Network Science

Javier Garcia-Bernardo

Utrecht University

Tabular data

	X_1	X_2	…	X_N
Observation 1				
Observation 2				
Observation 3				
Observation 4				
Observation 5				

Text data

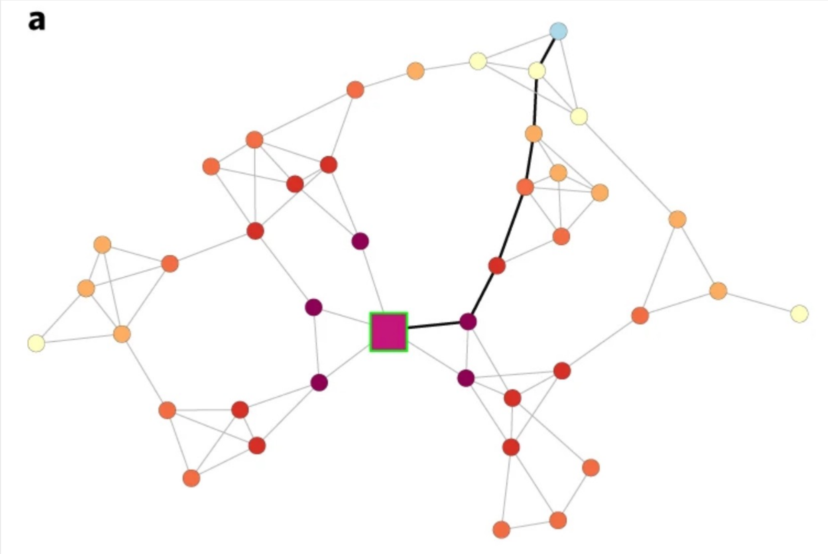
	word_1	word_2	…	word_N
Text 1				
Text 2				
Text 3				
Text 4				
Text 5				

Relational/transactional data

Our life is completely defined by networks: relationships, interactions, communications. Biological networks governing the interactions between genes in our cells determine our development, neural networks in our brain make us think, information networks guide our knowledge and culture, transportation networks allow us to move, and social networks sustain our life.

A First Course in Network Science, F Menczer, S Fortunato, C.A. Davis

Today: Relational data (networks)



	X_1 ?	X [illegible]	[illegible]	X_N ?
Agent 1				
Agent 2				
Agent 3				
Agent 4				
Agent 5				

If we were to study them using tabular data, how do we include connections?

Why do we care about the connections?

They reflect underlying patterns (e.g. differences in power/hierarchies/roles).

They constrain/facilitate future change.

Sometimes they are responsible of "emergent" phenomena that cannot be explain from looking at the actors.

1) They reflect underlying patterns (e.g. differences in power/hierarchies/roles).

RQ:

Which country has the most bargaining power?

Who is the most politically influential company?

What type of roles do employees of a company play (coordination/innovation/etc)?

What is the political ideology of media outlets?

2) They constrain/facilitate future change.

Block, P., Hoffman, M., Raabe, I. J., Dowd, J. B., Rahal, C., Kashyap, R., & Mills, M. C. (2020). Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. Nature human behaviour, 4(6), 588-596.

Which person would you vaccinate first?

app.wooclap.com/ADAV2024

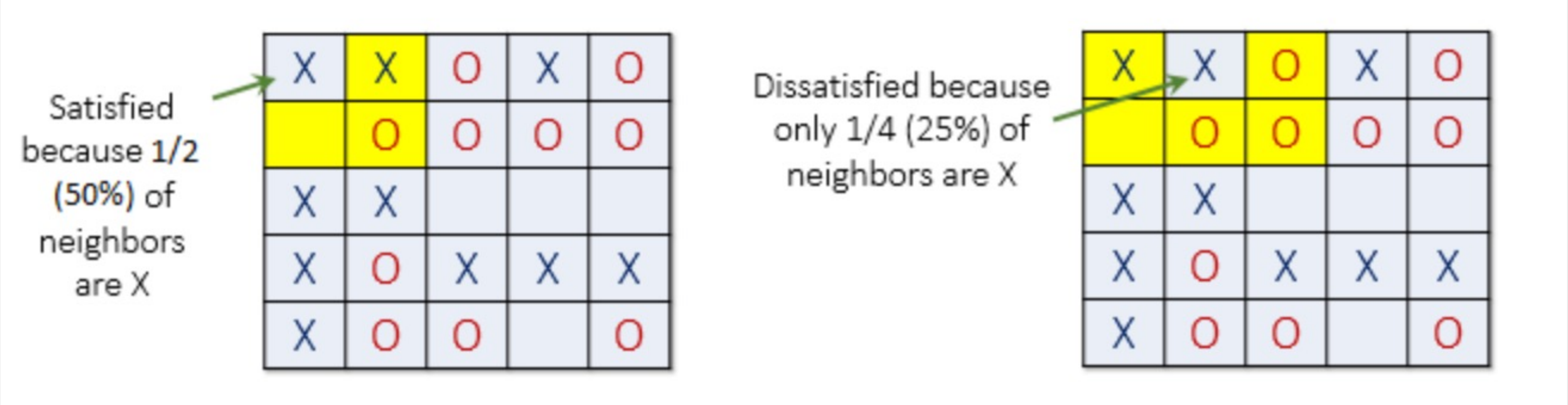
3) Emergence. e.g. Schelling model

Why do we see residential segregation?

Every actor lives in a house and is connected to its neighbors in a network.
Every actor is the same:

- They want to have 1/3 of their neighbors to be like them
- Otherwise, they move to a random house

Let's play!



http://nifty.stanford.edu/2014/mccown-schelling-model-segregation/

Two key concepts of today

- **Centrality:** Who are the key actors in the network?

- **Community structure:** What clusters of people are in the network?

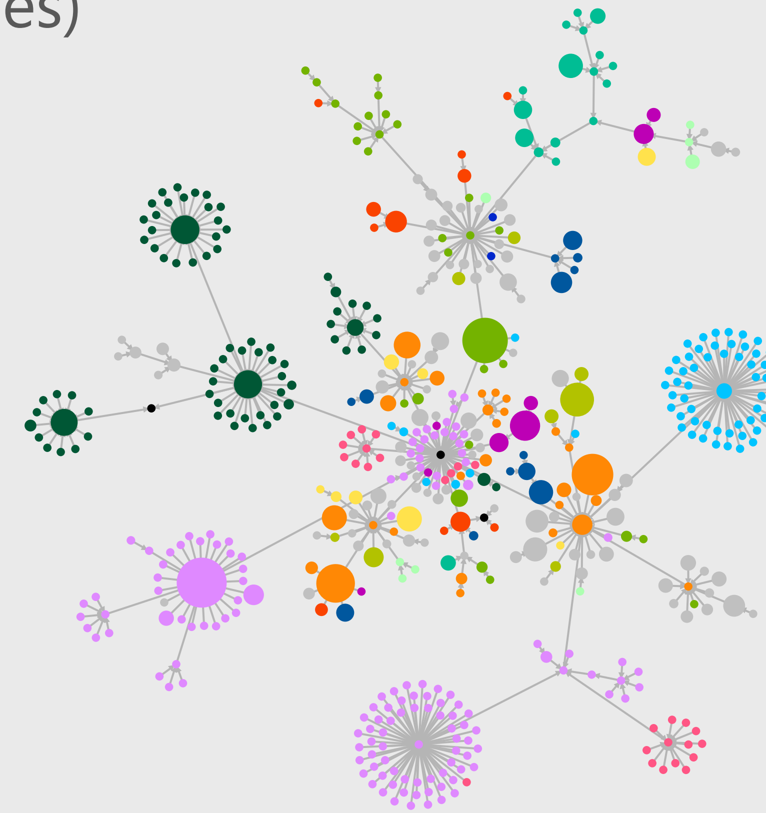
Introduction to networks

What is a network?

Mathematical representation of the relationships (edges) between entities (nodes)

The most important question to ask yourself is

What are the nodes and what are the edges?



Types of networks

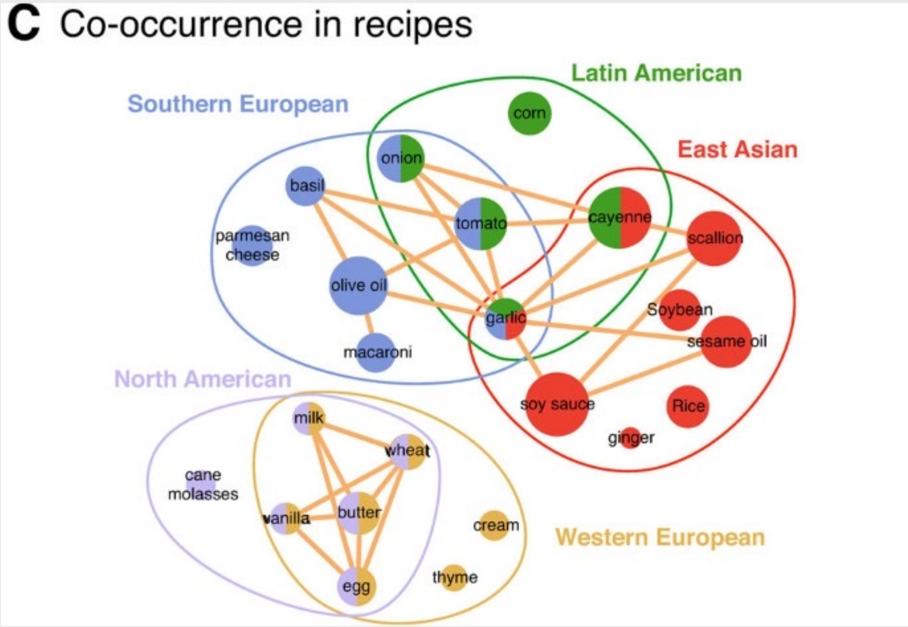
	Network	Nodes	Edges
Social/ Behavioral	Friendship	People	Friendships
	Instagram	Online accounts	Followers/likes
	Psychological	Symptoms	Co-ocurrence
Biology	Gene regularory	Genes	Activations/inhibations
	Food web	Animals	Predation
Economic	Trade	Countries/companies	Money flows
	Ownership	Companies	Ownership stakes
Intrastructure	Internet	Computers (IPs)	Data transmission
	Power grid	Power stations	Power lines
	Airplane network	Airports	Flights

Adapted from: https://aaronclauset.github.io/courses/5352/csci5352_F21_L1.pdf

Type of networks and characteristics

Type 1: Interaction and flow → "Real networks".

- Offline interactions
- Online interactions



Type 2: Affiliation → Node 1 is part of/related to node 2

- e.g., students in classrooms
- e.g., ingredients in recipes

Type 3: Co-occurrence → Node 1 is correlated with node 2

- e.g., stock market networks (the fluctuations in two stocks correlate)
- e.g., brain networks (the brain signals in two areas correlate)

Today we focus on the first type ("real networks")

Why do we care about networks?

Network structure and network dynamics reflect important information

Epidemiology: How to stop disease transmission in a social network?

Criminology: How to detect criminal actors in a network of money flows?

Biotechnology: Which genes to target to stop cancer in a gene regulatory network?

Ecology: Which animals we need to preserve to avoid ecosystem collapse?

Psychology: In a belief network, how does attitude change depend on the correlation between other attitudes?

Engineering: How to improve network performance and reliability in power grids?

Economics: How does country development depend on the type of products a country export?

Social science: How does social capital affect upward mobility?

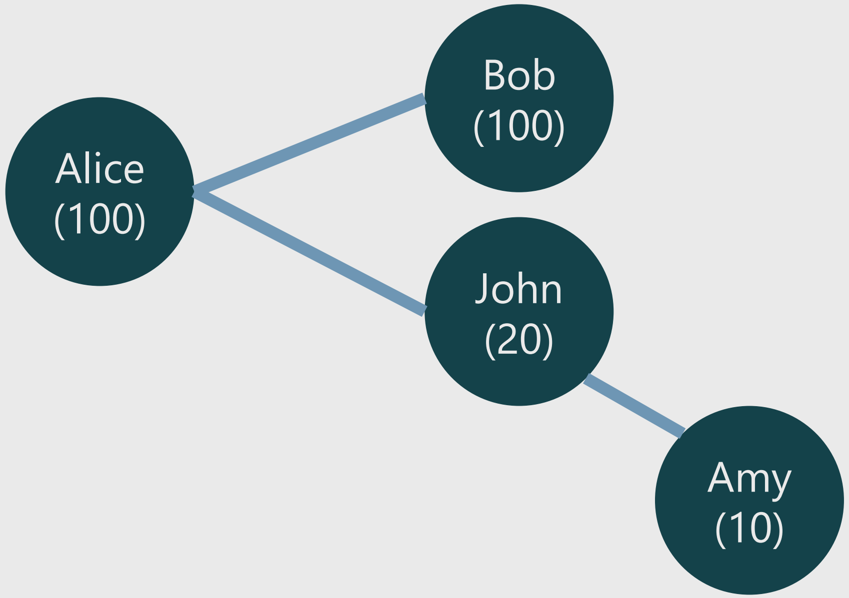
Physics view: Dependence on topology (reliability, dynamics, emergent behavior and phase transitions)

What other research questions could you answer using networks?

app.wooclap.com/ADAV2024

Basic definitions

Networks (graphs)



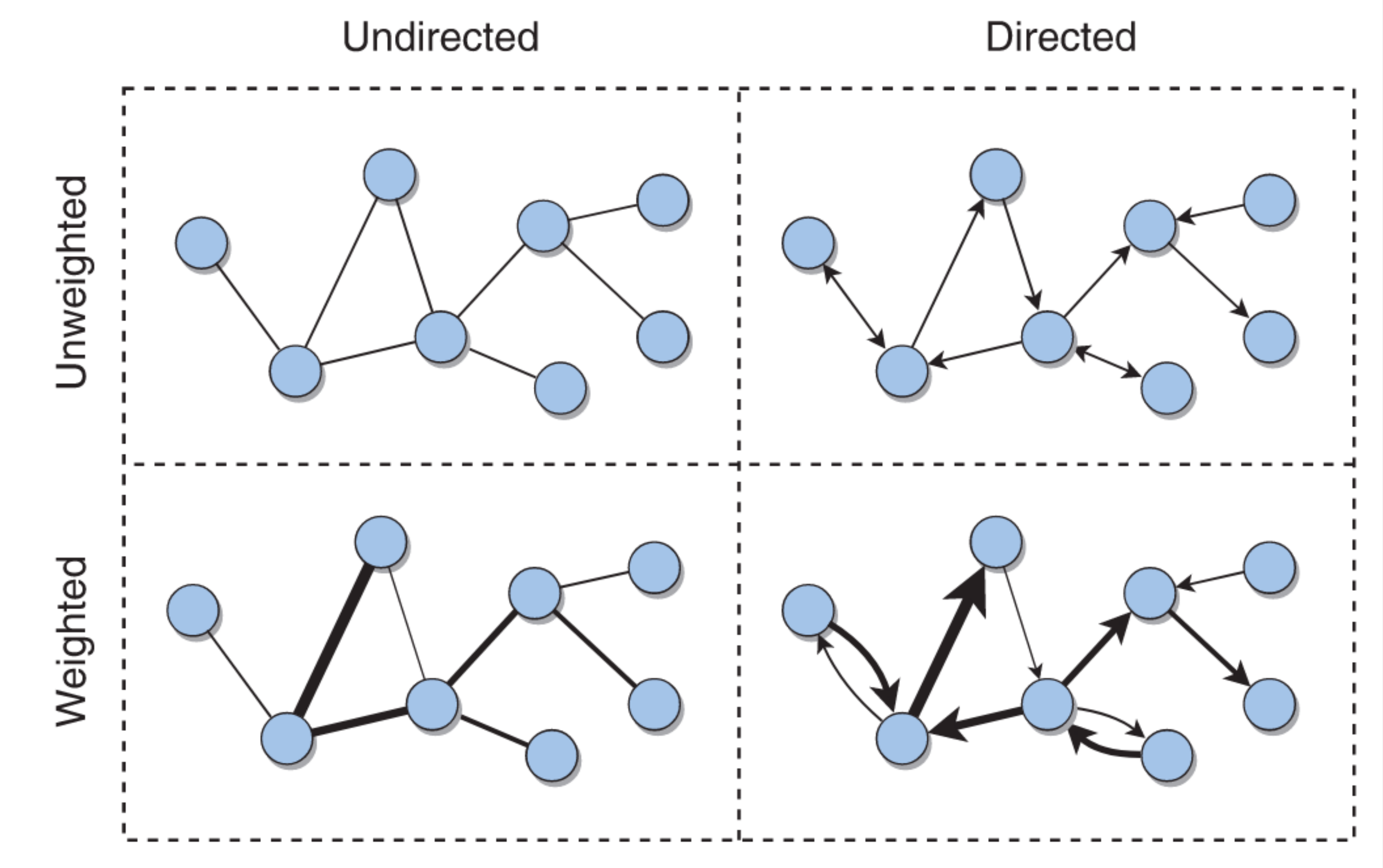
Nodes (vertices, actors) connected by
edges (links, connections, relationships)

N: **Nodes** = {Alice, Bob, John, Amy}
E: **Edges** = {(Alice, Bob), (Alice, John), (John, Amy)}

The edge (i,j) connects node i to node j

Nodes can have **attributes** (e.g. gender, income, etc)
Edges can have **attributes** (e.g. type, strength, etc)

Directed vs undirected; weighted vs unweighted



Source: A first course in network science (2020)

Undirected: The link (i,j) connects node i to node j in both directions

Directed: The link (i,j) connects node i (source) to node j (target)

Weighted: There is a weight associated to each edge

Degree in undirected networks

Definition: Number of neighbors in the network

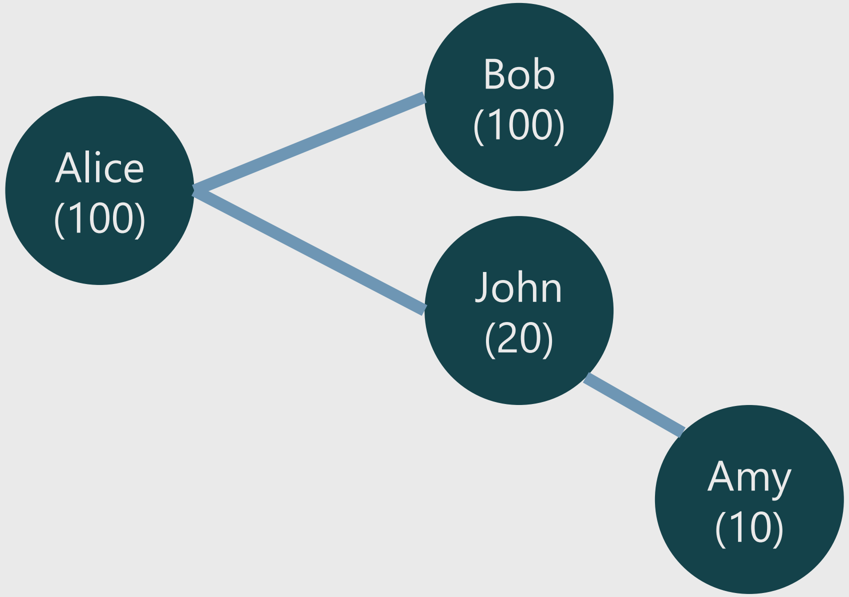
Node: degree

Alice: 2

Bob: 1

John: 2

Amy: 1



Degree in directed networks

Out-degree: Number of outgoing edges

In-degree: Number of incoming edges

Total degree: Sum of out and in degree

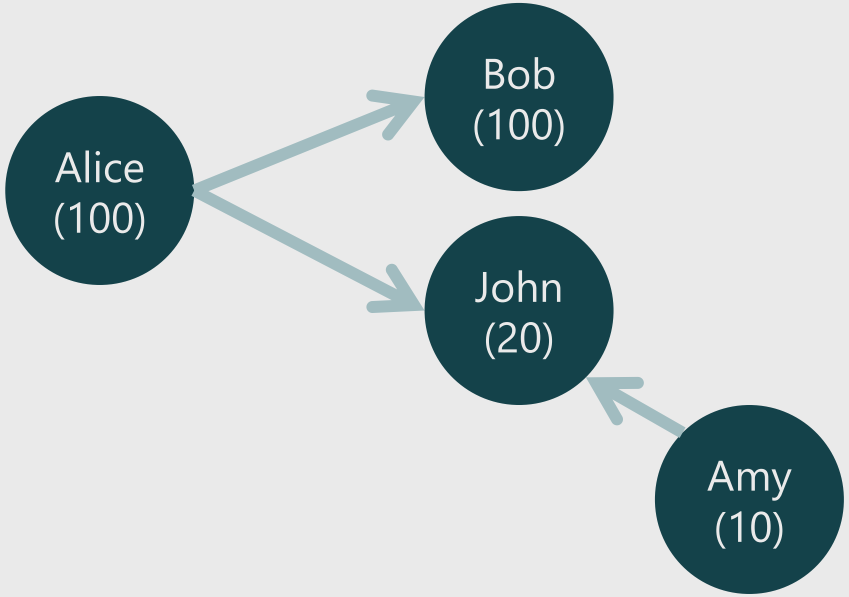
Node: (out, in, total)

Alice: (2, 0, 2)

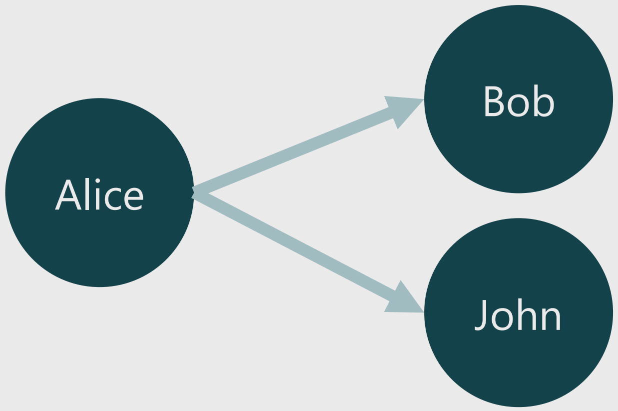
Bob: (0,1, 1)

John: (0,2, 2)

Amy: (1,0, 1)



Network representation



Adjacency list (edgelist):

- Adv: It is dense: Only keeping edges
- Disadvantage: Hard to work with

Origin	Target	Weigth
Alice	Bob	1
Alice	John	1

Adjacency matrix:

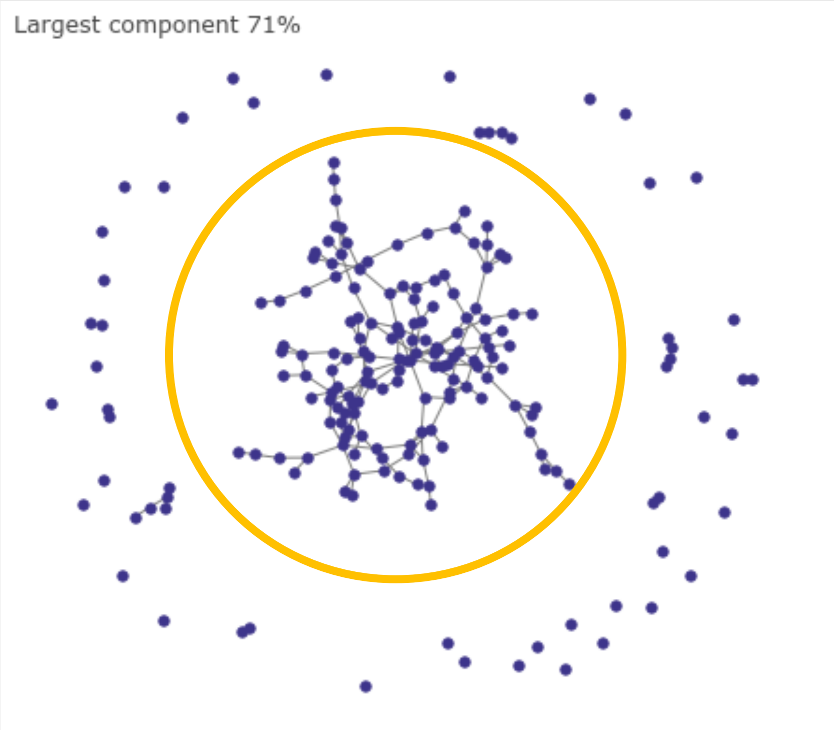
- Adv: Math is easy (matrix multiplication)
- Disadvantage: It is sparse (mostly zeros). 1 million nodes → 1 trillion numbers

Target → ↓ Source	Alice	Bob	John
Alice	0	1	1
Bob	0	0	0
John	0	0	0

In computer → Sparse matrices: Best of both worlds

Network metrics and characteristics

Connectedness



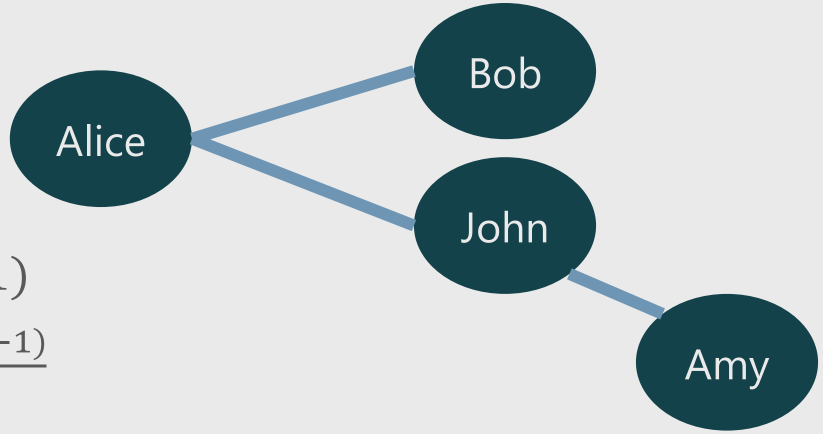
Real networks are typically connected, forming a **"giant component"**

If the average degree < 1 → many small components

If the average degree > 1 → suddenly the system becomes connected

Density

Definition: Number of edges present / potential number of edges



- Number of edges = 3
- Potential number of edges in directed network = (4*3) = $N \cdot (N-1)$
- Potential number of edges in undirected network = (4*3)/2 = $\frac{N\,(N-1)}{2}$

Density = 3/6 = 50%

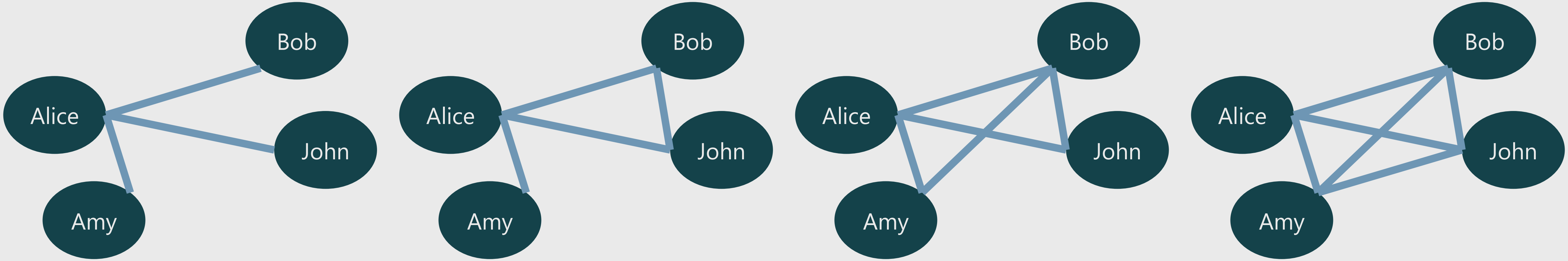
Real networks are typically **sparse** (out of the 8B people on earth, you have very few friends)

Clustering coefficient (~transitivity)

Strogatz, Watts (1998): The share of your neighbors who are connected to each other

Real networks have **high clustering**

Clustering of Alice:



0/3 1/3 2/3 3/3

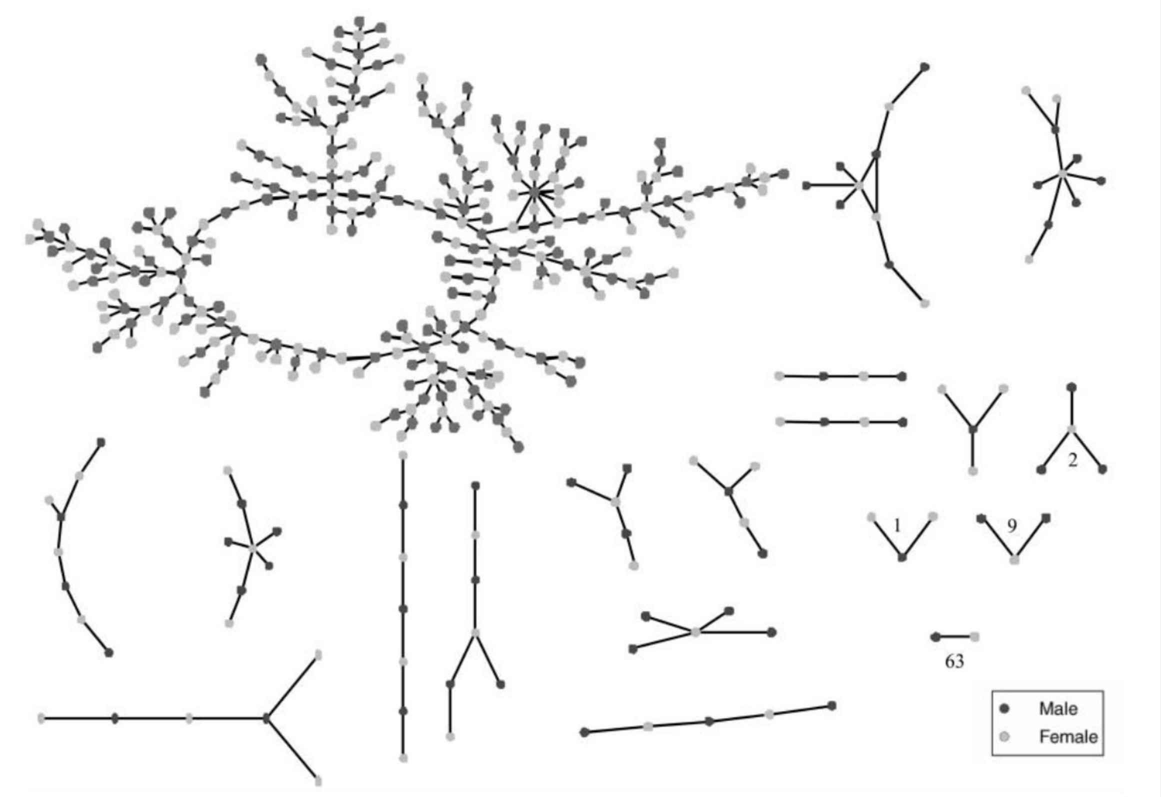
Assortativity (homophily)

Preference for nodes to attach to others that are similar in some way

Defined with respect of an attribute (e.g. gender)

Ranges from -1 (fully disassortative) to 1 (fully assortative)

Paraisópolis favela and Morumbi, in São Paulo
Photography by Tuca Vieira (the guardian)



Romantic links between teenagers
Bearman, Moody, Stovel (1991)

Small world: six degrees of separation

Milgram's experiment (1987)

Image source: Baek et al, 2021

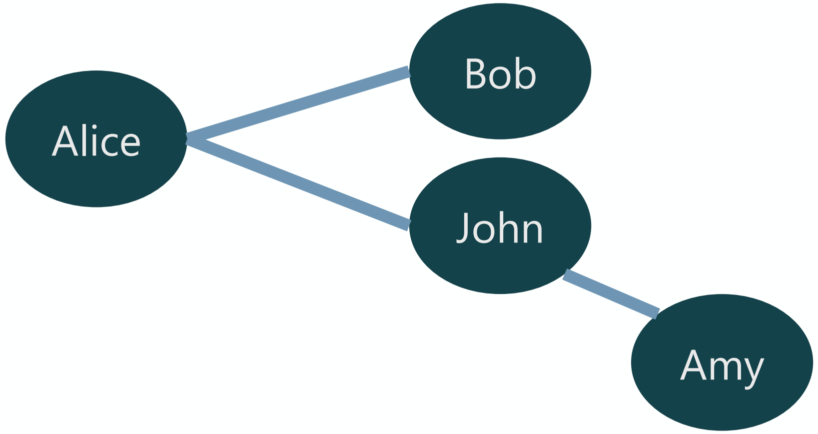
Shortest path between node 1 and node 2:

- Minimum number of steps requires to go from node 1 to node 2
- Between Alice, Amy → 2

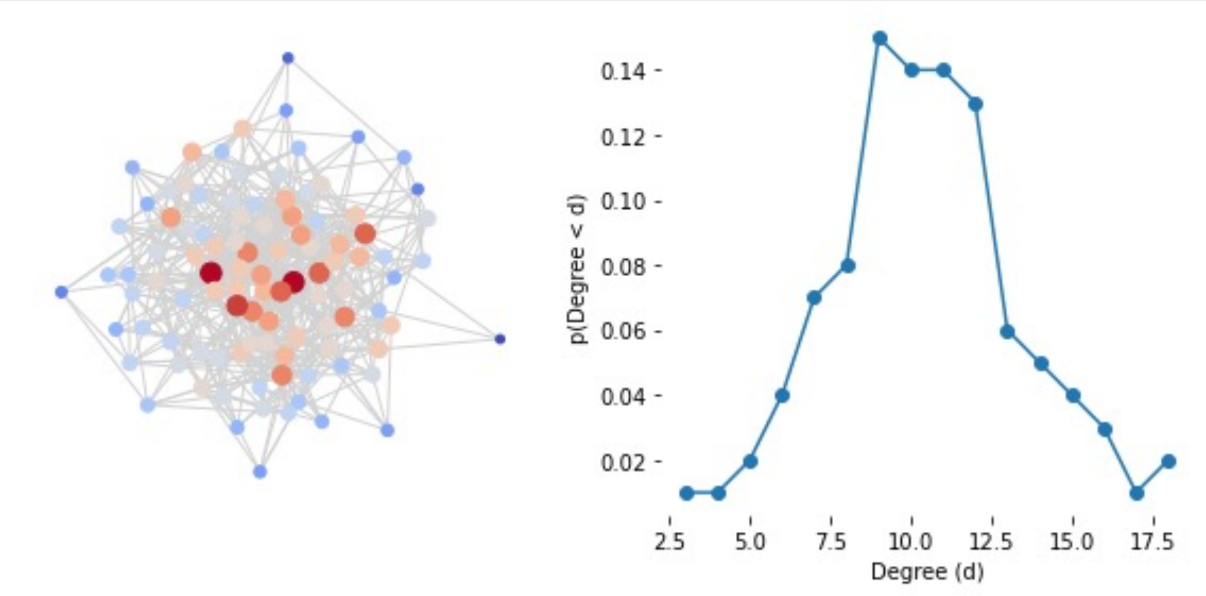
Diameter:

- Longest "shortest path" between two nodes
- In our network: 2 (Alice -> John -> Amy)

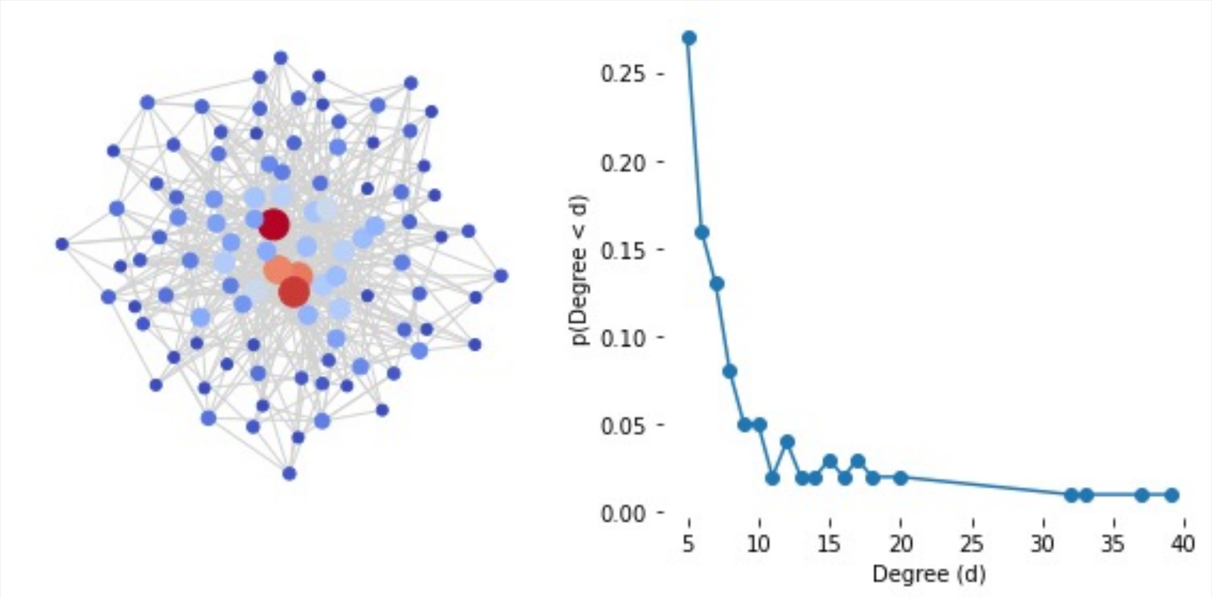
Real networks have **small diameters** because hubs connect diverse parts of the network



Skewed degree distributions



Random network



Real network

Network repository: *networks.sweked.de*

terrorists_911 — 9-11 terrorist network

Problems with this dataset? Open an issue.
You may also take a look at the source code.

The network in this dataset can be loaded directly from graph-tool with:

```
import graph_tool.all a
g = gt.collection.ns["t
```

Description

Network of individuals and their known social associations, centered around the hijackers that carried out the September 11th, 2001 terrorist attacks. Associations extracted after-the-fact from public data. Metadata labels say which plane a person was on, if any, on 9/11.[1]

1. Description obtained from the ICON project. ↩

Tags

Social | Offline | Unweighted | Metadata

Citation

V. Krebs, "Mapping networks of terrorist cells." Connections 24, 43-52 (2002)., https://doi.org/10.5210/fm.v7i4.941 [@sci-hub]

Upstream URL OK

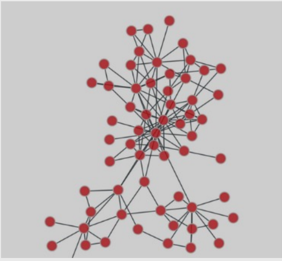
https://aaronclauset.github.io/datacode.htm

Networks

Tip: hover your mouse over a table header to obtain a legend.

Name	Nodes	Edges	$\langle k\rangle$	σ_k	λ_h	τ	r	c	$\oslash$	S	Kind	Mode	NPs
terrorists_911	62	152	4.90	4.00	7.25	19.05	-0.08	0.36	5	1.00	Undirected	Unipartite	id name group

Ridiculograms



swingers — Swingers and parties (2013)

Problems with this dataset? Open an issue.
You may also take a look at the source code.

The network in this dataset can be loaded directly from graph-tool with:

```
import graph_tool.all a
g = gt.collection.ns["s
```

Description

A bipartite sexual affiliation network representing "swing unit" couples (one node per couple) and the parties they attended.[1]

1. Description obtained from the ICON project. ↩

Tags

Social | Offline | Unweighted

Citation

A.-M. Niekampab et al., "A sexual affiliation network of swingers, heterosexuals practicing risk behaviours that potentiate the spread of sexually transmitted infections: A two-mode approach." Social Networks 35(2), 223-236 (2013), https://doi.org/10.1016/j.socnet.2013.02.006 [@sci-hub]

Upstream URL 404

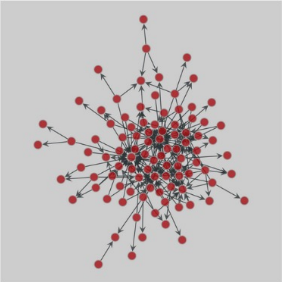
https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/swingers

Networks

Tip: hover your mouse over a table header to obtain a legend.

Name	Nodes	Edges	$\langle k\rangle$	σ_k	λ_h	τ	r	c	$\oslash$	S	Kind	Mode	NPs	EPs
swingers	96	232	2.42	5.19	7.46	5.19	-0.34	0.00	7	1.00	Directed	Bipartite	name	2

Ridiculograms



Recap

There is important information encoded in relationships
Modeling systems using networks allow us to study that information
We can represent networks using adjacencies matrixes or adjacencies lists

We can describe networks using:

- Number of nodes and edges
- Density
- Assortativity
- Clustering coefficient / Transitivity
- Diameter

http://javier.science/panel_network/

15 minutes break

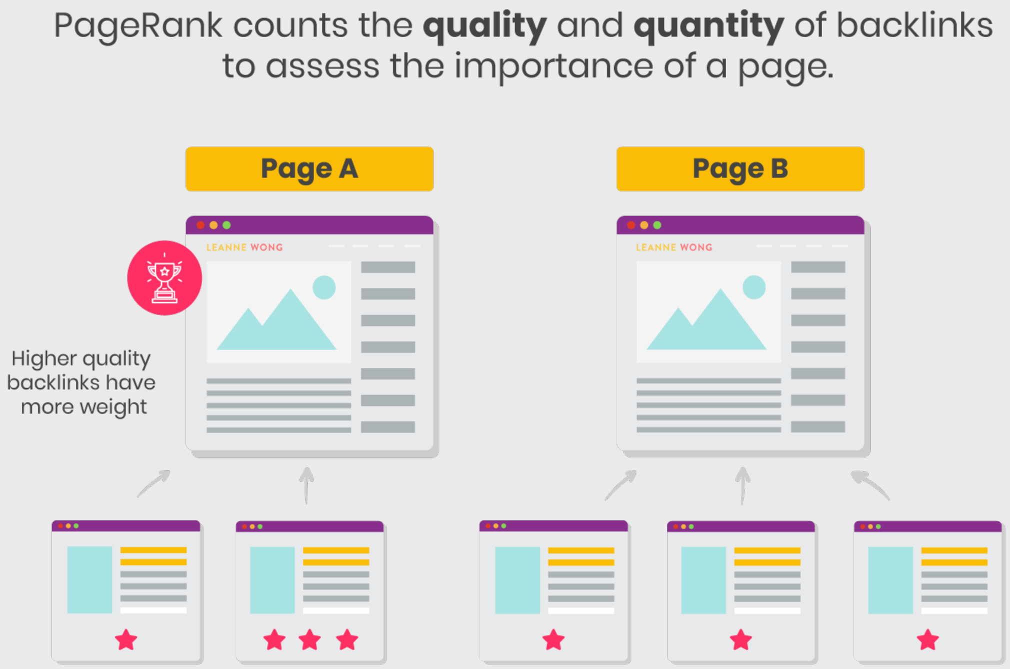
Centrality

Motivating examples

How to stop the spread of diseases?

Important nodes: bridges

How to sort Google results?



https://www.leannewong.co/google-pagerank/

Important nodes: those linked by important nodes

Centrality

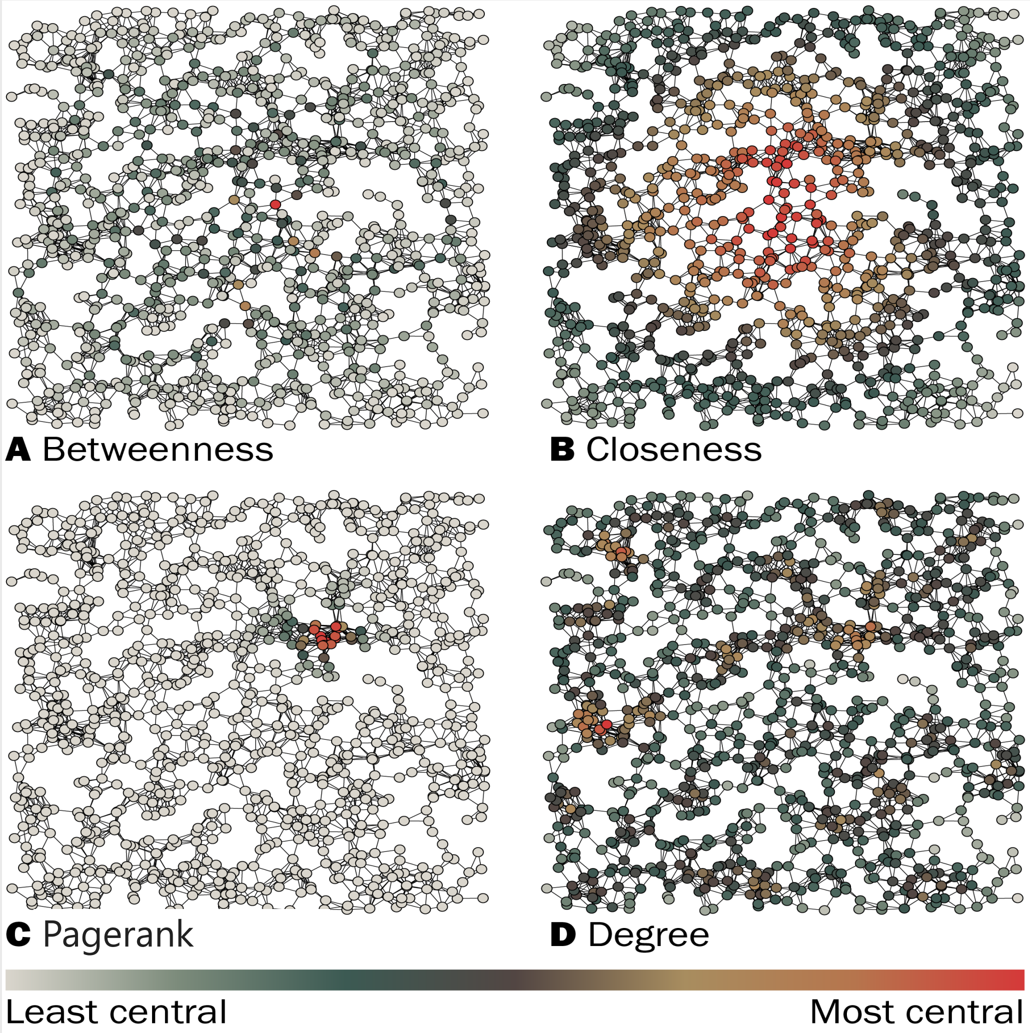
Who are the key actors in the network?

Centrality measures allow to answer this question.

Different centrality measures define importance in different ways :

- *Degree*: Connected to many nodes
- *Closeness*: Close to all other nodes
- *Betweenness*: In the middle of shortest paths
- *Pagerank*: Connected to important nodes

Centrality identify *the most important nodes.* It does not quantify the importance of nodes in general. The relative rankings of non-important nodes may be meaningless.

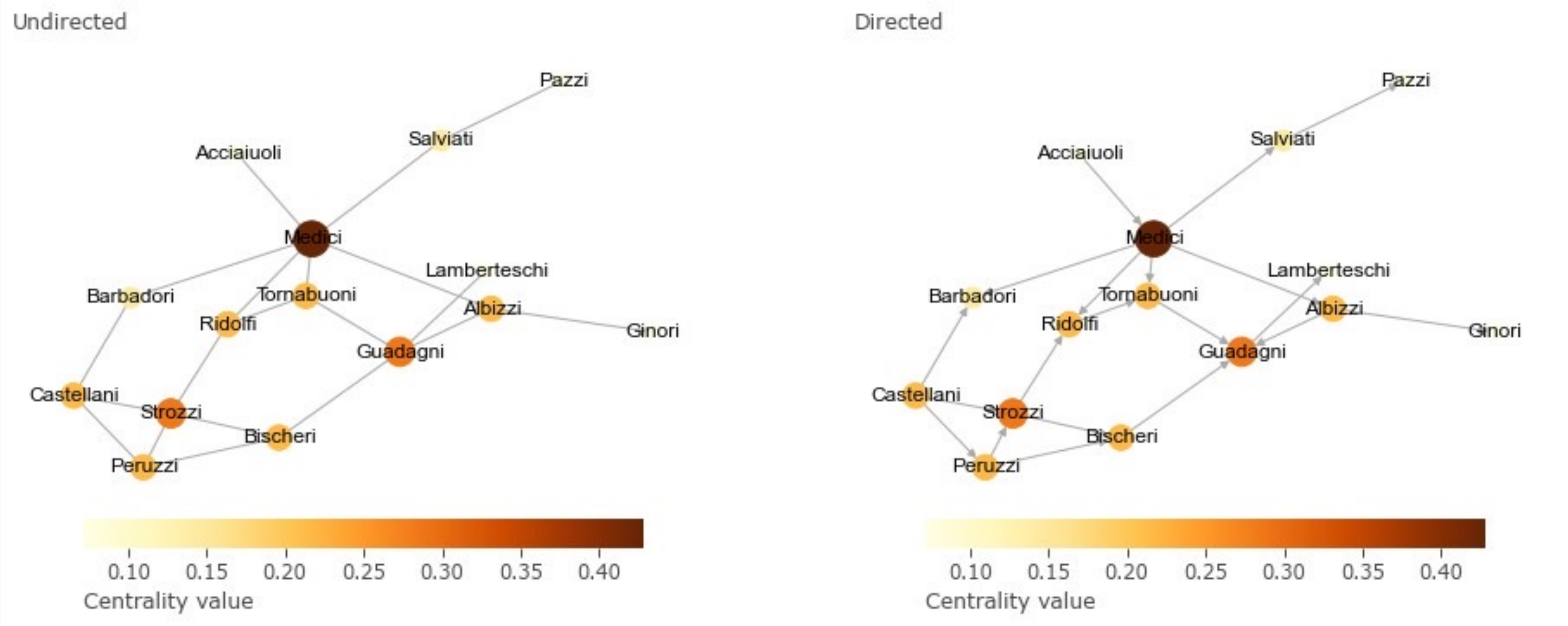


Degree centrality = $^{d_i}/_{N\text{-}1}$

d_i = degree of node i

$N - 1$ = number of nodes - 1 (max. potential number of partners without self-edges)

Measures the **local** influence of the node



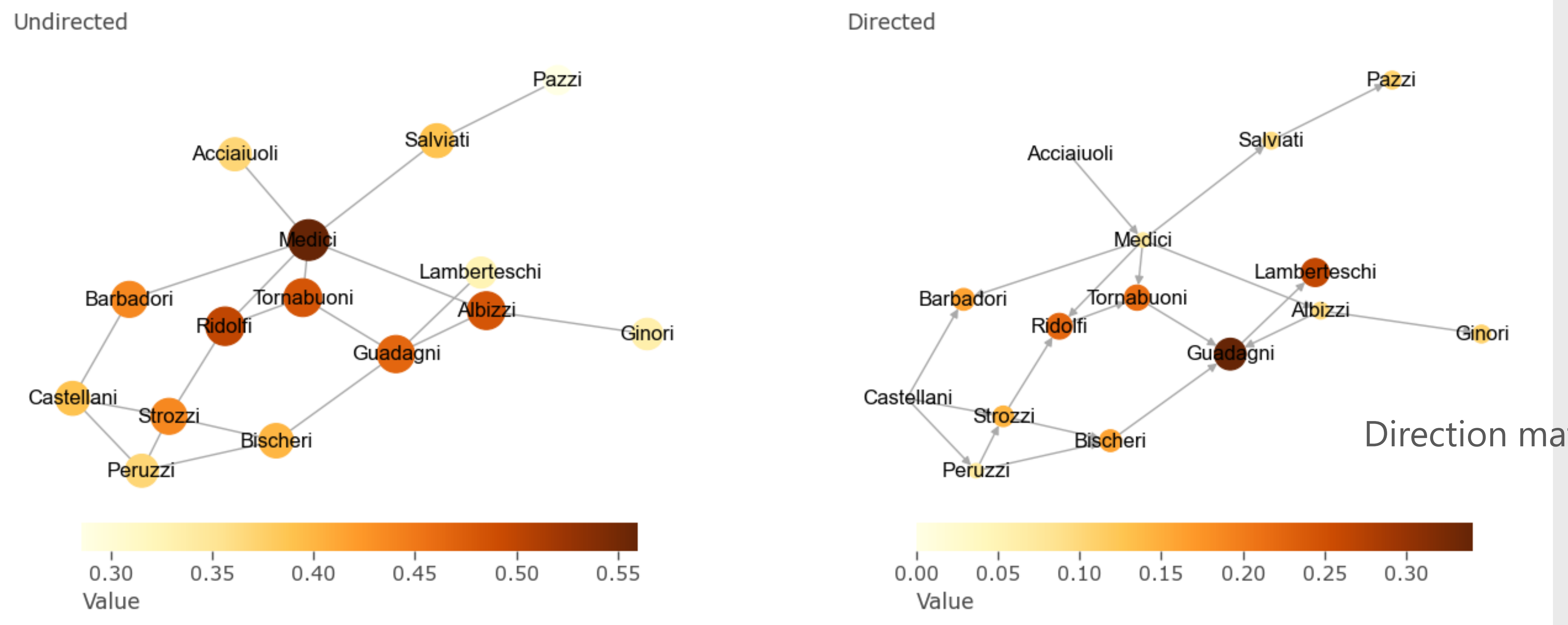
Closeness centrality = $^{1}/_{l_i}$

l_i = average distance of node *i* to all other nodes := $l_i = \frac{1}{N-1}\sum_j d_{ij}$

d_{ij} = shortest distance from node *i* to node *j*

Only useful in fully connected networks

Measures the **most central** node in the network (closest from all other nodes)



Direction matters!

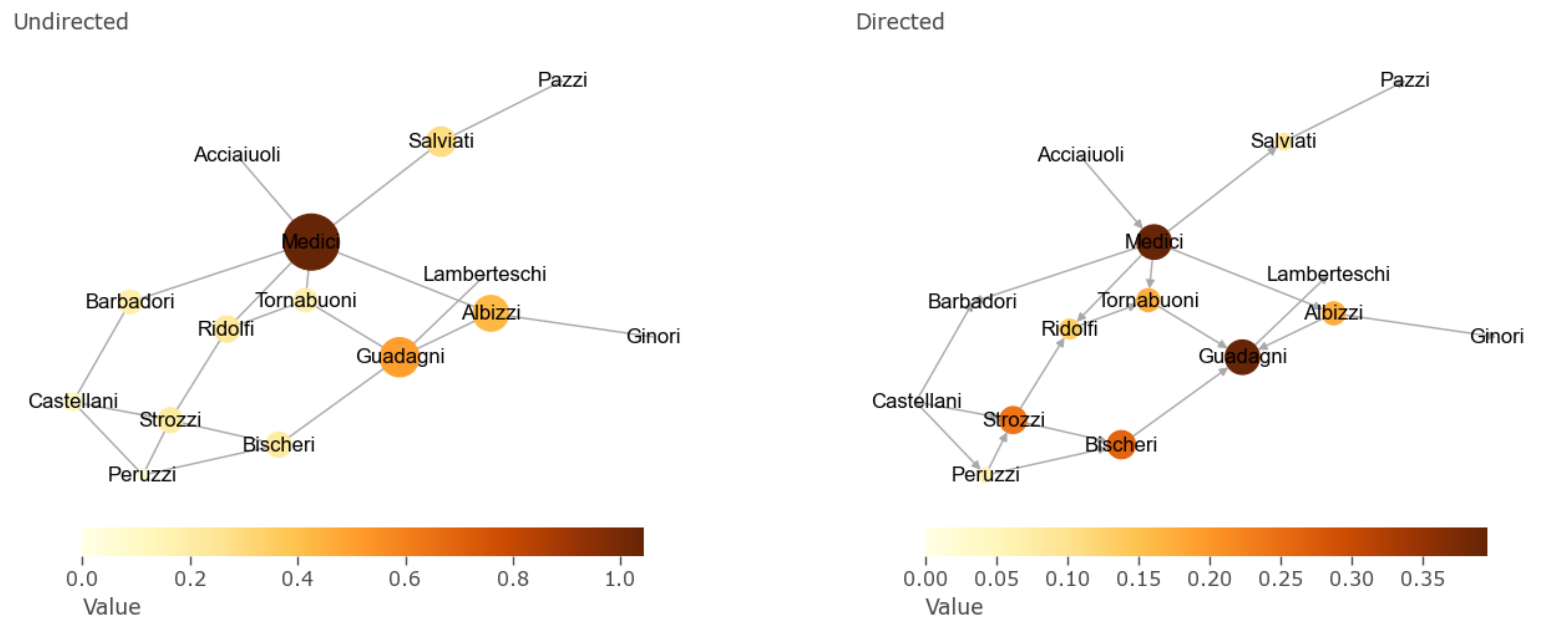
Betweeness centrality = $^1/_{n^2} \sum_{st} n_{st}^i$

n_{st}^i = 1/g if node *i* lies on the *g* shortest paths between nodes *s* and *t*

Assumptions:

- every pair of nodes in the network exchanges messages at the same average rate
- messages always take the shortest available path though the network

Measures **brokerage** in the network ➔ disruption of these nodes = disruption of communication



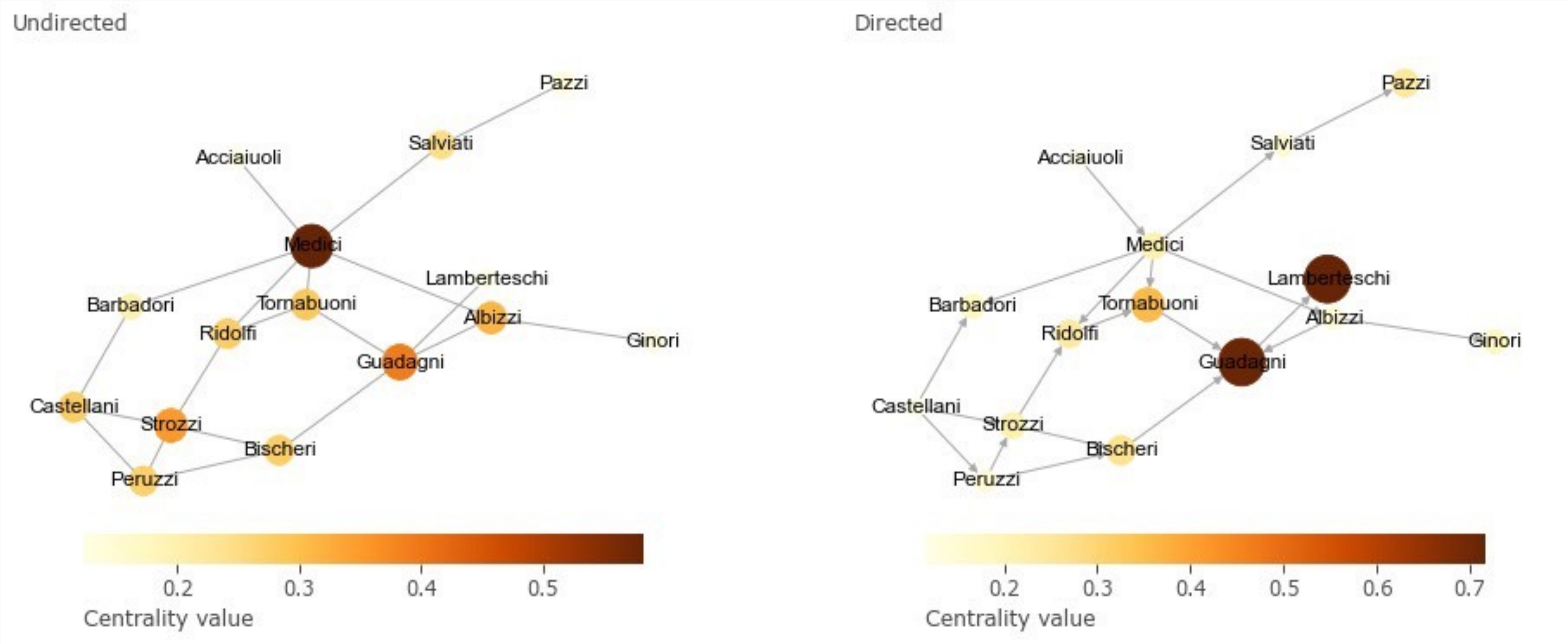
Freeman (1977), and Anthonisse (1971, unpublished)

Pagerank centrality = $(1-\alpha)\sum_j A_{ij}\, {}^{p_j}/_{d_j} + \alpha$

d_j = Degree of node *j*. p_j = Pagerank centrality of node j

Takes into account how central your neighbors are. Each node has a minimum value of α. The pagerank of a node is α plus **the pagerank of your neighbors** (normalized by their out-degree)

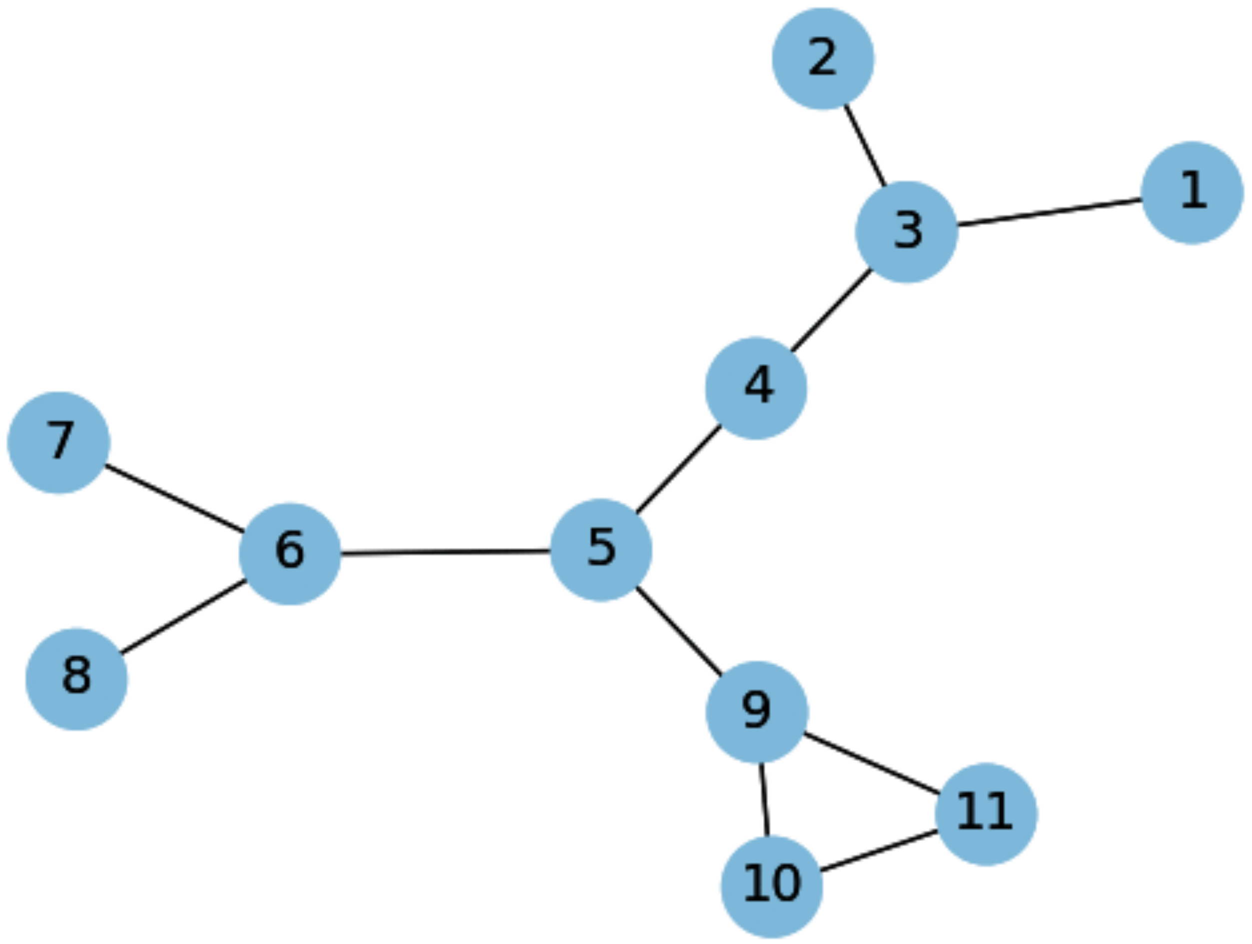
Measures total **influence** in the network (assuming all nodes are the same)



Bonacich, 1987

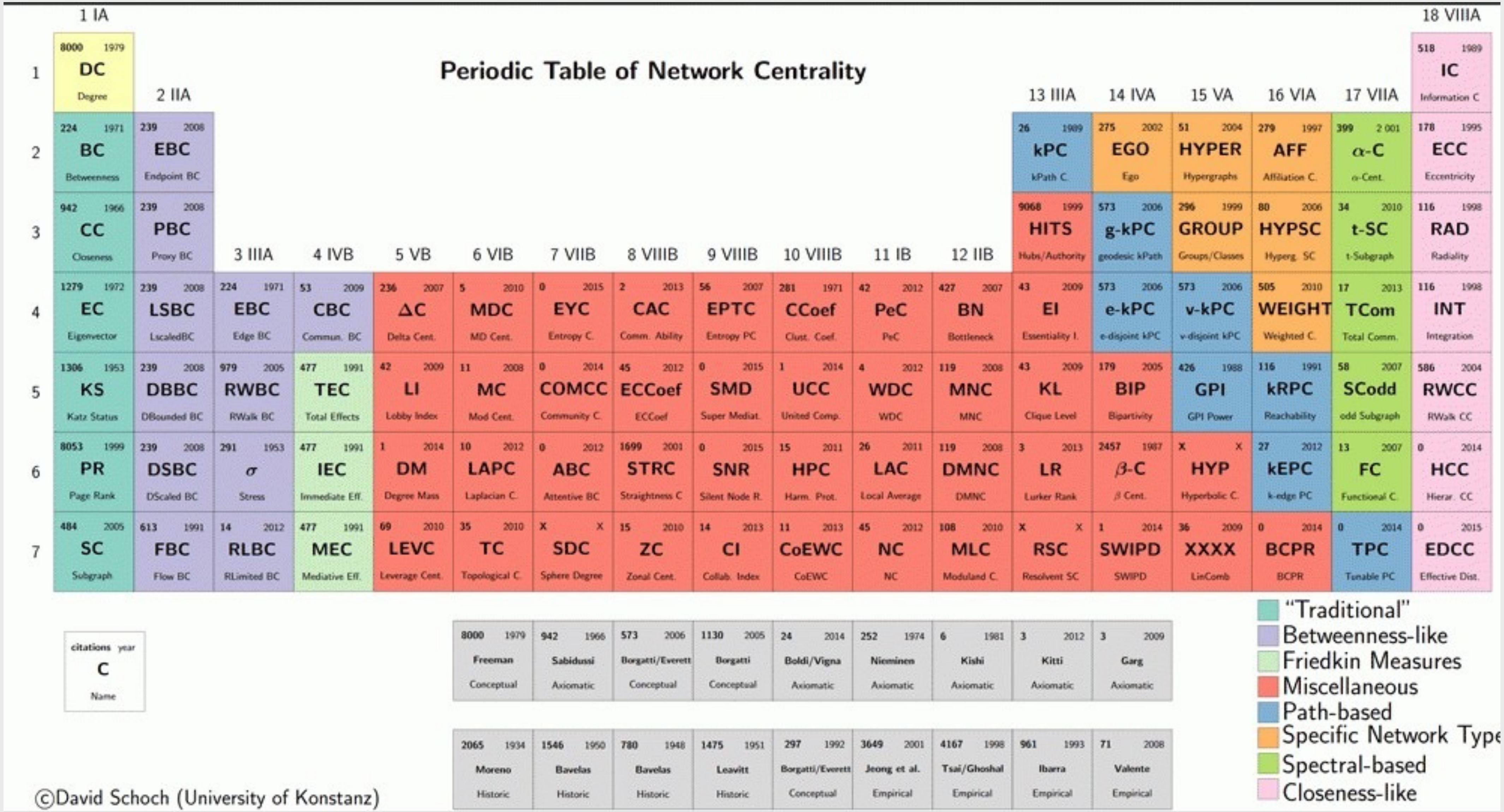
Page Rank

Iteration init



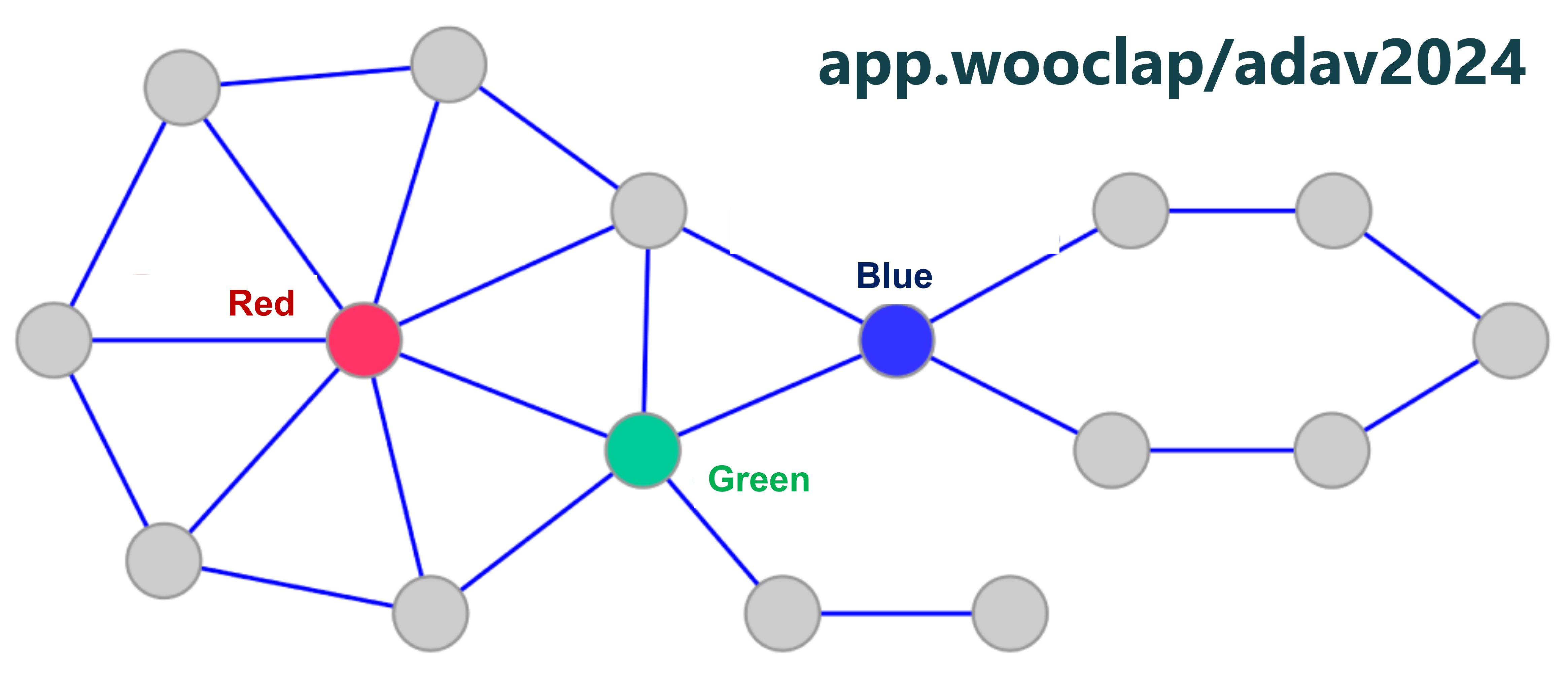
Use a centrality measure that fits your question, not the one that gives you the best results

Consider what is the real objective (e.g. is it to stop a disease or protect specific groups?) *(https://petterhol.me/2019/01/11/the-importance-of-being-earnest-about-node-importance/)*



javier.science/panel_network

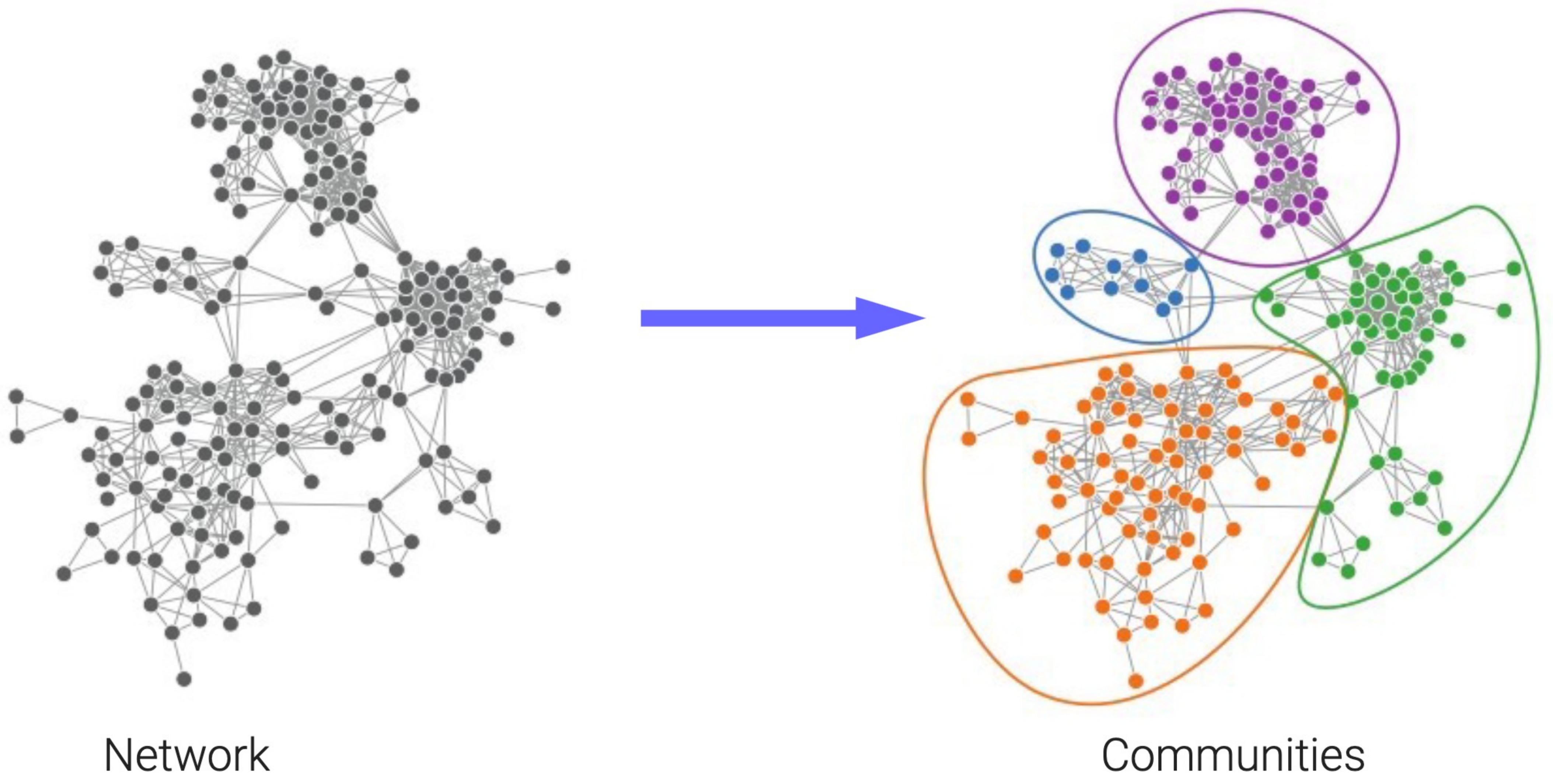
Which node has higher degree/betweenness/closeness?



https://cytoscape.org/cytoscape-tutorials/presentations/modules/network-analysis/index.html

Community detection

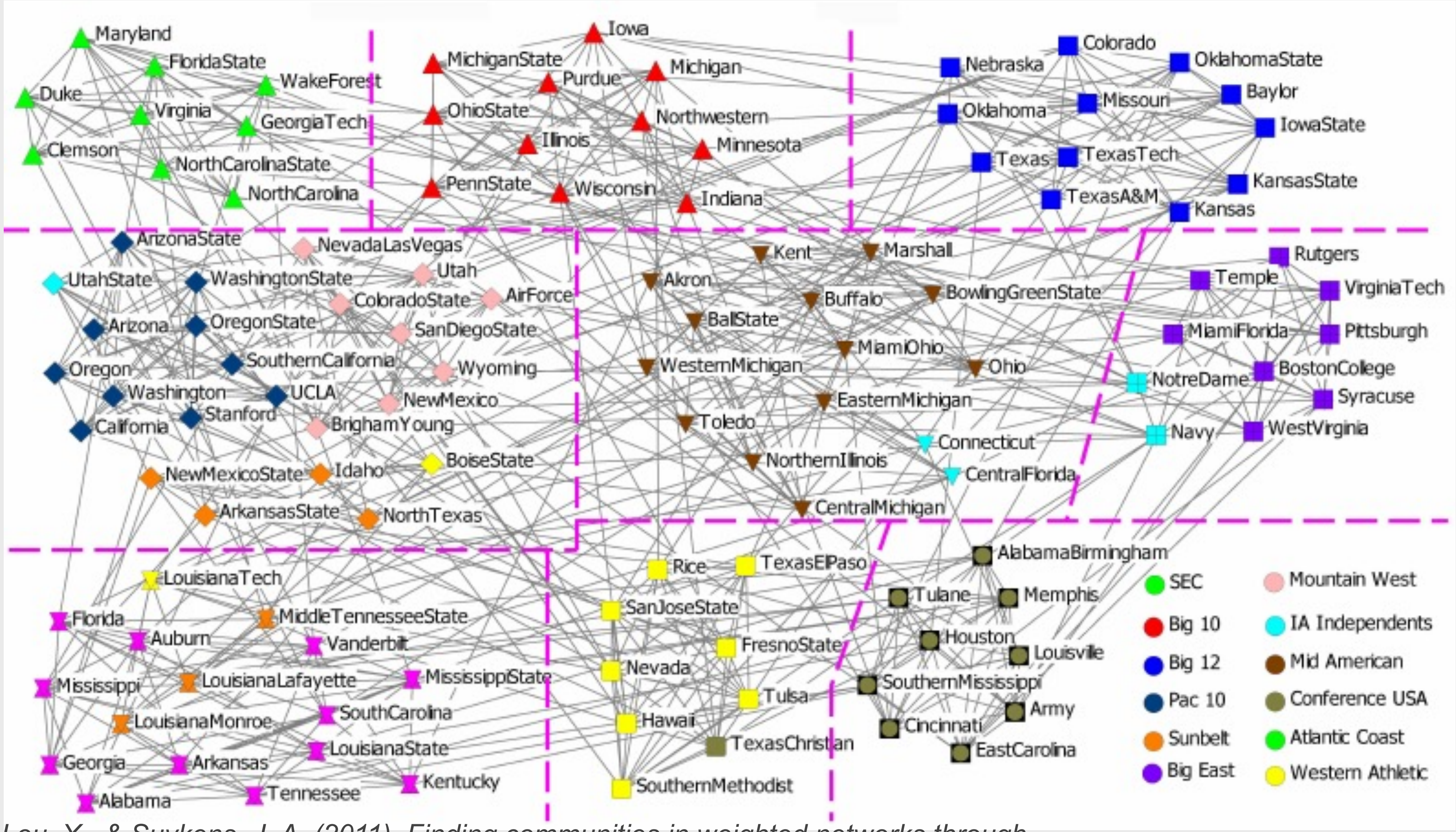
Adapted from the materials of Leto Peel, Network Science Summer School, https://net-science.github.io/


Network
Communities

Community detection

- It’s a type of **unsupervised learning:** We have the inputs (info on the nodes/edges), not the output (the community label)
- We want to learn the outputs with a model that uses some assumptions
- Typical assumption: nodes in the same community have the same type of connectivity pattern
  - E.g. many links within communities and few links across communities

Often we use node attributes to see if the method is working



Lou, X., & Suykens, J. A. (2011). Finding communities in weighted networks through synchronization. Chaos: An Interdisciplinary Journal of Nonlinear Science, 21(4).

But there can be multiple good ways to partition a dataset (e.g. a network)!

"Cluster" these objects

"Cluster" these objects

Red

Not Red

"Cluster" these objects

Cannot Fly

Can Fly

"Cluster" these objects

Transport

Not Transport

"Cluster" these objects

Not Alive

Alive

There may be many good ways to partition a network, some unrelated to the node attributes you have!

How to partition the network?

Many methods

Often: many links within communities and few links across communities

Main example: Create communities to maximize modularity (*Q*)

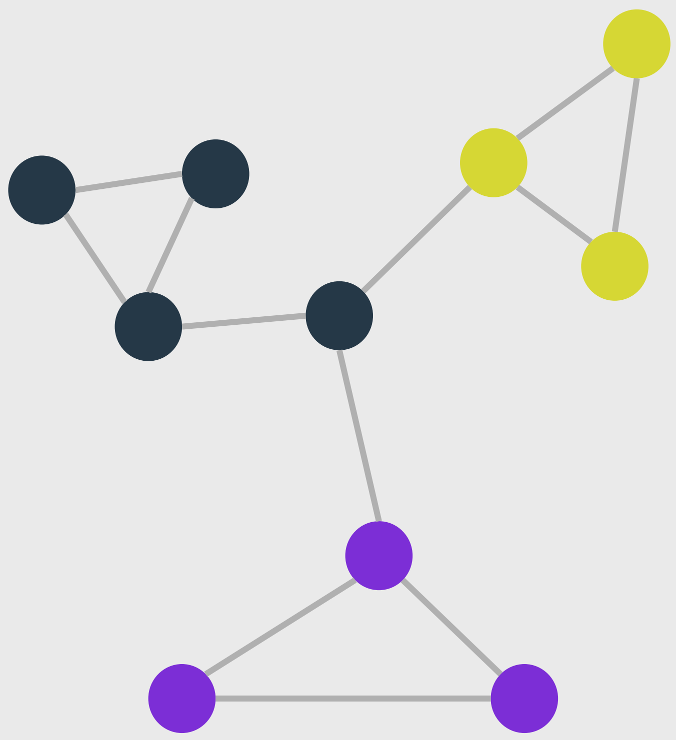
$$Q = \sum_c (e_{cc} - a_c^2)$$

Fraction of links inside community *c* (e_{cc})

Expected fraction of links within a community in a random network (a_c^2)

$$a_c = \sum_{i \in c} k_i / 2m$$

How to partition the network?



$$Q = \sum_c (e_{cc} - a_c^2)$$

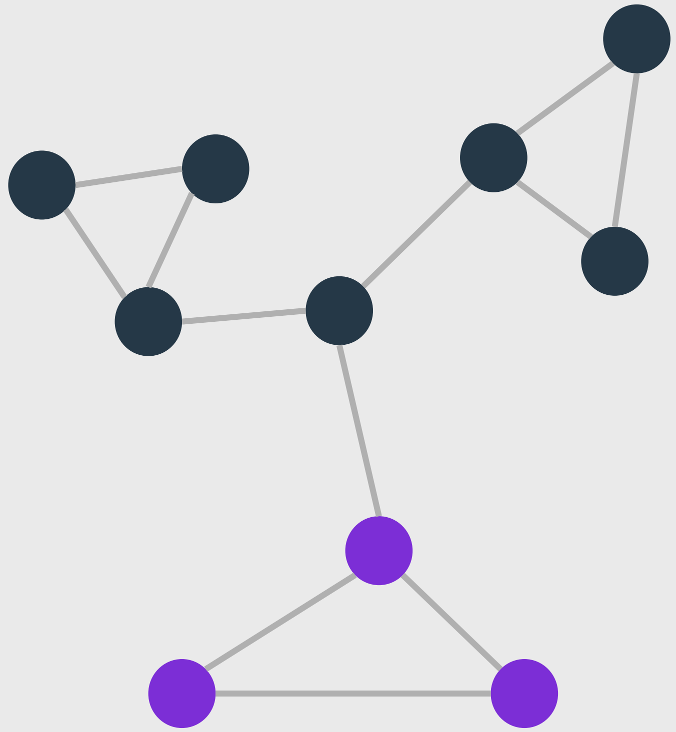
Fraction of links inside community *c*

Expected fraction of links within a community in a random network

$$a_c = \sum_{i \in c} k_i / 2m$$

	E_{cc}	a_c	$E_{cc} - a_c$^2
c=Black	4/12	10/24	0.160
c=Yellow	3/12	7/24	0.165
c=Purple	3/12	7/24	0.165
Modularity			**0.490**

How to partition the network?



$$Q = \sum_c (e_{cc} - a_c^2)$$

Fraction of links inside community c

Expected fraction of links within a community in a random network

$$a_c = \sum_{i \in c} k_i / 2m$$

	E_{cc}	a_c	$E_{cc} - a_c$^2
Black	8/12	17/24	0.165
Purple	3/12	7/24	0.165
Modularity			**0.310**

javier.science/panel_network

How to partition the network?

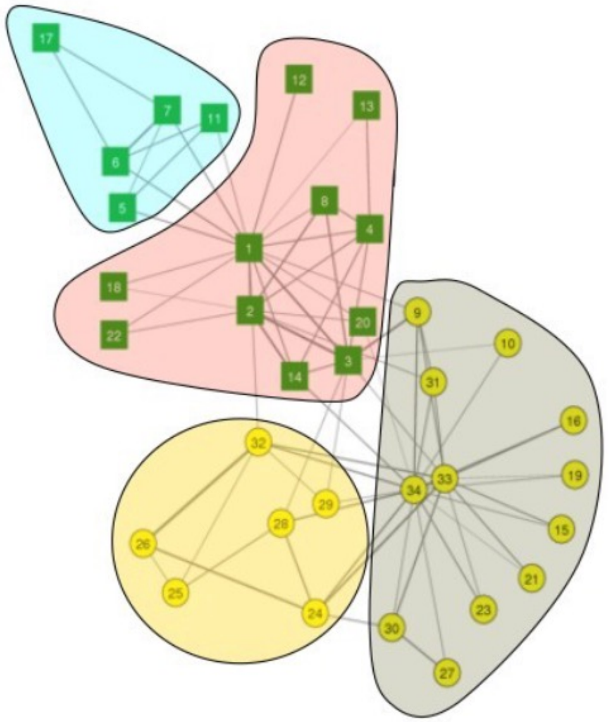
Algorithms for modularity maximization (and related methods)

- Louvain and Leiden algorithms
- Spinglass algorithm (allows to penalize existing and non-existing links differently)

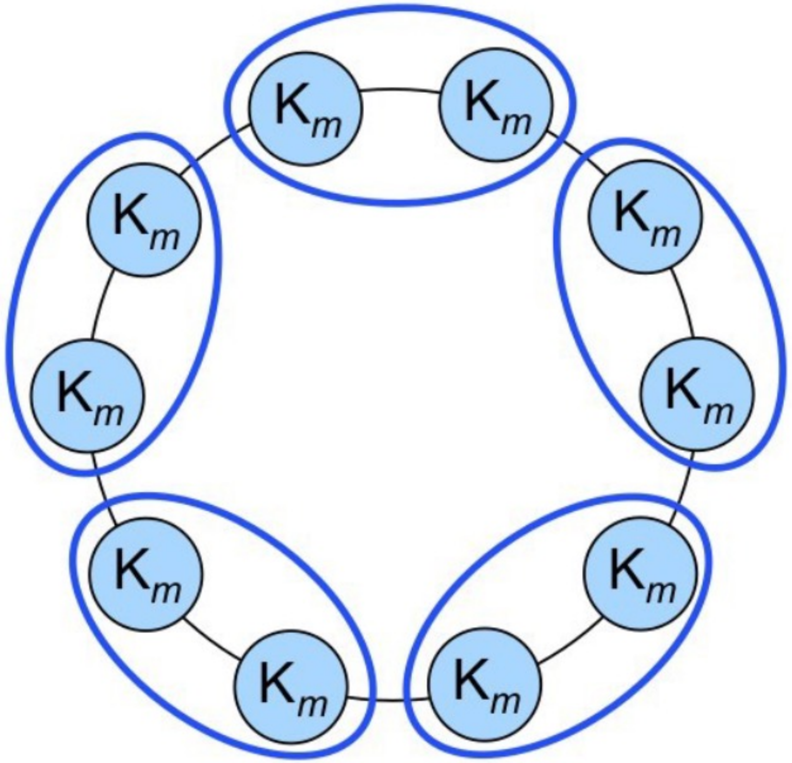
Other algorithms

- Walktrap: Drop many “random walkers” in the network and see how often they visits pairs of nodes in the same walk.
- Label propagation:
  - Each node is initialized with a unique label. Iteratively, each node adopts the label that most of its neighbors currently have.
  - We can add information on some pre-labelled nodes
- Statistical inference: the Stochastic Block Model

Problems with modularity



Finds spurious communities
(overfitting)



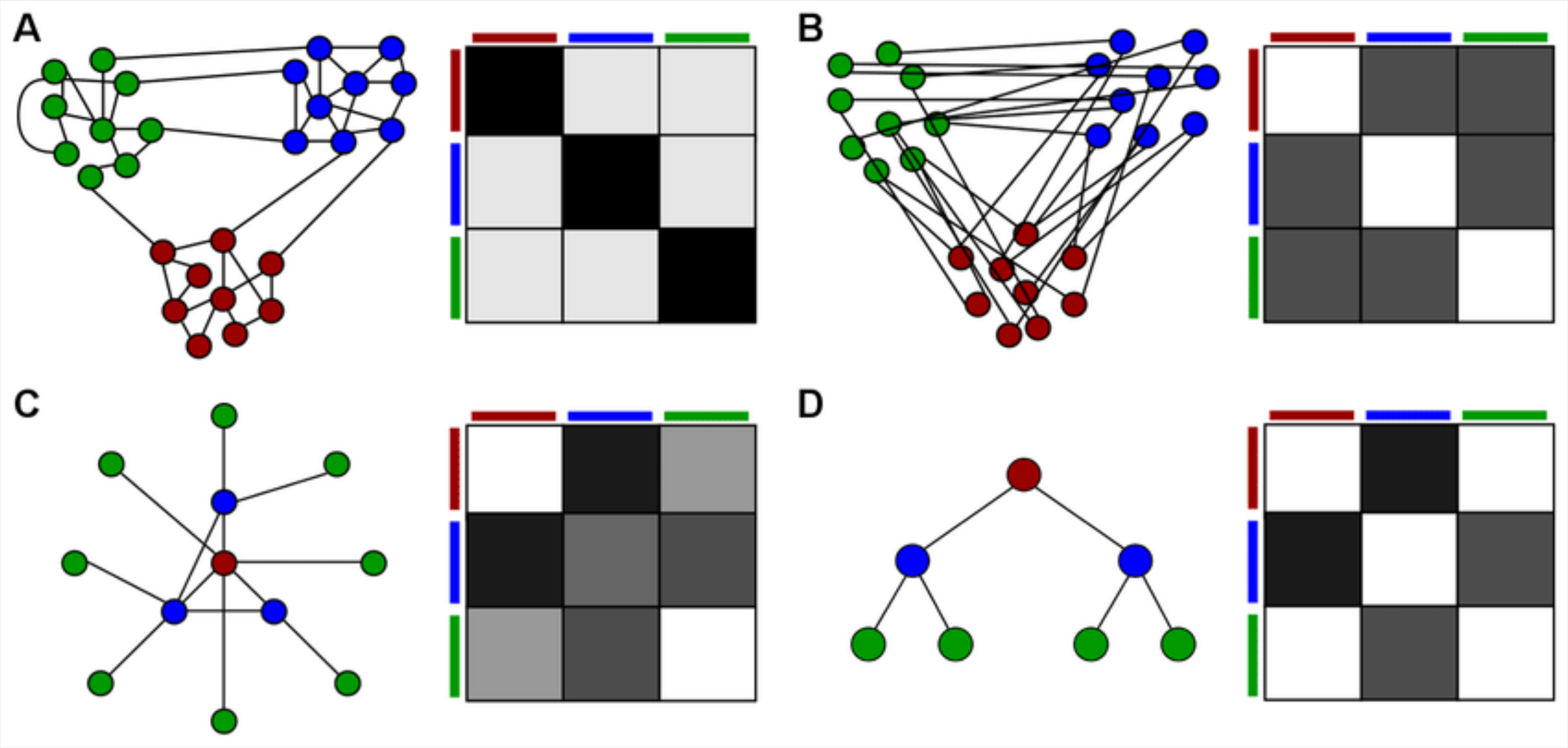
Resolution limit
(underfitting)

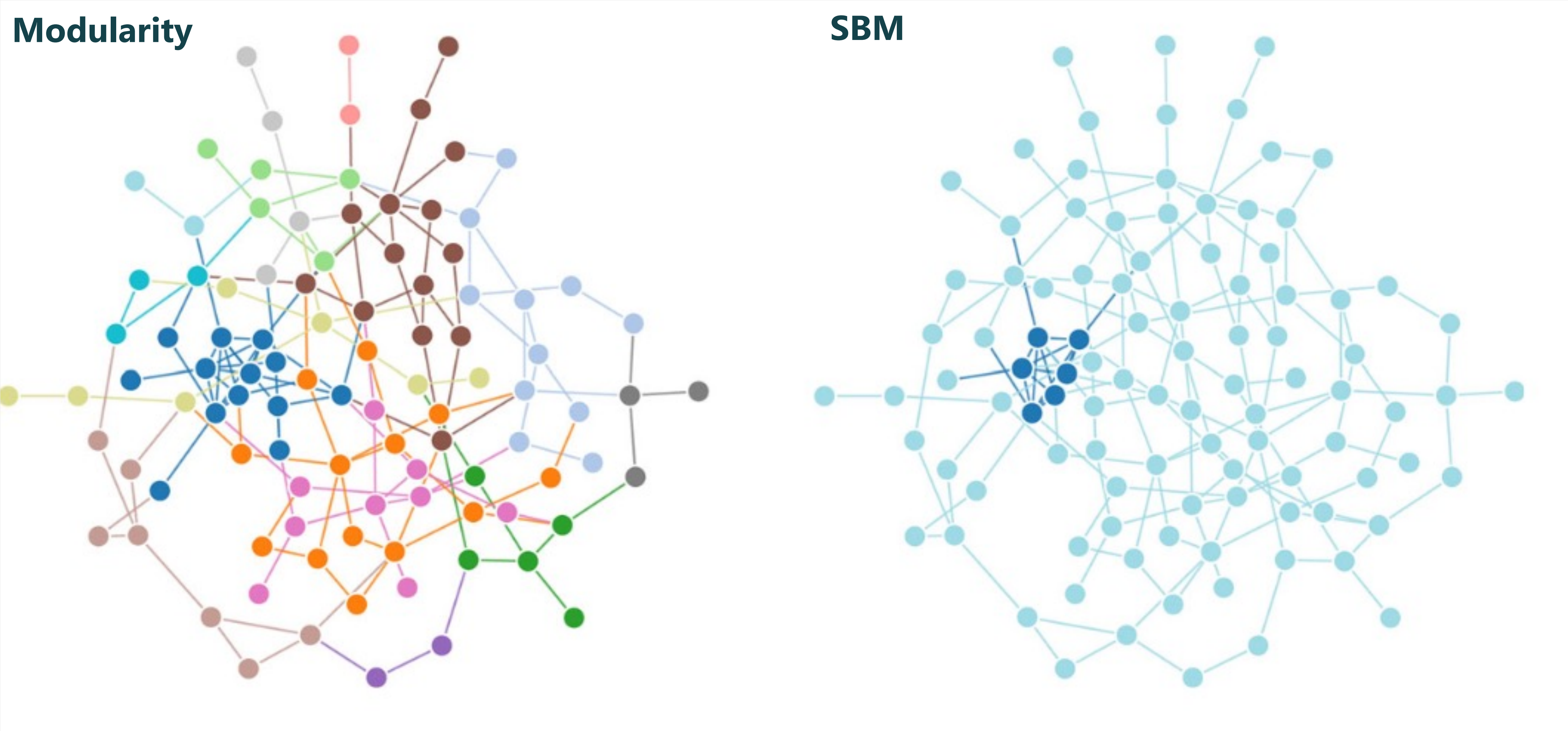
Stochastic Block Model (SBM)

Defines communities as *unique connectivity patterns*, represented in a block matrix.

Can find other connectivity patterns apart from the ones defined by modularity maximization (“many connections within communities, few between communities”)

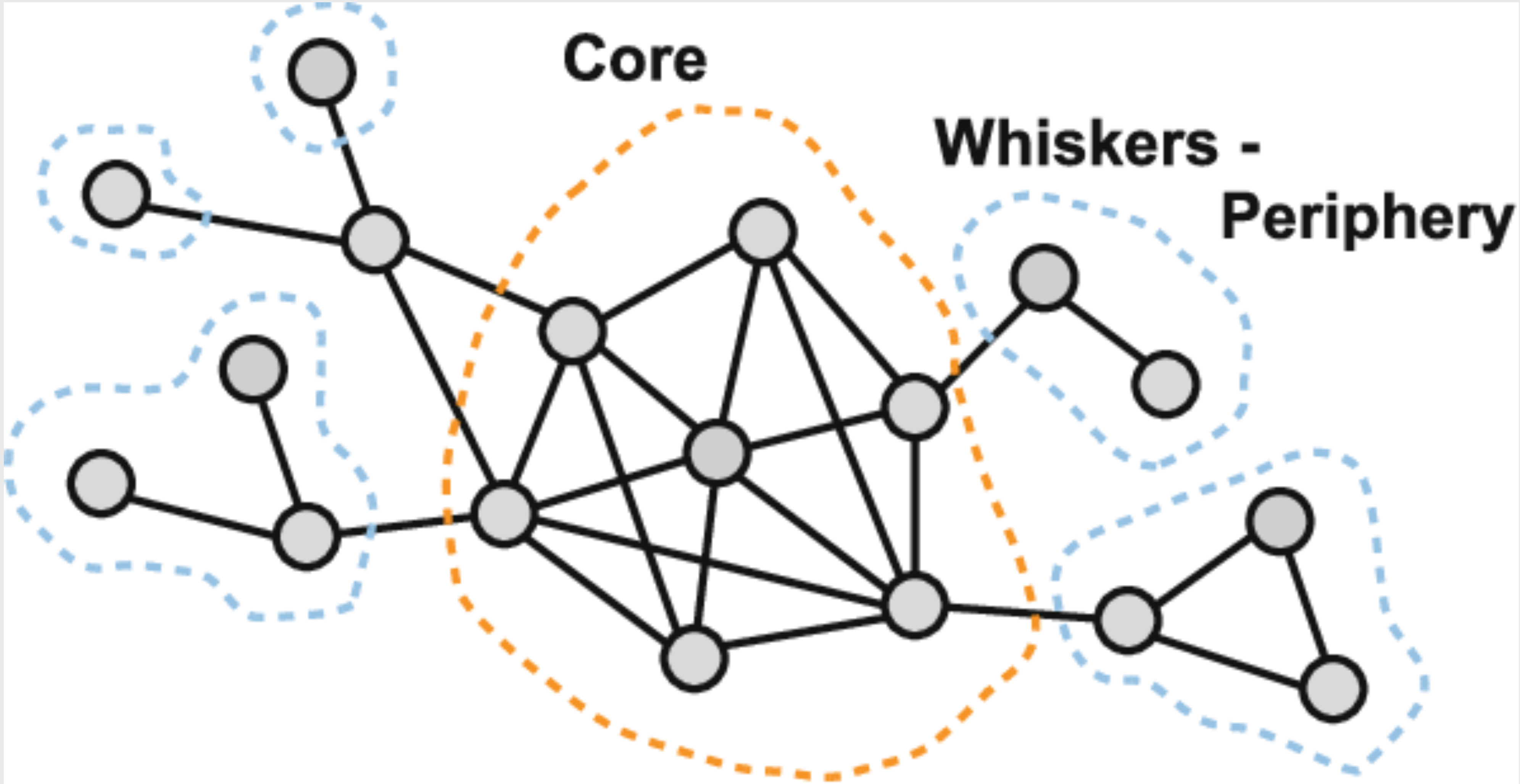
But very complex!





Peel, L., Peixoto, T. P., & De Domenico, M. (2022). Statistical inference links data and theory in network science. *Nature Communications*, *13*(1), 6794.

What communities would modularity maximization find in a core-periphery network? What about SBM?



(Leskovec et al. 2008)

Want to learn more about networks?

Network Science

2024 Utrecht University Summer School

Practical information

- **Dates**: July 15th — July 19nd, 2024
- **Location**: Utrecht University, Science Park
- **Instructors**: Javier Garcia-Bernardo, Leto Peel, Mahdi Shafiee Kamalabad, Elena Candellone, Jiamin Ou, Vincent Buskens
- **Preparation**: Install R, RStudio and Anacondas

Github repository

All slides, code and data can be found here. The lectures and code can also be explored using the links below.

Recap of today

There is important information encoded in relationships

Modeling systems using networks allow us to study that information

We can represent networks using adjacencies matrixes or adjacencies lists

We can **describe networks**: number of edges and nodes, density, assortativity, transitivity, diameter

We can find the most important nodes using **centrality measures**

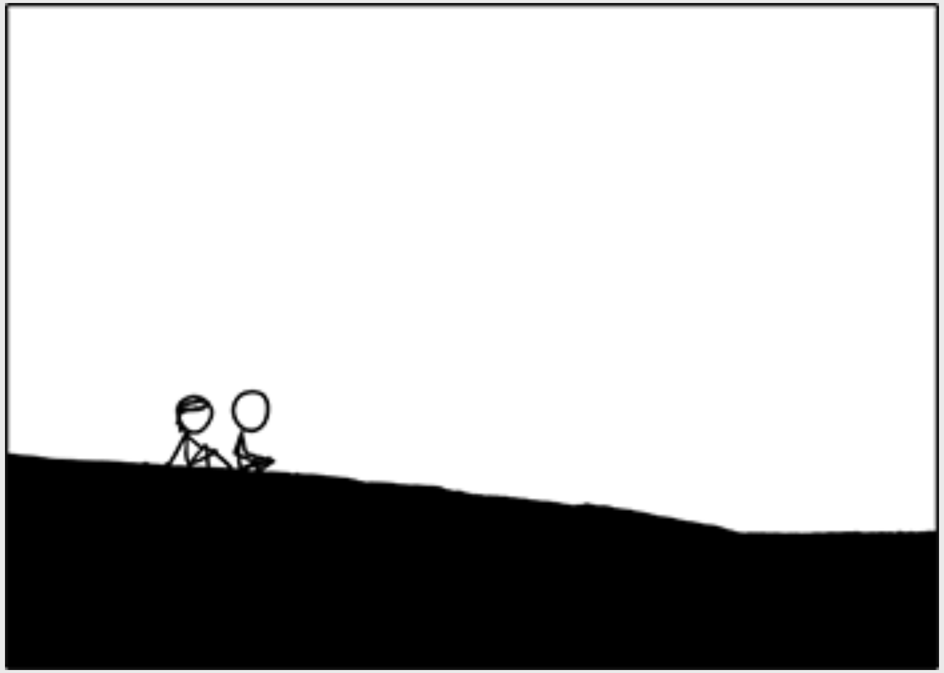
- Different measures define “importance” in different ways: degree, closeness, Pagerank and betweenness

We can find clusters of nodes using **community detection algorithms**

- Modularity maximization: Detect communities with many edges within communities, few edges between
- SBM: Detect communities with unique connectivity patterns

Recap of ADAV

- **Data visualization:**
  - Design principles
  - Interactive / RShiny
- **Building blocks of statistical learning:**
  - Bias vs variance trade-off, underfitting vs overfitting
  - Training/test split
- **Types of models:**
  - Regression: LASSO/Ridge
  - Classifications: Tree-based methods
- **Types of data:**
  - Tabular
  - Text: sentiment analysis
  - Networks: centrality and community detection



Last remarks

Exam: Tuesday June 25th 11.00-13.00 (BETA EDUC)

1. Test moment: Please make sure you can see it on Remindo. Otherwise email Dr. Giachanou.
2. Practice exam: Already available on Remindo. Otherwise email Dr. Giachanou.
3. Special provisions: if you have special provisions, you should have received an email from Dr. Giachanou. If not, please email her again.

Exam review: July 2nd 12.00-12.45 (RUPPERT 002)

Last 10 minutes: Fill the evaluation forms (link in your email inbox)