Network Science Summer School

Universiteit Utrecht

Instructors

Javier Garcia-Bernardo
Methodology & Statistics
Utrecht University

Leto Peel
Data Analytics & Digitalisation
Maastricht University

Mahdi Shafiee Kamalabad
Methodology & Statistics
Utrecht University

Vincent Buskens
Sociology
Utrecht University

Jiamin Ou
Sociology
Utrecht University

Participants

Why are you taking this summer school?

General program

Monday:

Introduction to networks

Network centrality

Tuesday:

Network models (Leto)

Statistical models (Mahdi)

Wednesday:

Community detection (Leto)

Link prediction (Javier & Leto)

Thursday:

Network reconstruction (Mahdi)

Friday:

Dynamics in networks (Vincent & Jiamin)

Day program

09:30–10:00:

Introductions

10:00–11:00:

Introduction to network science

11:00–13:00:

Practical + discussion:

Network tools in Python

13:00-14:00

Lunch

14:00–15:00:

Network representation

Centrality

15:00-17:00:

Practical + discussion:

Centrality measures

Introduction to networks

Network game

Introduce yourself, and find one thing you have in common:

- Countries (apart from the NLD) where you have lived
- Favorite cuisine
- Sports you practice
- Programming languages you use
- …

Draw a line in the whiteboard, write the names in this spreadsheet:
https://tinyurl.com/network-game

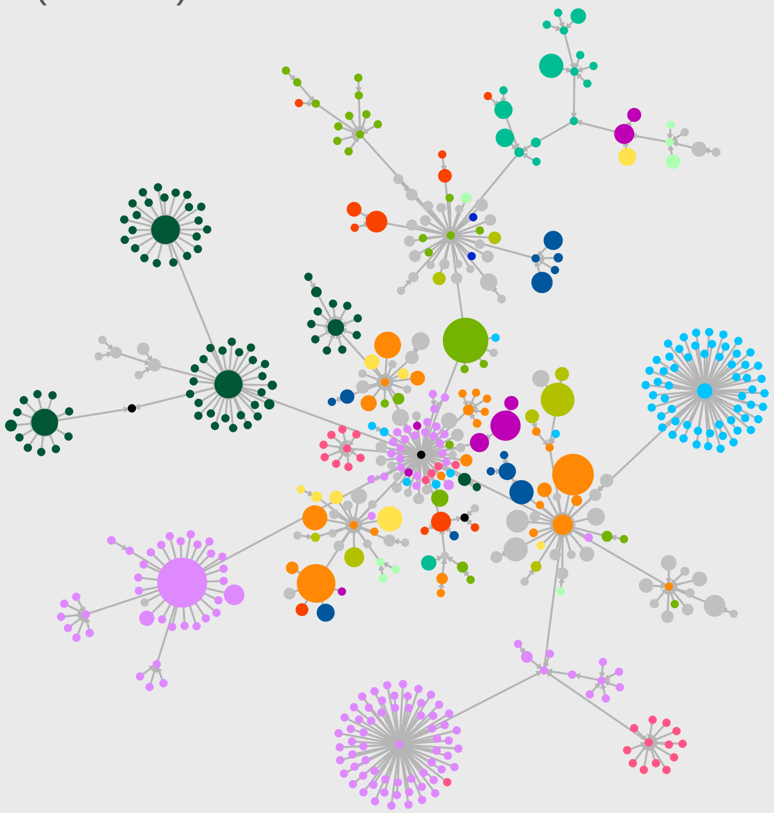
What is a network?

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

The network is not the system, only an abstraction.

The most important question to ask yourself:

What are the nodes and what are the edges?



Types of networks

	Network	Nodes	Edges
Social	Friendship	People	Friendships
	Follower	Online accounts	Followers/likes
	Psychological	Symptoms/Behaviors	Co-ocurrence
Biology	Gene regularory	Genes	Activations/inhibations
	Food web	Animals	Predating
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

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

Type of networks and characteristics

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

- Infrastructure
- Offline interactions
- Online interactions

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

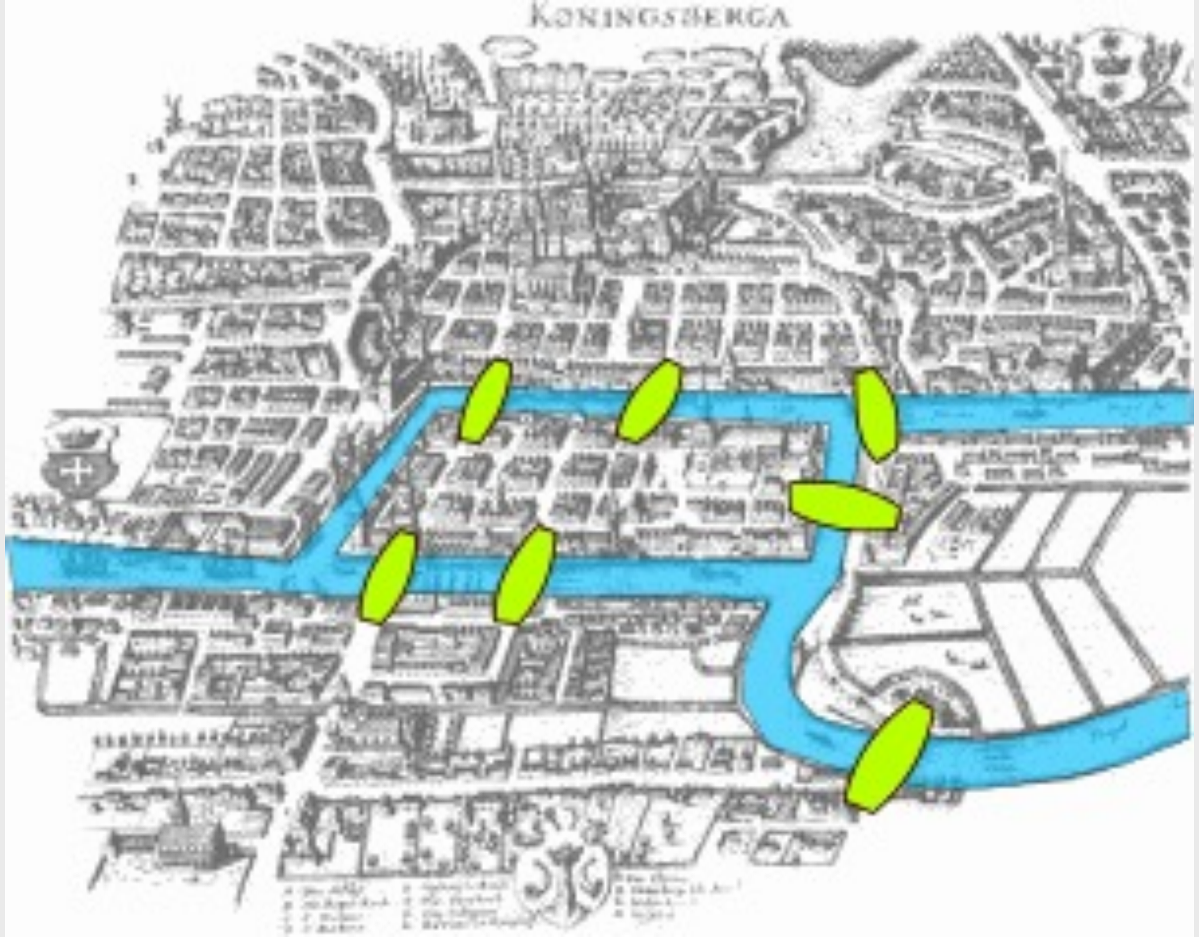
- Formal ties, e.g. most administrative data: e.g., students in classrooms
- Bipartite networks

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

- Stock market networks
- Brain networks
- Proximity networks

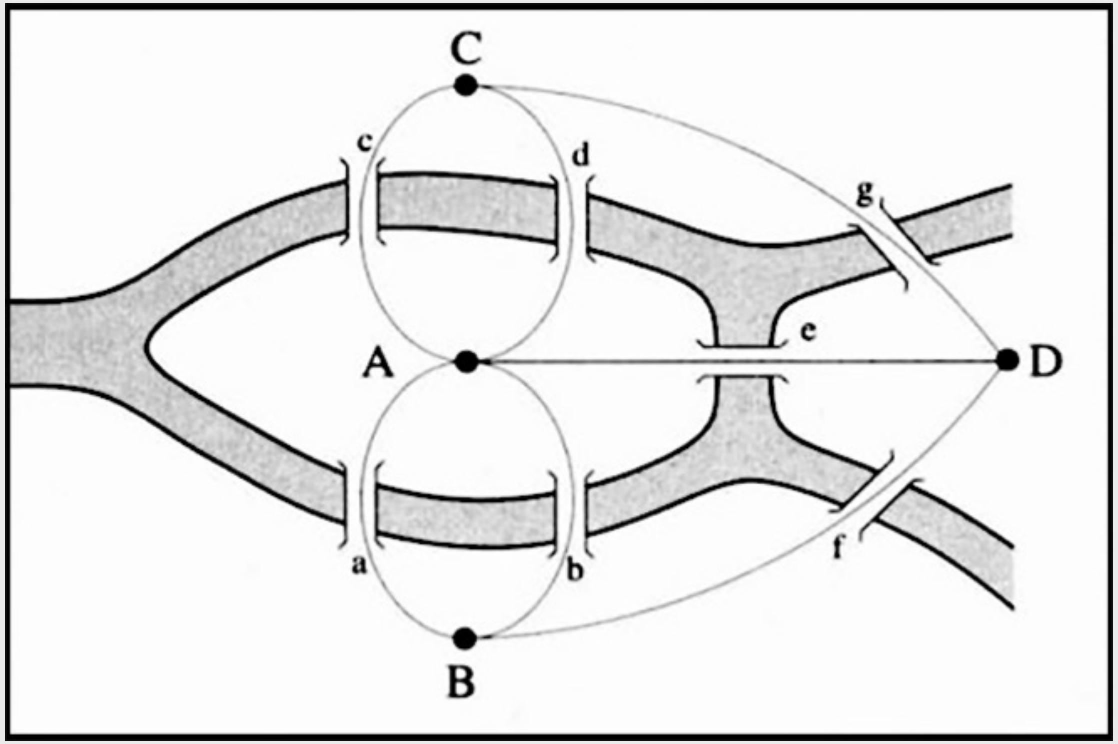
What about family trees?

Bridges of Königsberg

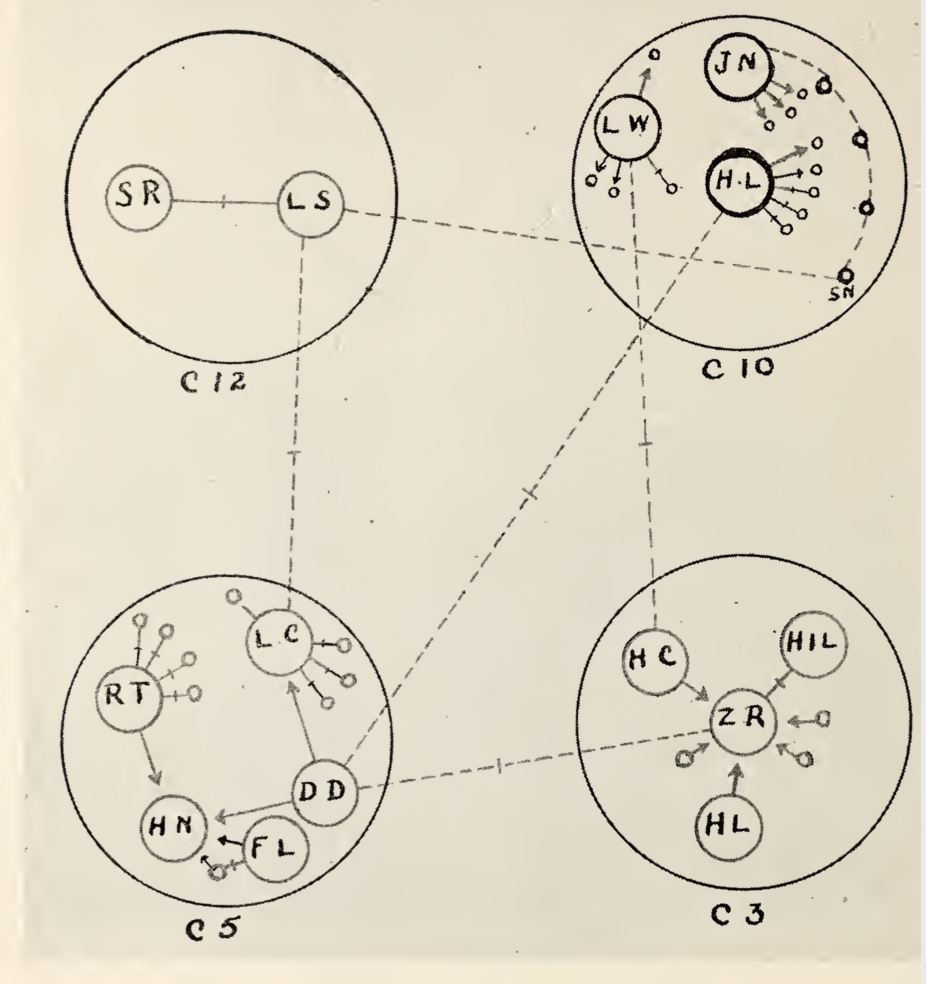


Is there a way you crosses each bridge exactly once and returns to the starting point

Euler (1736)



Brief history of social network science:



Moreno. Who shall survive?

Network science: Social and behavioral scientists in the XX century (e.g. Jacob Moreno & Hellen Hall Jennings, Harrison White, Mark Granovetter)

- Hellen Hall Jennings and Jacob Moreno (1930s): Hudson School for girls: Sociometry. Networks can represent the systems and how information spreads
- Jeffrey Travers and Stanley Milgram's (1969): Small-world studies
- Nancy Howell (1969): *The Search for an Abortionist*, women acquired scarce information through short chains of weak ties.
- Mark Granovetter (1973) *The Strength of Weak Ties*. Diffusion of information takes place primarily through bridges (weak ties). Strong links are redundant.
- Harrison White (1976): Blockmodels for networks
- Duncan Watts, Steven Strogatz (1998): Next wave of network science

Why do we care?

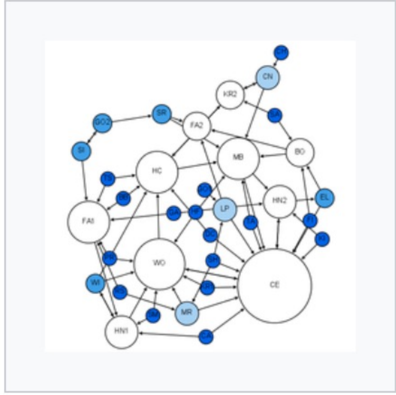
Theoretical links to social science (dangerous generalizations below!):

- Social capital: The position of an individual in their social network (embeddedness) presents opportunities and imposes outcomes.
- Network measures map to social theories: e.g. structural holes and network closure (Burt, 2001)
  - **Structural holes**: social capital is created by a network in which people can broker connections between otherwise disconnected segments ~ betweenness centrality
  - **Network closure**: social capital is created by a network of strongly interconnected element ~ transitivity

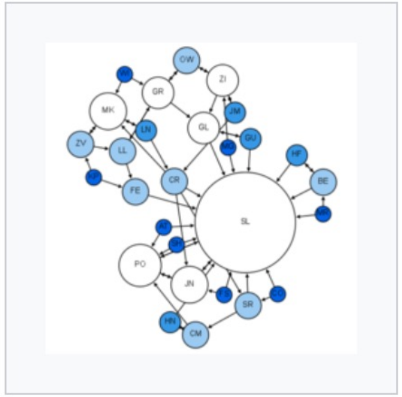
Networks:

- More than the aggregation of dyadic ties
- Reflect preferences (**selection**)
- **Influence** us: spread of information, diseases, opportunities

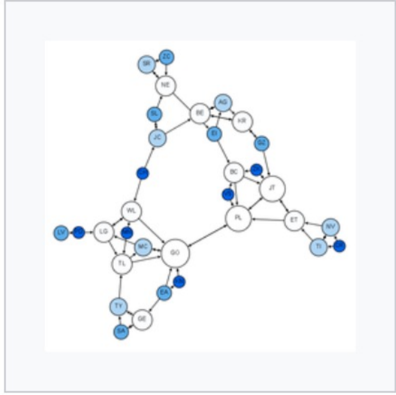
Why do we care?



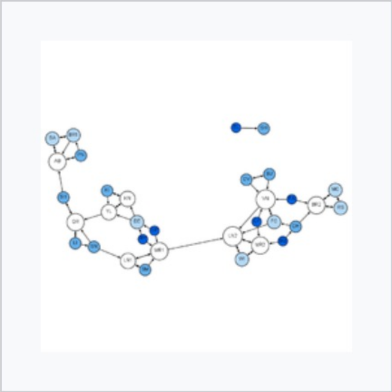
1st Grade



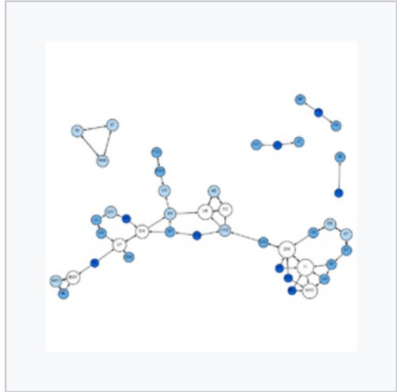
2nd Grade



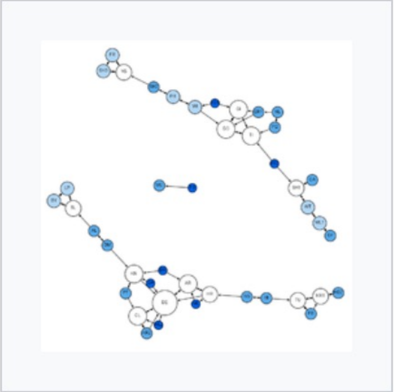
3rd Grade



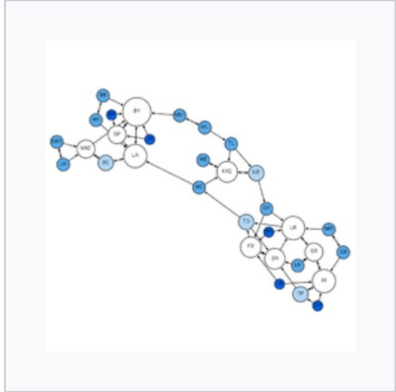
4th Grade



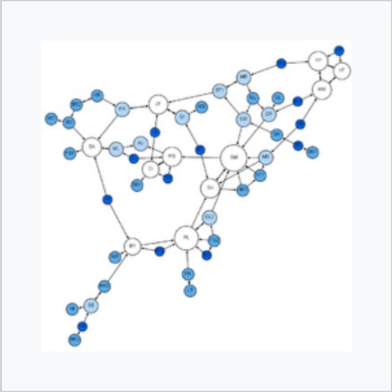
5th Grade



6th Grade



7th Grade



8th Grade

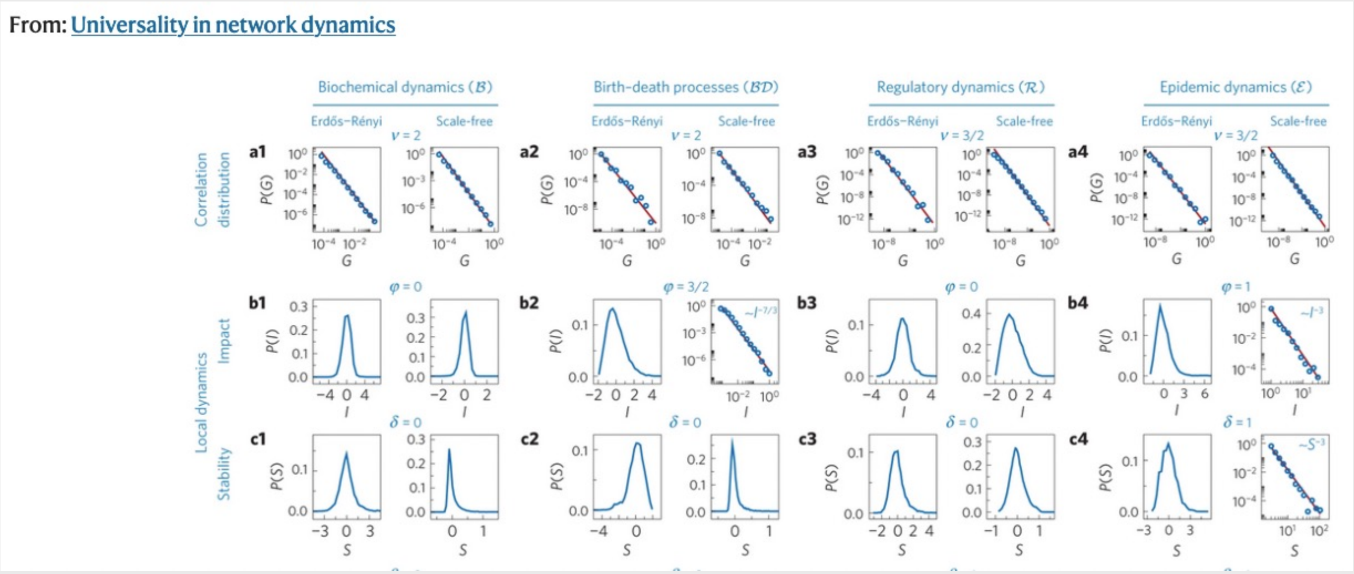
Moreno, source: wikipedia

Why do we care?

Find insights that we would miss if we would study the nodes independently (one person != society)

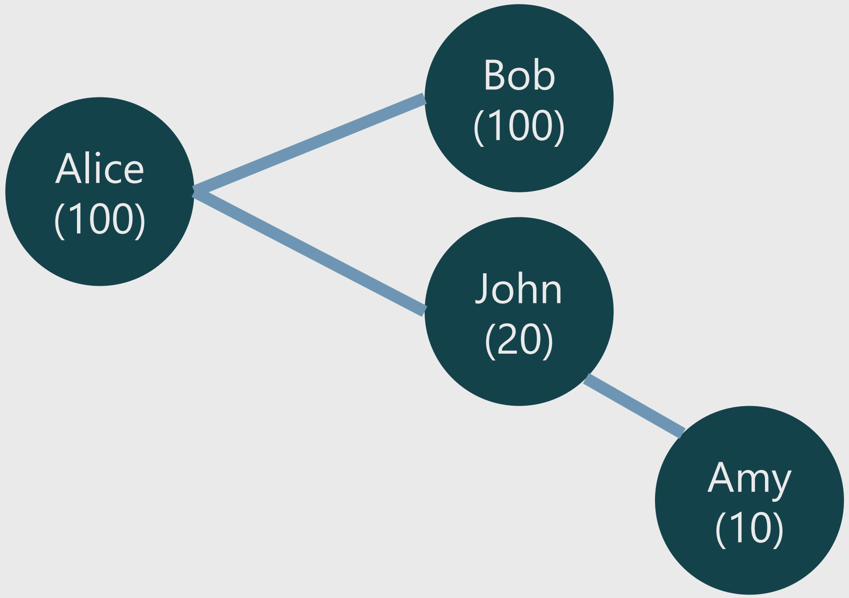
Complex systems view (dangerous generalizations below!):

- Network structure determines how information/epidemics spread (contagion)
- Interested in emergent behaviours:
  - Universality / scale-invariance (heavy tails) / fractals
  - Phase transitions and percolation



Basic definitions

Networks (graphs)



Nodes (vertices) connected by **edges** (links)

N: **Nodes** = {Alice, Bob, John, Amy}

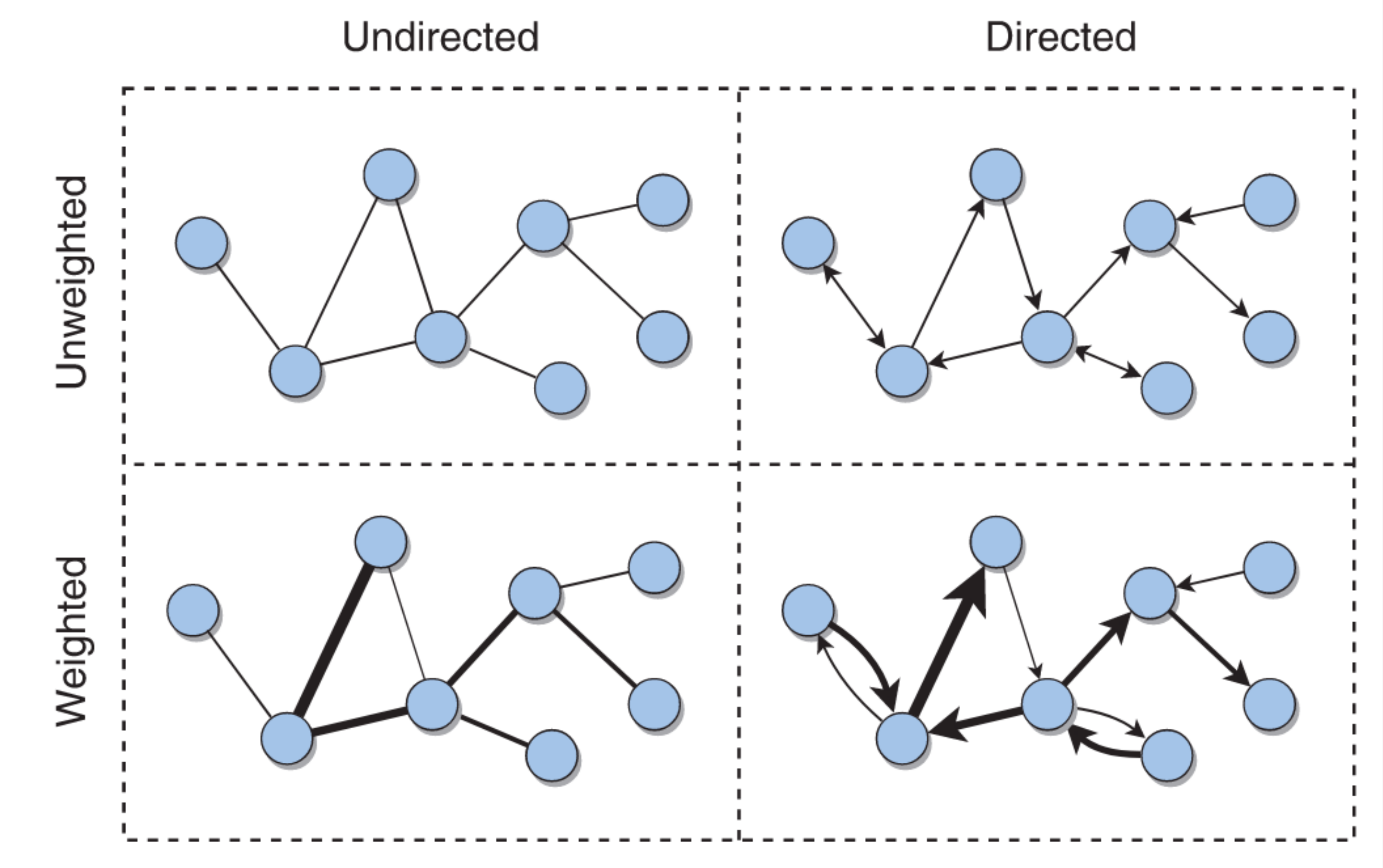
E: **Edges** = {(Alice, Bob), (Alice, John), (John, Amy)}

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

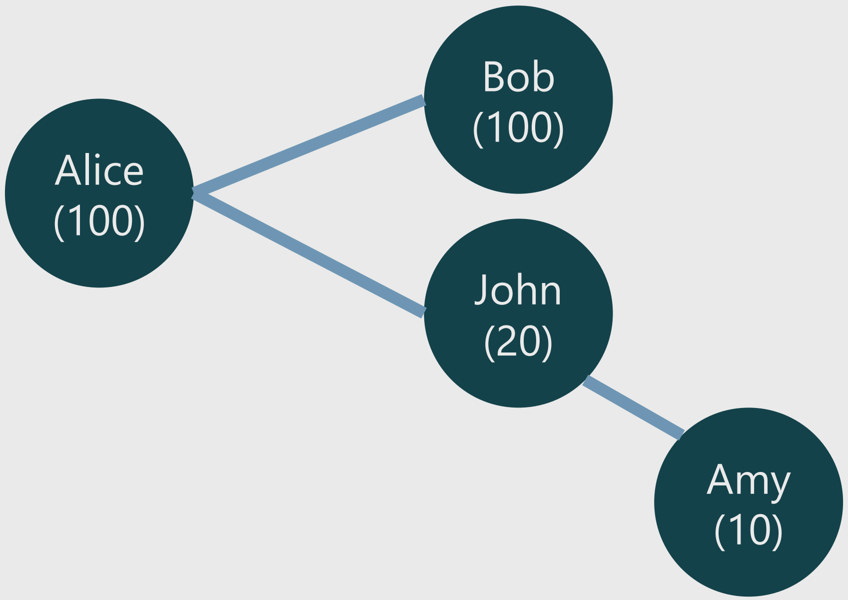


Source: A first course in network science (2020)

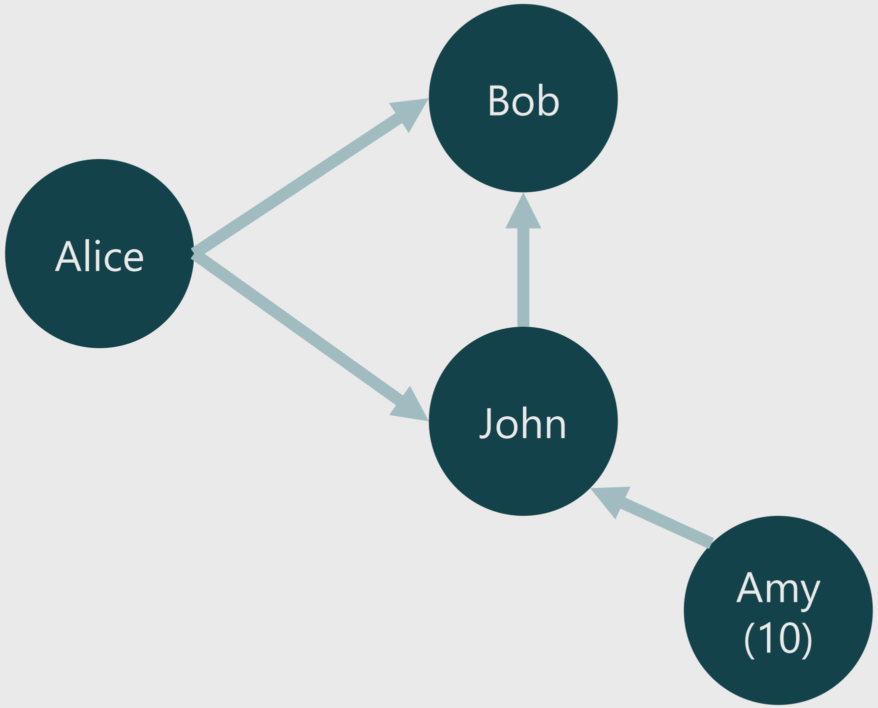
Undirected: The link (i,j) connects node i to node j in both directions

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

Weighted: There is a weight associated to each edge



Undirected Acyclic Graphs (Trees)



Directed Acyclic Graphs (DAGs)

Degree in undirected networks

Definition: Number of neighbors in the network

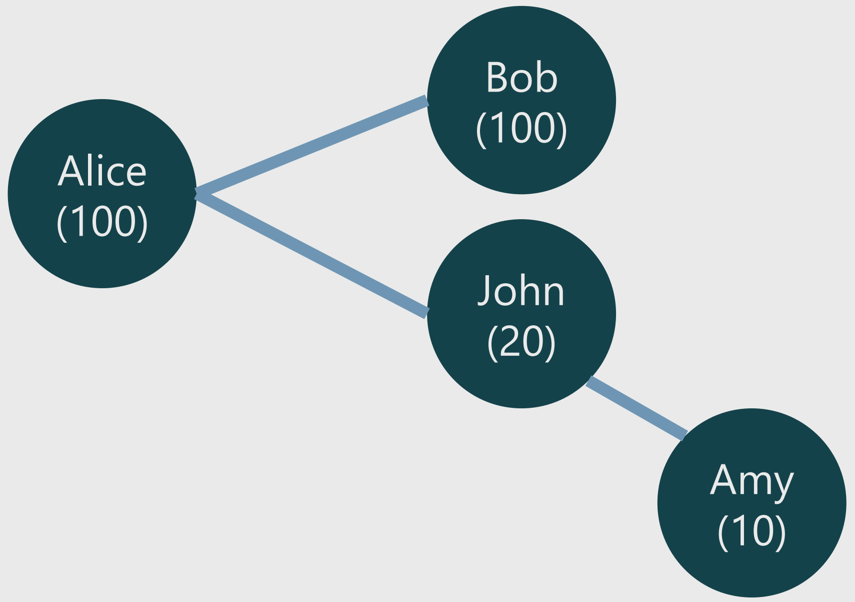
Node: degree

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

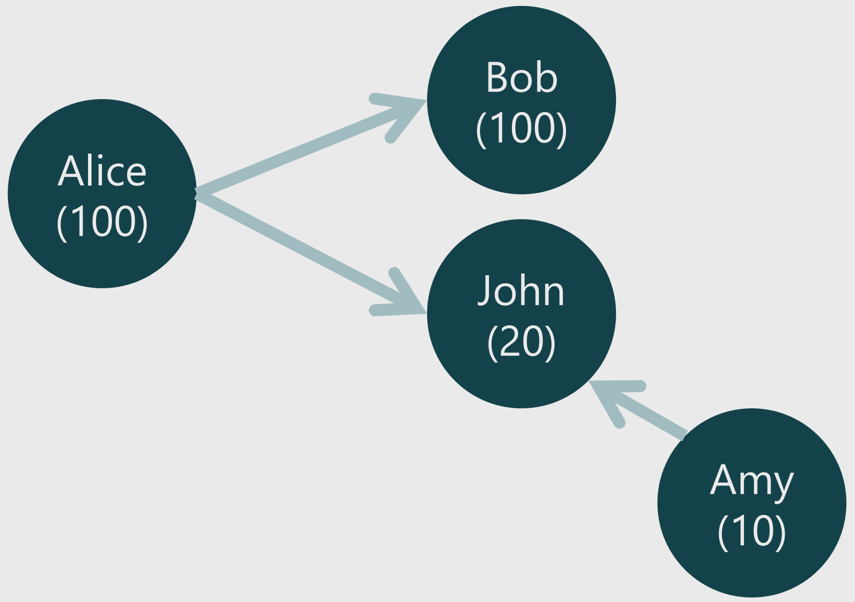
Node: (out, in, total)

Alice: (2, 0, 2)

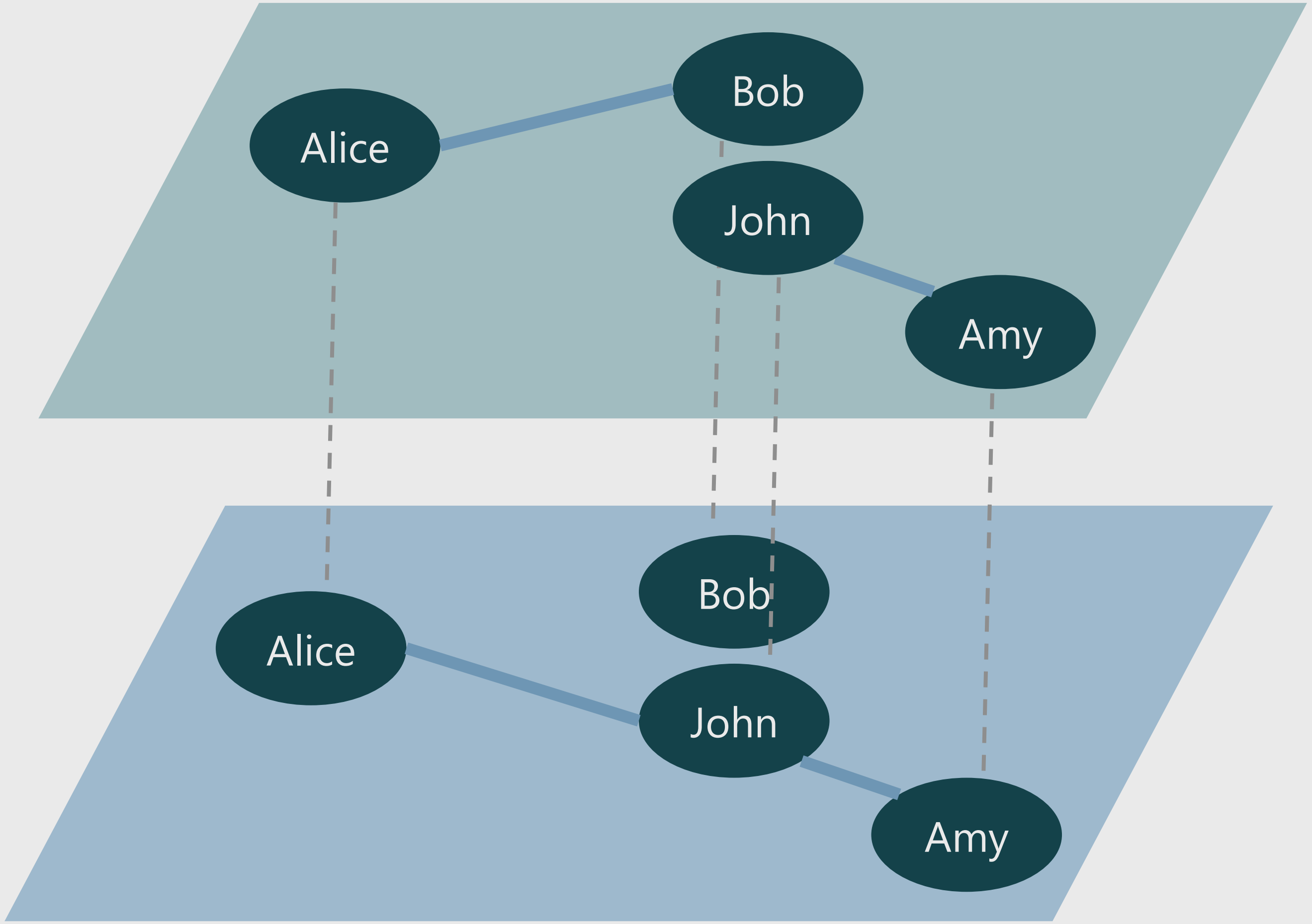
Bob: (0,1, 1)

John: (0,2, 2)

Amy: (1,0, 1)

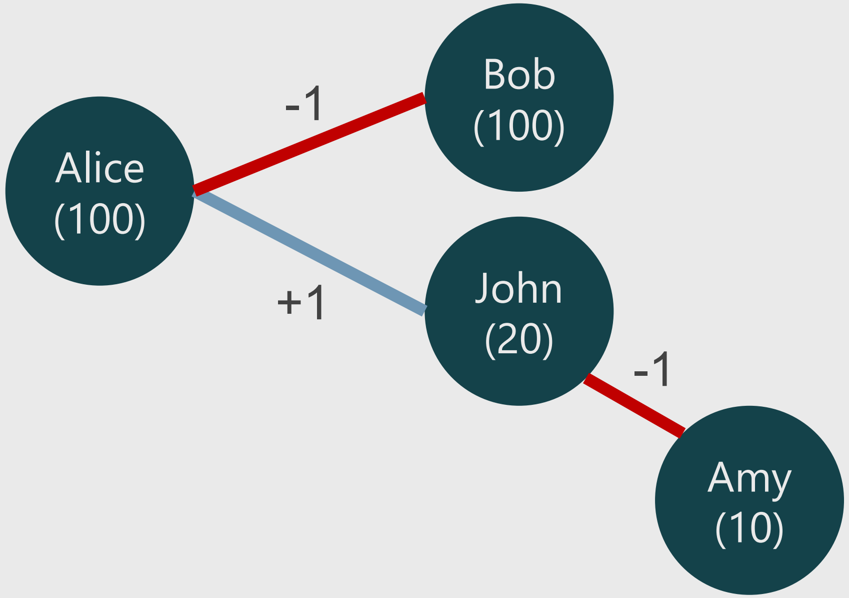


Other types of networks: Multiplex



Other types of networks: Signed

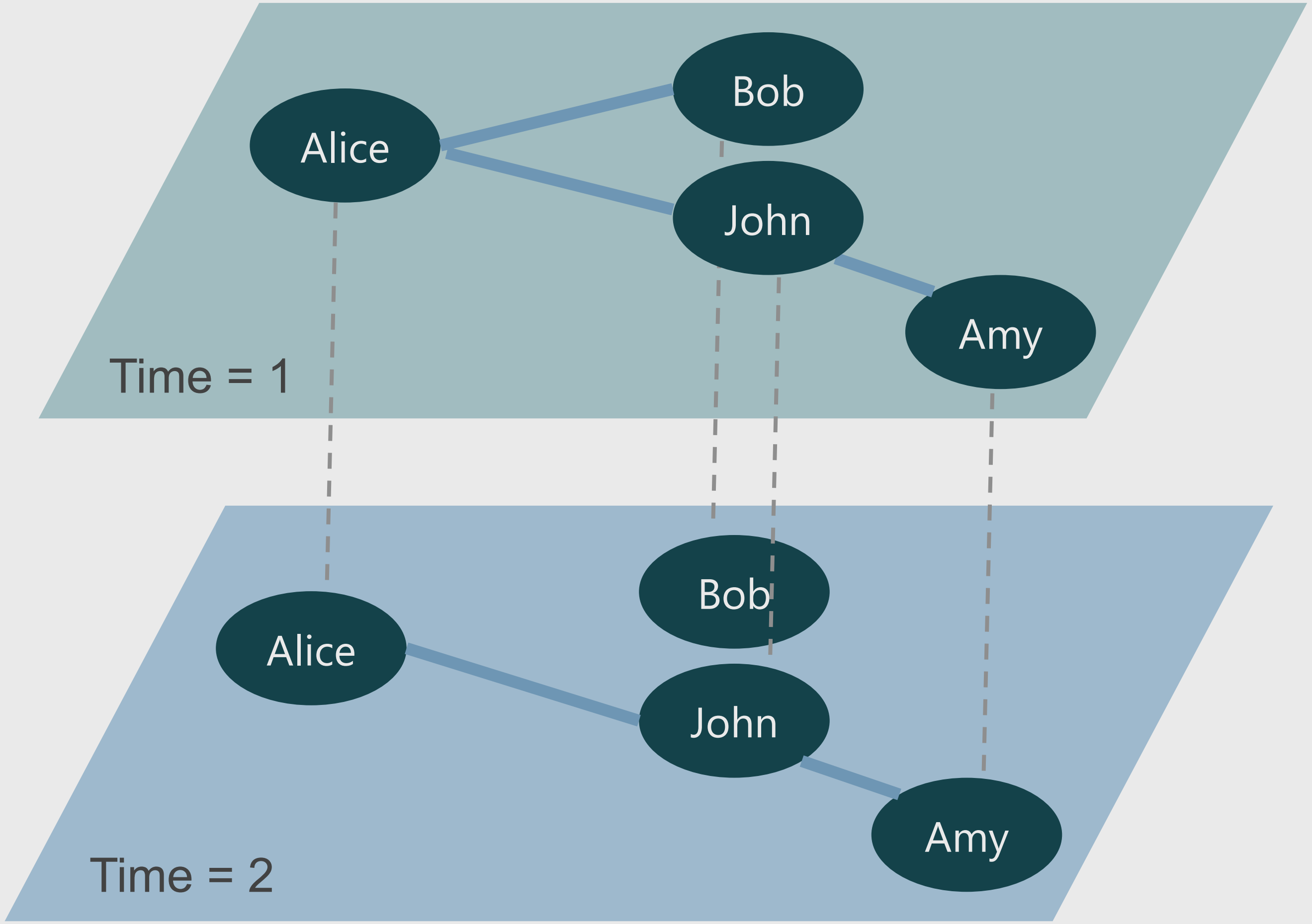
Structural balance / cognitive theory of link creation



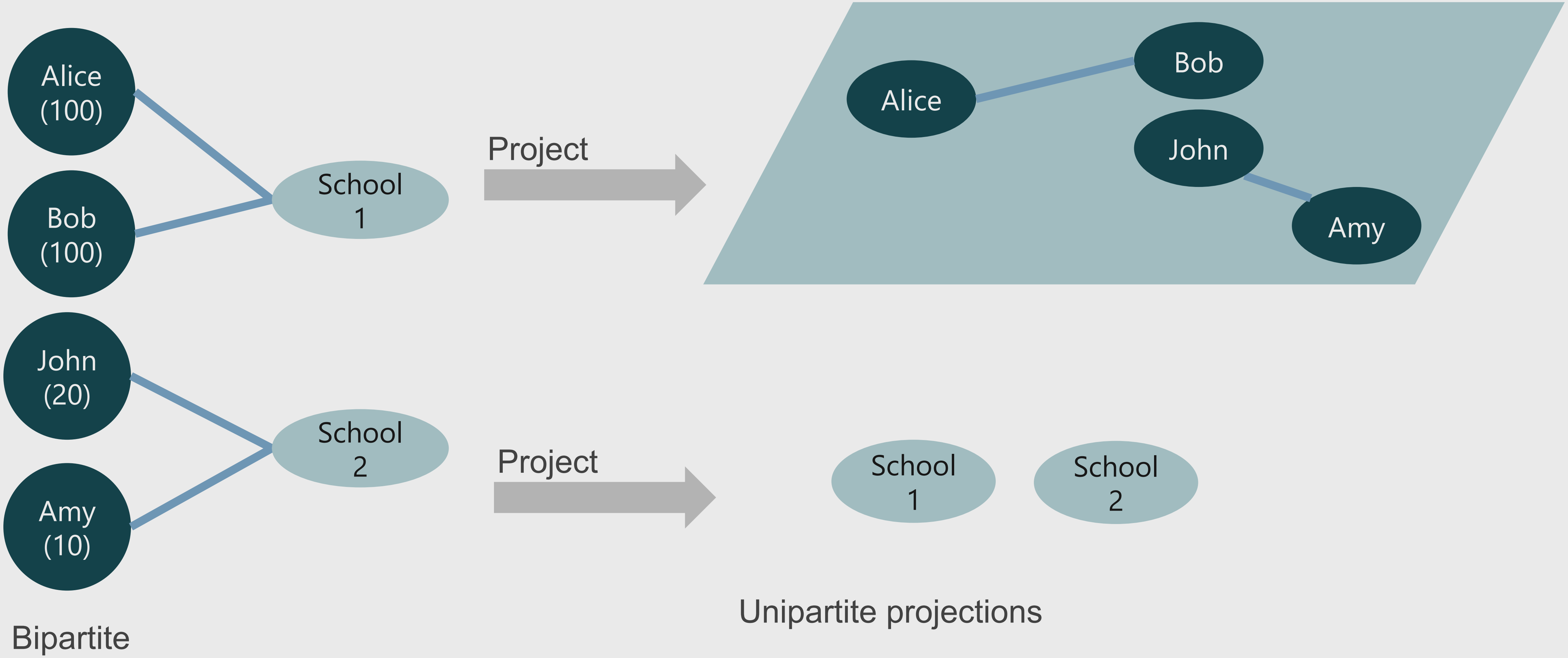
Other types of networks: Temporal

Either:

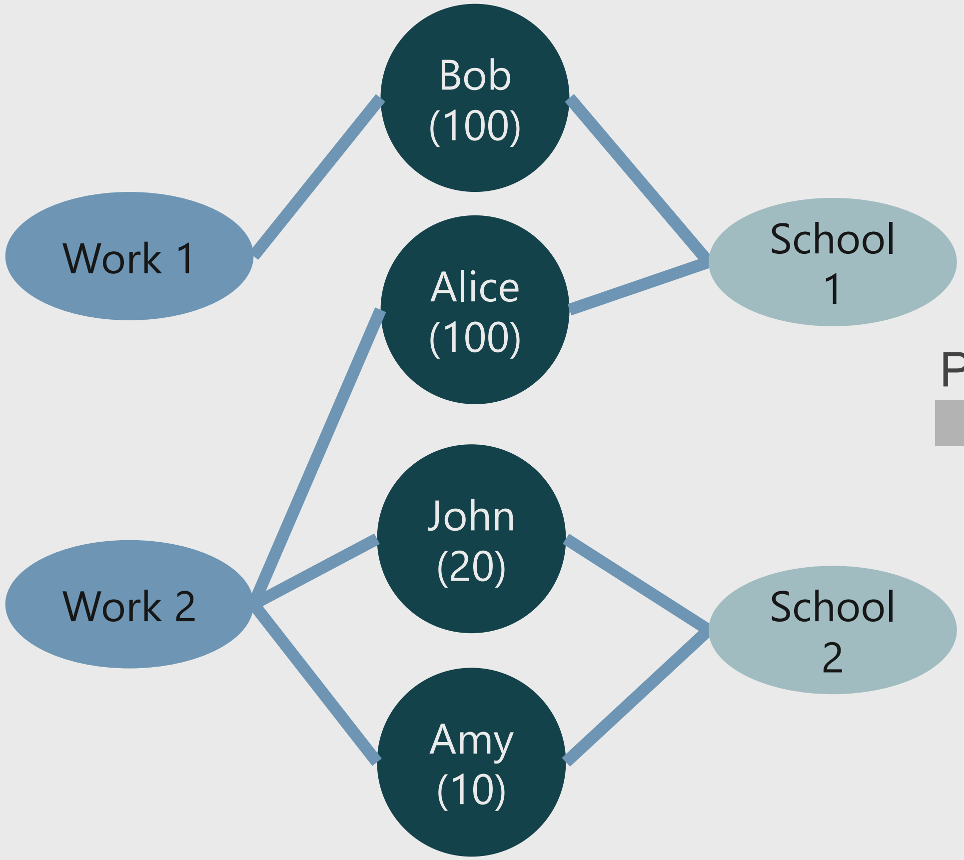
- Snapshots
- Time of events



Other types of networks: Bipartite

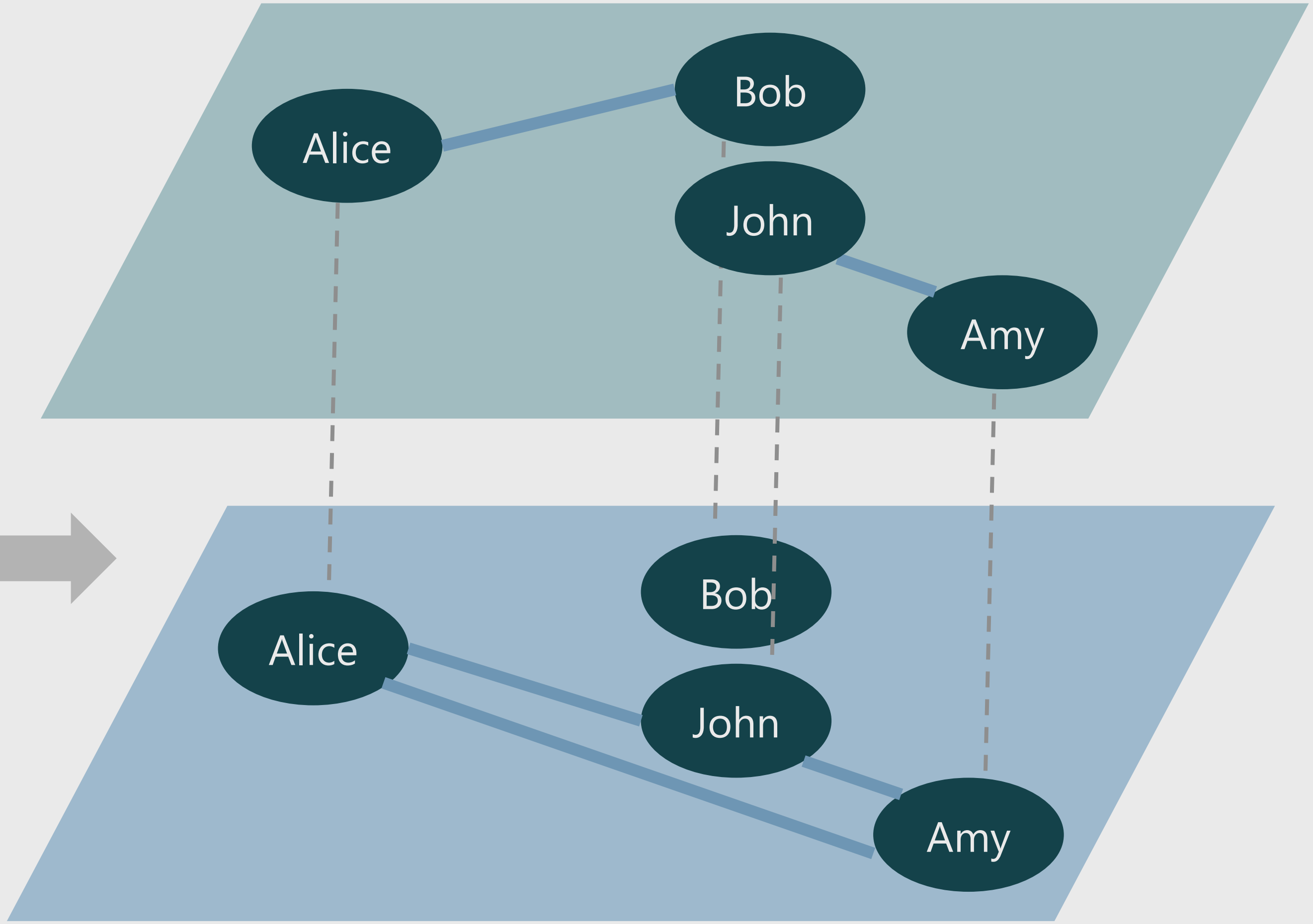


Other types of networks: Multipartite



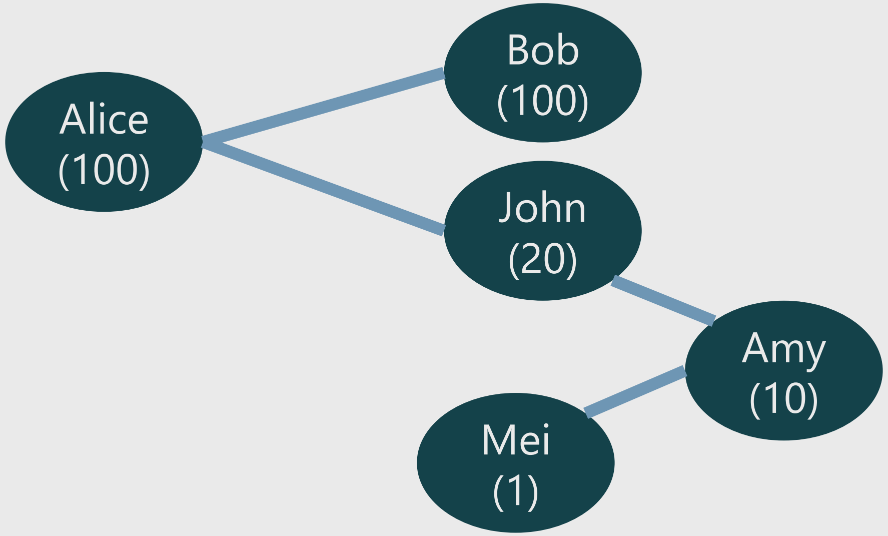
Multipartite network

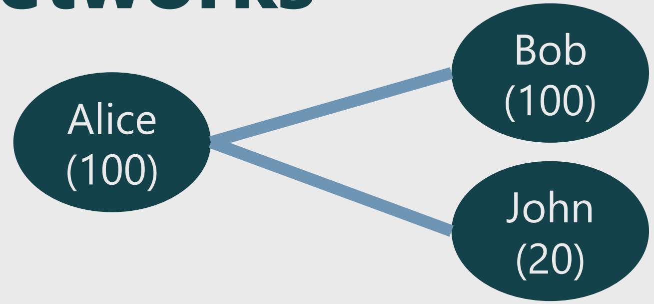
Project



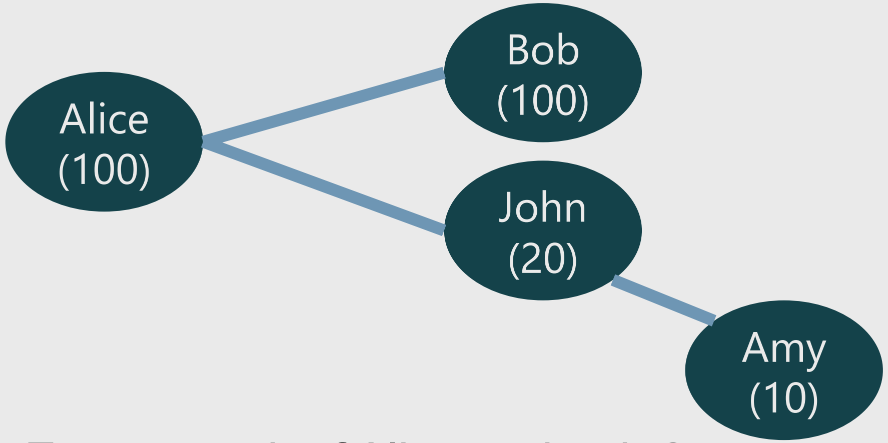
Multiplex projection

Other types of networks: Ego-networks



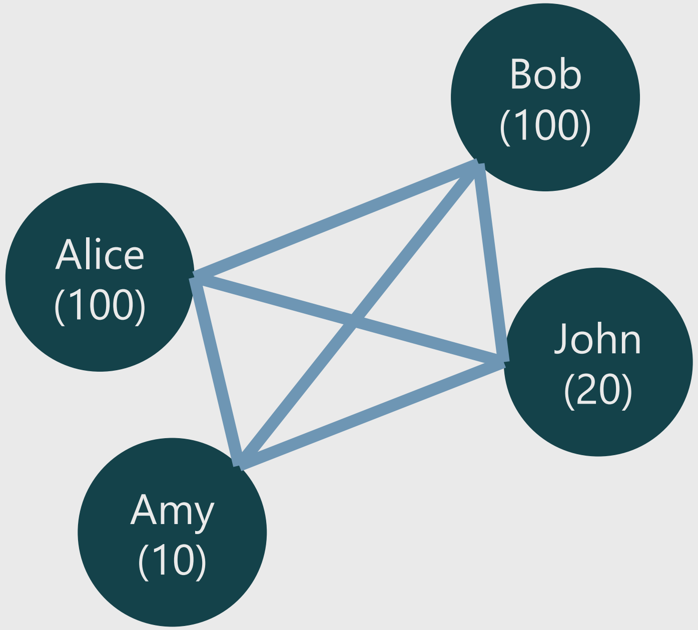


Ego network of Alice at depth 1



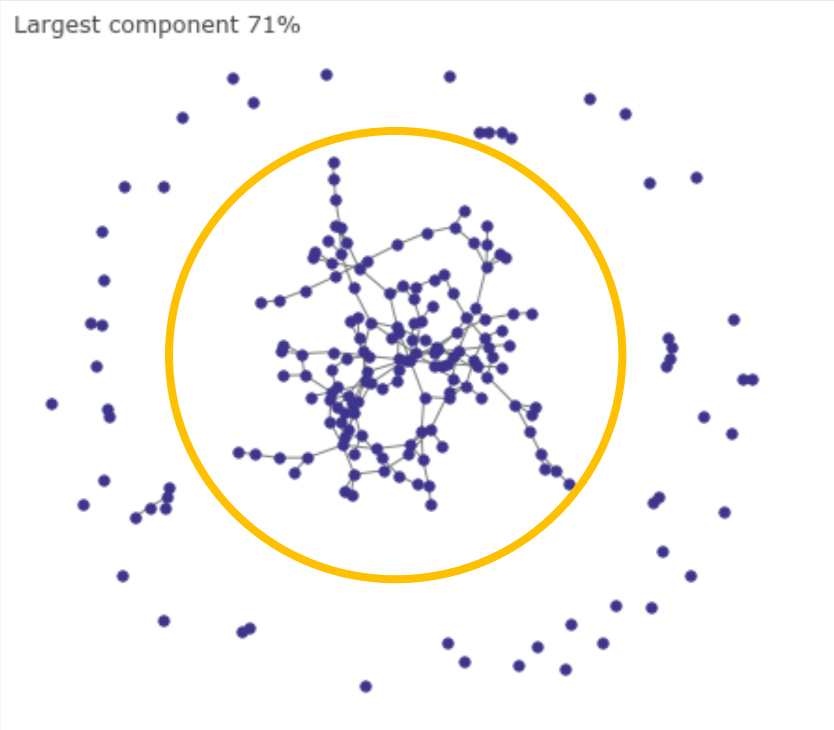
Ego network of Alice at depth 2

Other types of networks: Clique



Network characteristics

Connectedness



Real networks are typically connected, forming a **“giant component”**

In random graphs:

- If the average degree < 1 → many small components
- If the average degree > 1 → suddenly the system becomes connected

In “real” graphs:

- Graph has a giant component even for degree == 1

Let’s try this!

Small world: six degrees of separation

Milgram's experiment: six degrees of separation

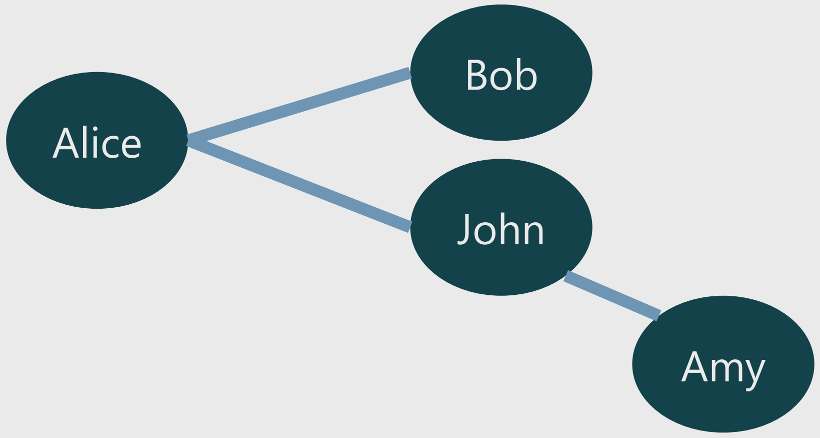
Strogatz, Watts: small number of random links are enough to create small world networks

Shortest path between node 1 and node 2:

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

Diameter:

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

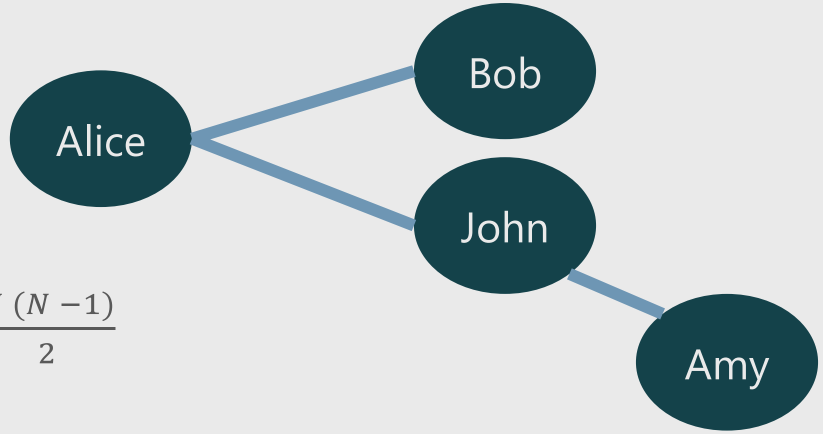


Real networks have **small diameters**

Density

Definition: Number of edges present / potential number of edges

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



Density = 3/6 = 50%

Real networks are typically **sparse**

As size increases density decreases (average degree is usually fixed)

Local clustering (~transitivity)

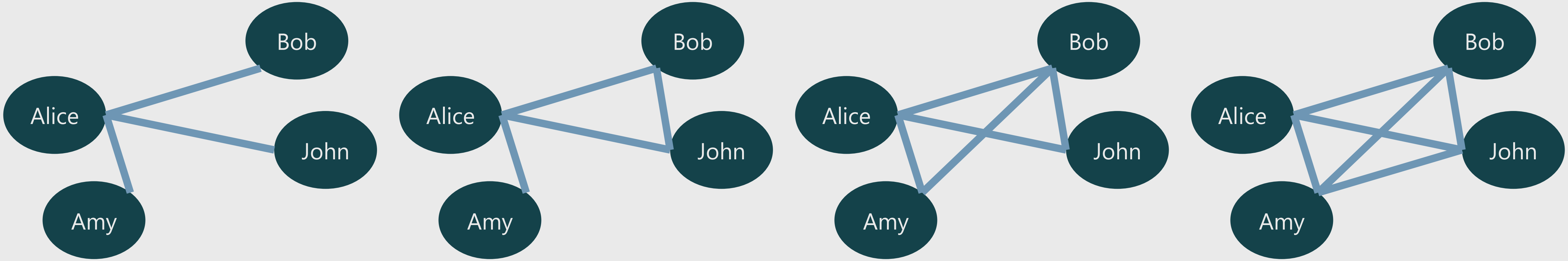
Strogatz, Watts (1998): How many of your neighbors are connected to each other

Local clustering coefficient: Number of triangles/number of triplets at the node level

Transitivity (global clustering) of a network: Number of triangles/number of triplets

Real networks have **high clustering**

Clustering of Alice:

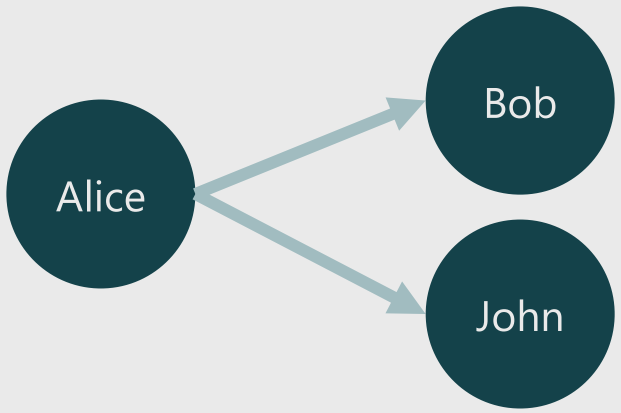


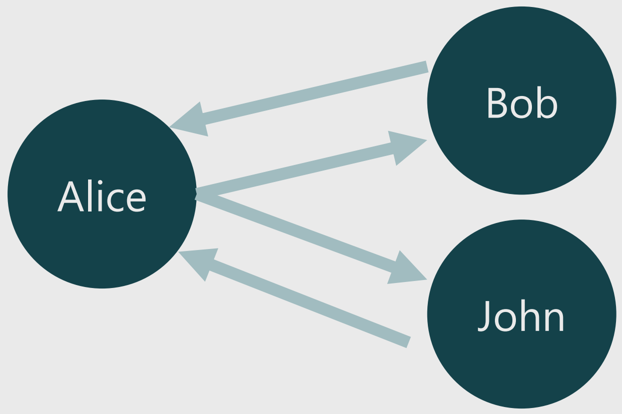
0/3 1/3 2/3 3/3

Reciprocity

Directed networks

Ratio of the number of edges pointing in both directions to the total number of edges in the graph.

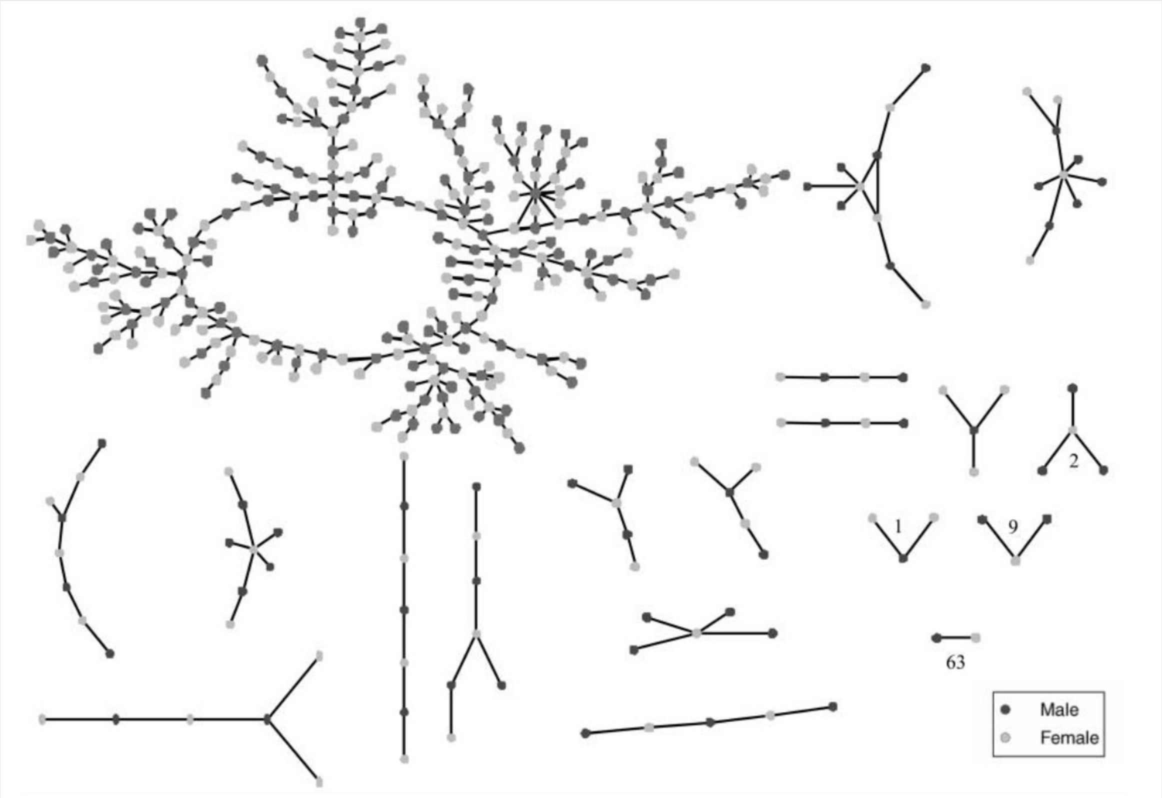




Assortativity (homophily)

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

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



Romatic links between teenagers
Bearman, Moody, Stovel (1991)v

Assortativity (homophily)

At the network level:

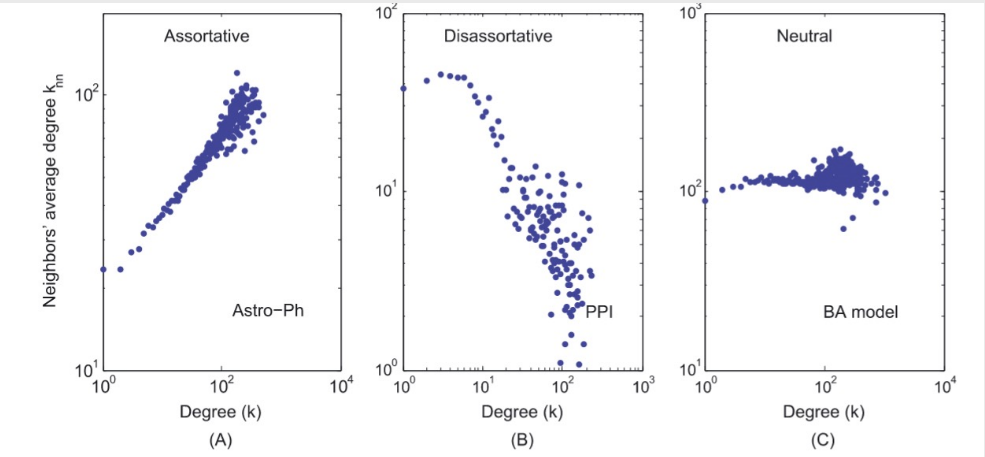
- Categorial unordered variable = Modularity
  - (Actual links between edges between nodes of same type – expected number of links between nodes of same type)/number of links
- Continuous variable: Pearson's correlation across edges.

Mixing patterns in networks, Newman, Physical Review E, 67 026126, 2003

At the local level:

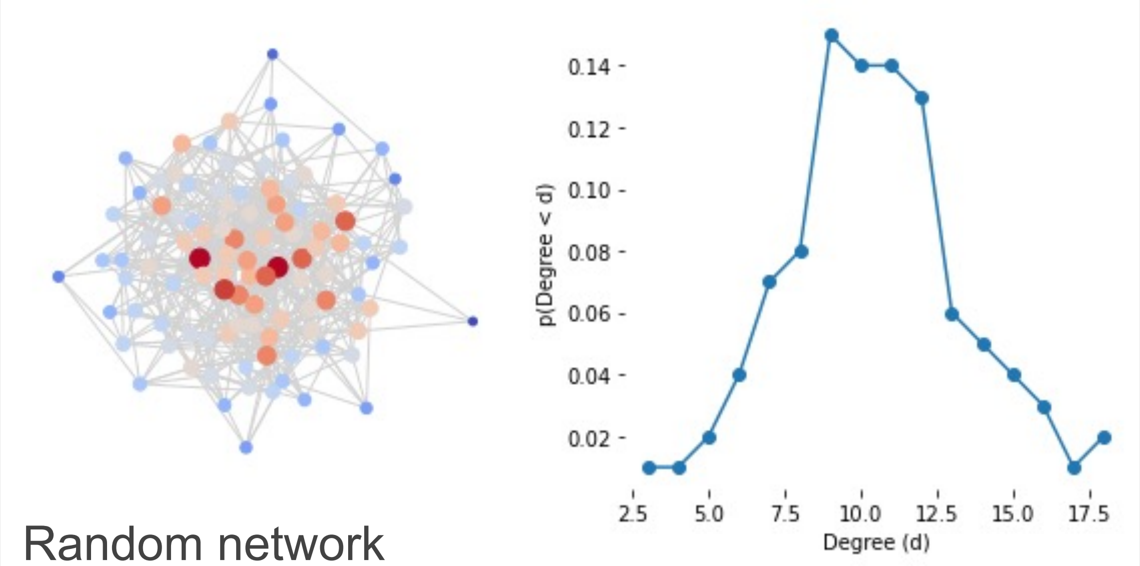
- Real networks can be locally assortative or disassortative
- Exercise: Draw a degree-assortative network

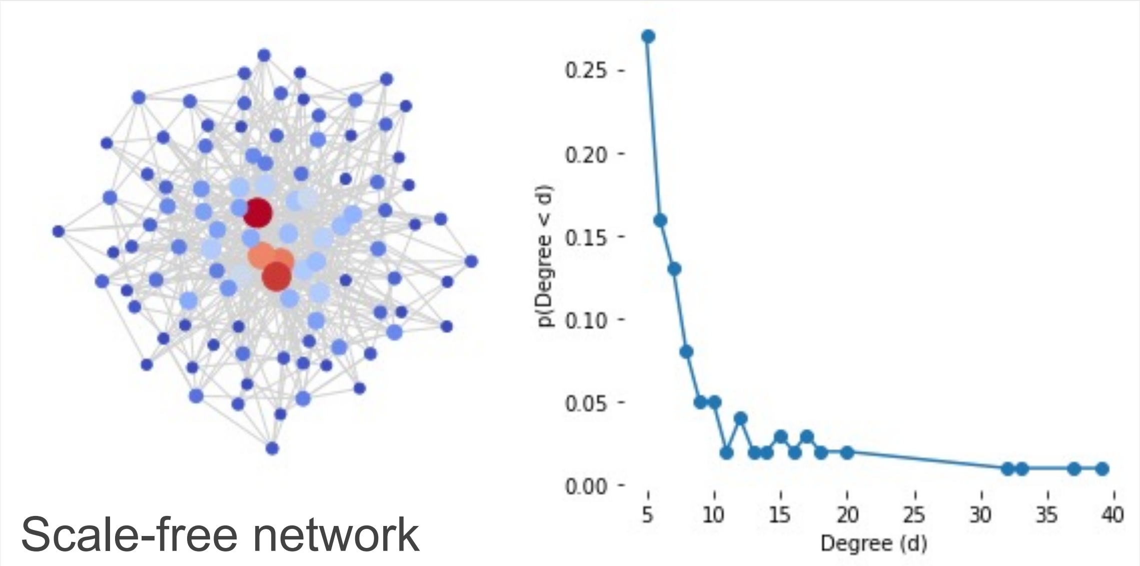
Multiscale mixing patterns in networks, Peel, Delvenne and Lambiotte (2018)



Jiang et al (2016)

Heavy tails / scale-free





Networks are not random, they have heavy degree distributions

PDF (probability density function)

→ Degree vs probability of degree

→ Represented by histogram

Many possible mechanisms:

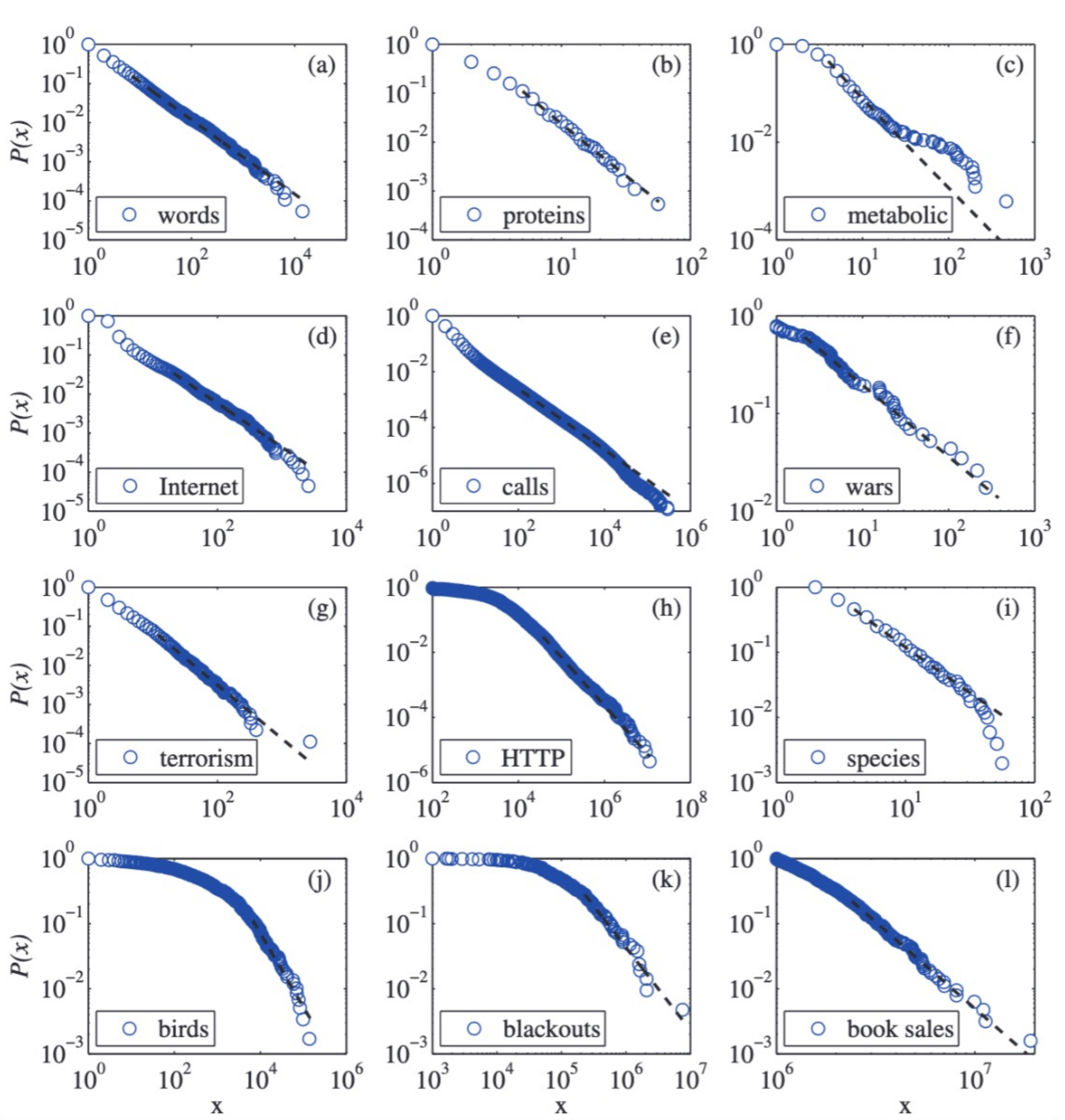
- Multiplicative growth
- Preferential attachment (Rich gets richer, Matthew effect, Fit gets richer)
- Copying models

Growing networks: https://www.stat.cmu.edu/~cshalizi/networks/16-1/lectures/08/li.pdf

Heavy tails

Most complex systems have **heavy tail distributions**

Most real networks have heavy tail degree distributions



Clauset, Shalizi & Newman (2009)

Random networks don’t have heavy tails

PDF (probability density function)

→ Degree vs probability of degree

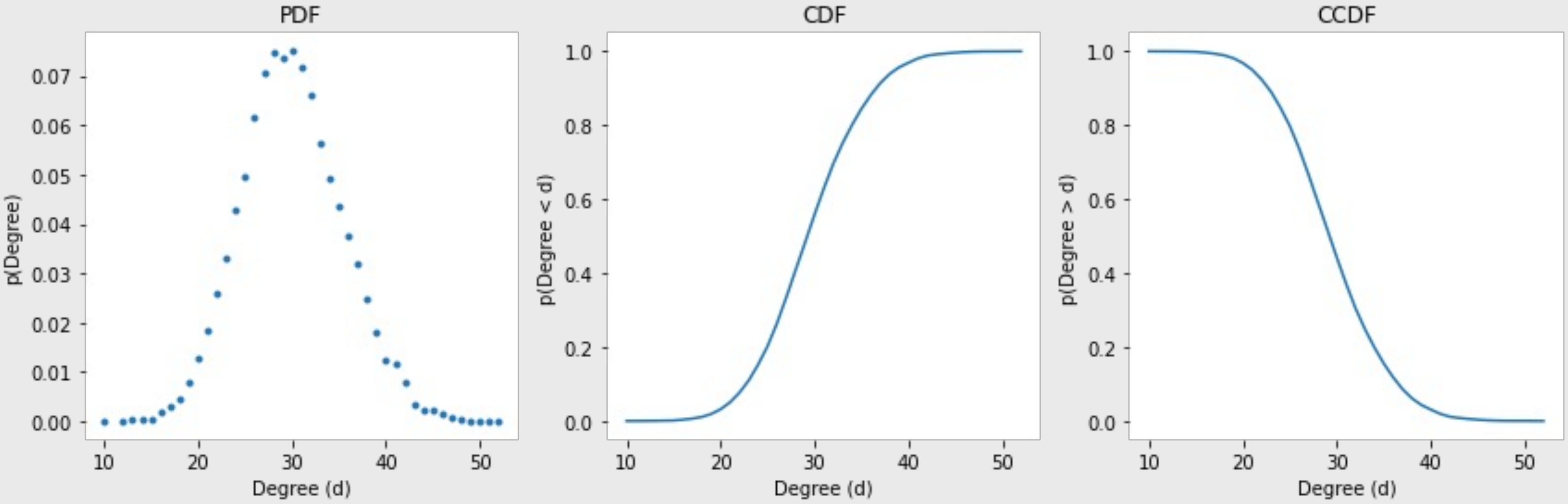
→ Represented by histogram

CDF (cumulative density function)

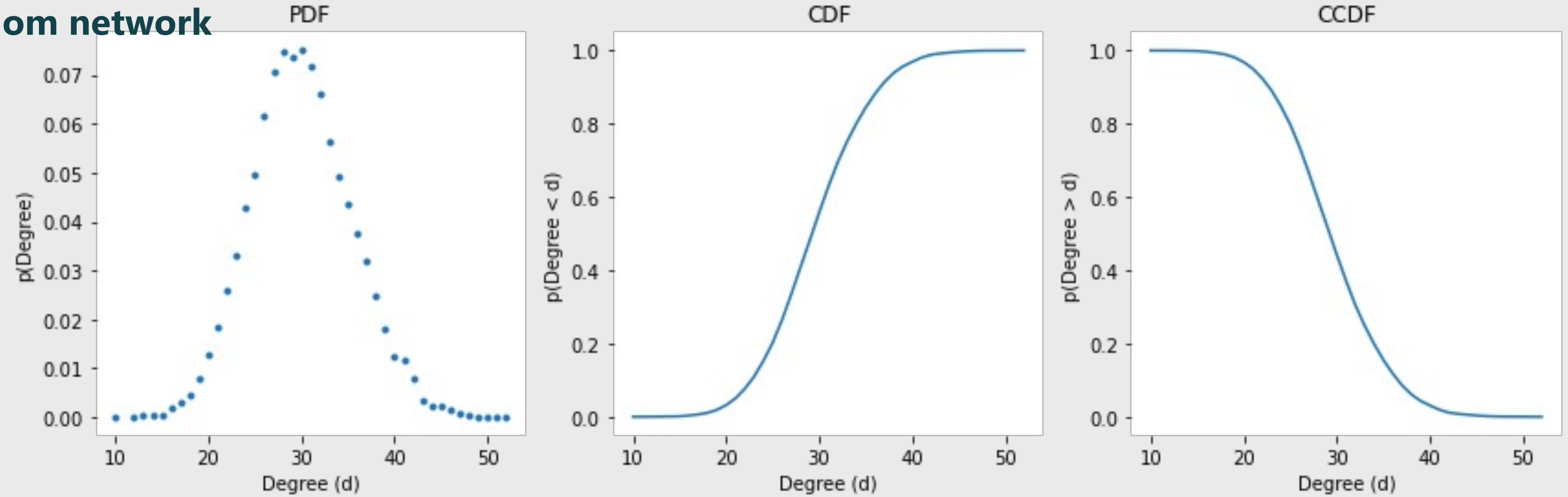
→ Degree s vs probability degree < s

CCDF: Complementary CDF

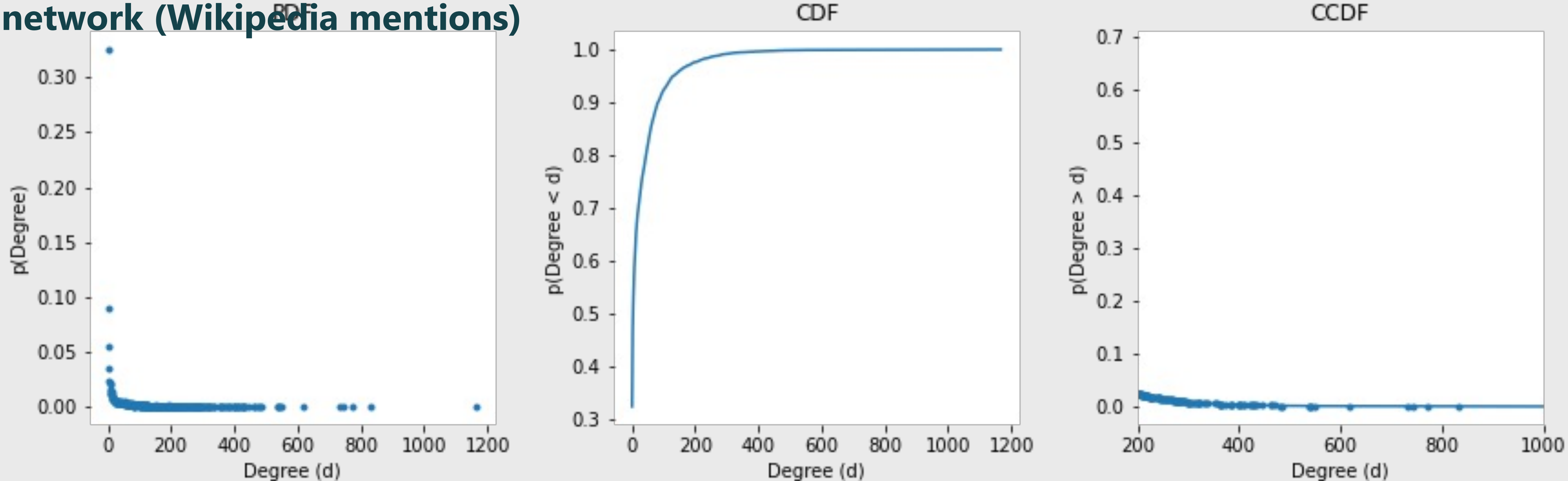
→ Degree s vs probability degree > s



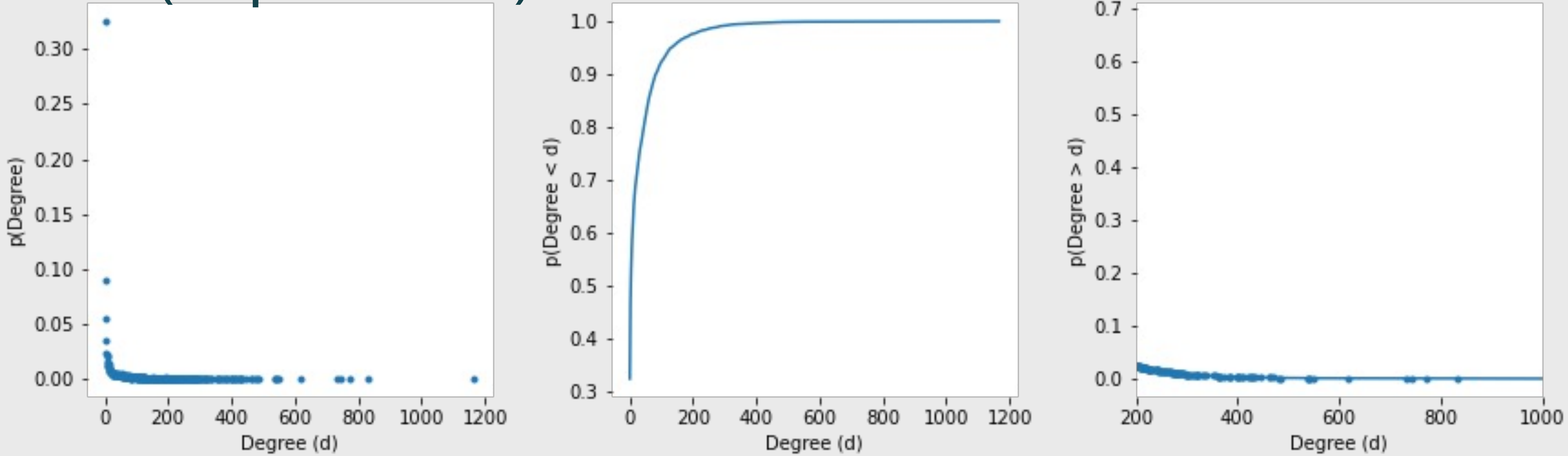
Random network



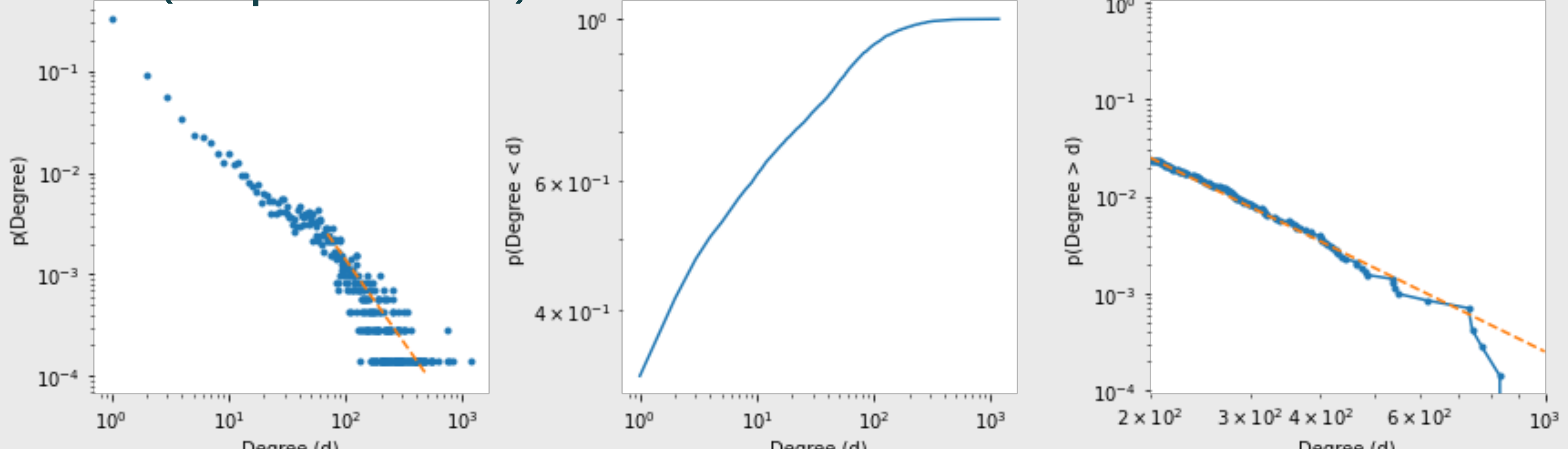
Real network (Wikipedia mentions)



Real network (Wikipedia mentions)



Real network (Wikipedia mentions)



Is it a power-law? $P(d) \sim d^{-\alpha}$

Critical Truths About Power Laws

Most reported power laws lack statistical support and mechanistic backing.

MICHAEL P. H. STUMPF AND MASON A. PORTER

SCIENCE • 10 Feb 2012 • Vol 335, Issue 6069 • pp. 665-666 • DOI: 10.1126/science.1216142

Scale-Free Networks

Scientists have recently discovered that various complex systems have an underlying architecture governed by shared organizing principles. This insight has important implications for a host of applications, from drug development to Internet security

BY ALBERT-LÁSZLÓ BARABÁSI AND ERIC BONABEAU

50 SCIENTIFIC AMERICAN MAY 2003

Article | Open Access | Published: 04 March 2019

Scale-free networks are rare

Anna D. Broido & Aaron Clauset

Nature Communications **10**, Article number: 1017 (2019)

Love is All You Need

Clauset's fruitless search for scale-free networks

by Albert-László Barabási, March 6, 2018

Comment | Open Access | Published: 04 March 2019

Rare and everywhere: Perspectives on scale-free networks

Petter Holme

Nature Communications **10**, Article number: 1016 (2019) | Cite this article

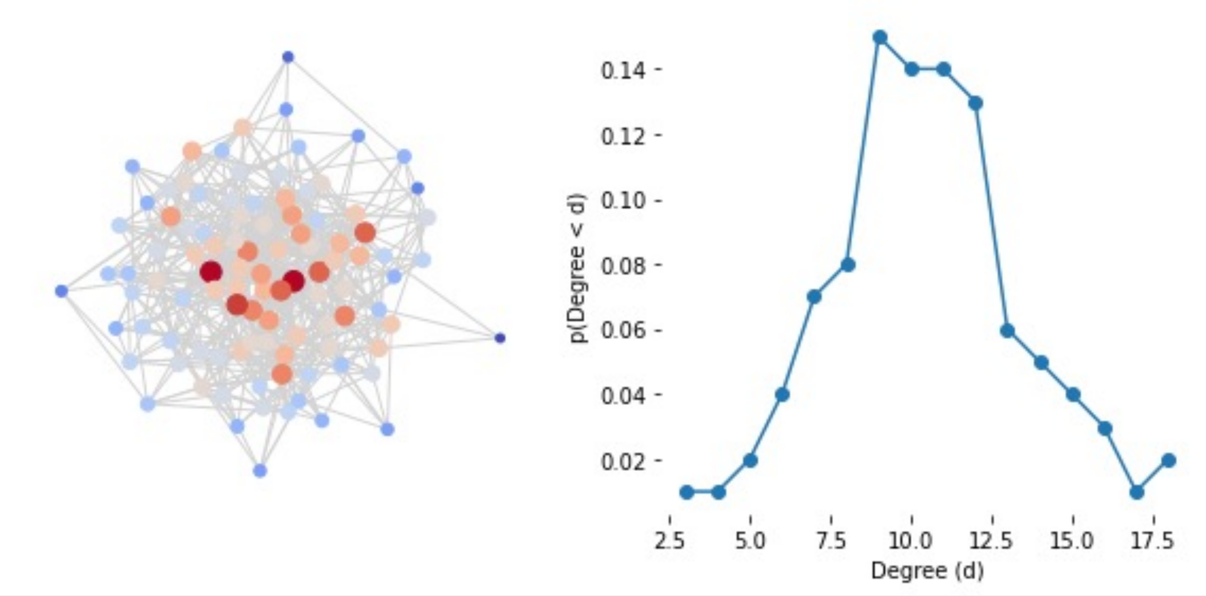
True scale-free networks hidden by finite size effects

Matteo Serafino, Giulio Cimini, Amos Maritan, +3 , and Guido Caldarelli Authors Info & Affiliations

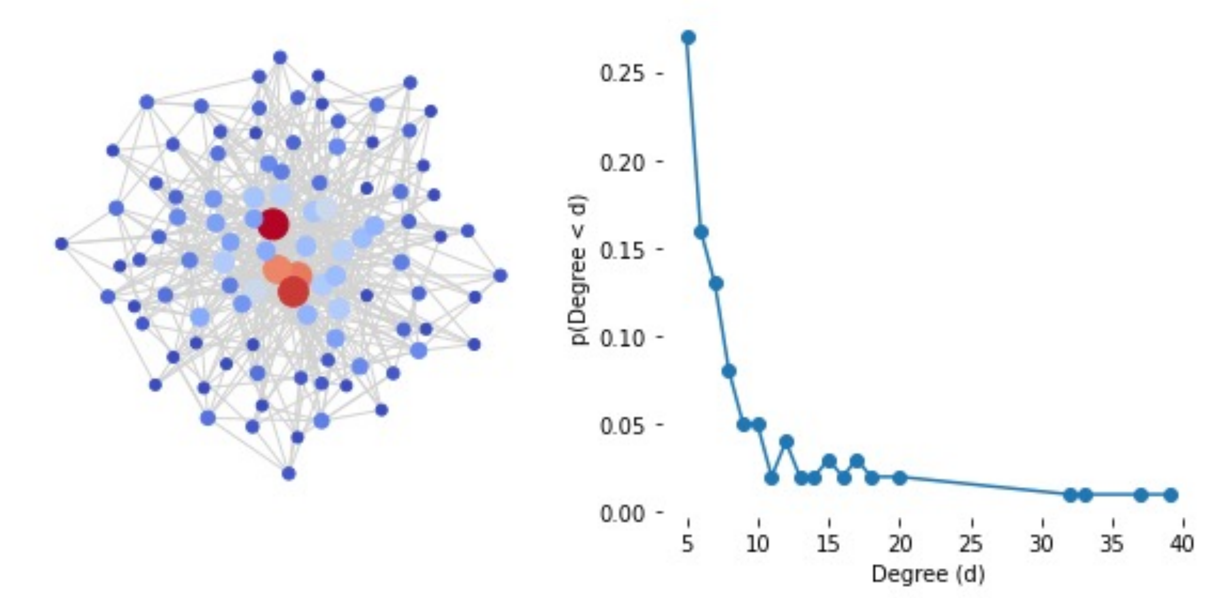
Edited by Lai-Sang Young, New York University, New York, NY, and approved November 2, 2020 (received for review July 3, 2020)

December 30, 2020 | 118 (2) e2013825118 | https://doi.org/10.1073/pnas.2013825118

Robustness to failures
Fragility to targeted attacks



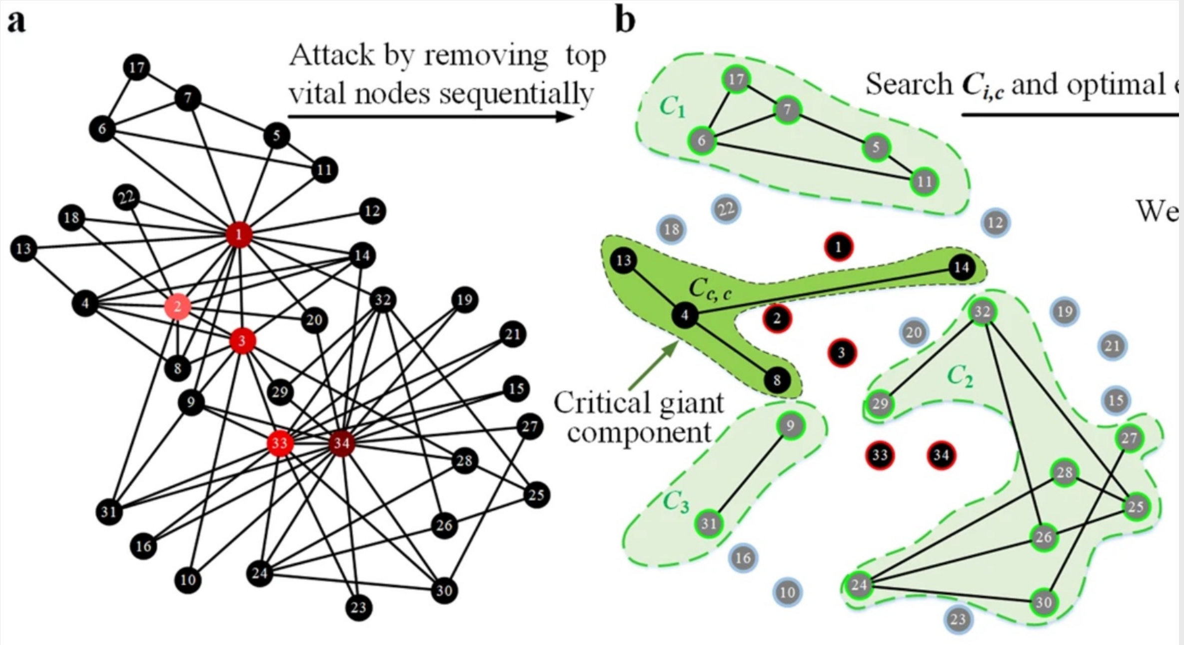
Random network



Power-law network

Robustness to failures
Fragility to targeted attacks

Albert, Jeong, Barabasi (2000) Attack and error tolerance of complex networks



Li et al (2011)

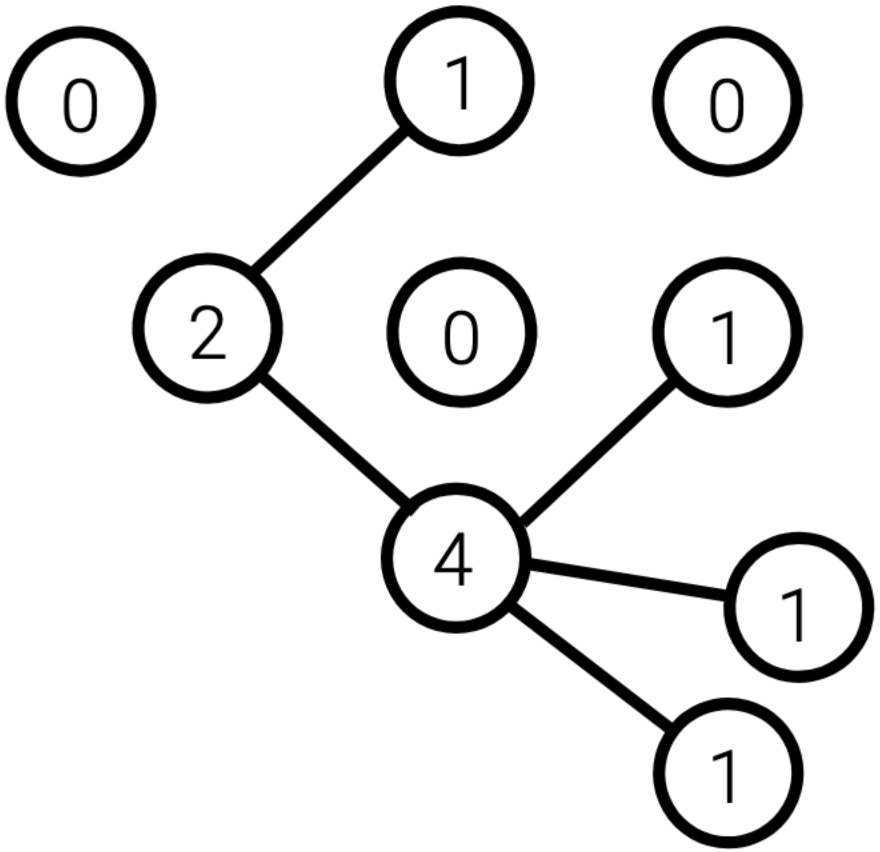
Friendship paradox

Your friends are more popular than you are

Some people have no friends.

But because they appear in nobody's friendship circles, they're not making anyone else feel popular.

The same applies to other people: the more friends you have, the likelier you are to be represented in people's friendship circles.



Feld, S. L. (1991). Why your friends have more friends than you do. American Journal of Sociology, 96(6), 1464-1477.

Friendship paradox

Your friends are more popular than **you** are

average friend
(count node proportional
to their degree)

average person
(count each node once)

Feld, S. L. (1991). Why your friends have more friends than you do. American Journal of Sociology, 96(6), 1464-1477.

Types of analysis

They should fit your research question

Types of analysis: Descriptive statistics

Describe the network characteristics (density, diameter, average degree, clustering, etc)

Types of analysis: Centralities

What are the most important nodes in the network?

- The one with more connections → **Degree centrality**
- The one linked to more important neighbors → **Pagerank / Eigenvector / Katz centrality**
- The one closest to all other nodes → **Closeness centrality**
- The ones that act as brokerage? → **Betweeness centrality**

More on this this afternoon

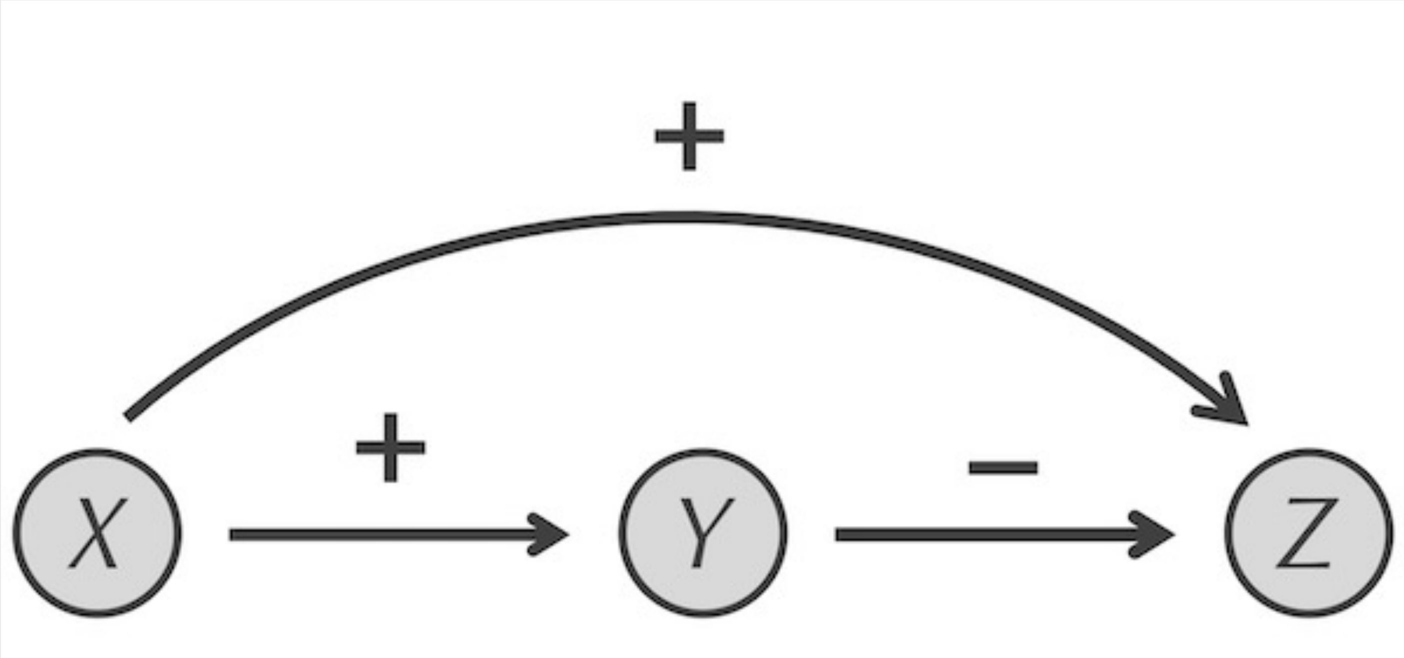
Types of analysis: Node-level regression

Calculate node-level features:

- Centrality
- Local clustering (transitivity / embeddedness)
- Local reciprocity
- Local assortativity (homophily)
- …
- Include in your model (e.g. a regression)

Types of analysis: Motif mining

Find overrepresented patterns



Feed-forward loop (https://biologicalmodeling.org/motifs/feedforward)

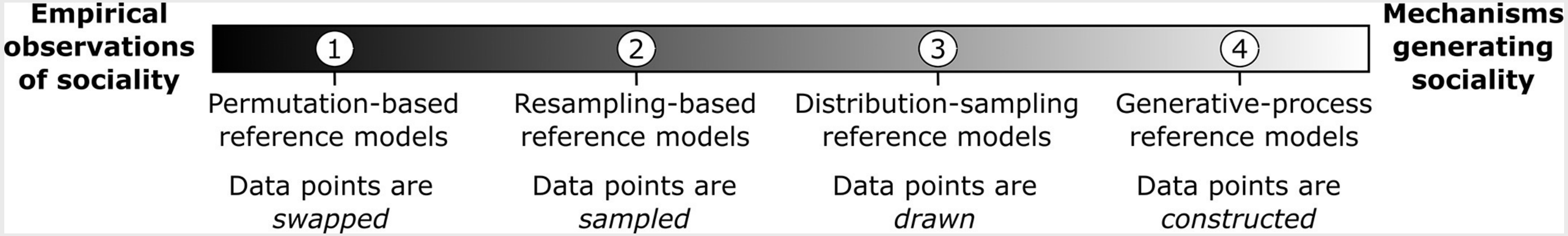
Types of analysis: Testing hypothesis

We observe some behavior in the network (e.g. the assortativity is 0.5). Is this relevant?

(1) Create a **reference model** to compare with it

- Permuting attributes holding the network constant
- Configuration model (permuting edges maintaining the degree distribution)
- Generative models (e.g. rich gets richer model)
- ERGM (which features of dyads affect the presence or strength of edges.)
- ABM

(2) Which factors affect link creation/disruption? Quadratic assignment procedure (QAP); Relational event model (REM)



Hobson 2021

More on this on Tuesday

Permutation of attributes (QAP)

Calculate significance by resampling.
Avoids running a regression (which is difficult with all the interdependences)

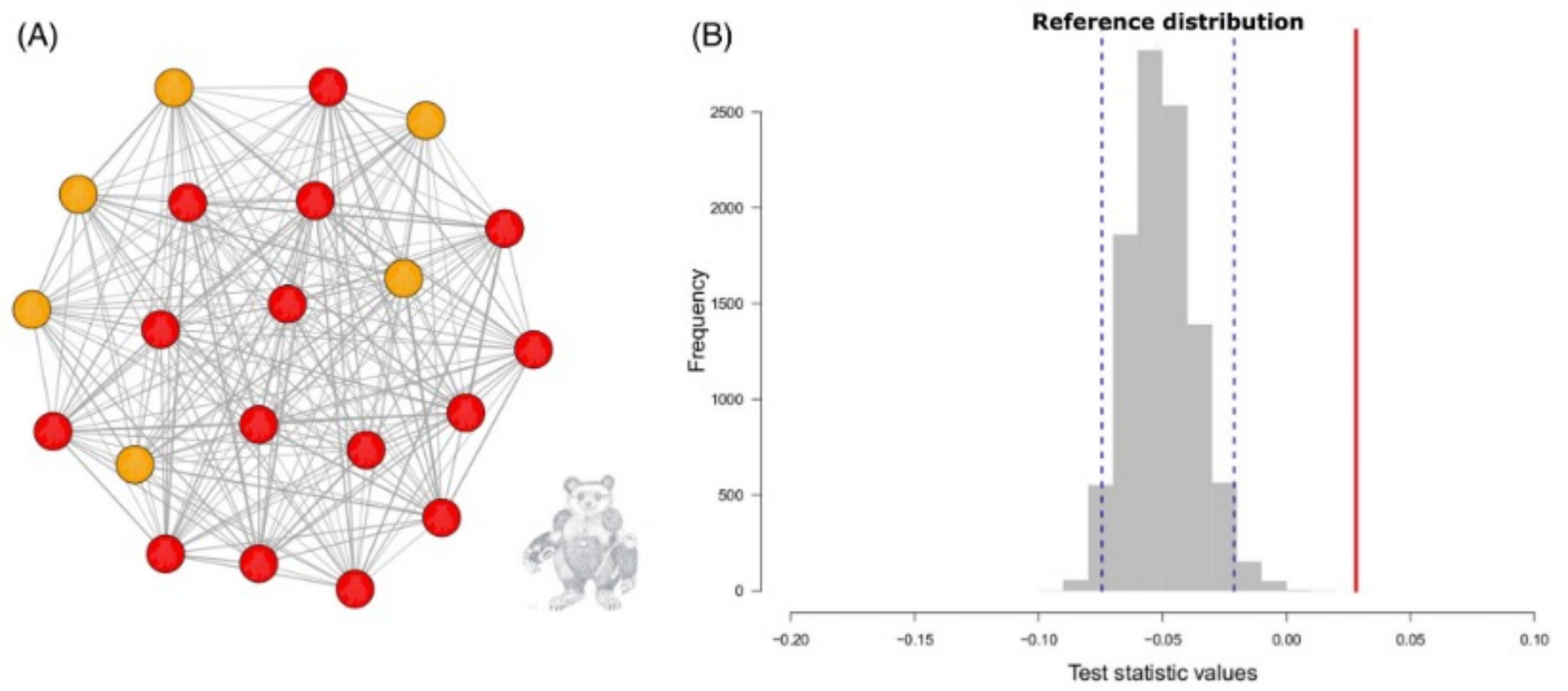
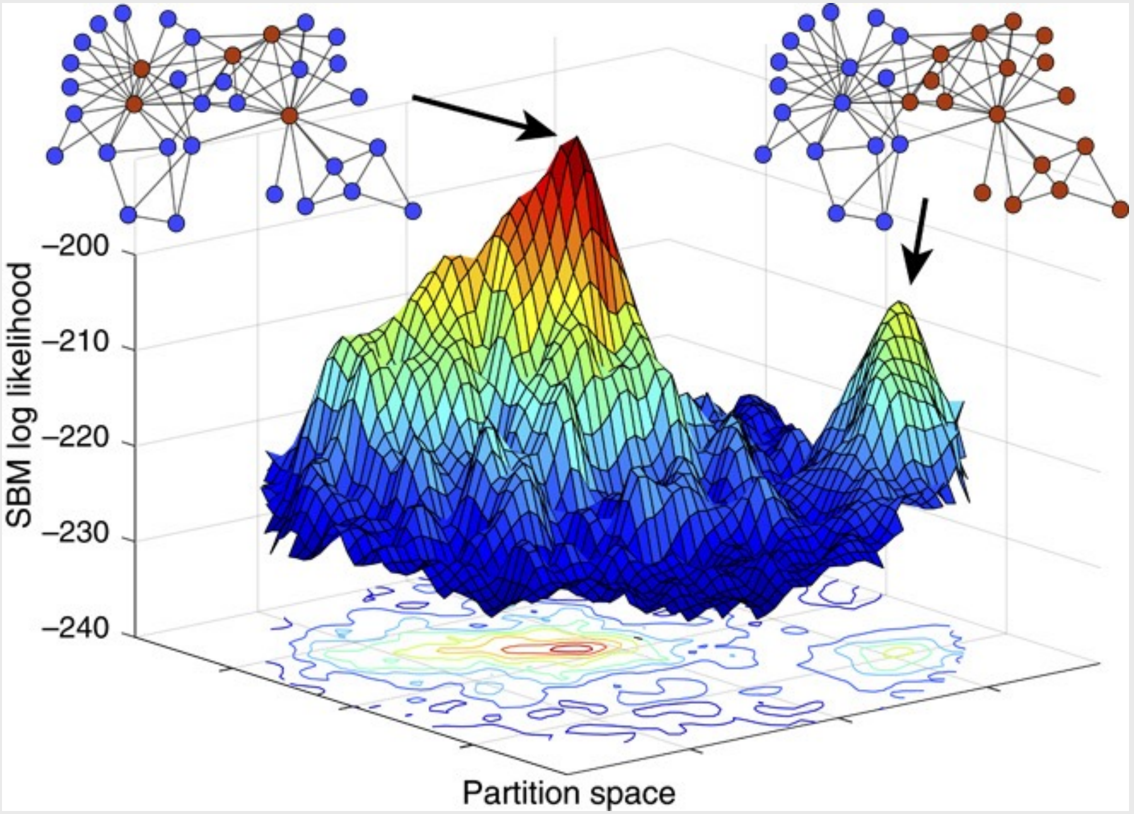


Fig 2. An example of study approaches: do burbils socially assort by nose colour? (A) Association network of burbils, with nodes colour-coded by nose colour and (B) distribution of values based on the permutation procedure of Team 1; observed value of the test statistic shown as a red solid line and the 2.5 and 97.5% quantiles of the reference distribution as blue dashed lines.

Hobson 2021

Types of analysis: Community detection



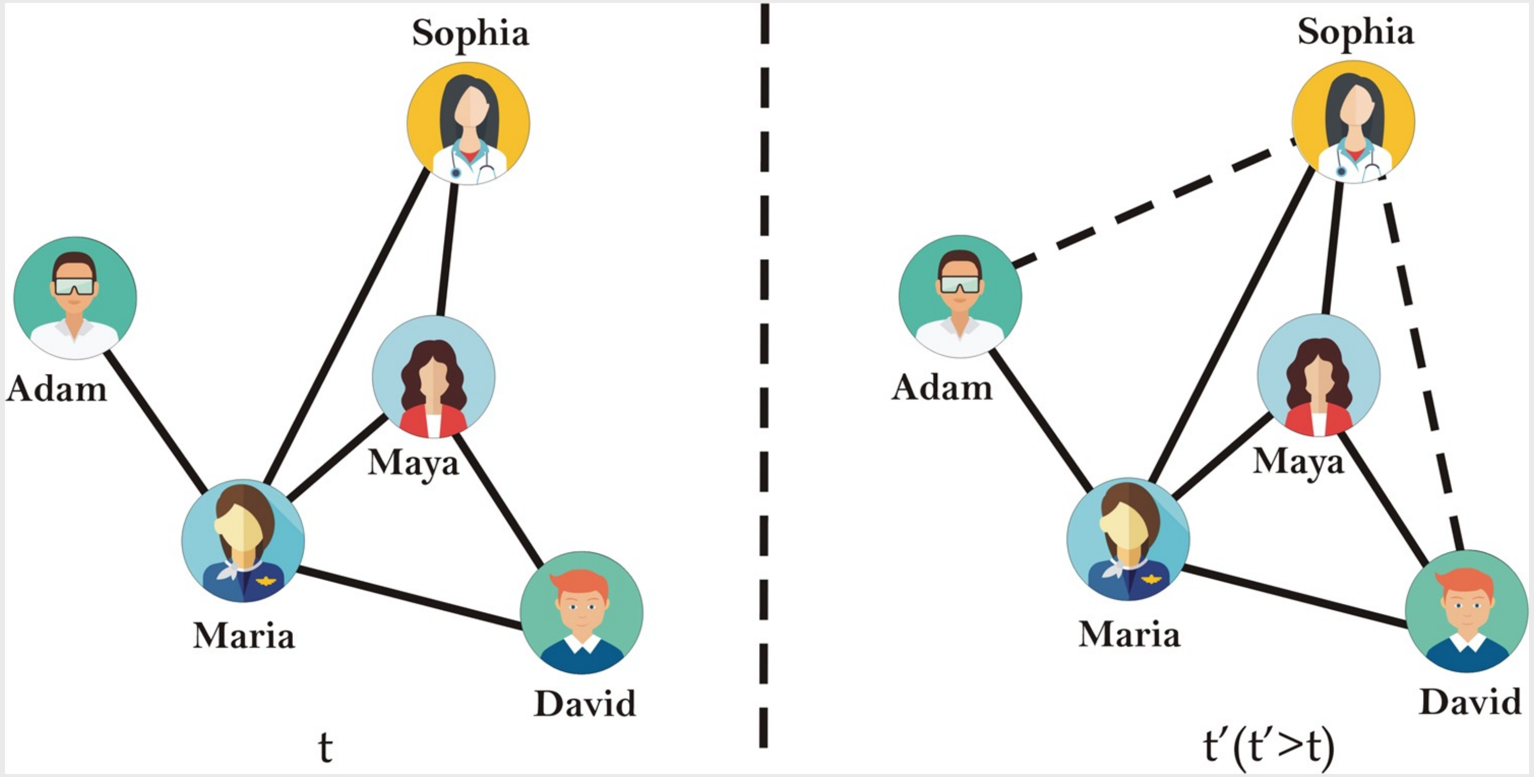
Peel et al

What clusters of nodes can we find in the network?

- **Stochastic Blockmodels** (Harrison White, structural equivalence, core-periphery)
- **Modularity minimization**

More on this on Wednesday

Types of analysis: Link/metadata prediction



Ahmad et al 2020

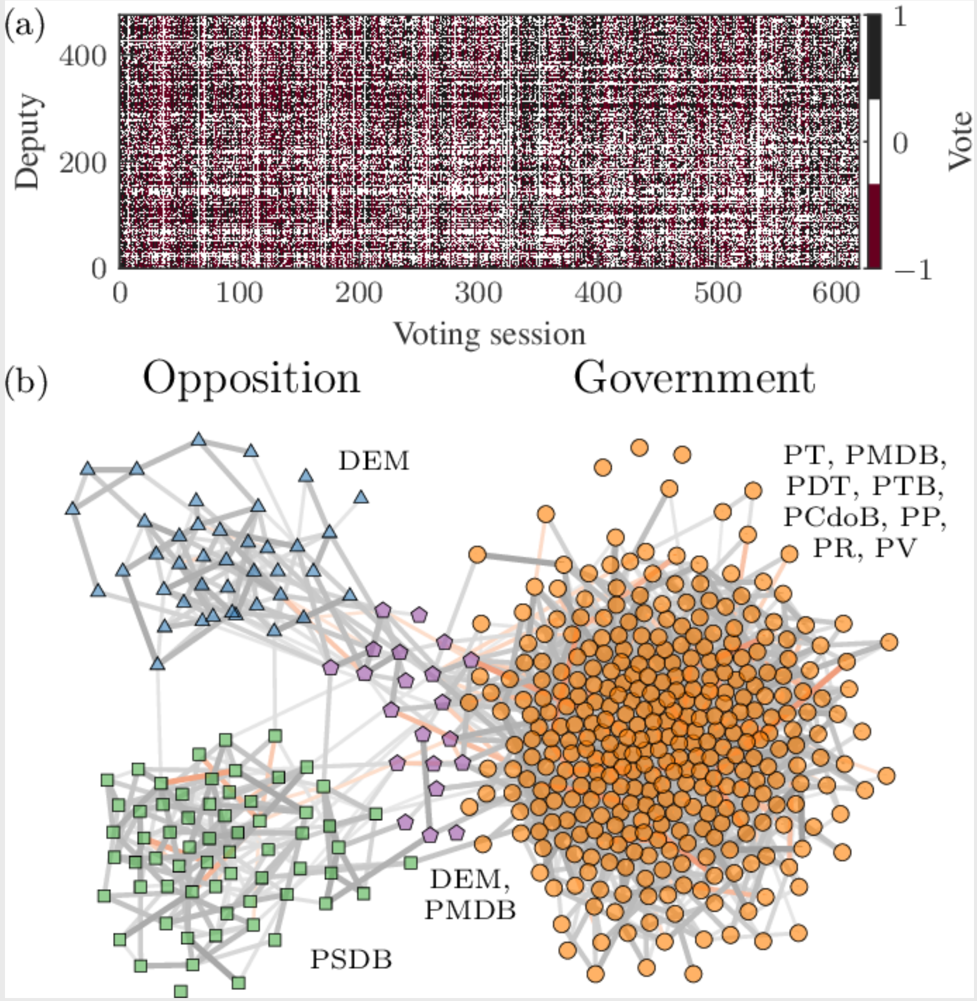
Networks are rarely complete

Approaches such as triangle closure, SBM or node embeddings

More on this on Wednesday

Types of analysis: Network reconstruction

Network from co-occurrences



Network Reconstruction and Community Detection from Dynamics, Peixoto 2019

More on this on Thursday

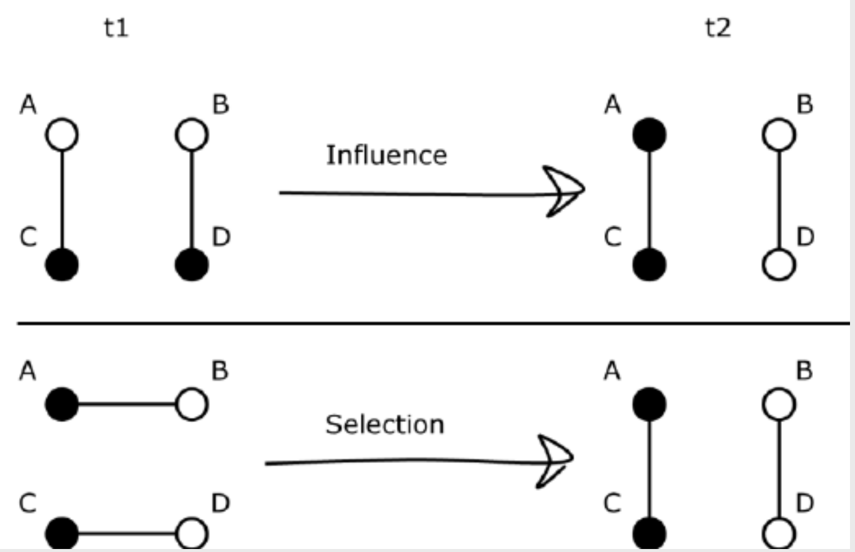
Types of analysis: Dynamics

How does behavior/diseases/information spread?

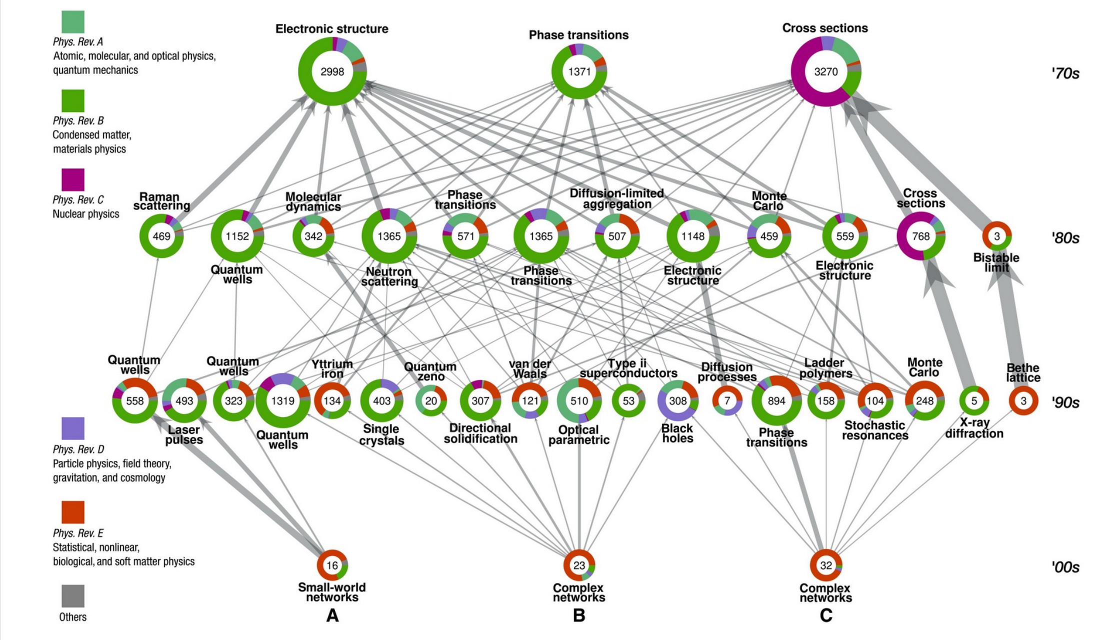
Model matters: Simple contagion vs complex contagion

Longitudinal data: Allow to test selection vs influence, evolution of communities over time, co-evolution of network topology and ideas

Simulations: Contagion, gene expression



Friemel, 2015



Bovet et al, 2022

More on this on Friday

Resources

Tools

Libraries:

- igraph (C, Python & R wrappers)
- Networkx (Pyhon)
- graph-tool (Python, UNIX)
- statnet (R)

Gephi: open-source network analysis and visualization software package

Interactive network visualization:

- visNetwork (R)
- pyVis (Python)

Data

Stanford Large Network Dataset Collection:
https://snap.stanford.edu/data/

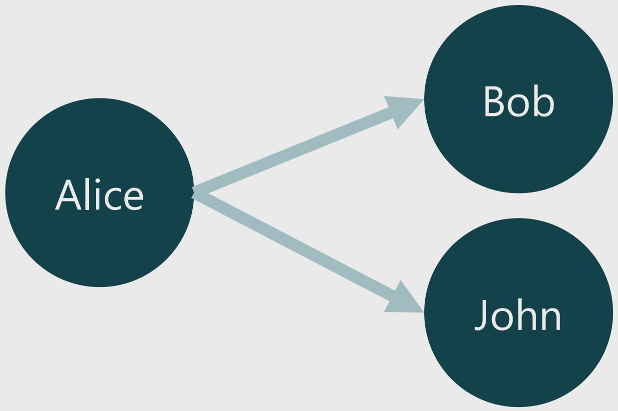
Network repository:
https://networkrepository.com/networks.php

Index of Complex Networks:

https://icon.colorado.edu

Matrix representation

Network representation



Adjacency list:

A. It is dense: Only keeping edges

Origin	Target	Weigth
Alice	Bob	1
Alice	John	1

Adjacency matrix:

- A. Linear algebra is easy
- Sparse: Many zeros → 1E6 nodes/10 million edges → 1 trillion options

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

Practical 1

1. Download materials:

- https://github.com/jgarciab/NetworkScience
- (click on code -> Download Zip)

2. Extract ZIP

3. Set up Python. On Windows & Mac using the graphical interface:

- Open Anaconda
- Go to “Environments” (left menu)
- Click on “Import” and specify the file “environment.yml” (it’s one of the files that you downloaded)
- Activate environment by clicking in the “play” button next to the environment.

On Linux (also works for Windows/Mac):

- Open a conda terminal (open a terminal)
- Navigate to the directory with the code using dir (ls) to list the files and cd XXX (cd XXX) to enter directory XXX.
- Create a new environment: *conda env create -f environment.yml*
- Activate environment: *conda activate networks*
- Launch jupyter notebook: *jupyter notebook*

4. Open and complete Python notebook: day*1a_intro_networks.ipynb*

5. Open and complete R notebook: *day1a_intro_networks.Rmd*