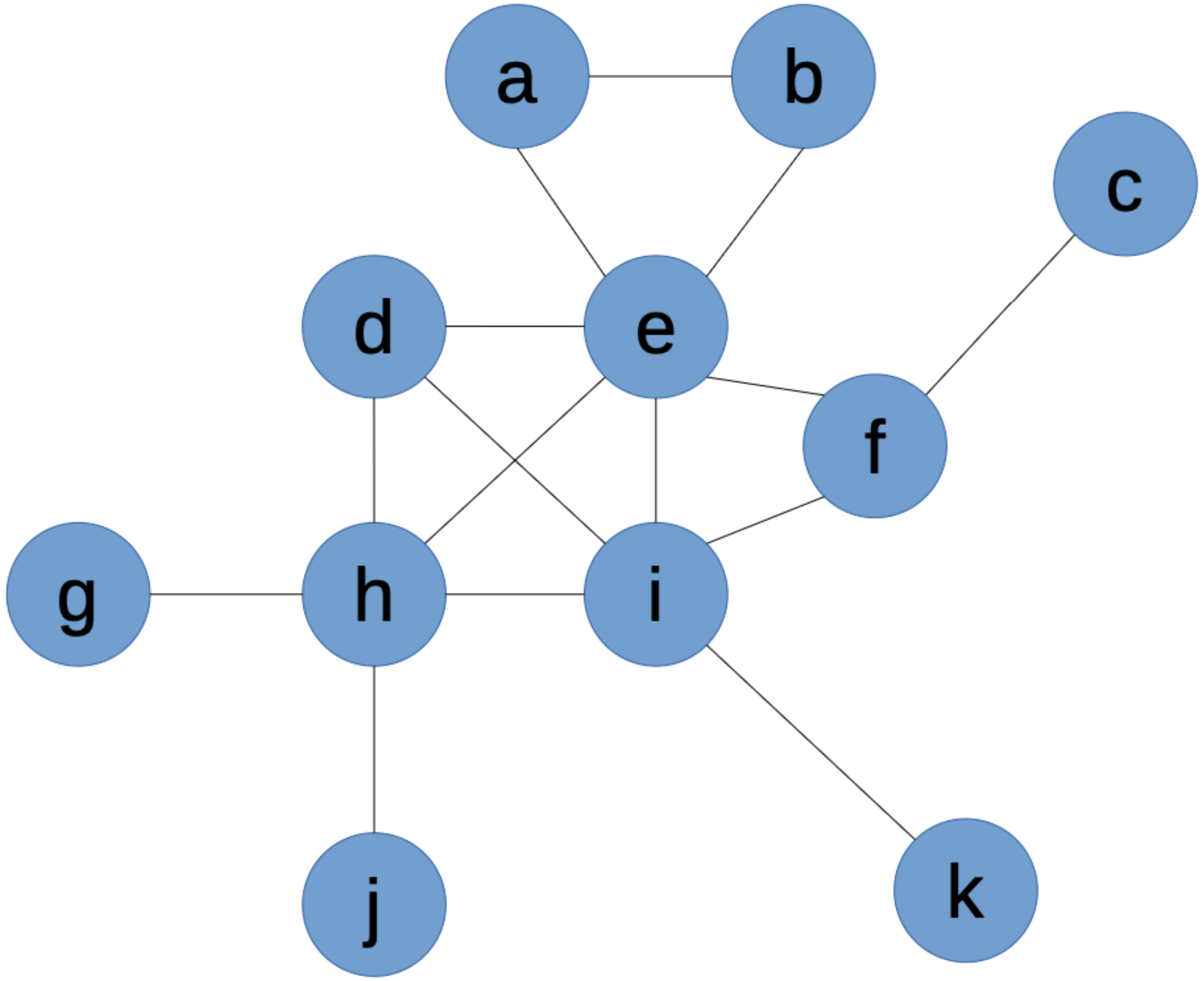






Source: Menczer, Fortunato, Davis: A First Course on Networks Science. Cambridge, 2020

Exercise



For each node in the graph, indicate the max k-core to which it belongs

(Mark each node with a number, upload an image of the result.)

[Source]

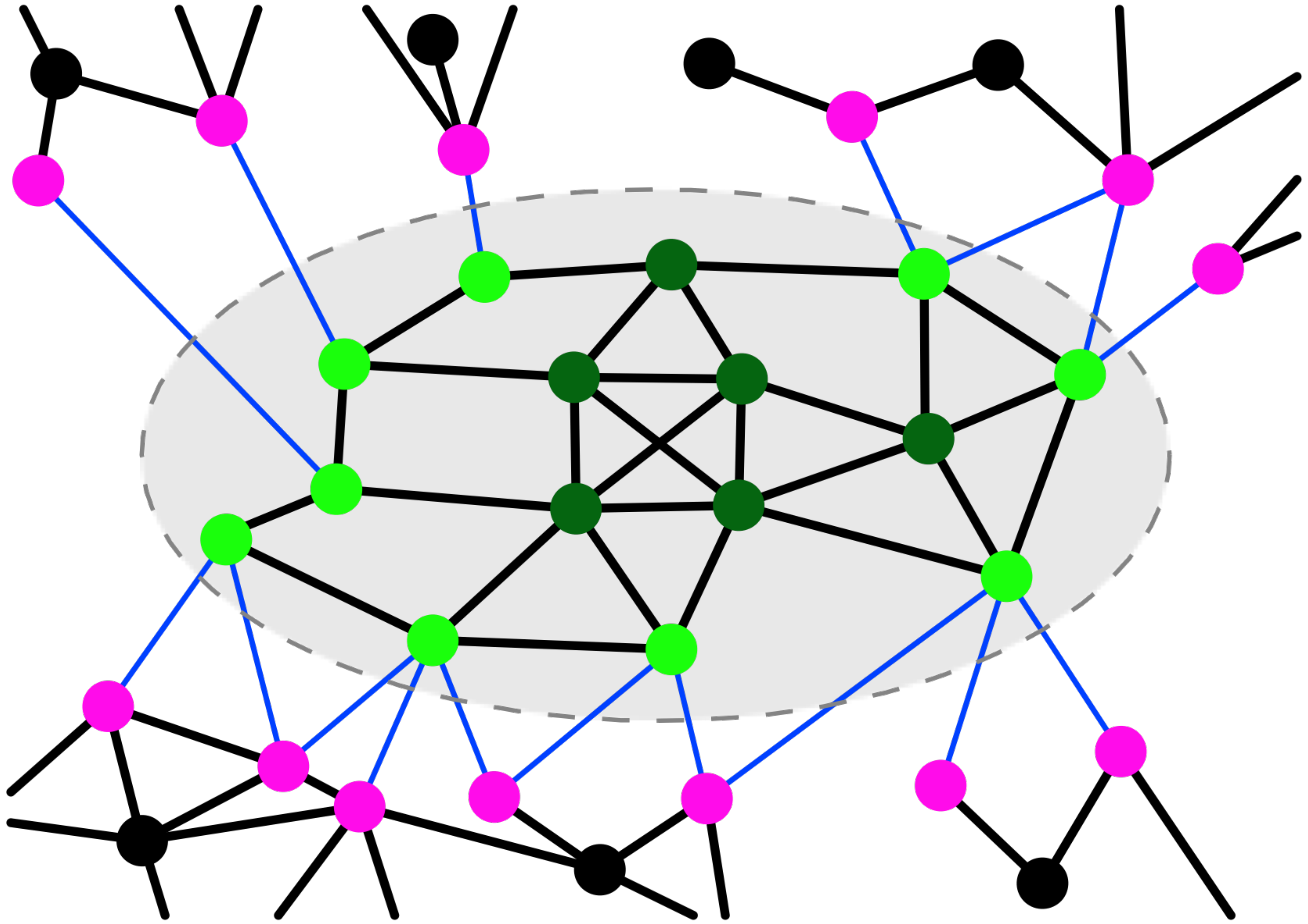
Pin board: https://upfbarcelona.padlet.org/chato/tt19-k-core-decomposition-575mnnke44c5ybyl

Basic definitions: variables

- **Internal degree of a node**: number of neighbors of the node in its community
- **External degree of a node**: number of neighbors of the node outside of its community
- **Community degree**: sum of the degrees of the nodes in the community
- **Internal link density**: ratio between the number of links L_C inside a community *C* and the maximal possible number of links that can lie inside *C:*

$$\delta_C^{int} = \frac{L_C}{L_C^{max}} = \frac{L_C}{\binom{N_C}{2}} = \frac{2L_C}{N_C(N_C-1)}$$

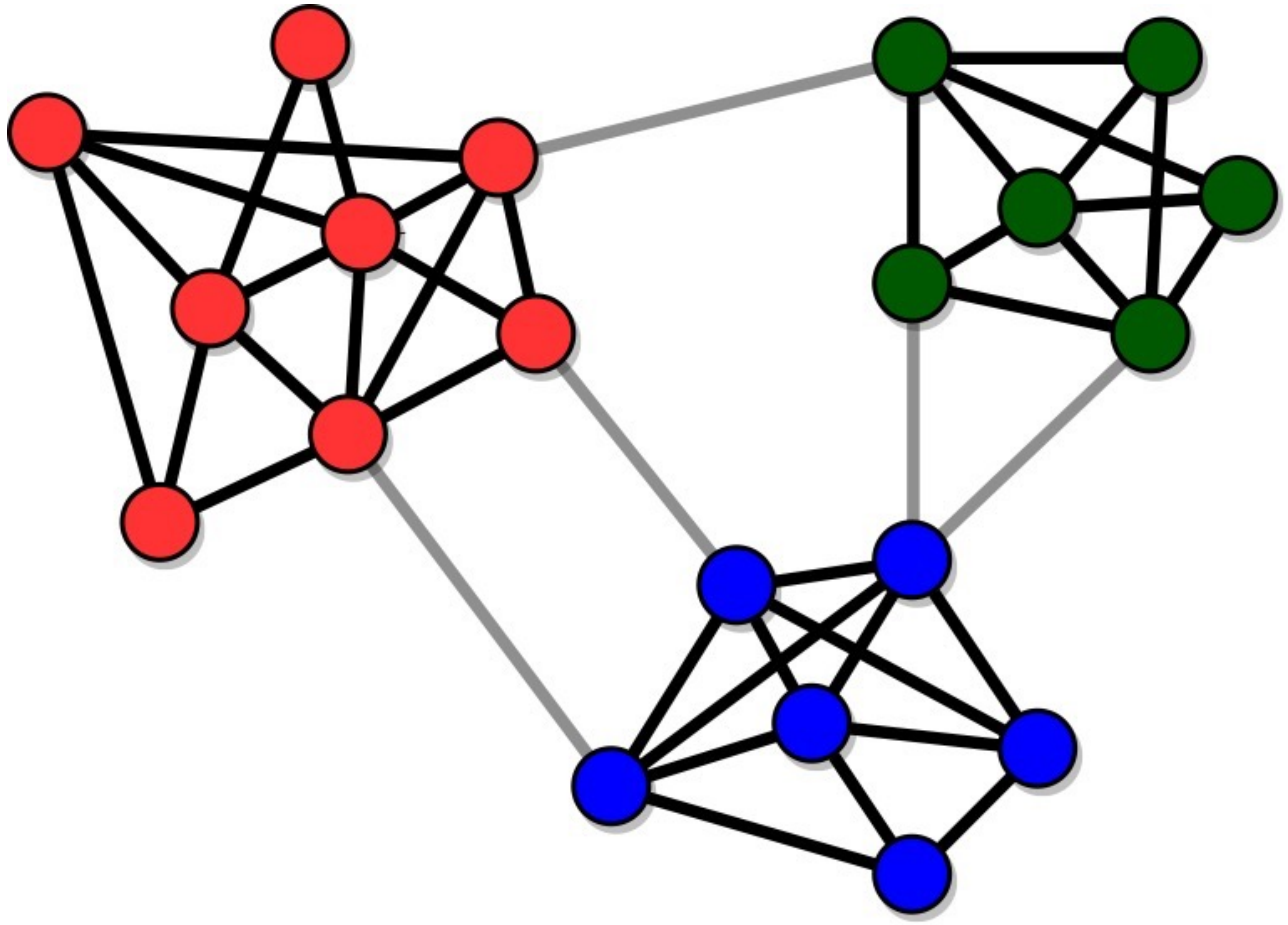
where N_C is the number of nodes in *C*



Basic definitions: community

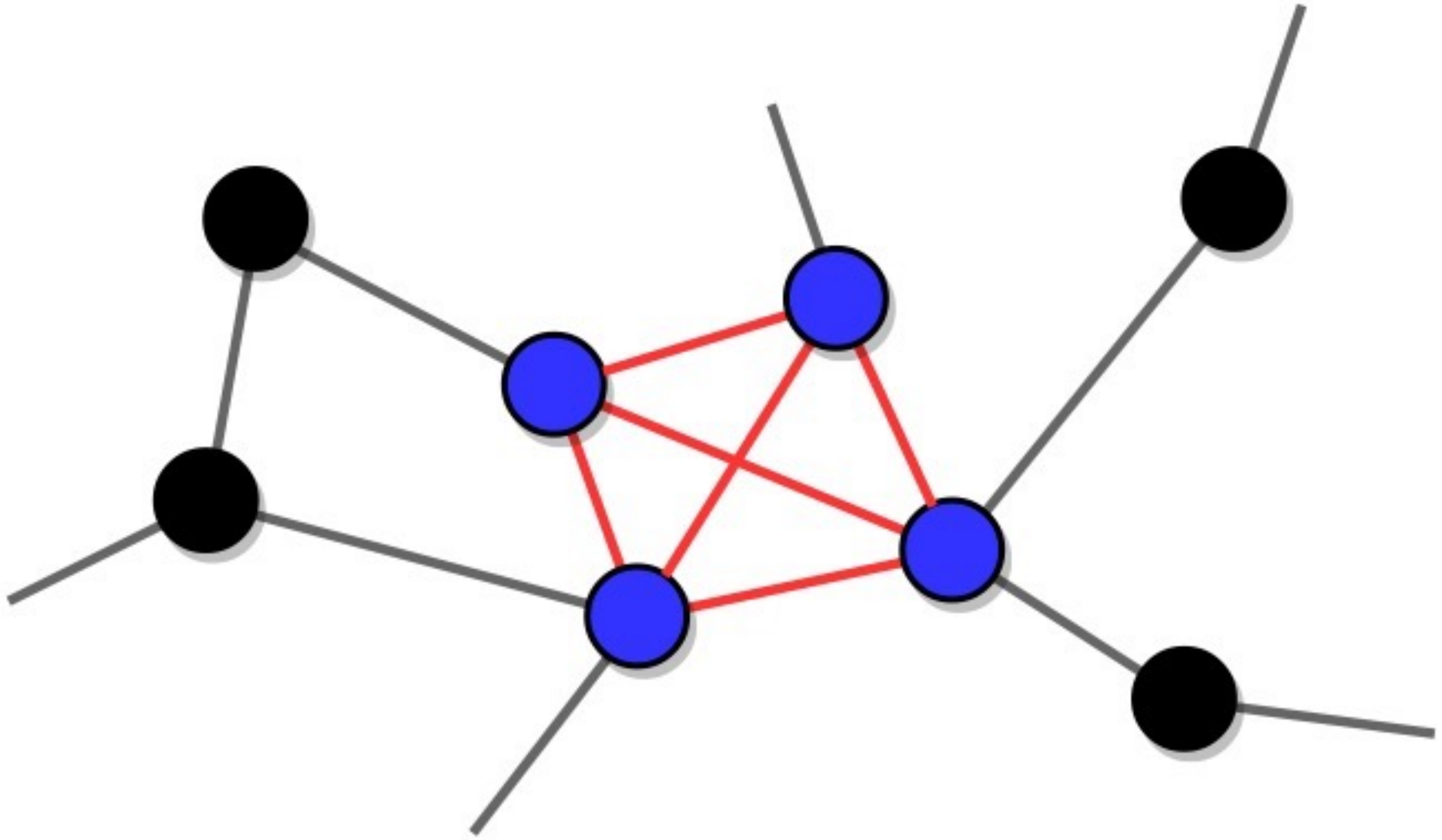
Two main features:

- **High cohesion:** communities have many internal links, so their nodes stick together
- **High separation:** communities are connected to each other by few links



Community definitions based on cohesion

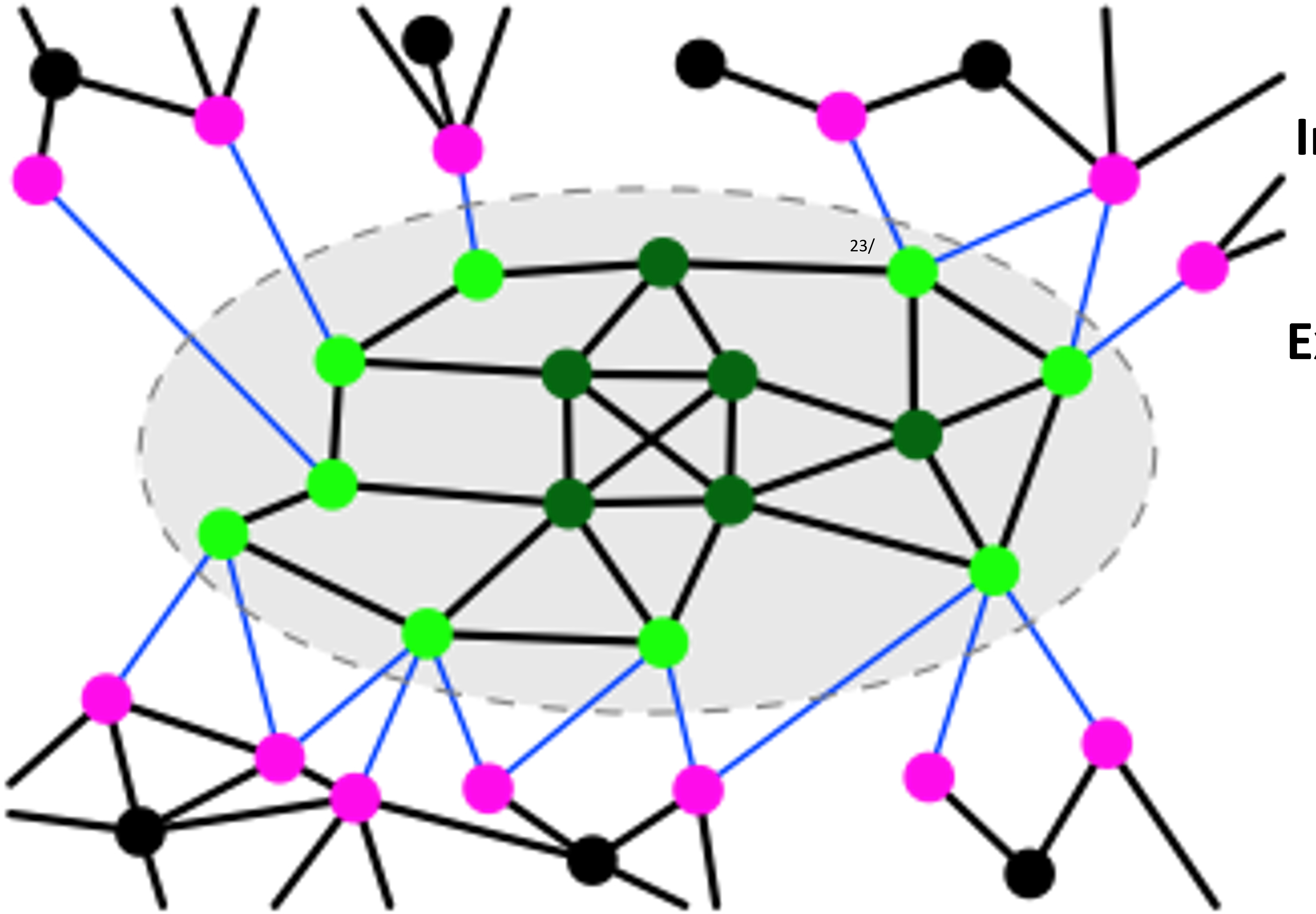
- **Principle:** focus on the cluster's properties, disregarding the rest of the network
- **Example: clique —** all internal links are there (maximal cohesion)
- **Problem:** nodes are connected to all others in the cluster, whereas in real communities they have different roles, which is reflected in heterogenous linking patterns



Community definitions based on cohesion versus separation

- **Principle:** definition tends to achieve high cohesion and high separation
- **Popular idea:** the number of internal links exceeds the number of external links
- Two concepts:
  - **Strong community:** subnetwork such that the internal degree of each node is greater than its external degree
  - **Weak community:** subnetwork such that the sum of the internal degrees of its nodes is greater than the sum of their external degrees

Communities: connected and dense



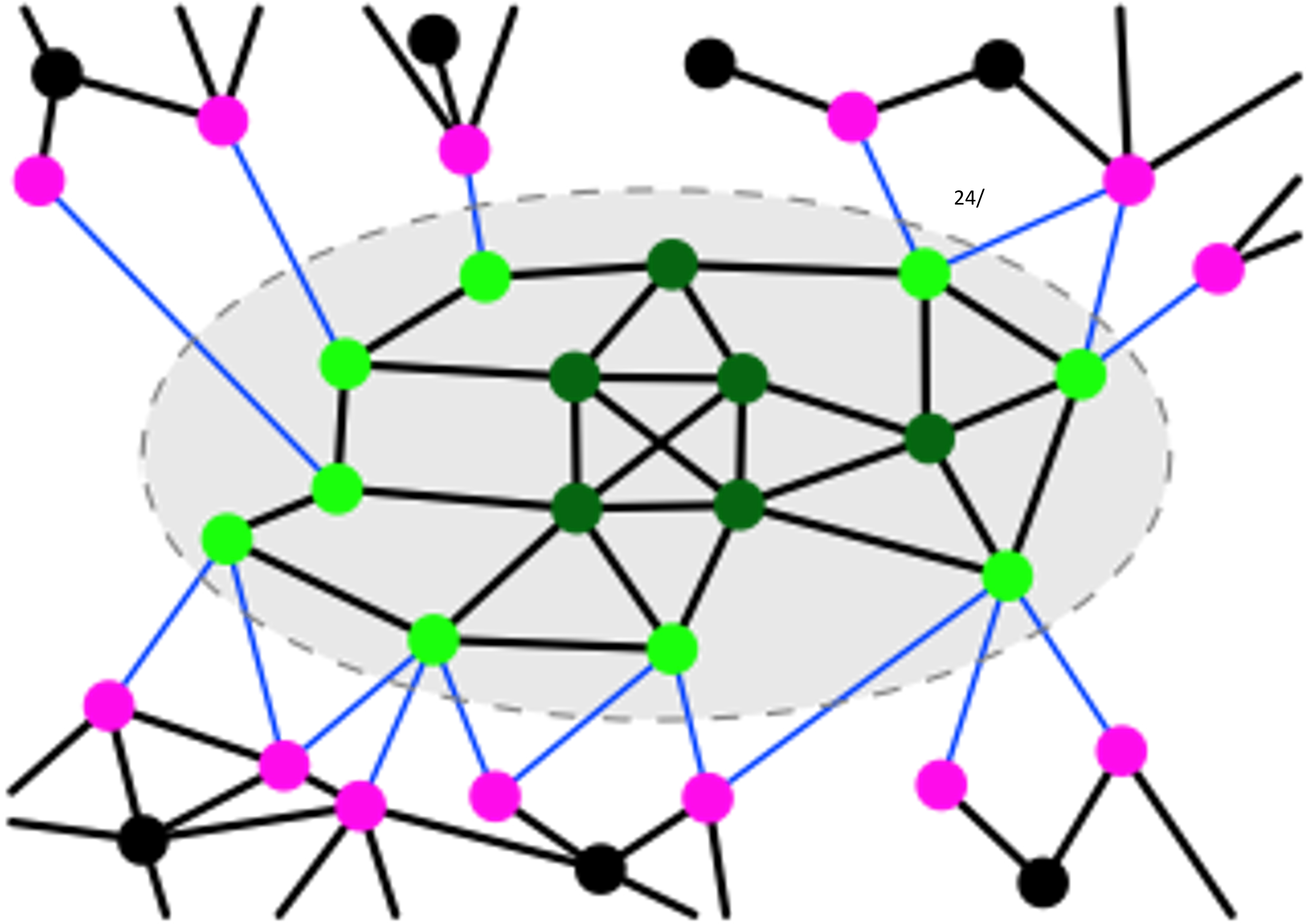
Given a community C

Internal degree $k^{int}(C)$ considers only nodes inside the community

External degree $k^{ext}(C)$ considers only nodes outside the community

$$k_i = k_i^{\text{int}}(C) + k_i^{\text{ext}}(C)$$

Strong community

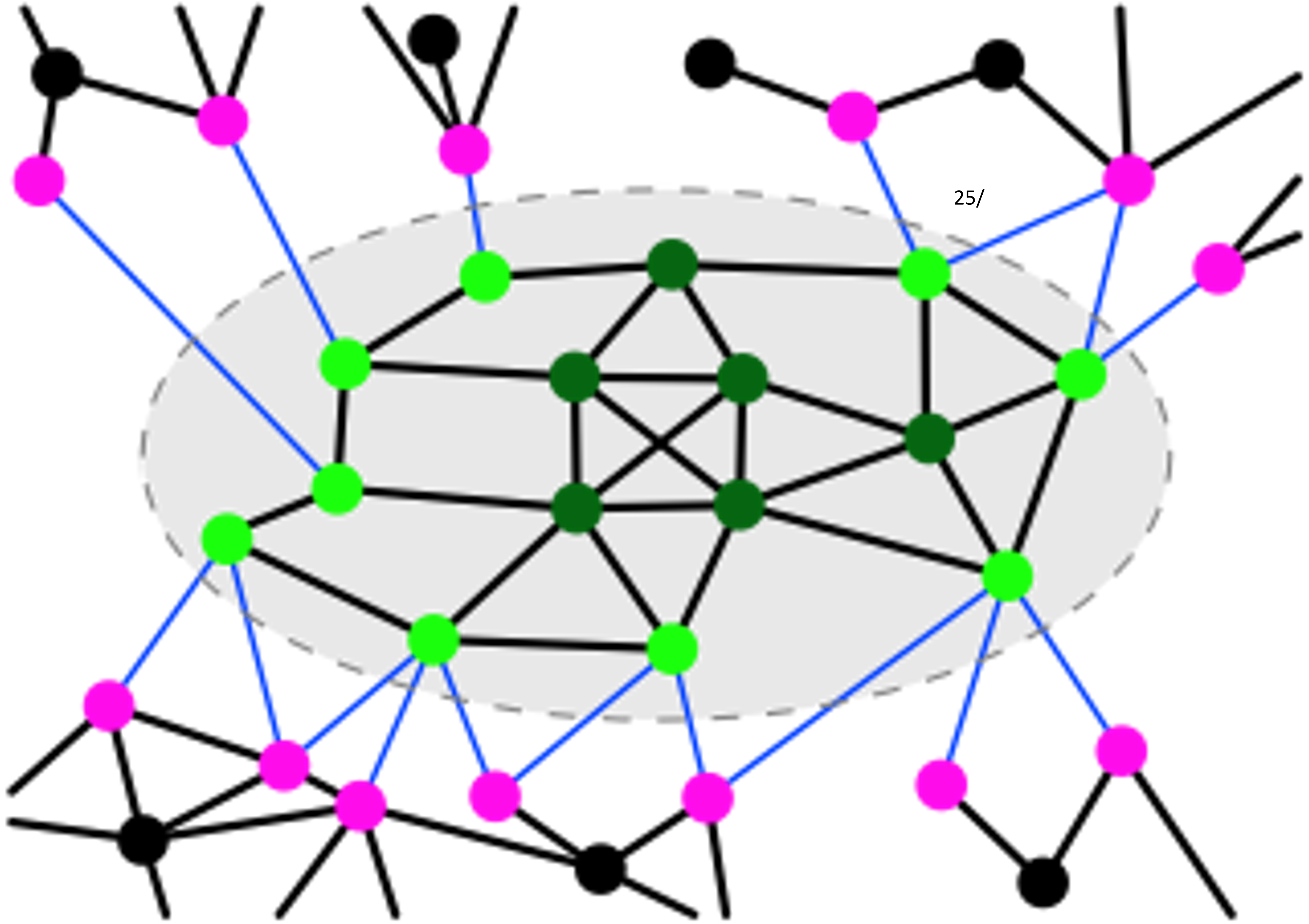


A community C is **strong** if **every node *i*** within the community satisfies:

$$k_i^{\text{int}}(C) > k_i^{\text{ext}}(C)$$

- Is the community of green nodes (dark green and light green) a strong community?
- What is the difference between dark green and light green nodes?

Weak community



A community C is **weak** if **on aggregate** nodes satisfy:

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

- All communities satisfying the strong property satisfy the weak one

Exercise

Is community A strong, weak, both?

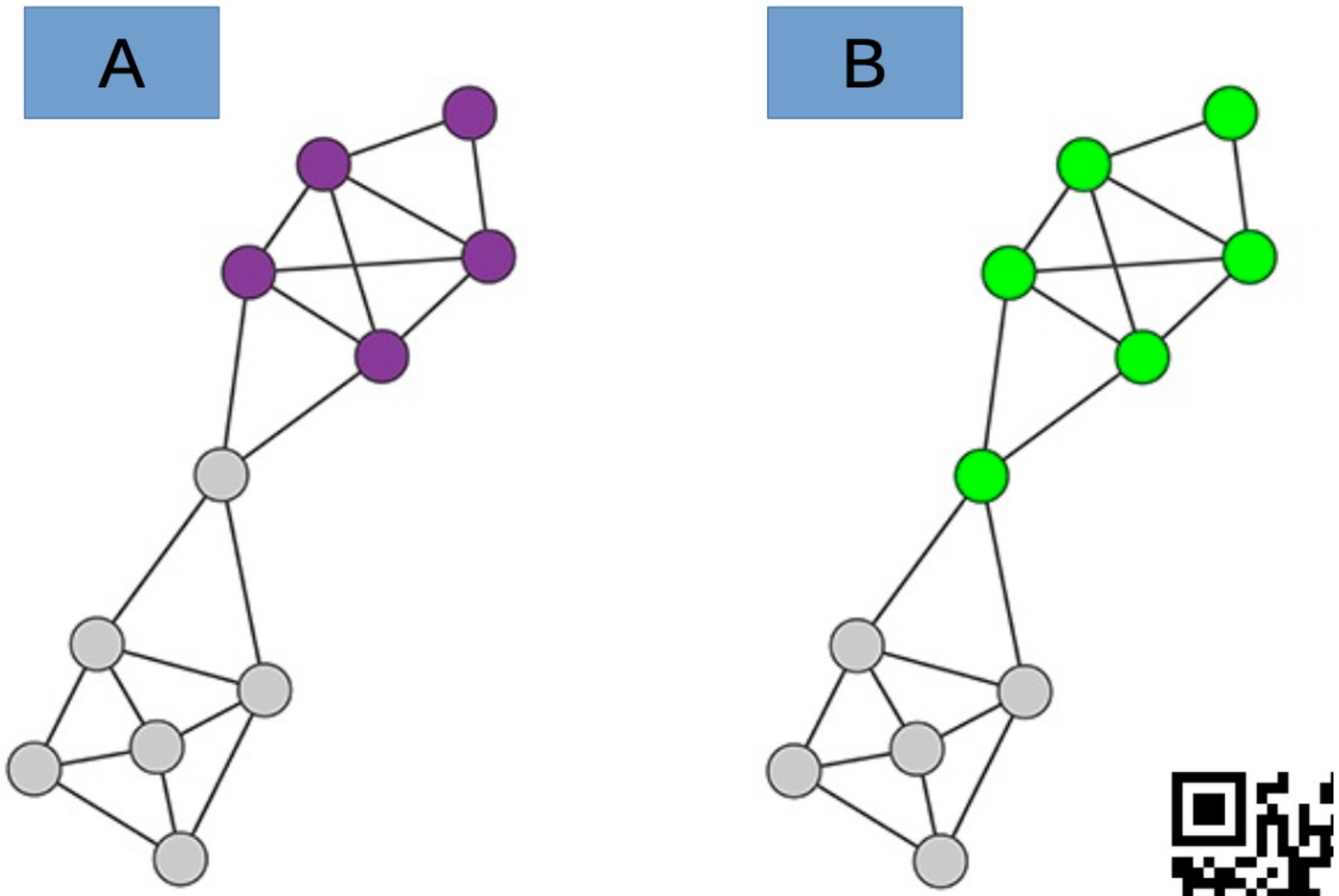
Is community B strong, weak, both?

A community C is **strong** if,
for all nodes *i* within the community:

$$k_i^{\text{int}}(C) > k_i^{\text{ext}}(C)$$

A community C is **weak** if:

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



Pin board: https://upfbarcelona.padlet.org/chato/tuxhfgcan59orweh

Community definitions based on cohesion versus separation

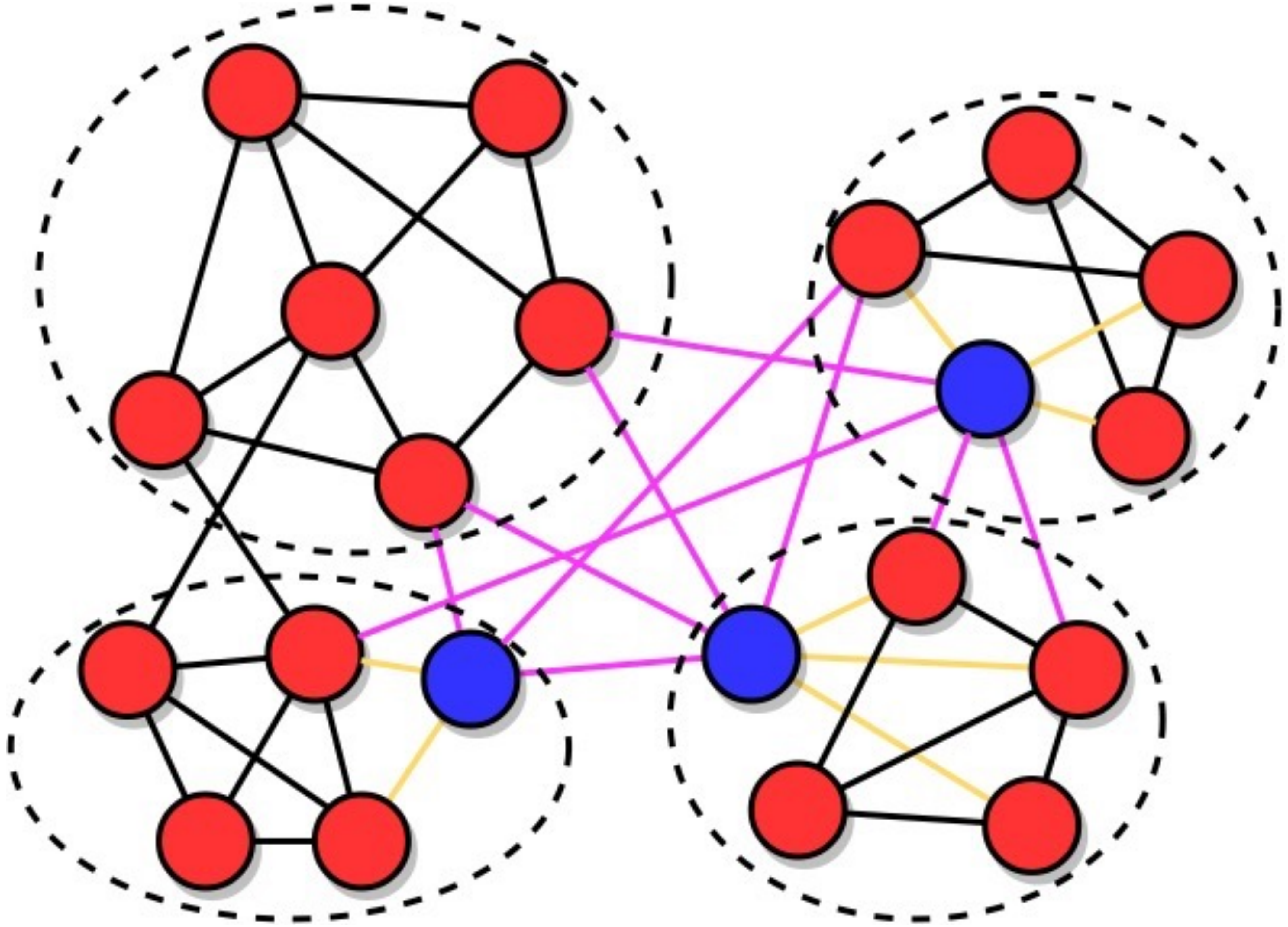
- **A strong community is also a weak community:** if the inequality between internal and external degree holds for each node, then it must hold for the sum over all nodes

- **A weak community is not a strong community, in general:** if the inequality between internal and external degree holds for the sum, it may be violated for one or more nodes

- **Problem:** in the definition of strong and weak community one compares a subnetwork with the rest of the network. It makes more sense to **compare subnetworks to each other!**

Community definitions based on cohesion versus separation

- **Less stringent definitions** of strong and weak community:
  - **Strong community:** subnetwork such that each node has more neighbors inside it than in any other community
  - **Weak community:** subnetwork such that the sum of the internal degrees of its nodes exceeds the total number of neighbors that the nodes have in any other community

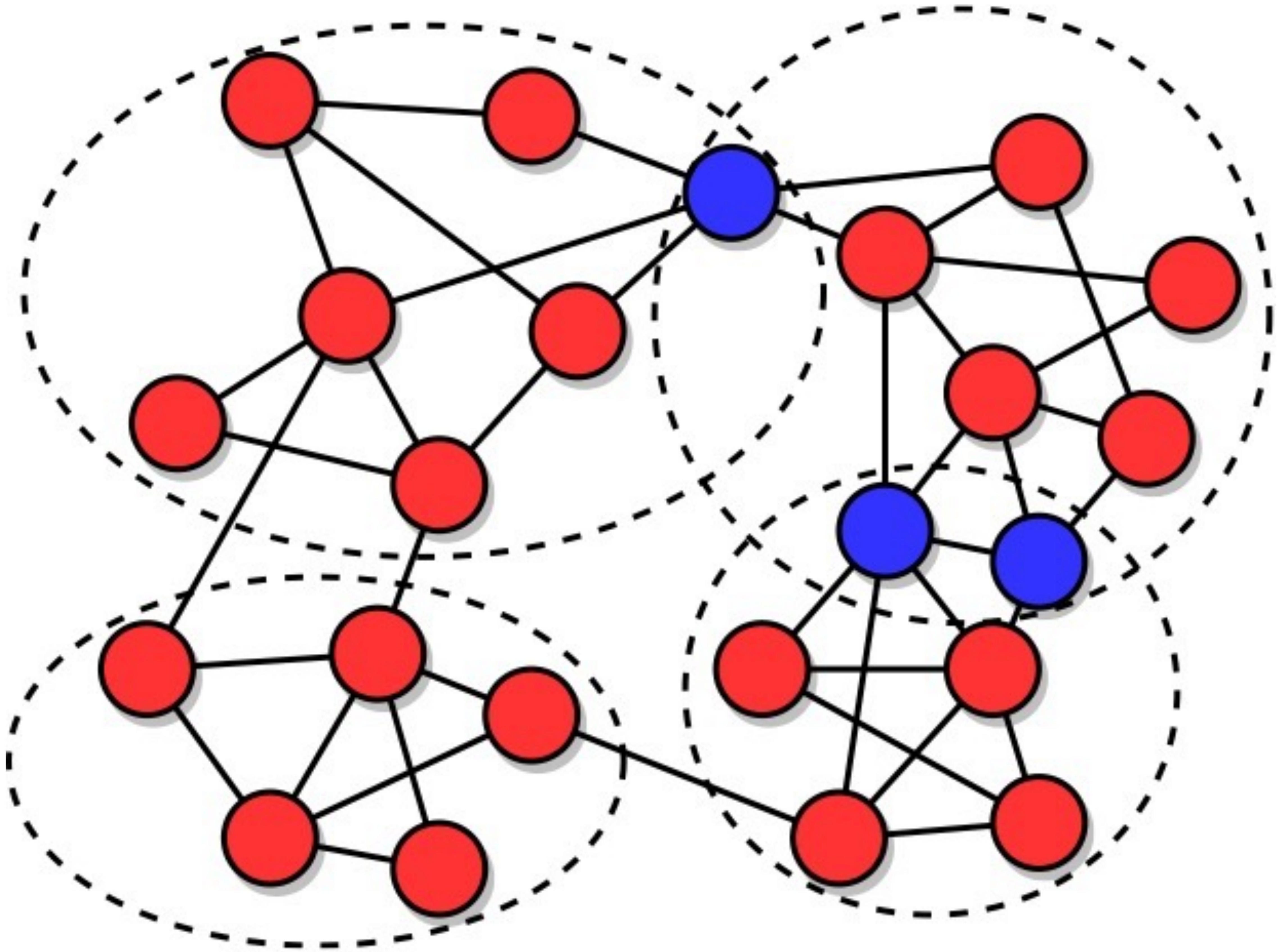


Overlapping communities

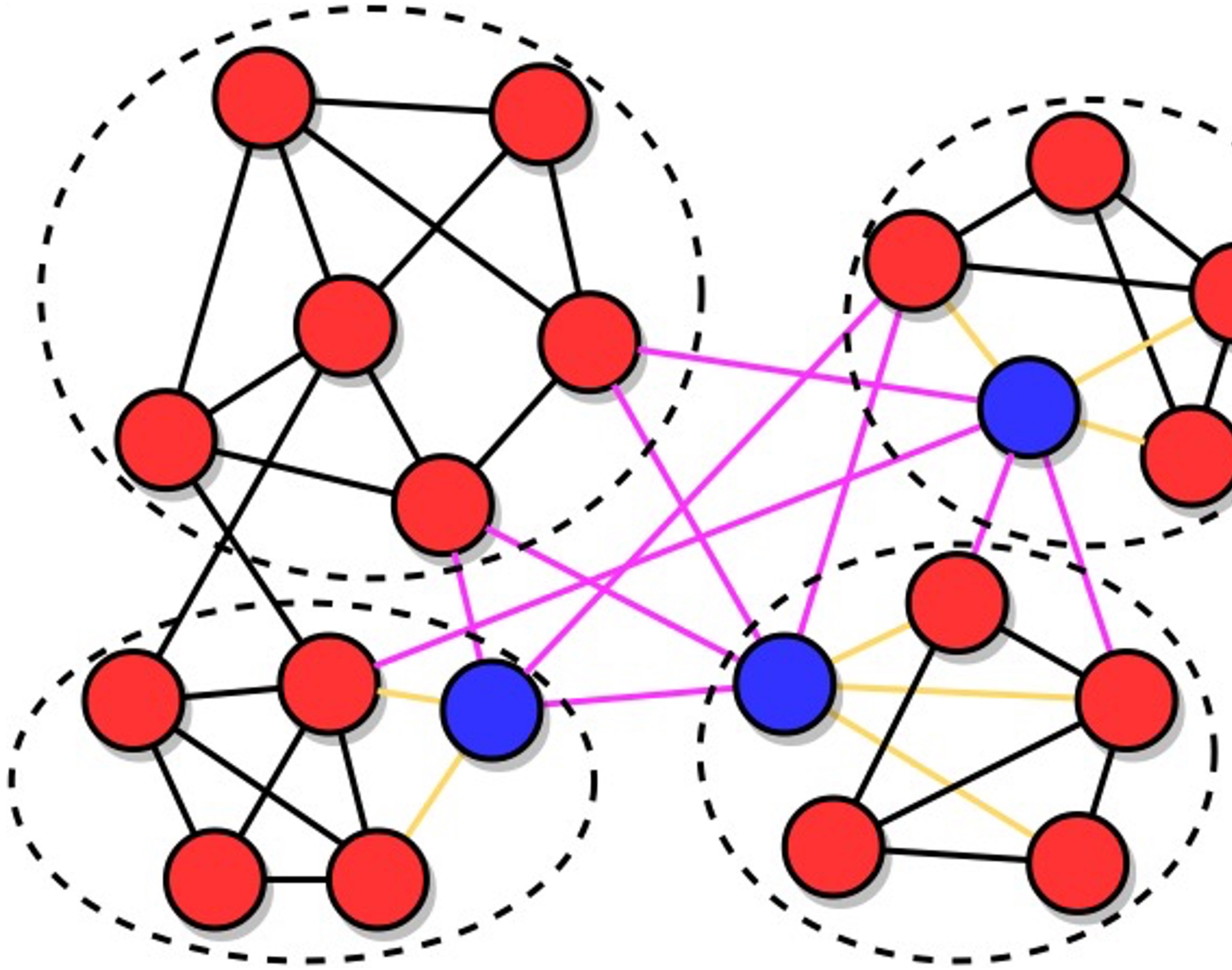
- Communities in many real networks **overlap**

- A division of a network into overlapping communities is called **cover**

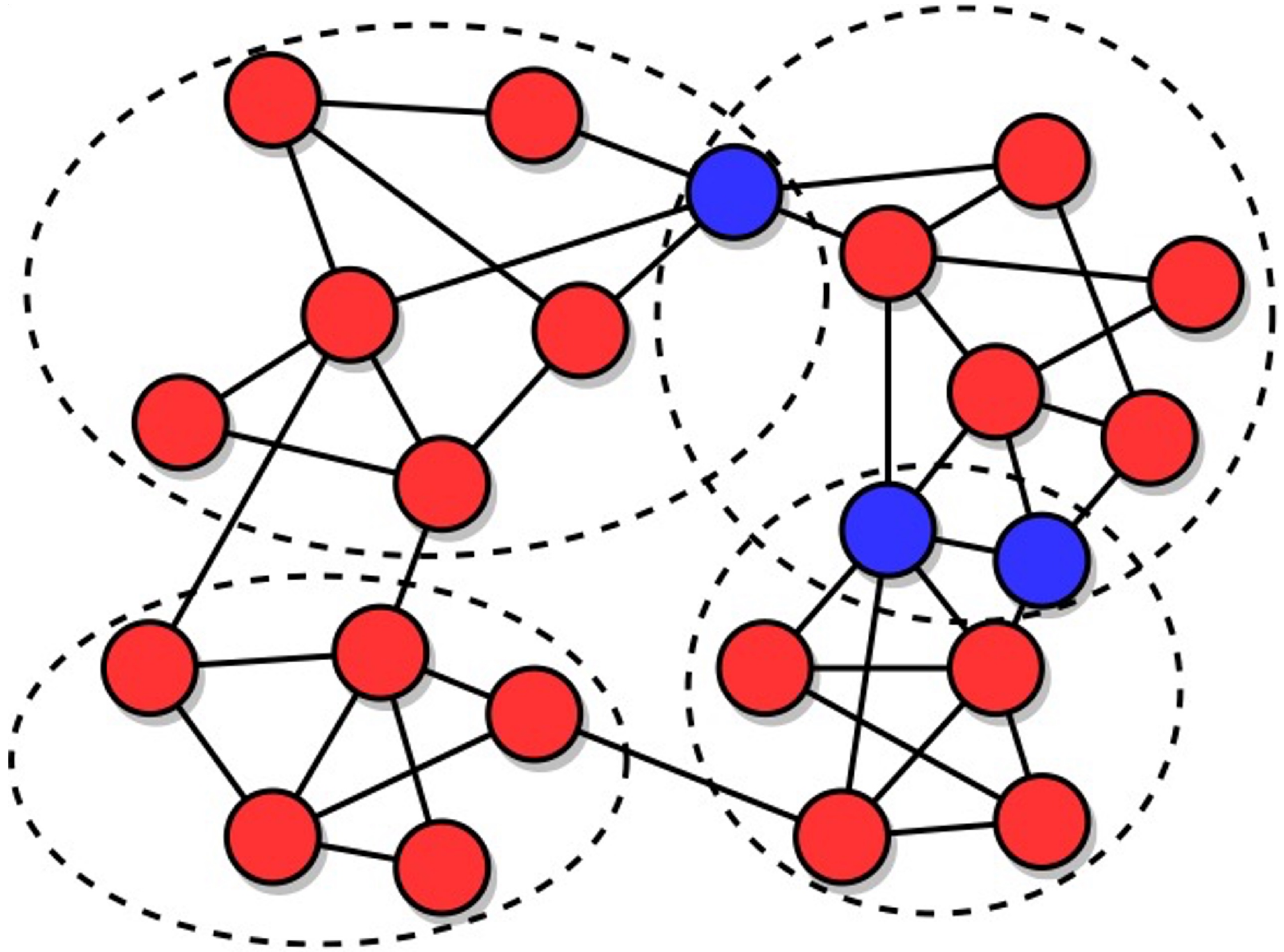
- The number of possible covers of a network is **far higher** than the number of partitions, due to the many ways clusters can overlap



Partition vs Overlapping communities



Partition, or *hard* clusters

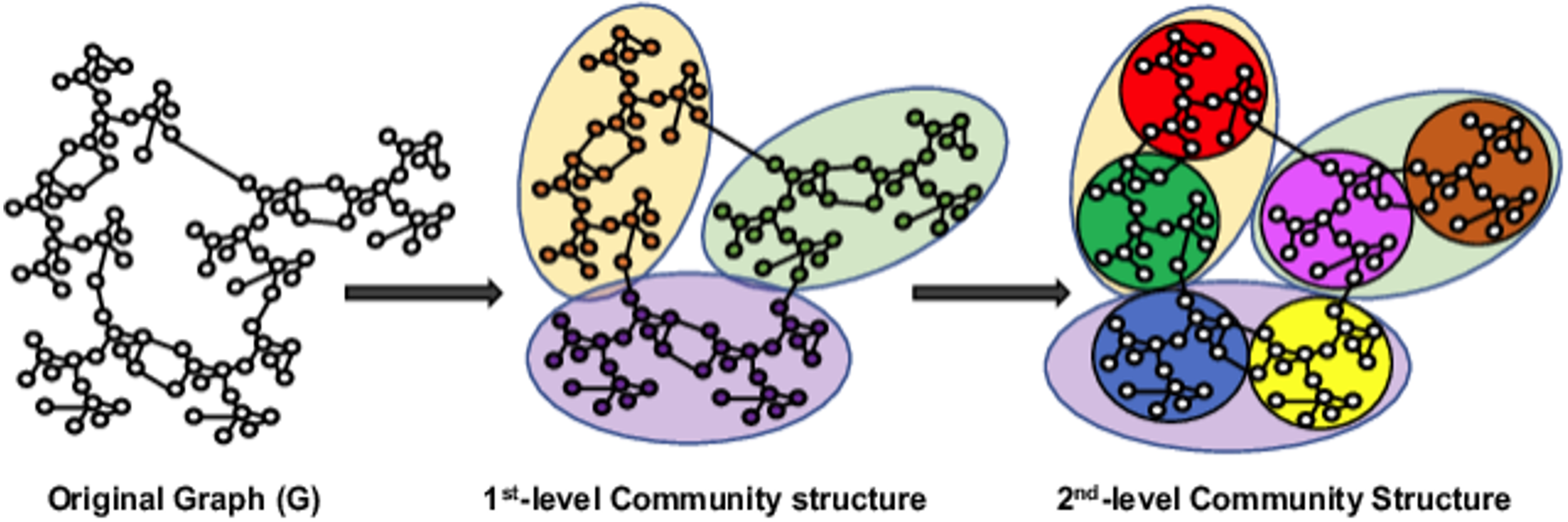


Overlapping communities, or *soft* clusters

What's special about blue nodes?

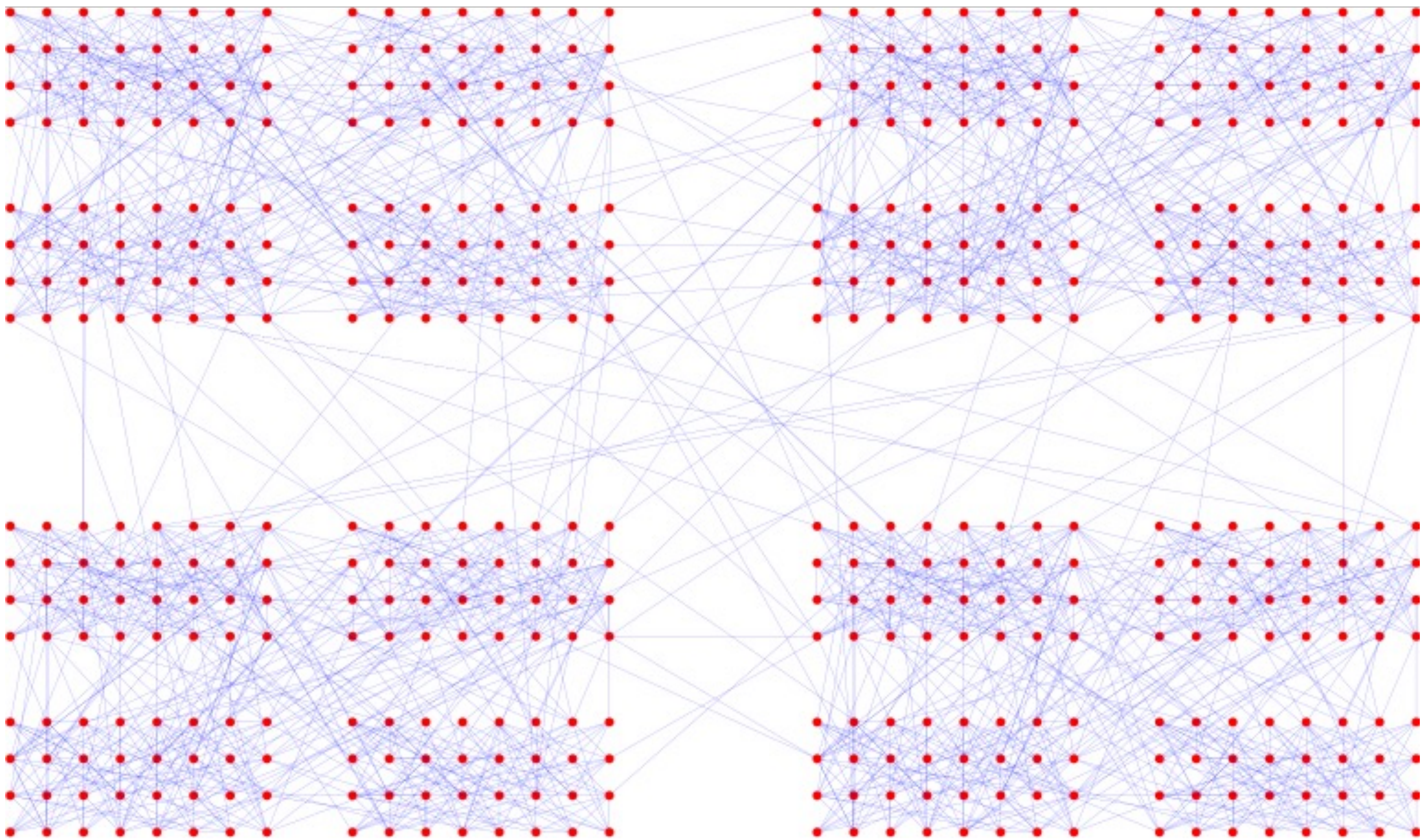
Blue nodes are in more than one community

Hierarchical communities



Hierarchical communities

- If the network has multiple levels of organization, its communities could form a **hierarchy**, with small communities within larger ones
- **Example:** branches in a company, in turn divided into departments
- **All hierarchical partitions are meaningful:** a good clustering algorithm should detect all of them

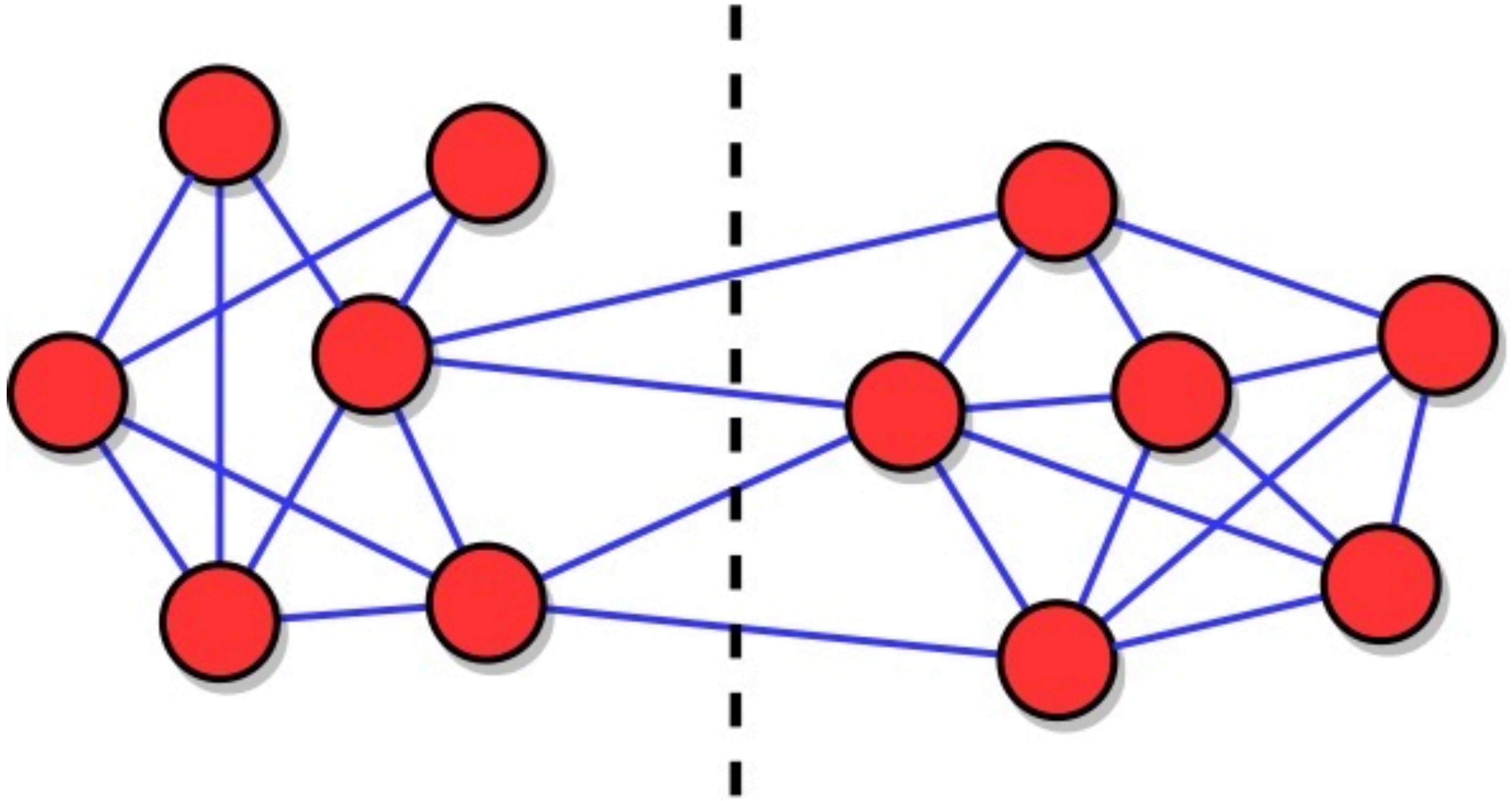


Finding communities: network partitioning

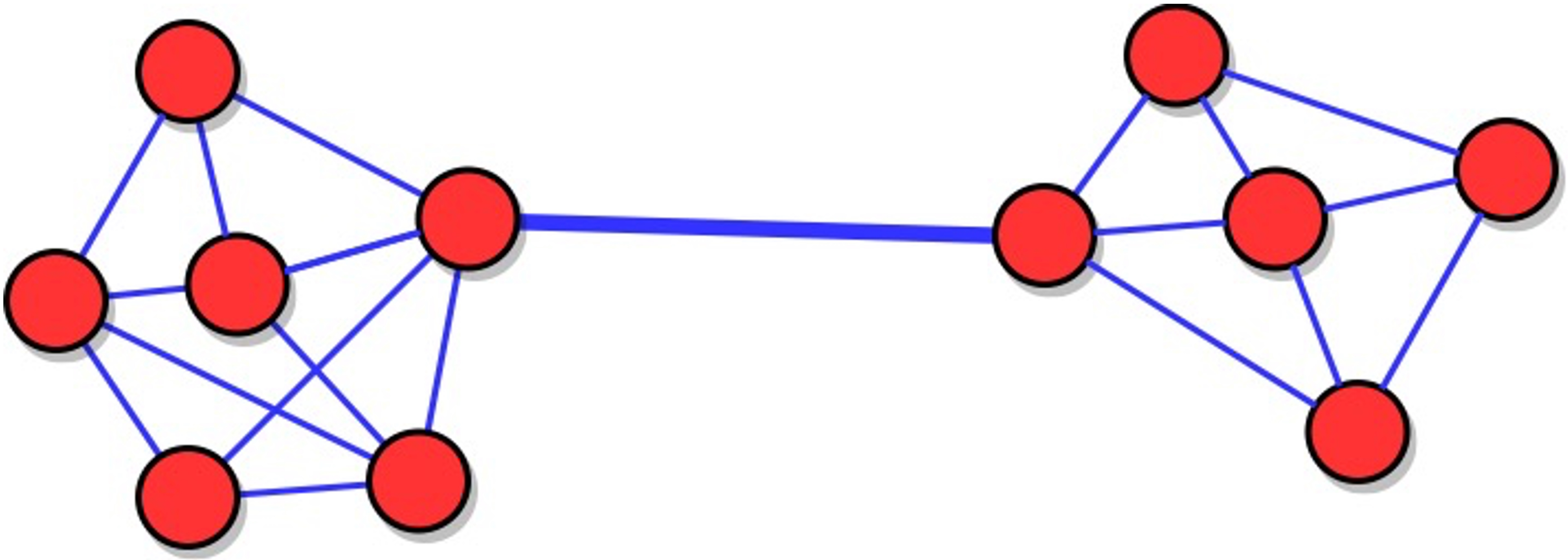
Network partitioning

- **The problem:** dividing the nodes of a network into a number of groups of predefined size, such that the number of links between the groups is minimal

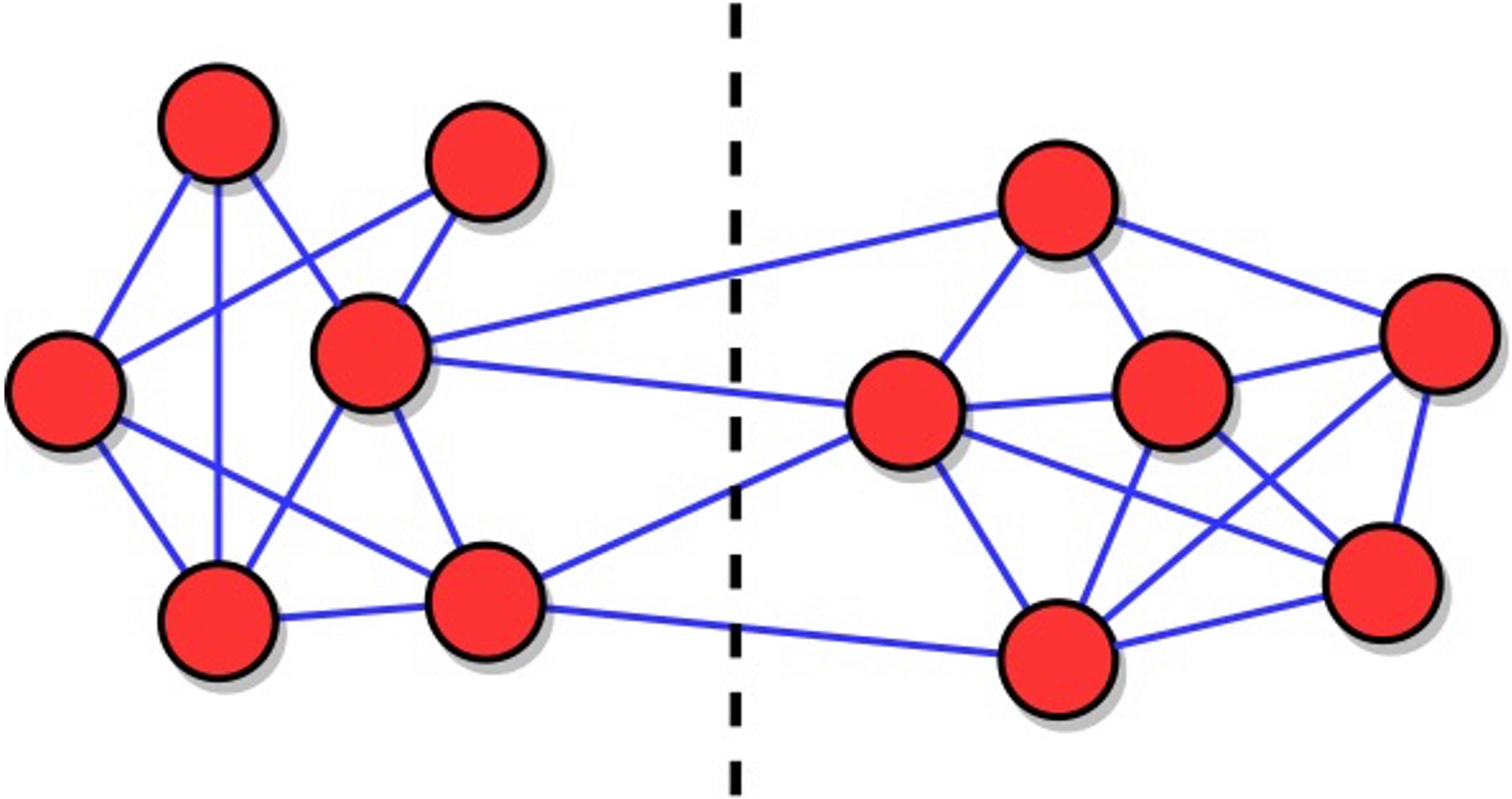
- The number of links between groups is called **cut size**



A graph that is easy to bisect



Graph bisection: finding a minimal "cut"

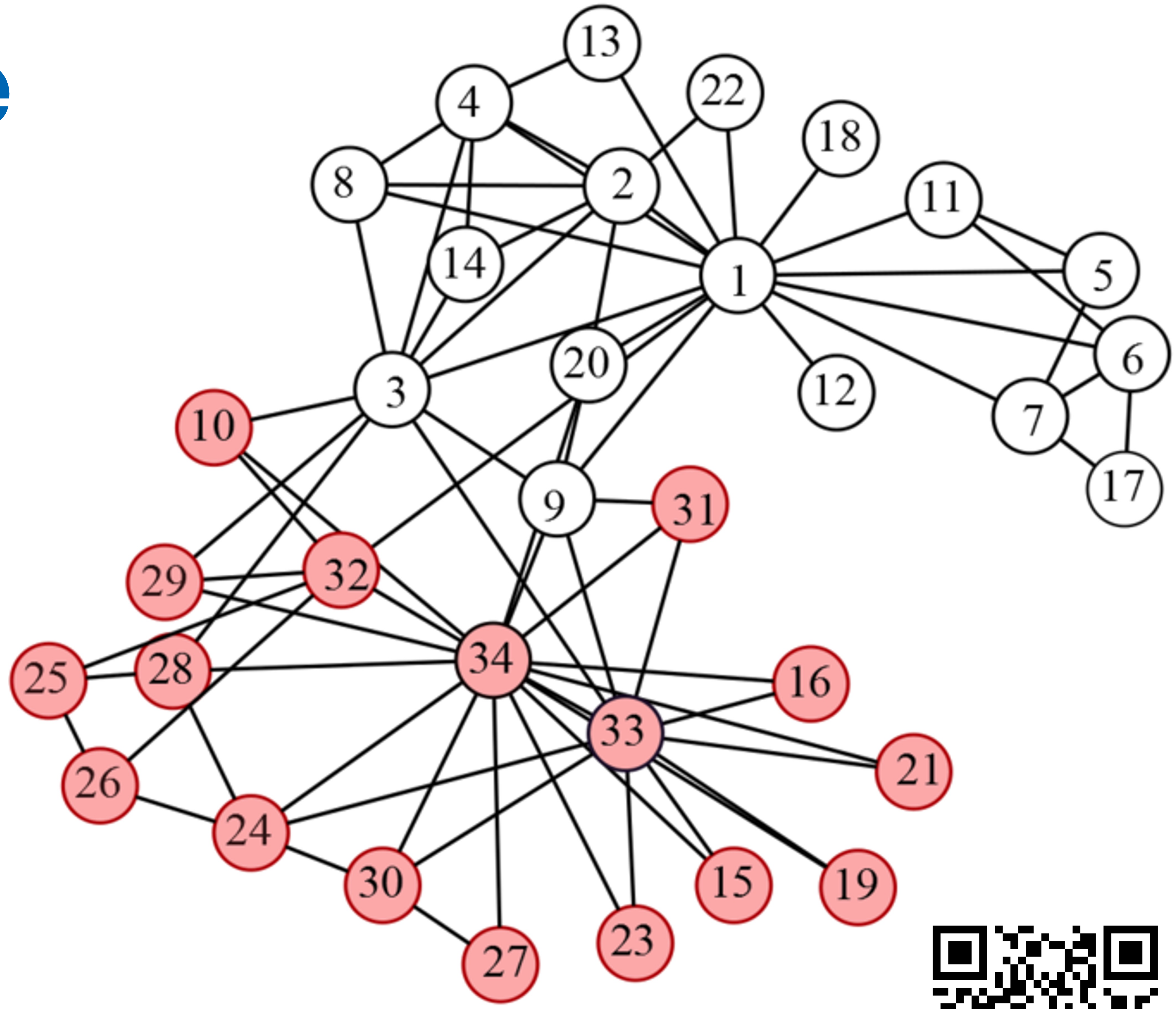


Simple exercise

Cut size under bisection

39/

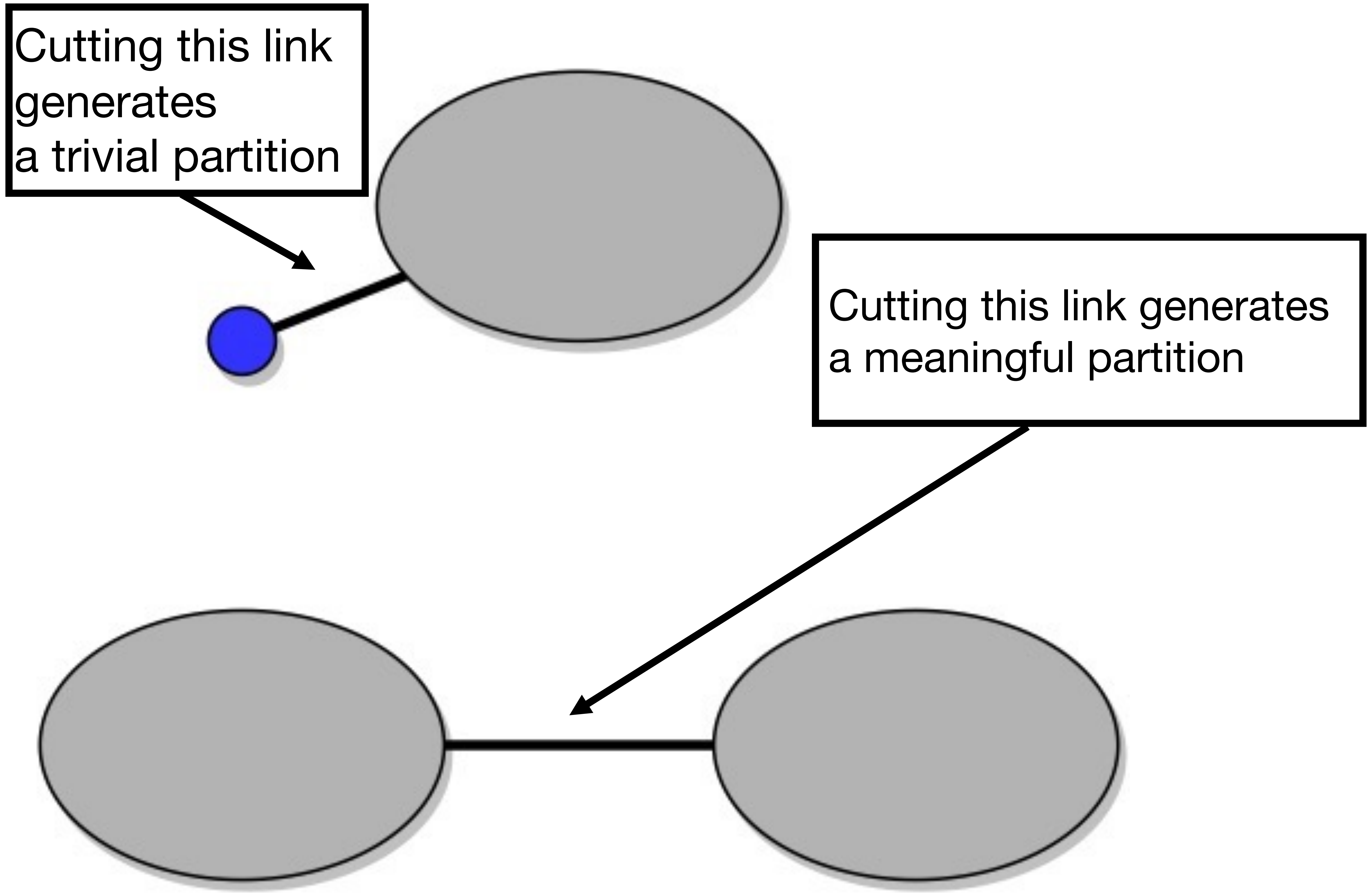
- What is the size of the white-red cut?
- If node 9 goes to the red component, what is the size of the white-red cut?



Pin board: https://upfbarcelona.padlet.org/chato/4qz0k8ro0zquen1

Network partitioning

- Problems:
  - If the number of clusters is not given beforehand, the trivial solution is a single cluster including everything
  - If the size of the clusters is not indicated, there may be trivial solutions by removing the nodes with lowest degree



Network partitioning: limits

- Clusters have to be well-separated, but they do not need to have high internal link density —> **clusters found via graph partitioning are not communities, in general**

- The number of clusters must be given as input, but it is usually unknown

```
# minimum cut bisection: returns a pair of sets of nodes
partition = nx.community.kernighan_lin_bisection(G)
```

Finding communities: Hierarchical clustering

Hierarchical clustering

- Hierarchical clustering delivers a nested set of partitions
- Main ingredient: **similarity measure**
- Examples:
  - In a social network it could indicate how close the profiles of two people are based on their interests
  - If nodes are embedded in space (i.e., they are points in a metric space), the (dis)similarity between two nodes can be expressed by their distance
  - If nodes are not embedded in space, similarity measures can be derived from the network structure

Similarity: structural equivalence

- **Concept:** nodes are similar if their neighbors are similar

$$S_{ij}^{SE} = \frac{\text{number of neighbors shared by } i \text{ and } j}{\text{total number of nodes neighboring only } i\text{, only } j\text{, or both}}$$

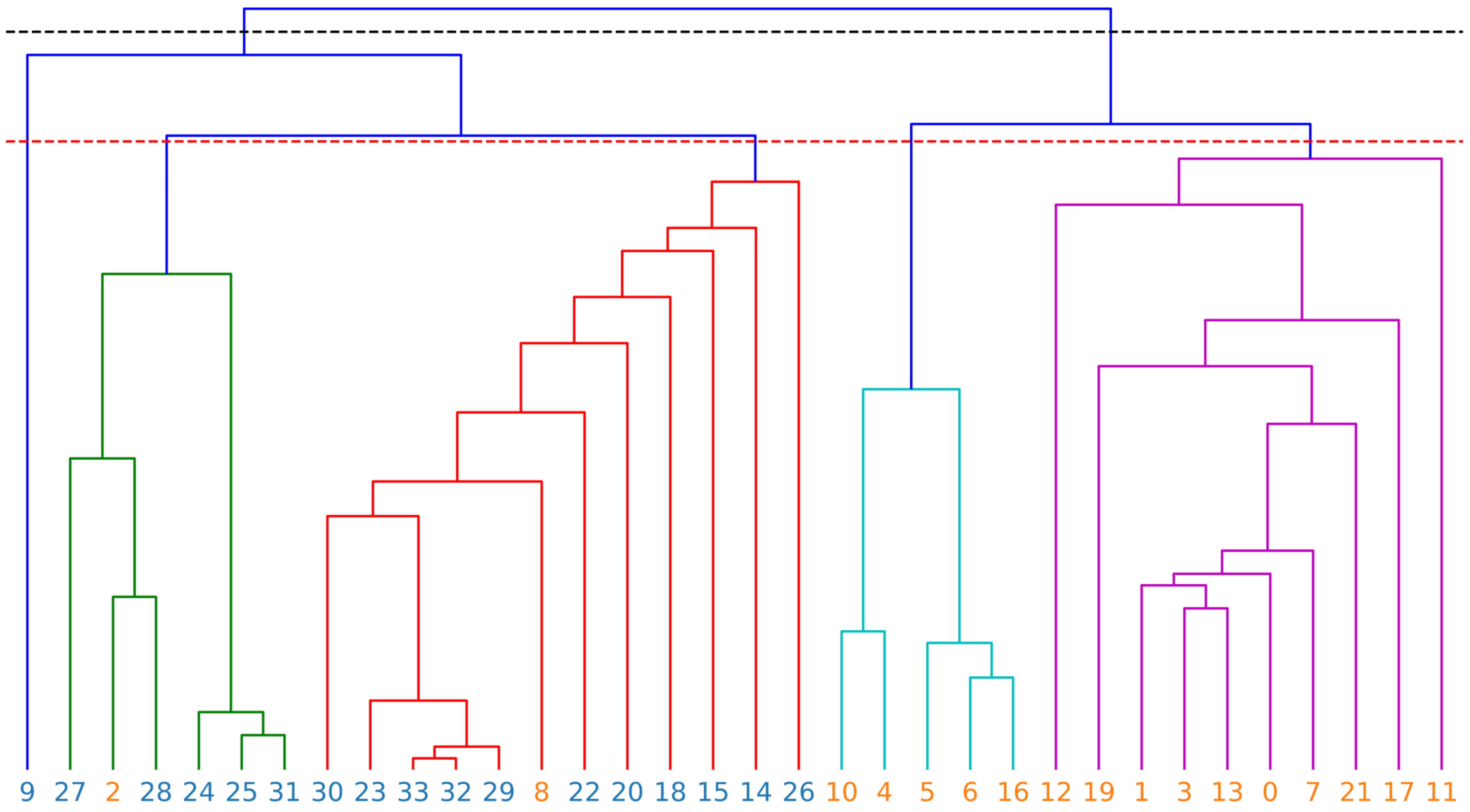
- **Examples:**
  - If the neighbors of *i* and *j* are (v_1, v_2, v_3) and (v_1, v_2, v_4, v_5), respectively, S_{ij} = 2/5 = 0.4, because there are two common neighbors (v_1 and v_2) out of five distinct neighbors in total (v_1, v_2, v_3, v_4, v_5)
  - If *i* and *j* have no neighbors in common, S_{ij} = 0
  - If *i* and *j* have the same neighbors, S_{ij} = 1

Hierarchical clustering

- **Two approaches:**
  - **Agglomerative hierarchical clustering:** partitions are generated by iteratively merging groups of nodes
  - **Divisive hierarchical clustering:** partitions are generated by iteratively splitting groups of nodes

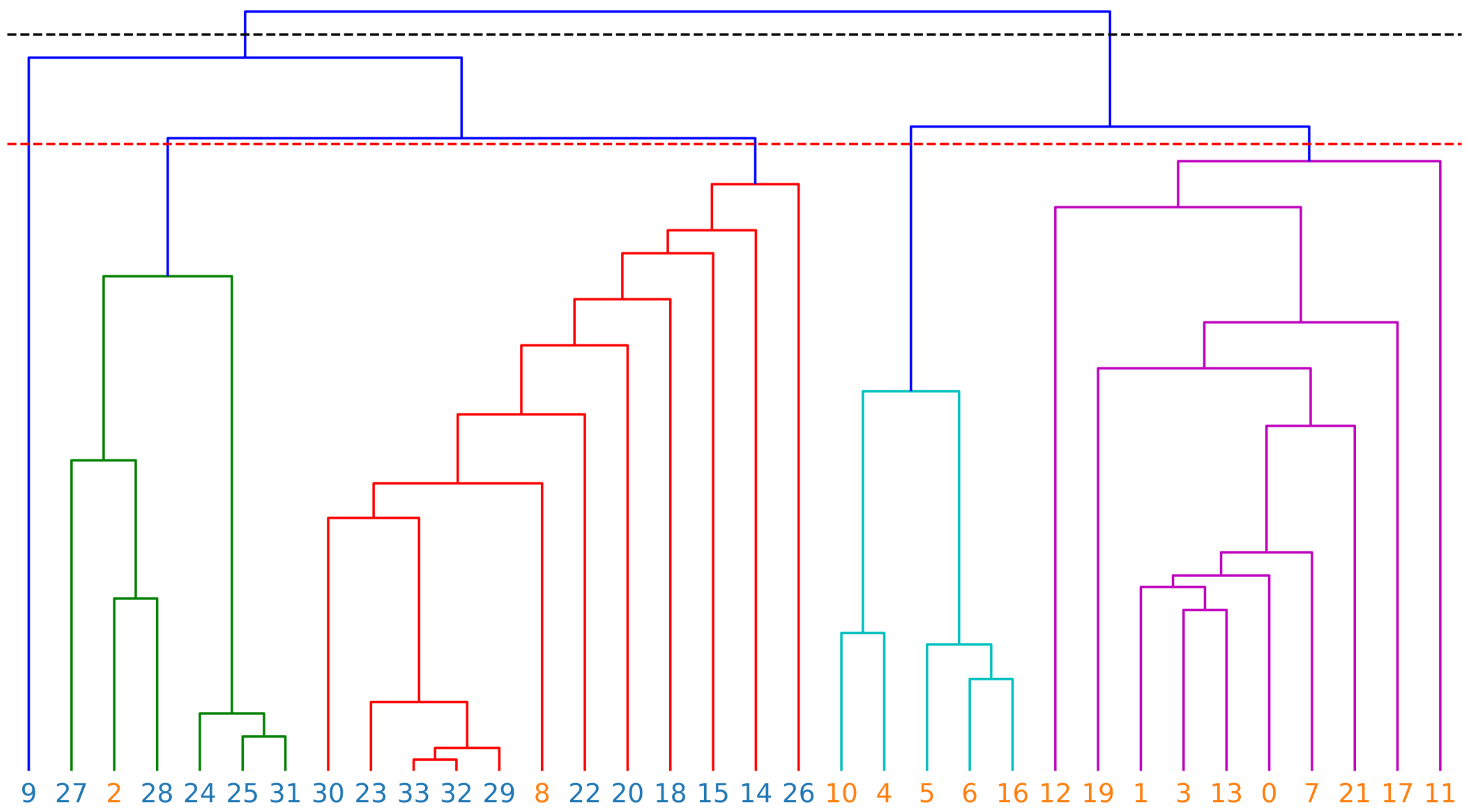
Dendrograms

- Outcome: **dendrogram (hierarchical tree)**
- A dendrogram is a compact summary of all partitions created by hierarchical clustering



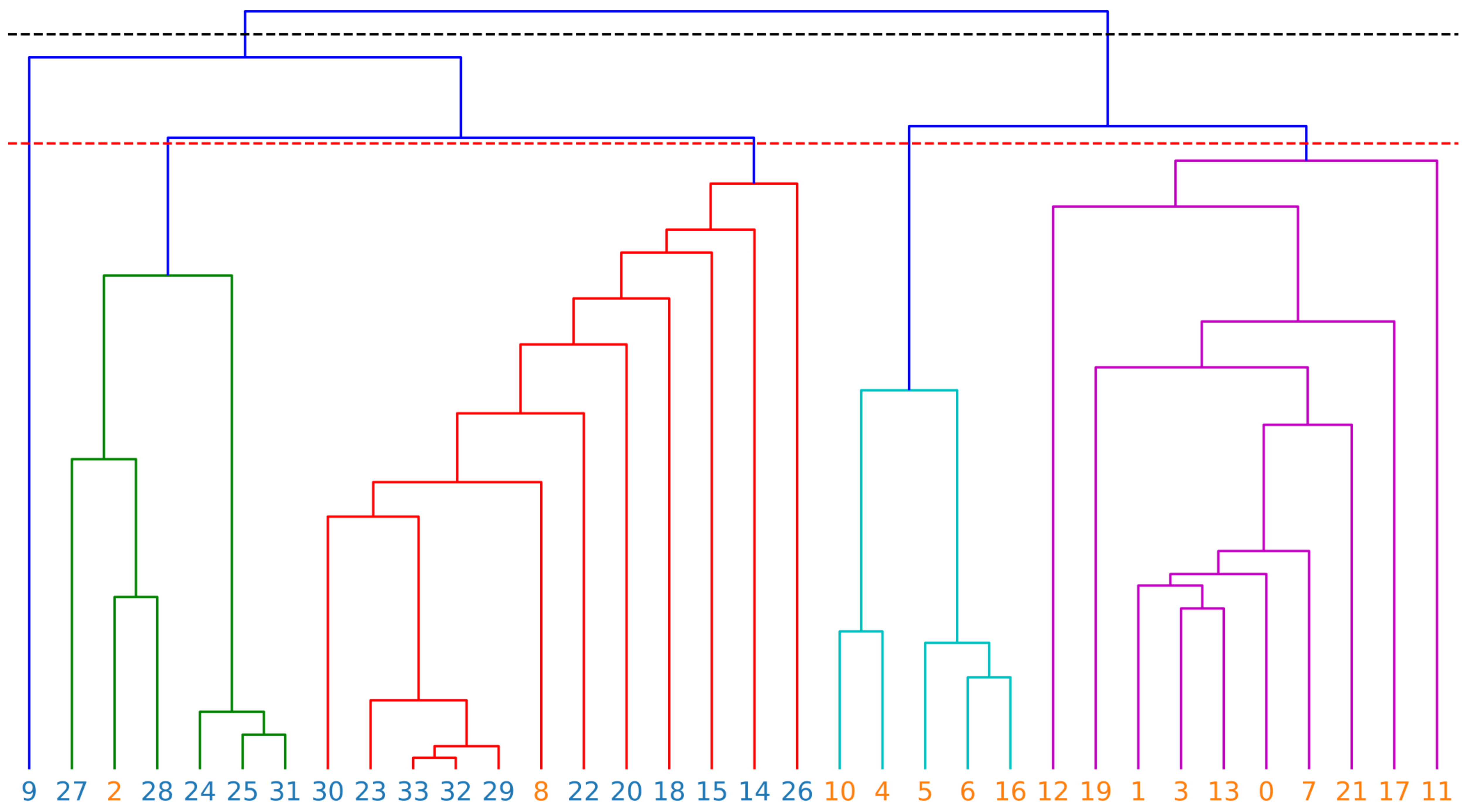
Dendrograms

- **Features:**
  - Bottom: leaves of the tree, indicated by the labels of the nodes
  - Going upwards, pairs of clusters are merged. Mergers are illustrated by horizontal lines joining two vertical lines, each representing a cluster
  - The nodes of each cluster can be identified by following the vertical line representing the cluster all the way down



Dendrograms

- Partitions are selected via **horizontal cuts** of the dendrogram: the clusters are the ones corresponding to the vertical lines severed by the cut
- High cuts yield partitions into a few large clusters, low cuts yield partitions into many small clusters
- **Hierarchy:** each partition has clusters including clusters of all partitions lying lower in the dendrogram



Hierarchical clustering: limits

- It delivers as many partitions: which one(s) shall we choose?
- Results usually depend on the similarity measure and on the criterion adopted to compute the similarity of the groups
- It is rather slow; networks with millions of nodes are out of reach

Summary

Things to remember

- Community definitions (strong, weak, overlapping, etc)
- K-cores decomposition algorithm
- Network partitioning
- Hierarchical clustering

Sources

- A. L. Barabási (2016). Network Science – Chapter 09
- D. Easly and J. Kleinberg (2010). Networks, Crowds, and Markets – Chapter 03
- F. Menczer, S. Fortunato, C. A. Davis (2020). A First Course in Network Science – Chapter 06
- URLs cited in the footer of slides