Networks and CBS

Javier Garcia-Bernardo

ODISSEI Social Data Science group

Utrecht University

Tabular data

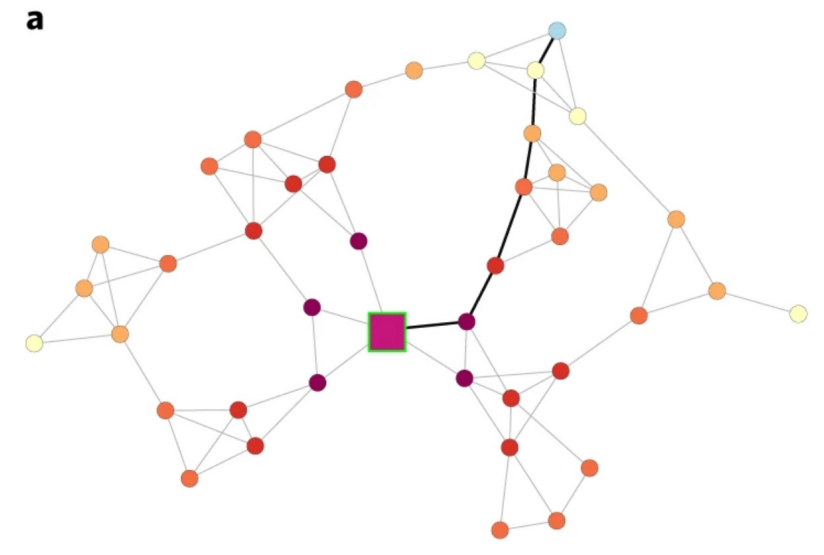
	child_post2020	Age	Income	...
Person 1	1	30	50000	
Person 2	0	25	30000	
Person 3	1	32	40000	
Person 4	0	40	25000	
Person 5	0	28	100000	

Relational data (networks)

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

Relational data (networks)



	X_1 ?	X_2 ?	...	X_N ?
Person 1				
Person 2				
Person 3				
Person 4				
Person 5				

If we were to study networks using tables, how do we include connections?

- Create individual-level variables
- Deep-learning: create “person embeddings”

Why do we care about the connections?

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

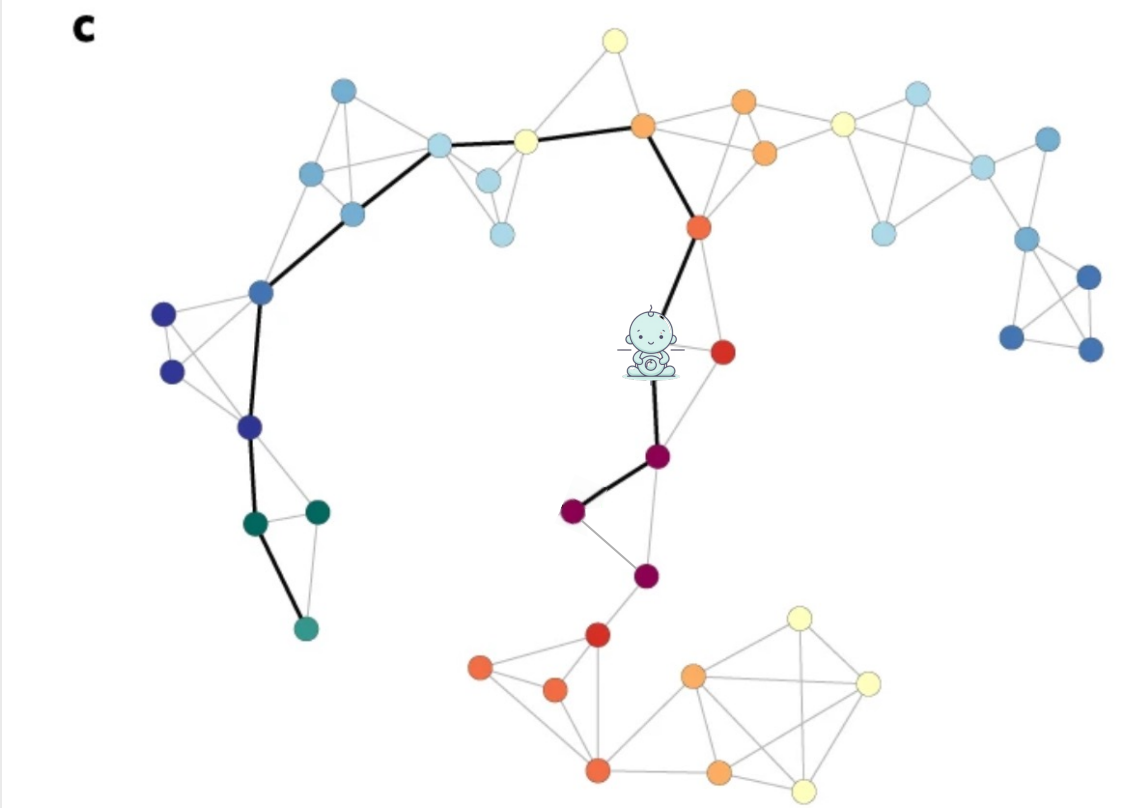
Network of countries trading with each other: Which country has the most bargaining power?

Network of connections between politicians and companies: Who is the most politically influential company? Who is corrupt?

Network of email communication within a company: What type of roles do employees of a company play (coordination/innovation/etc)?

Network of online media referencing each other: What is the political ideology of media outlets?

2) They constrain/facilitate future change.

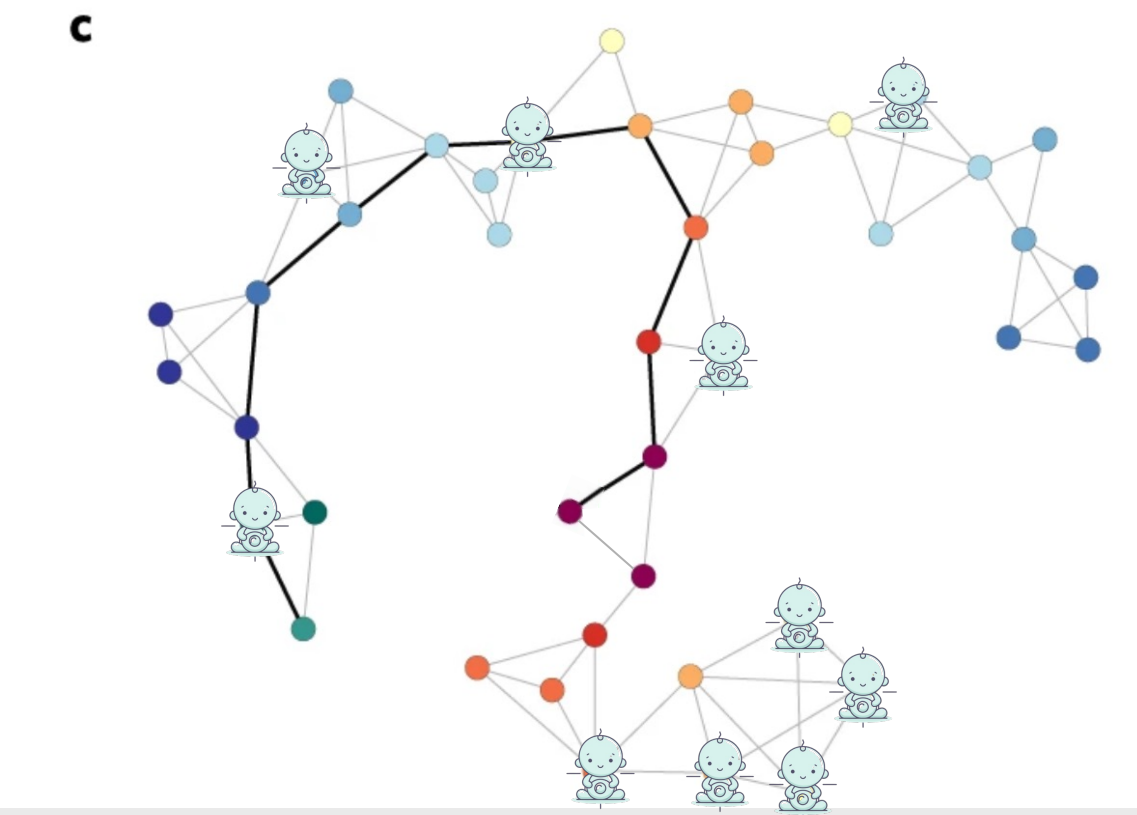


Block, P., Hoffman, M., Raabe, I. J., Dowd, J. B., Rahal, C., Kashyap, R., & Mills, M. C. (2020). Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. Nature human behaviour, 4(6), 588-596.

Which person would you vaccinate first?

Networks allow us to understand dynamics: epidemics, contagion, development

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 will become a parent next?

3) Interactions can create *emergence*

Sometimes a system (e.g. a society) has properties that the individual parts do not have. These properties are called "*emergent*" properties, and the system a *complex* system.

"There's no love in a carbon atom, no hurricane in a water molecule, no financial collapse in a dollar bill." *Peter Sheridan Dodds*

In social science: Connect micro-behavior to macro outcomes

e.g. Schelling model: why do we see urban 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

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

Examples of research questions that can be answered with networks

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 within-person attitude 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)

Today

- Introduction to Networks

- Networks at CBS and the *netCBS* package

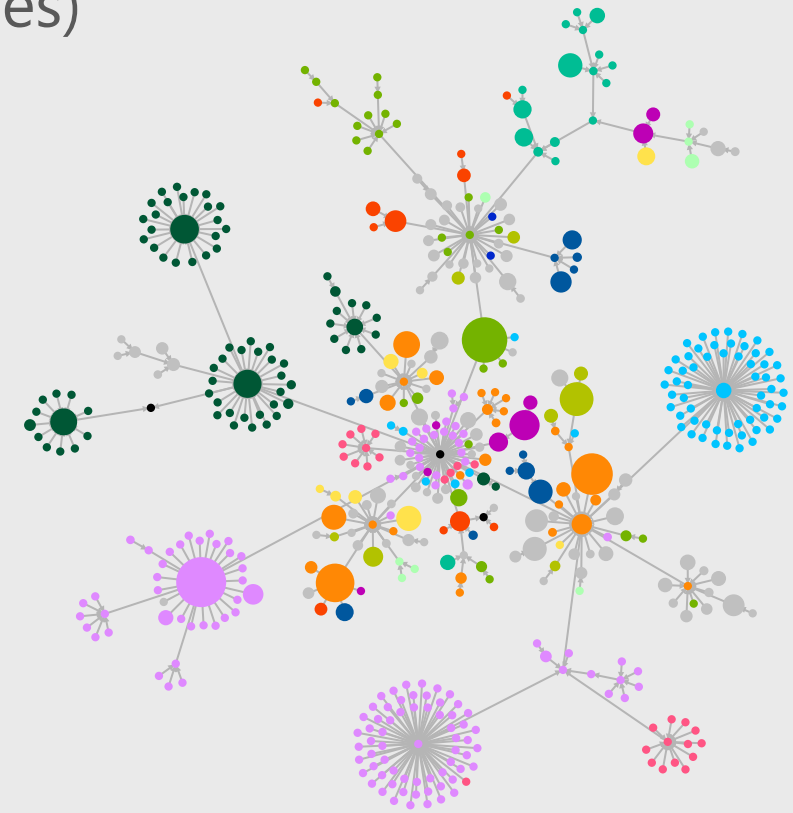
Introduction to networks

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	Attitudes or behaviors	Co-ocurrence
Biology	Gene regulatory	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

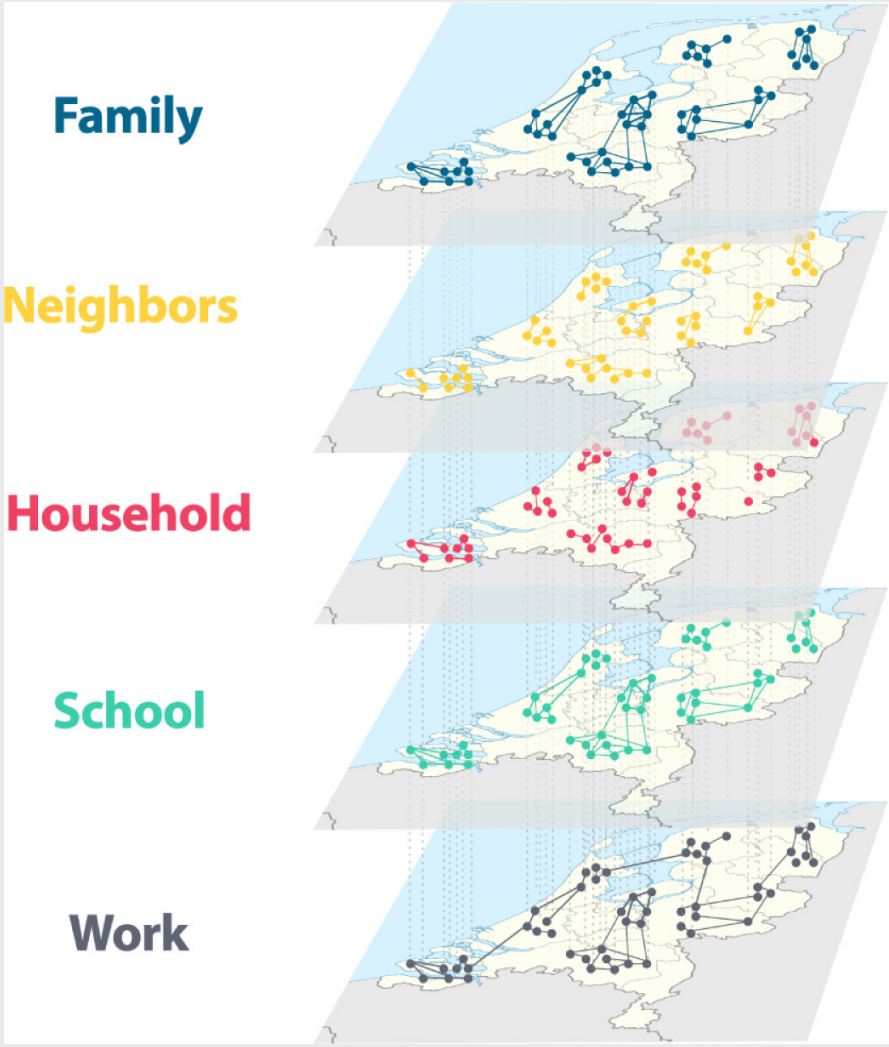
Type 2: Affiliation → Node 1 is part of/related to node 2

- Bipartite networks, e.g. students in classrooms

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

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

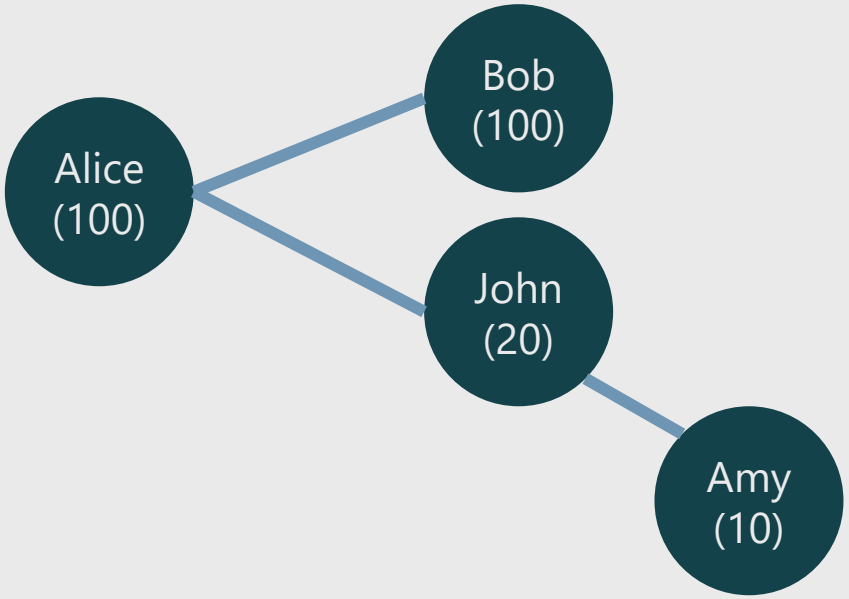
Today we focus on the first type ("real networks")



Kazmina et al, 2024

Basic definitions

Networks (graphs)



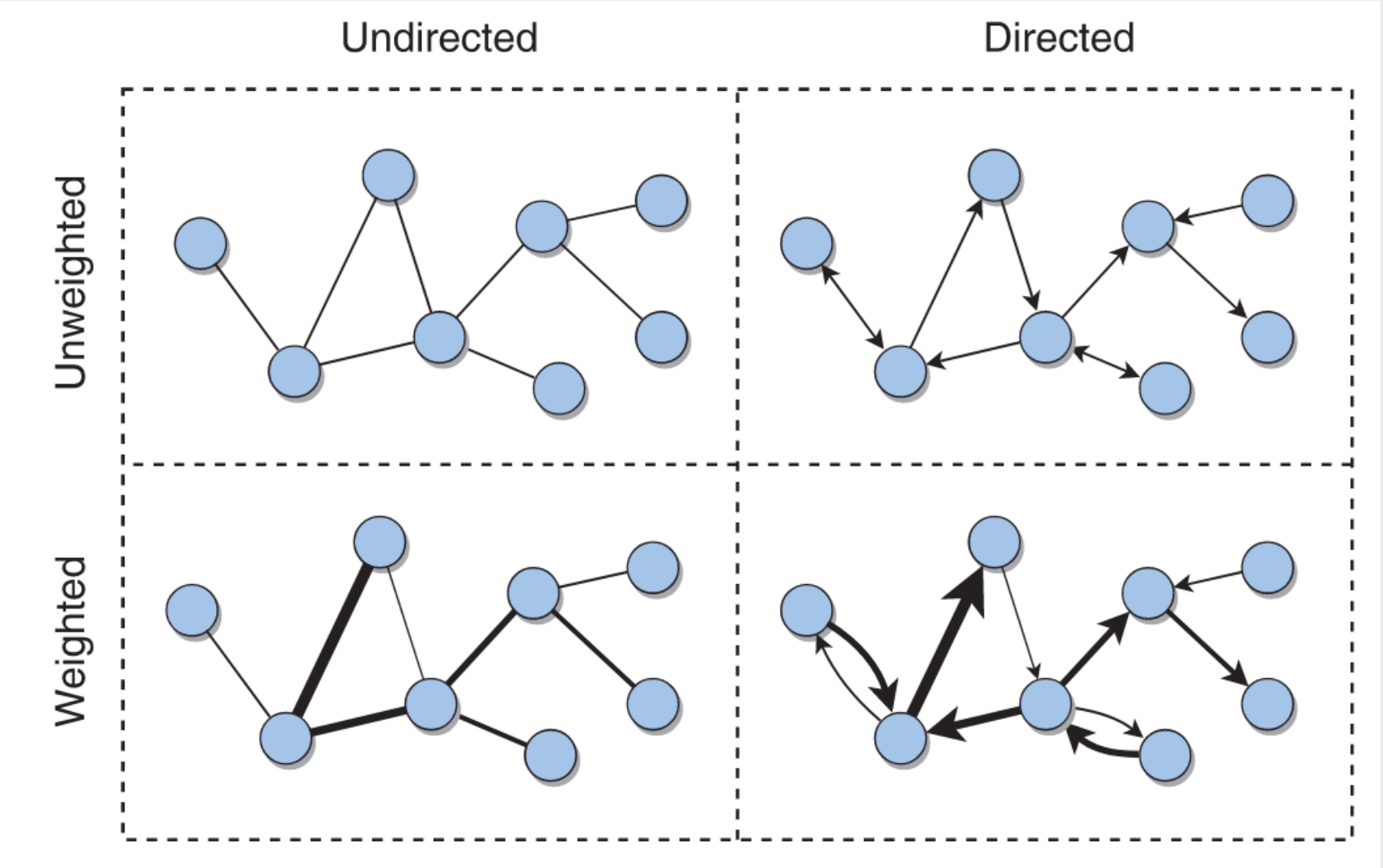
Nodes (vertices, actors) connected by
edges (links, connections, relationships)

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



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

Source: A first course in network science (2020)

Degree in undirected networks

Degree of a node: Number of neighbors

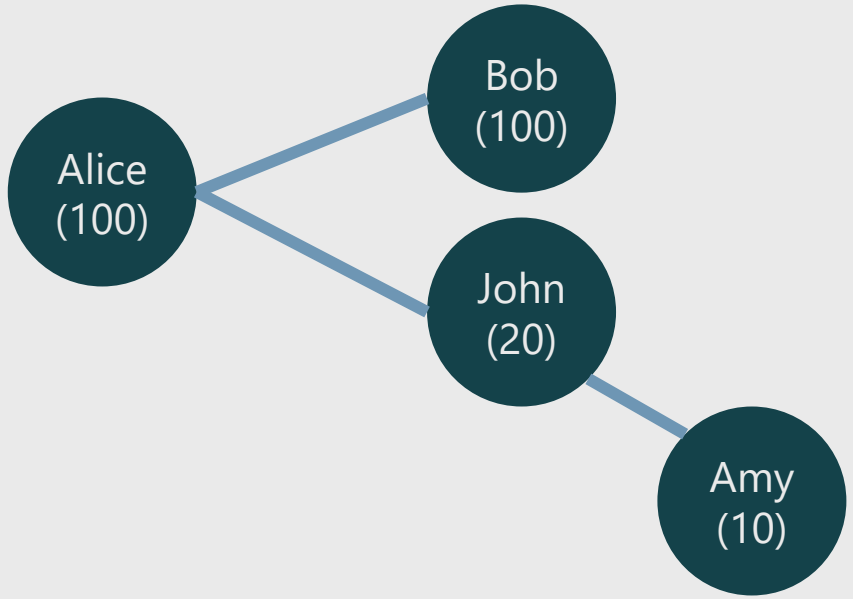
Node: degree

Alice: 2

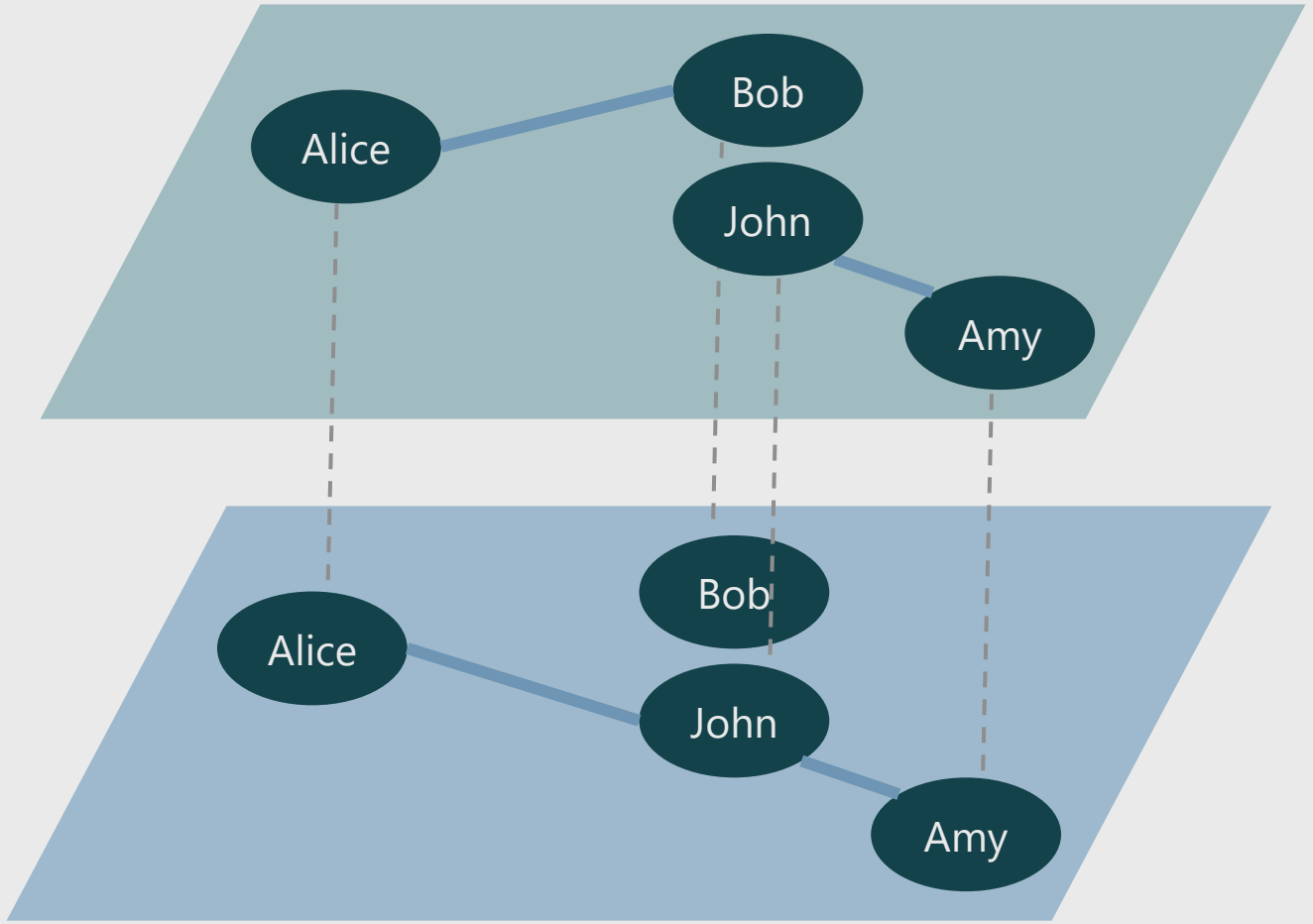
Bob: 1

John: 2

Amy: 1

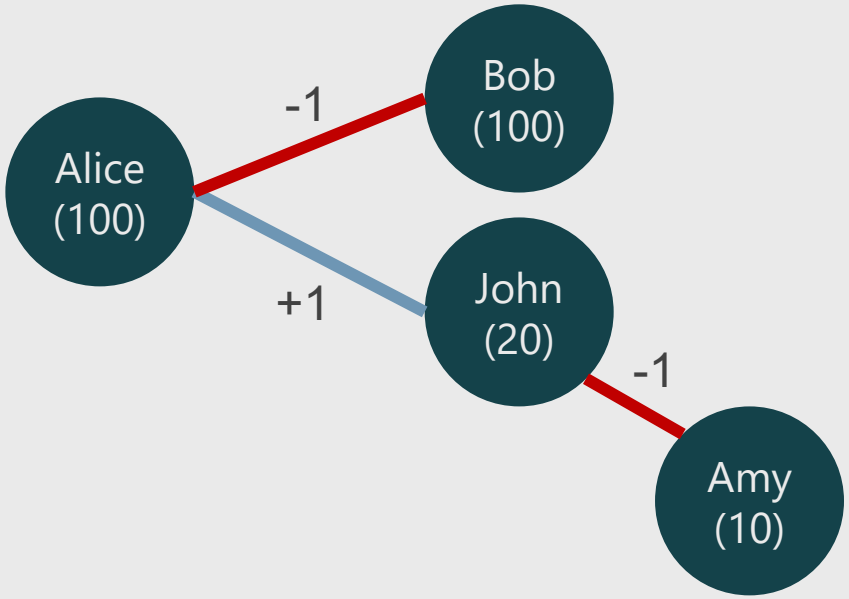


Other types of networks: Multiplex



Other types of networks: Signed

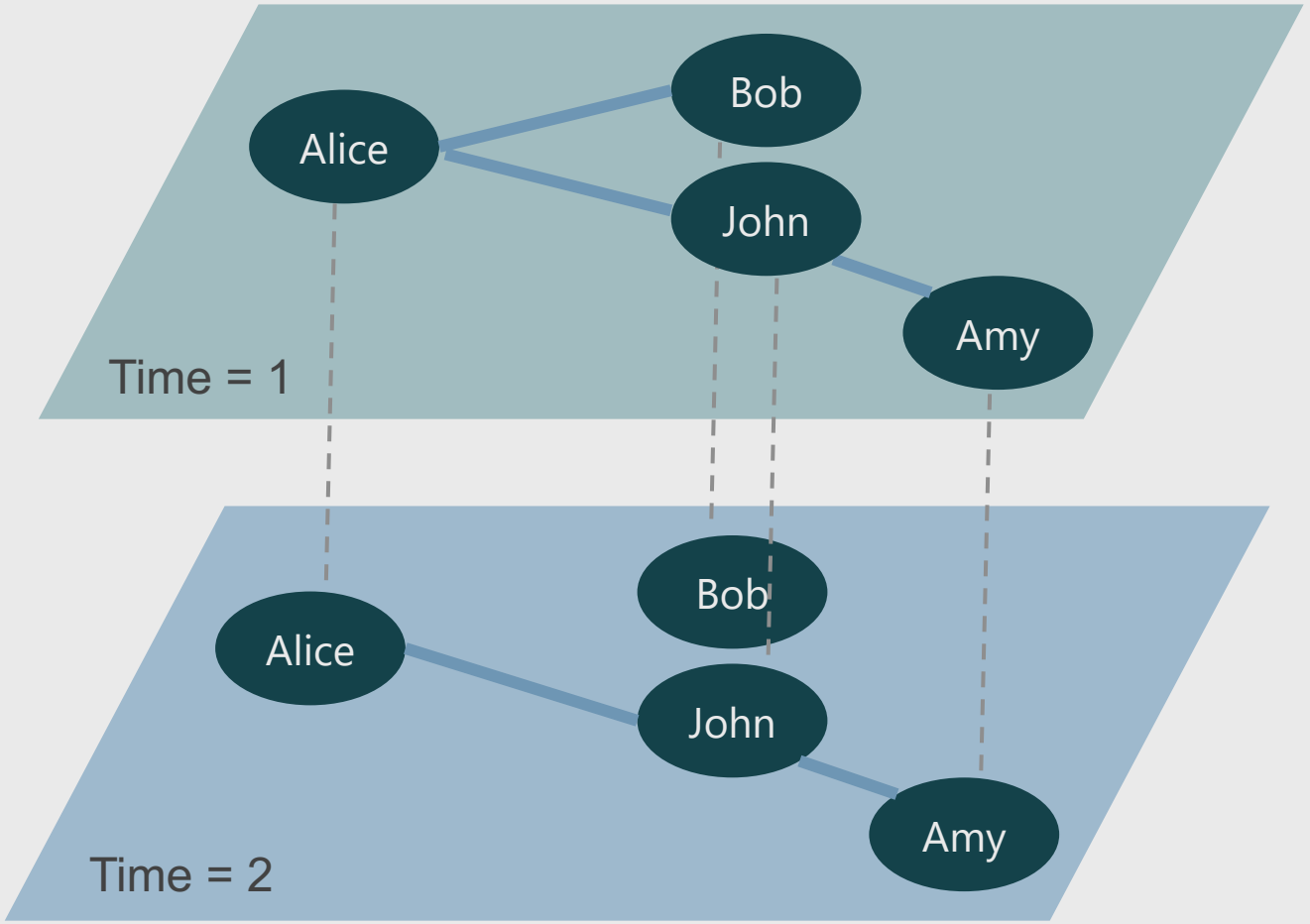
Structural balance



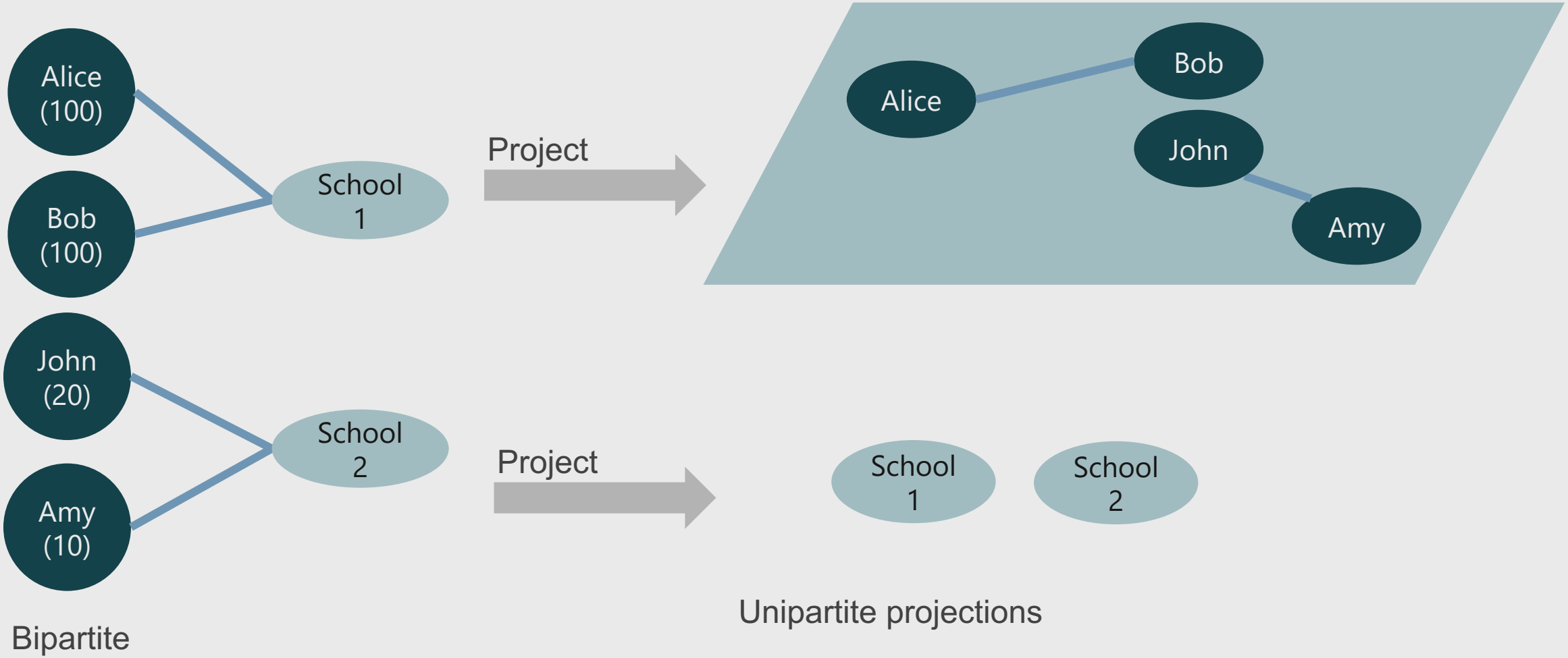
Other types of networks: Temporal

Either:

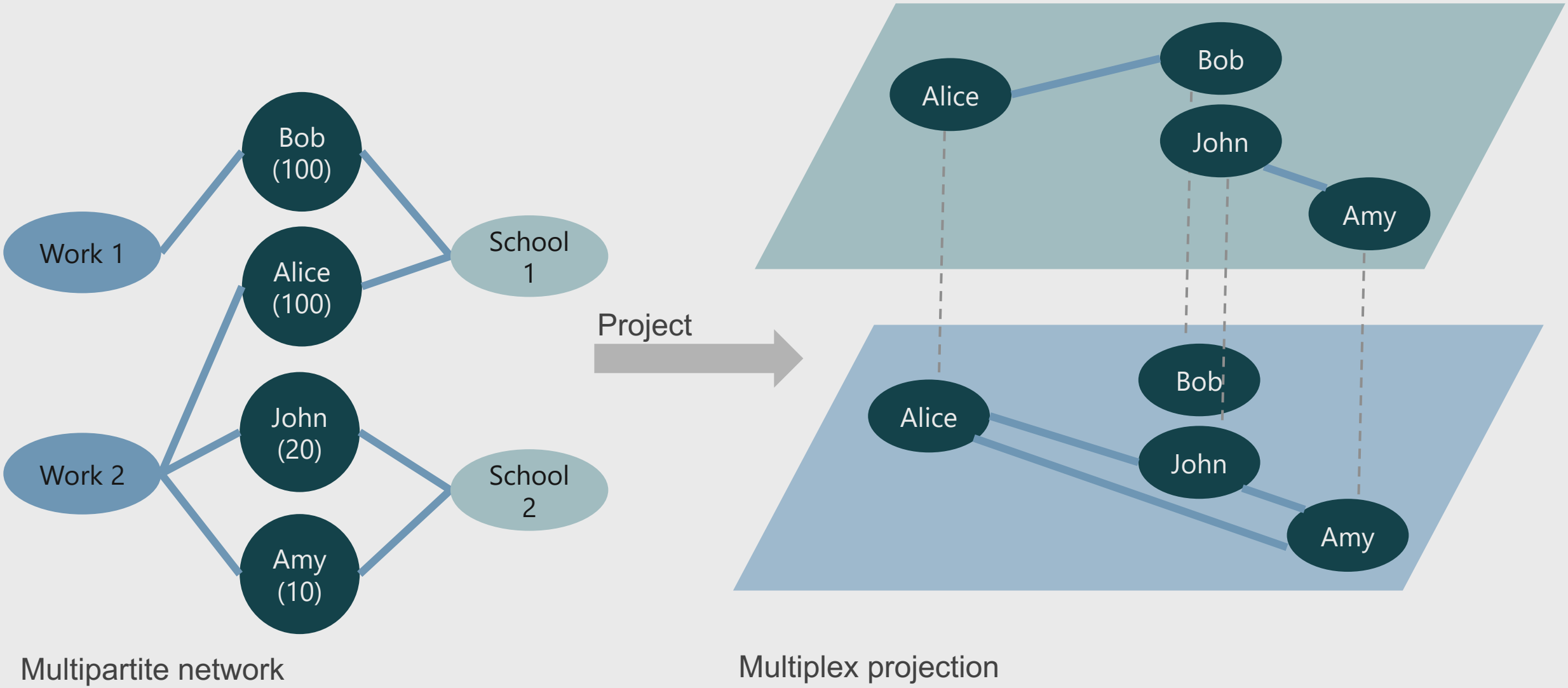
- Snapshots
- Time of events



Other types of networks: Bipartite



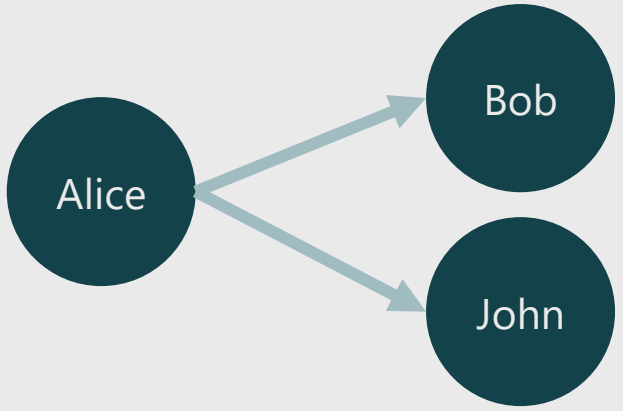
Other types of networks: Multipartite



Multipartite network

Multiplex projection

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:

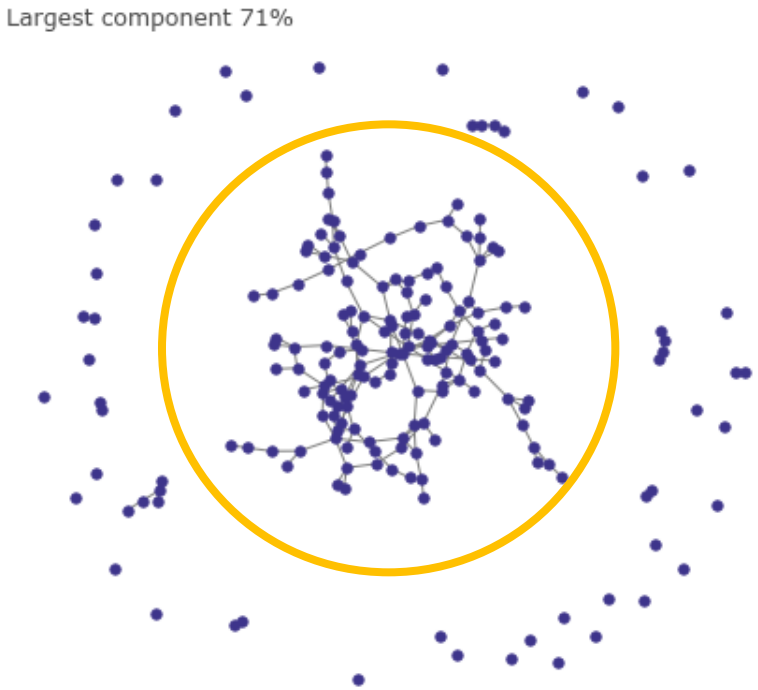
- Adv: Math is easy (matrix multiplication)
- Disadvantage: It is sparse (mostly zeros). 1 million nodes → 1 trillion numbers

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”**

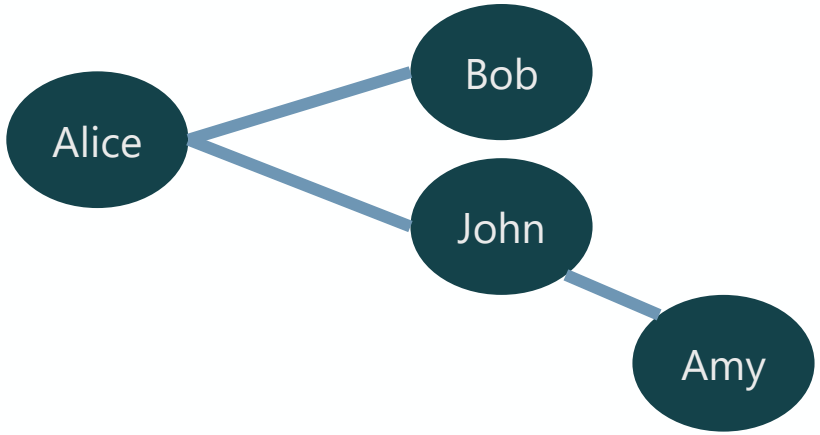
If the average degree < 1 → many small components

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

Density

Definition: Number of edges present / potential number of edges

Density = 3/6 = 50%



Larger networks are usually **sparser**

- Everybody on earth has 200 friends: edges = 200*8e12 = 1.6e15
- Possible ties: (8e12 ^ 2)/2 = 3.2e25
- Density = 1.6e16 / 3.2e25 = 0.00000005%

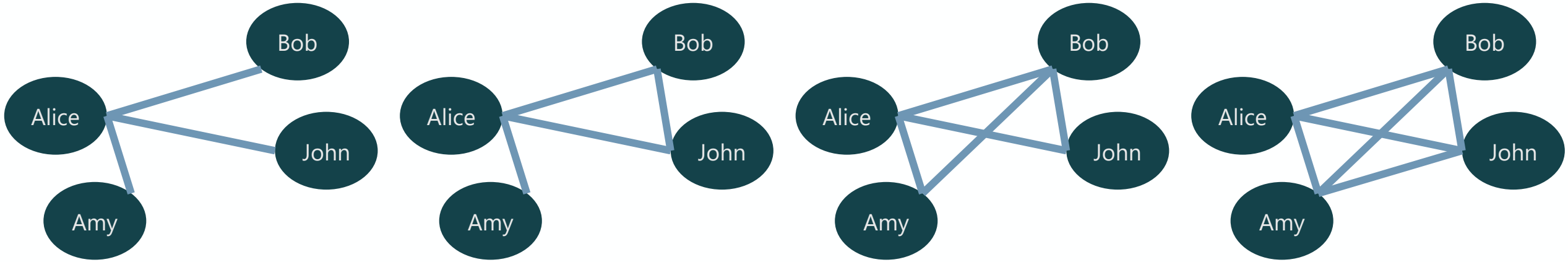
Clustering coefficient (~transitivity)

Local clustering (from the point of view of each node):

- Amongst all possible pairs of friends, how many are friends with each other?
- The share of triads (connected sets of 3 nodes) that are closed (form a triangle)

Real networks have **high clustering**

Clustering of Alice:



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

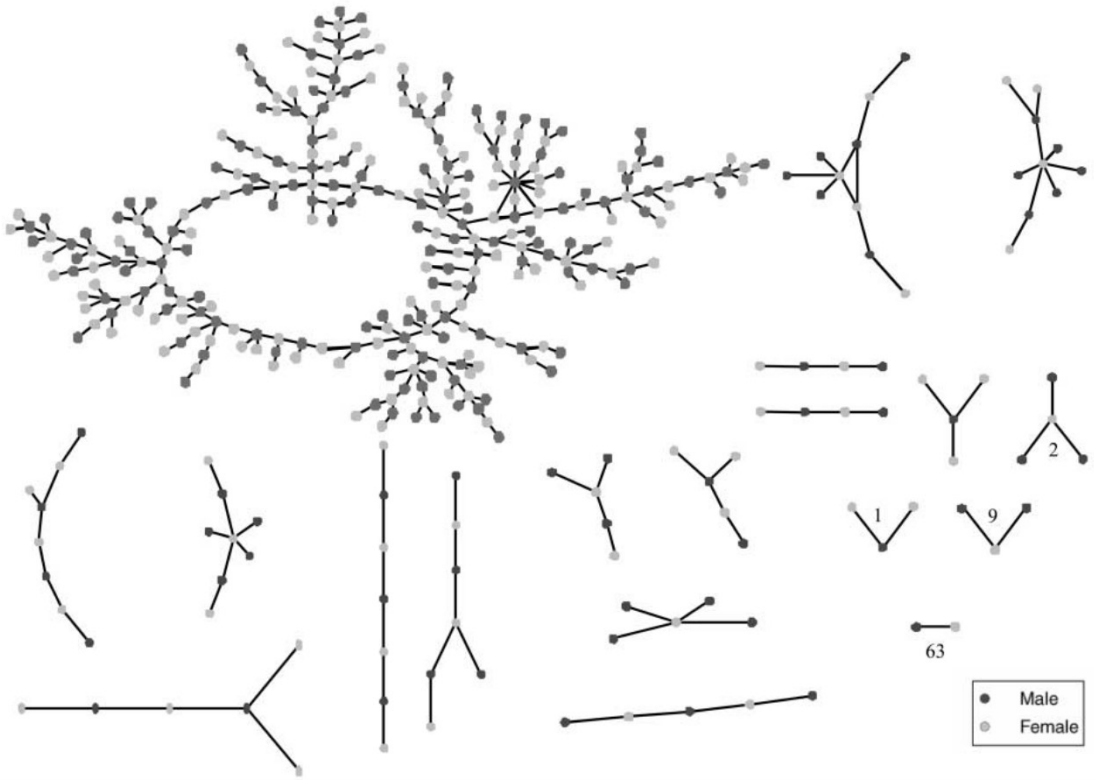
Assortativity (homophily)

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

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

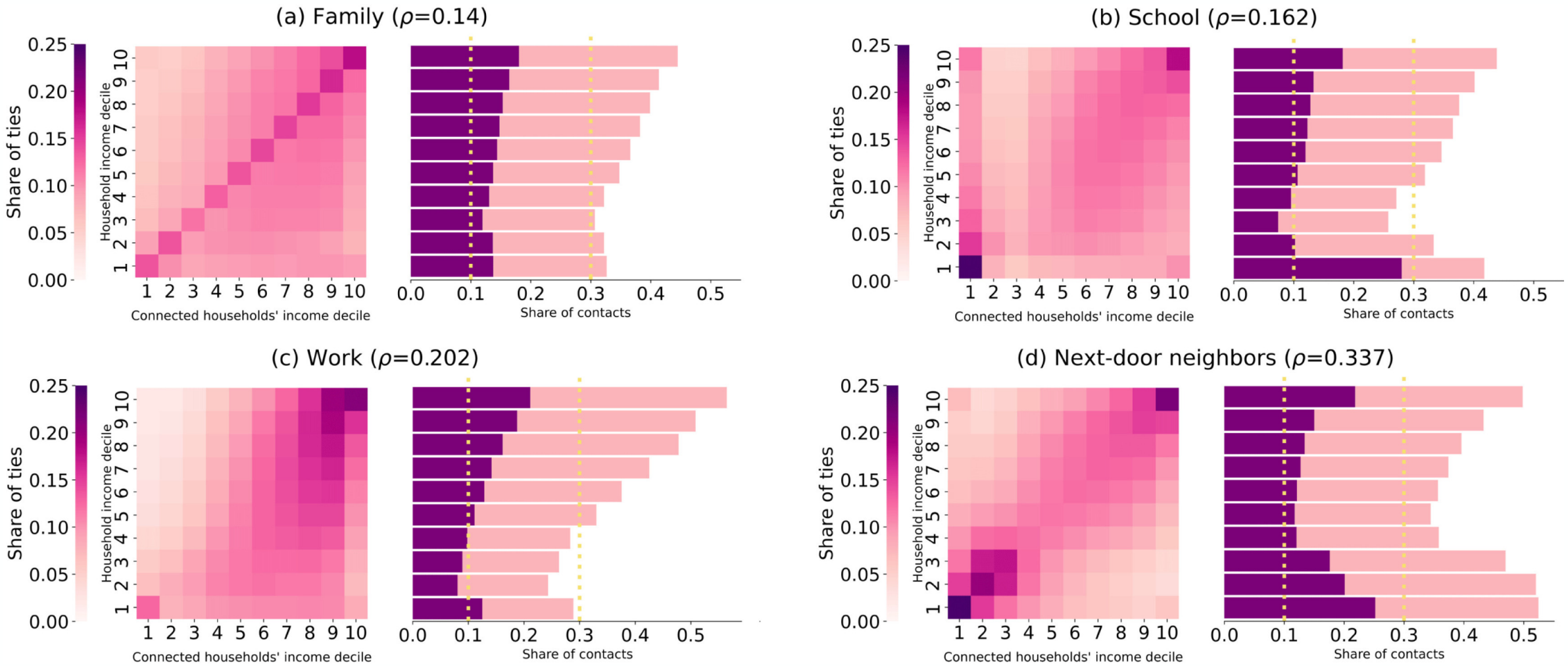
Ranges from -1 (fully disassortative) to 1 (fully assortative)

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



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

Assortativity in CBS



Socio-economic segregation in a population-scale social network, Kazmina, EM Heemskerk, E Bokányi, FW Takes Social Networks 78, 279-291

Small world: six degrees of separation

Milgram's experiment (1967)

Image source: Drewonwiki, Wikipedia

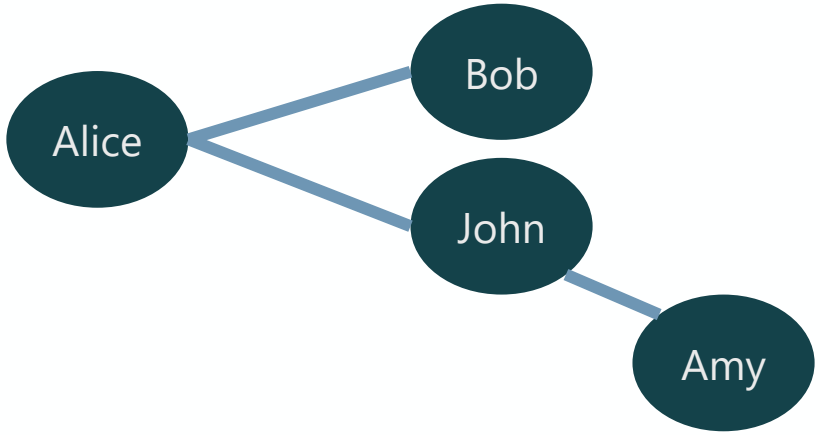
Shortest path between node 1 and node 2:

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

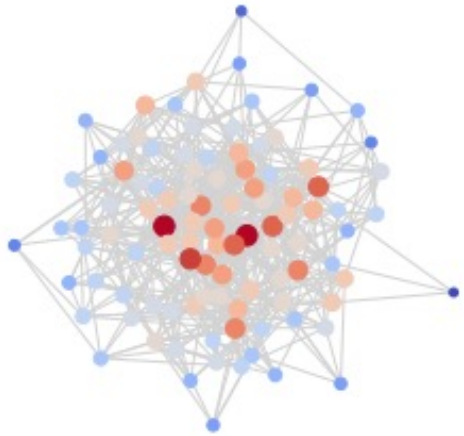
Diameter:

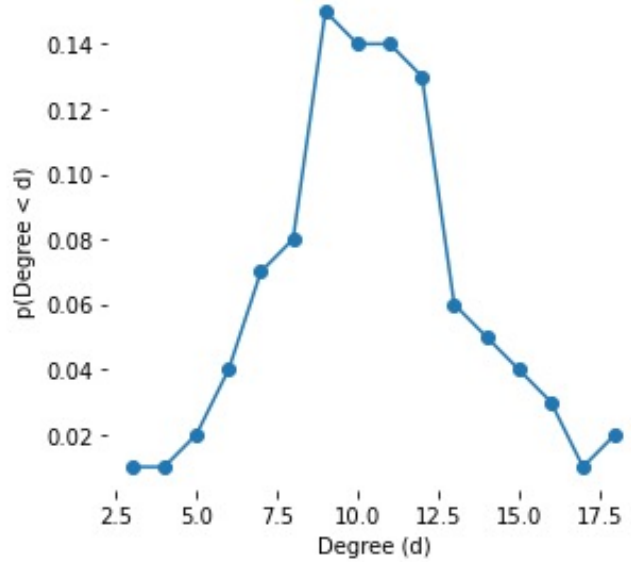
- Longest "shortest path" between two nodes
- In our network: 2 (Alice -> John -> Amy)

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

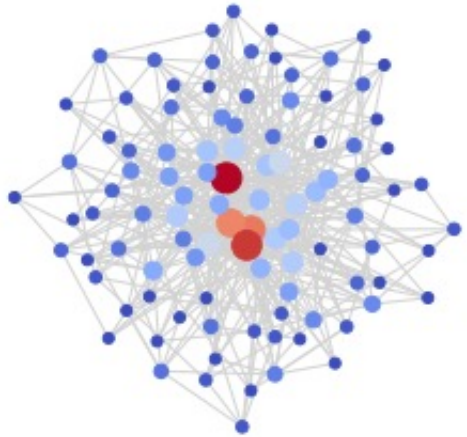


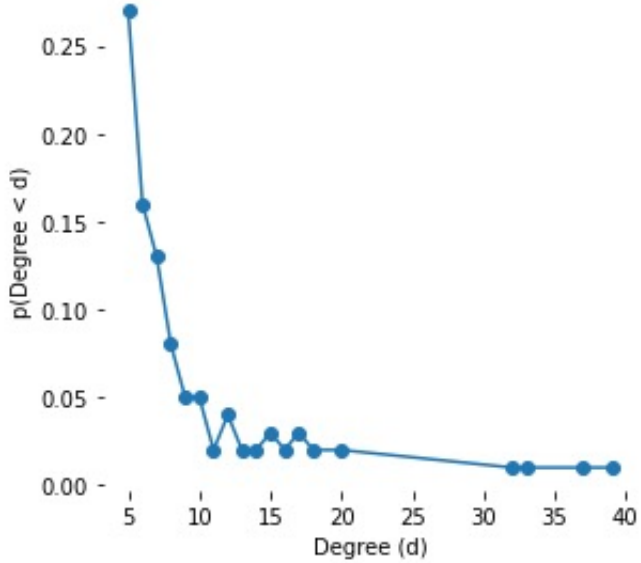
Skewed degree distributions





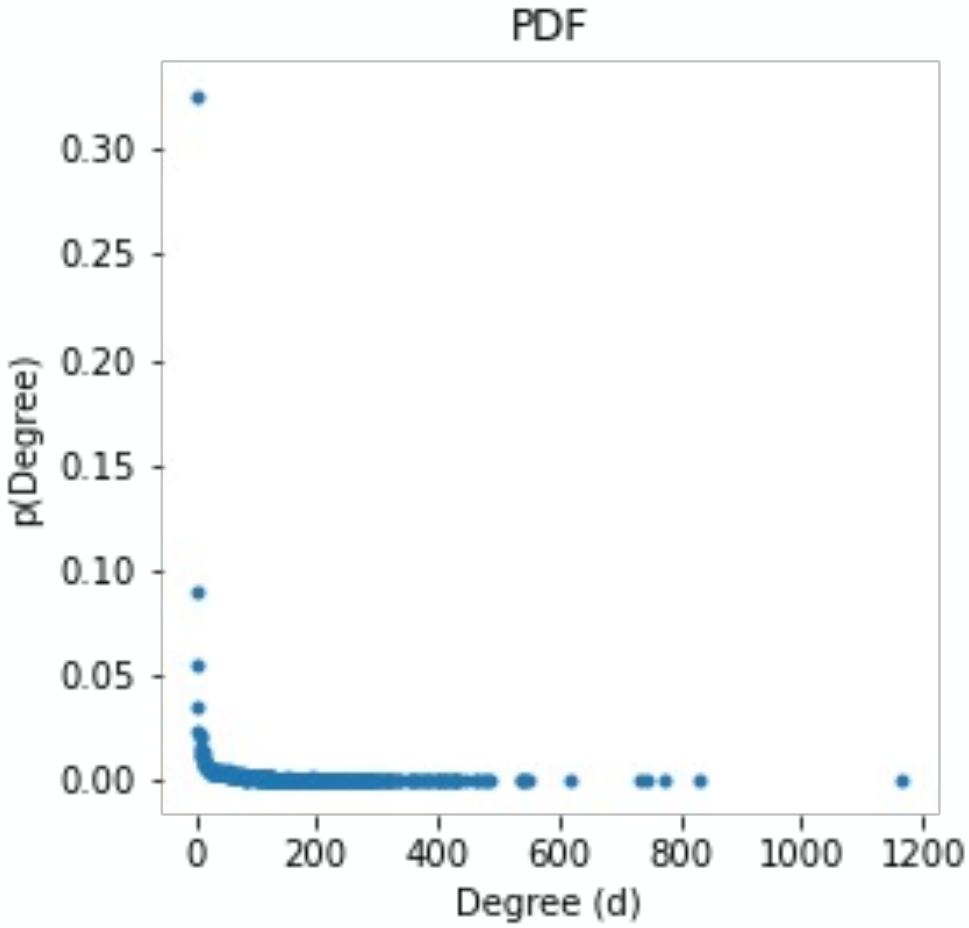
Random network

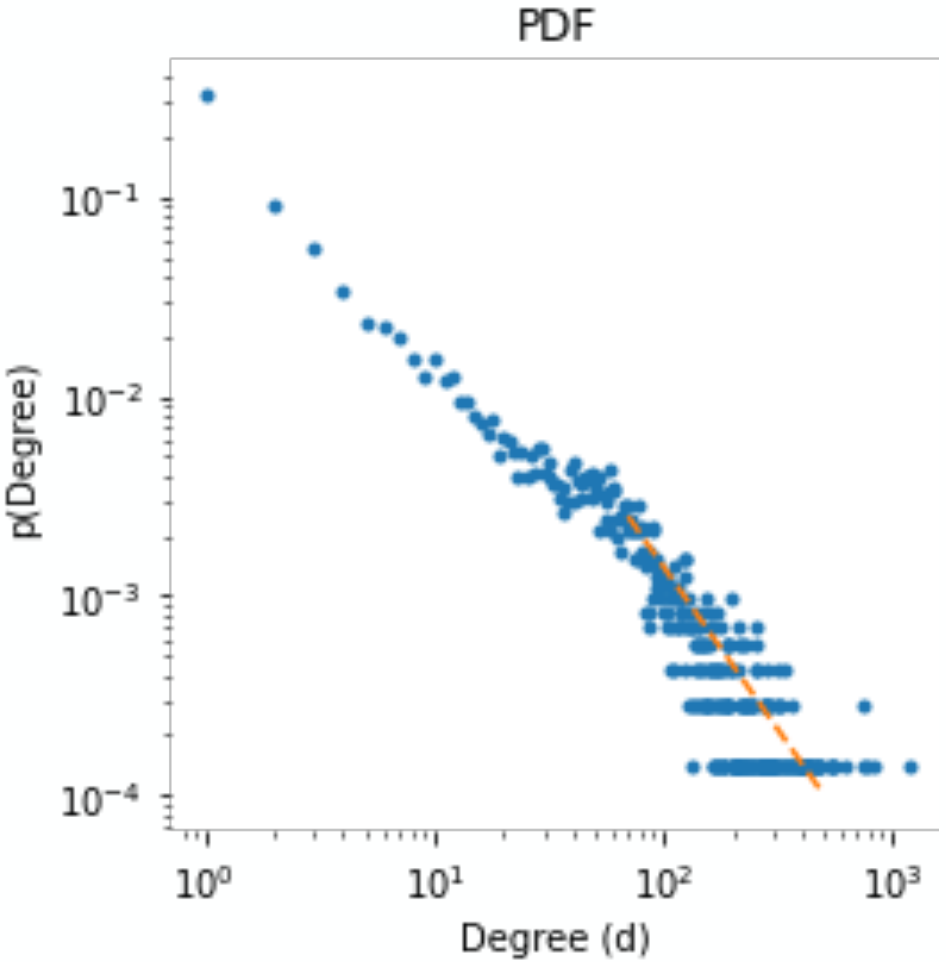


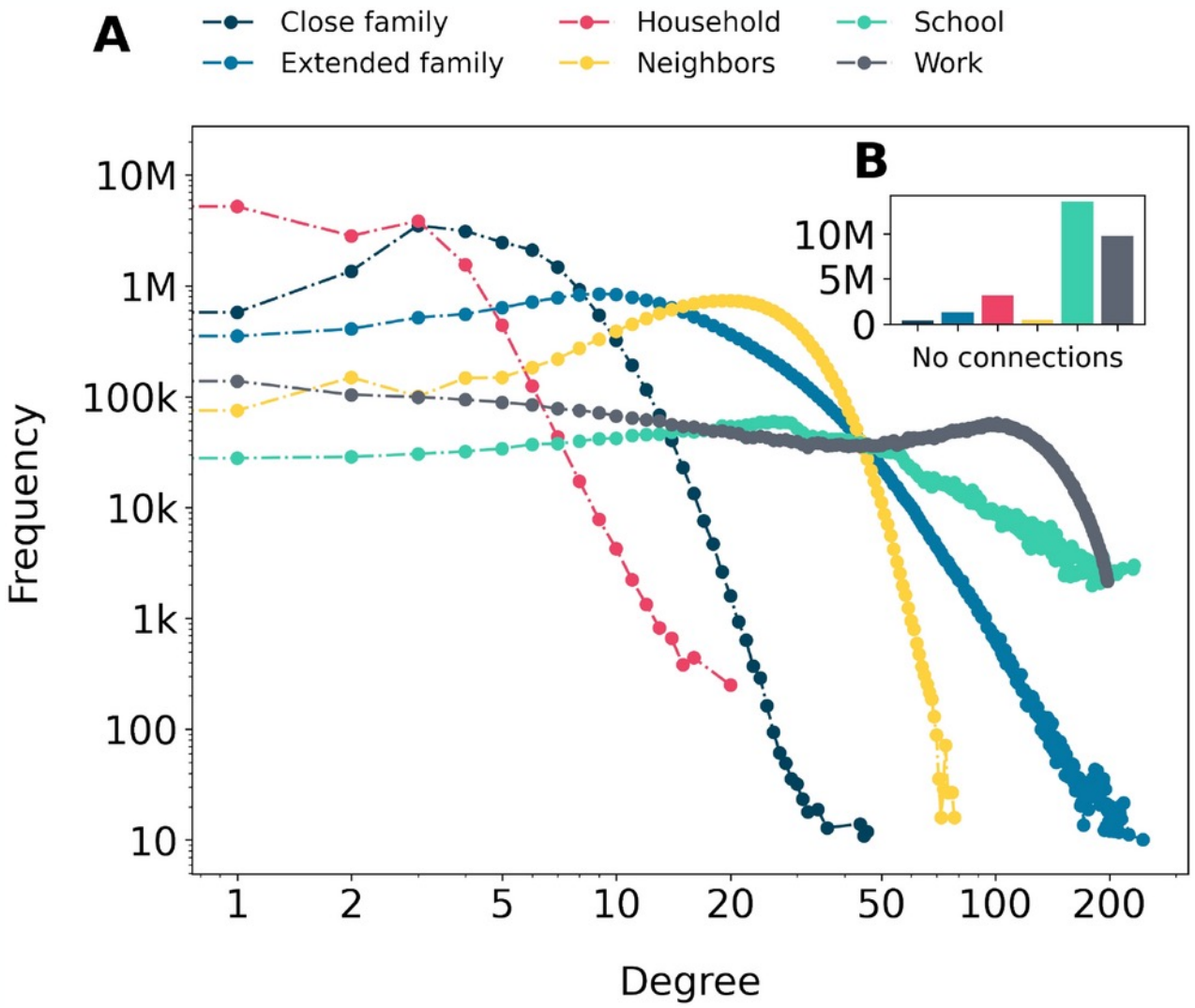


Realistic network

Real network (Wikipedia mentions)







Careful: The data is a bit different now!
Especially the neighbors

The anatomy of a population-scale social network, E Bokányi, EM Heemskerk, FW Takes
Scientific Reports 13 (1), 9209

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

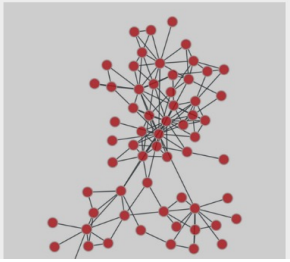
https://aaronclauset.github.io/datacode.htm

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

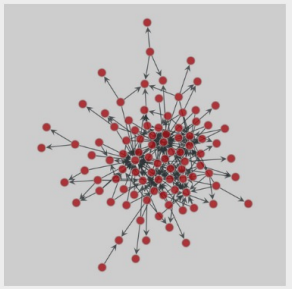
https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/swingers

Networks

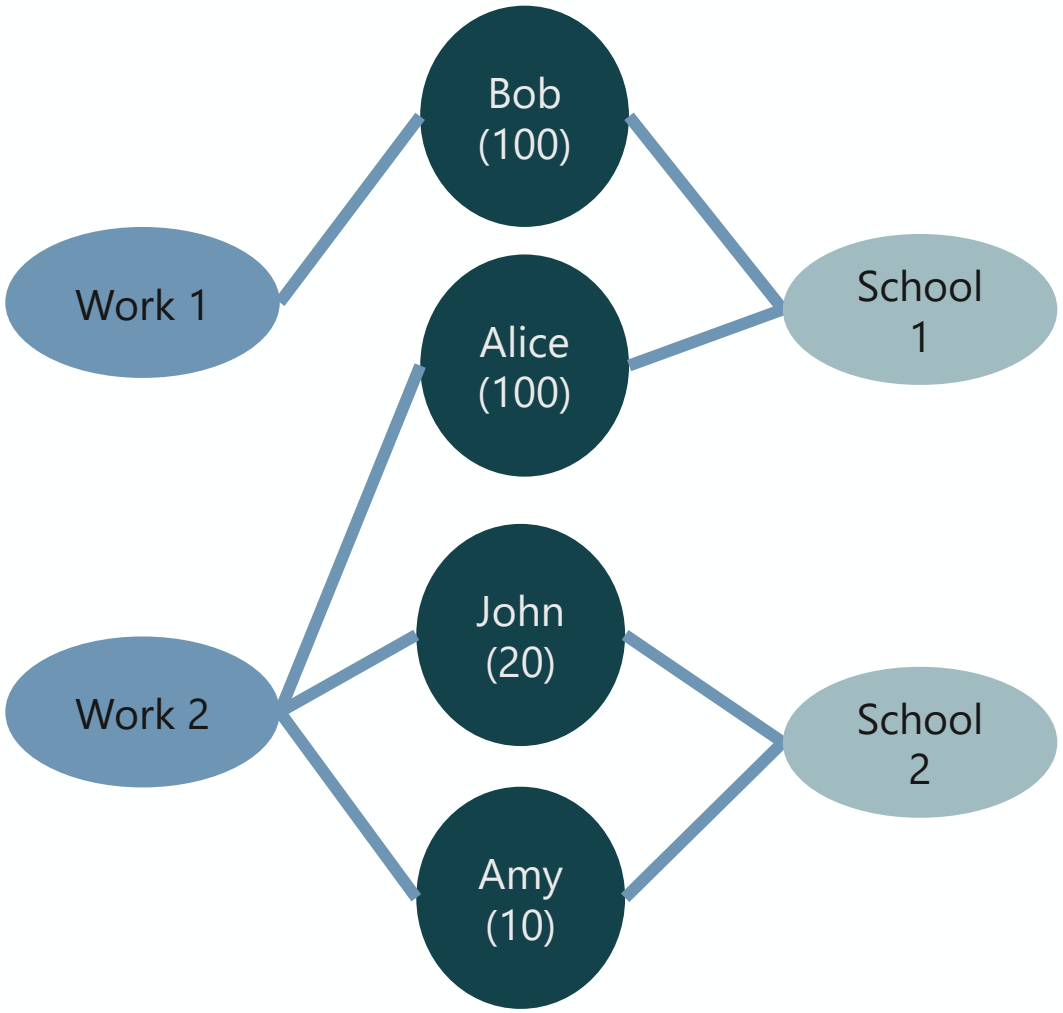
Tip: hover your mouse over a table header to obtain a legend.

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

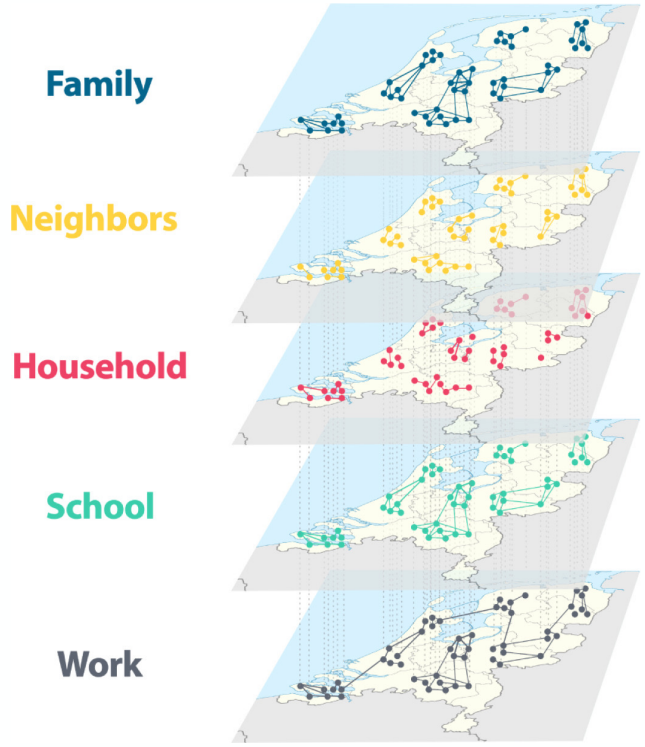


Affiliation networks at CBS



Multipartite network

Available networks:



How does the CBS newtorks look like?

→ Giant component → Most nodes are connected

→ Small world → Small diameter

→ Low density

→ Form cliques! (high clustering/transitivity)

→ Assortative (homophilic)

→ Heavy tail distributions

javier.science/panel_network

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: Centrality

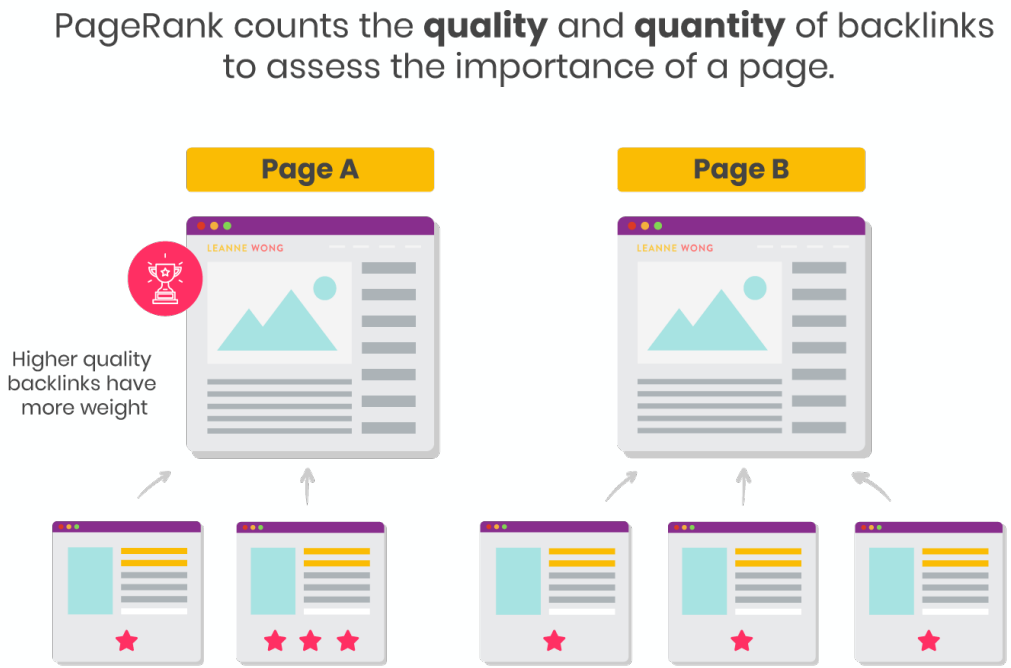
Who are the key actors in the network?

Centrality measures provide answers to this question.

How to stop the spread of diseases?

Important nodes: bridges

How to sort Google results?



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

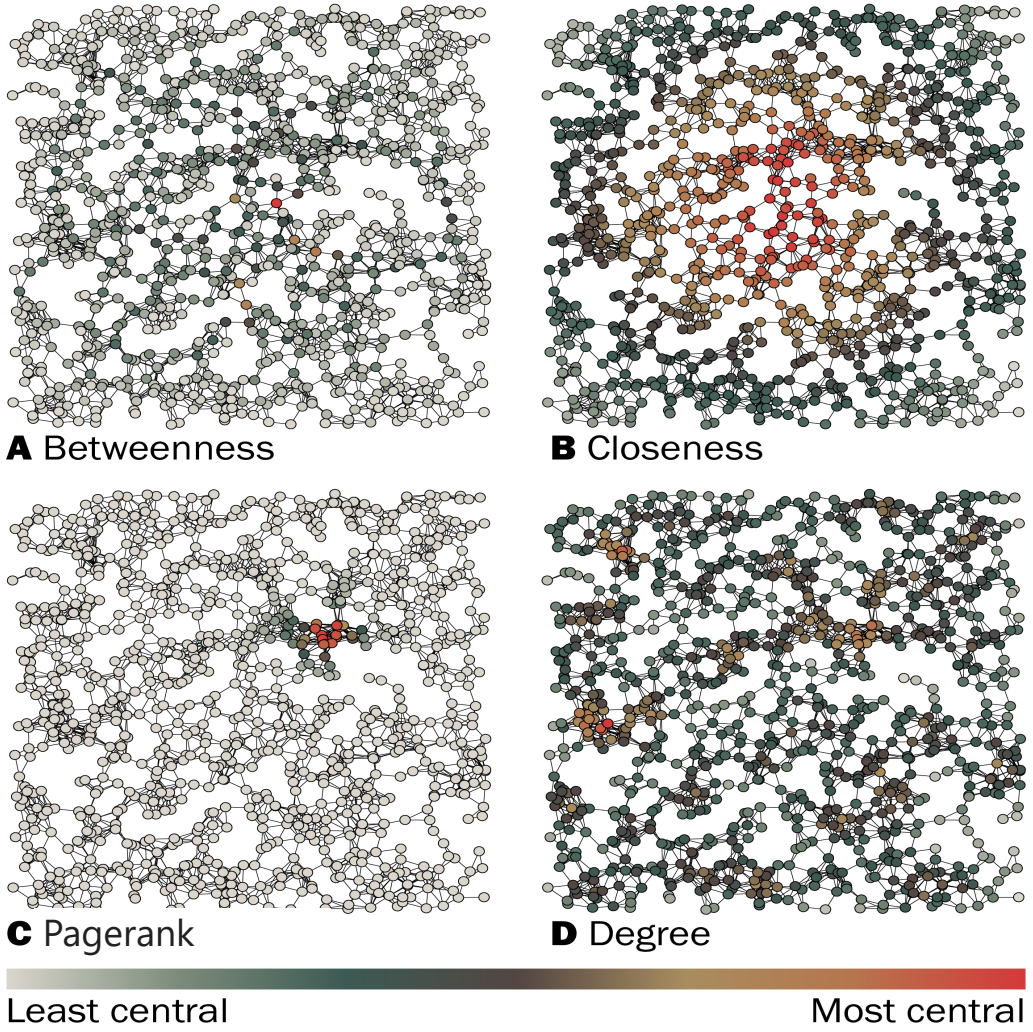
Important nodes: those linked by important nodes

Centrality

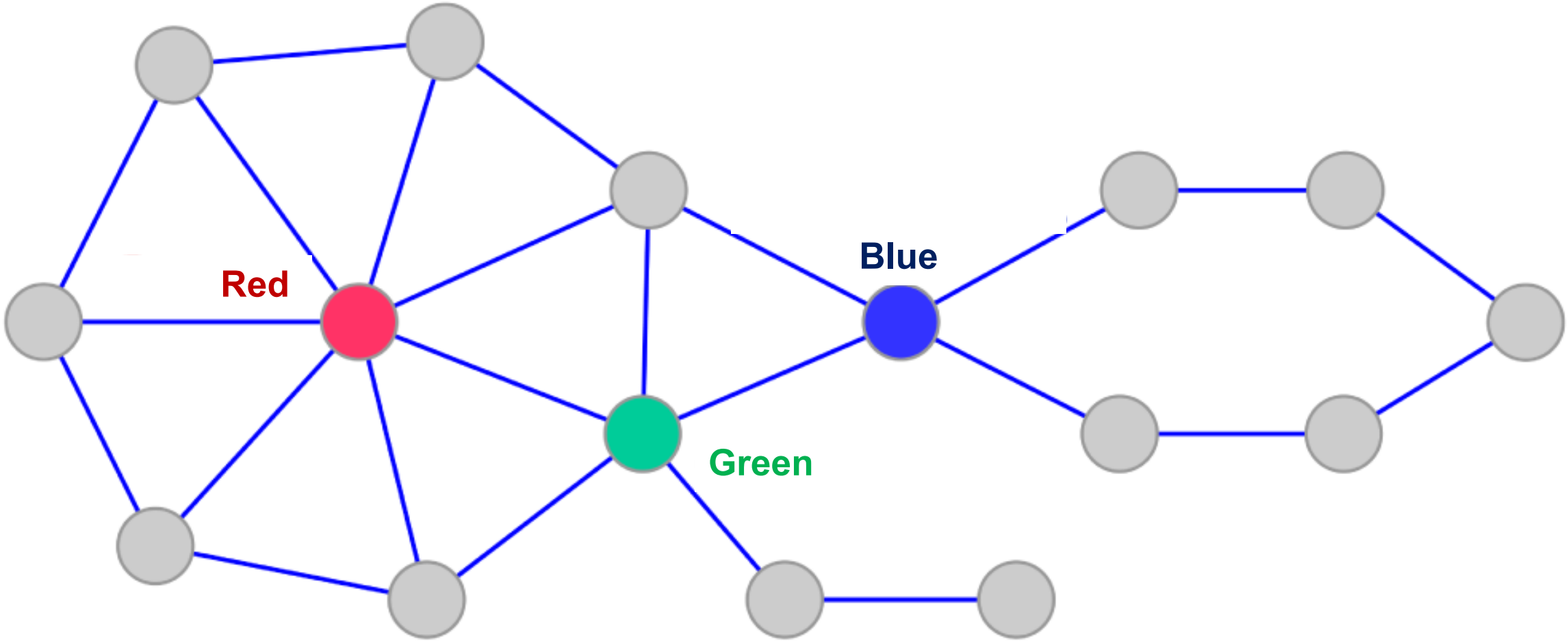
Different centrality measures define importance in different ways :

- *Degree*: Connected directly to many nodes
- *Closeness*: Close to all other nodes (on average)
- *Betweenness*: In the middle of many shortest paths
- *Pagerank/eigenvector/katz*: 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.



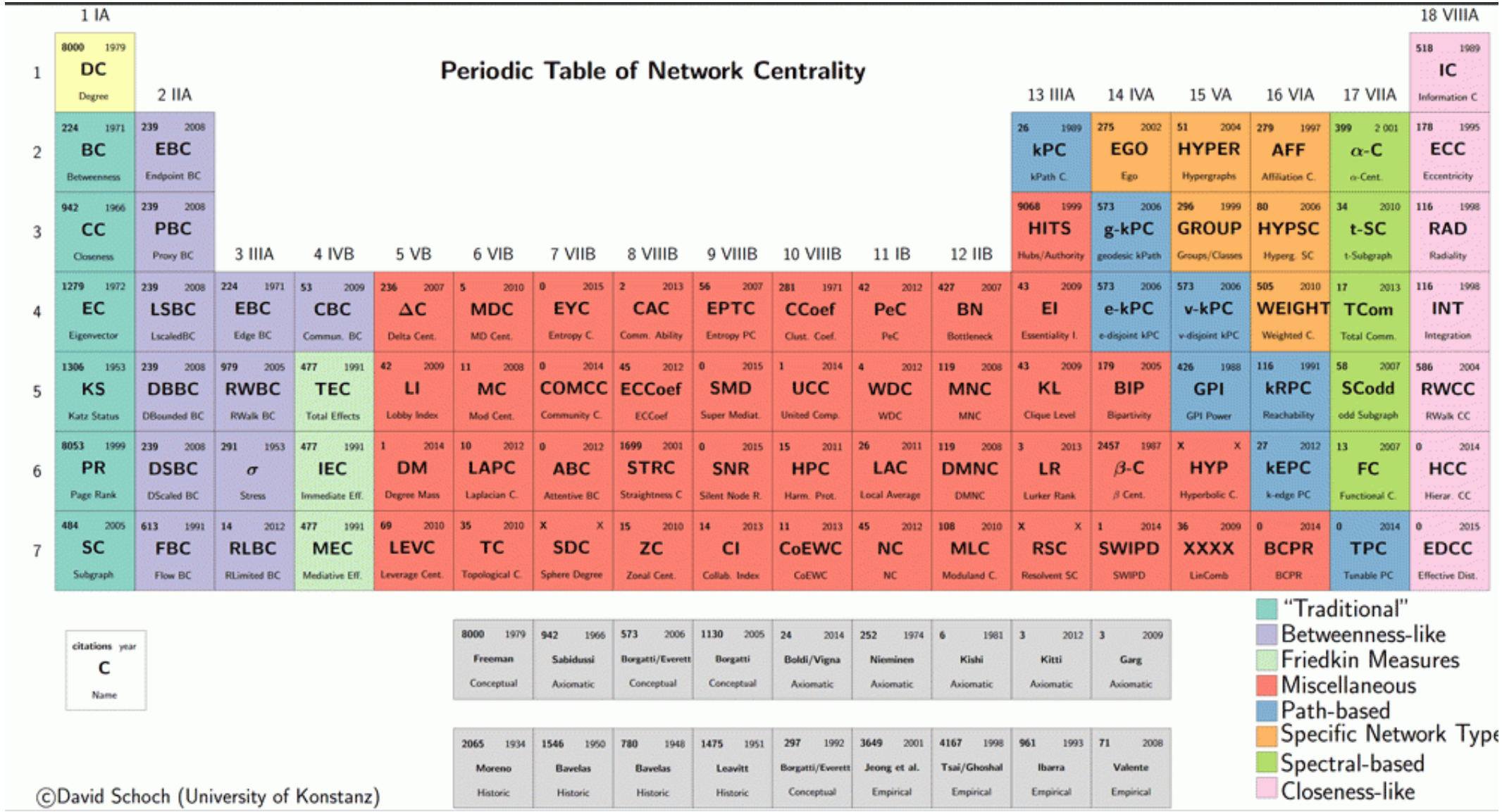
Which node has higher degree/betweenness/closeness?



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

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

Consider what is the objective (e.g. is it to enable low-income individuals to increase their social capital?) *(https://petterhol.me/2019/01/11/the-importance-of-being-earnest-about-node-importance/)*



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: Community detection

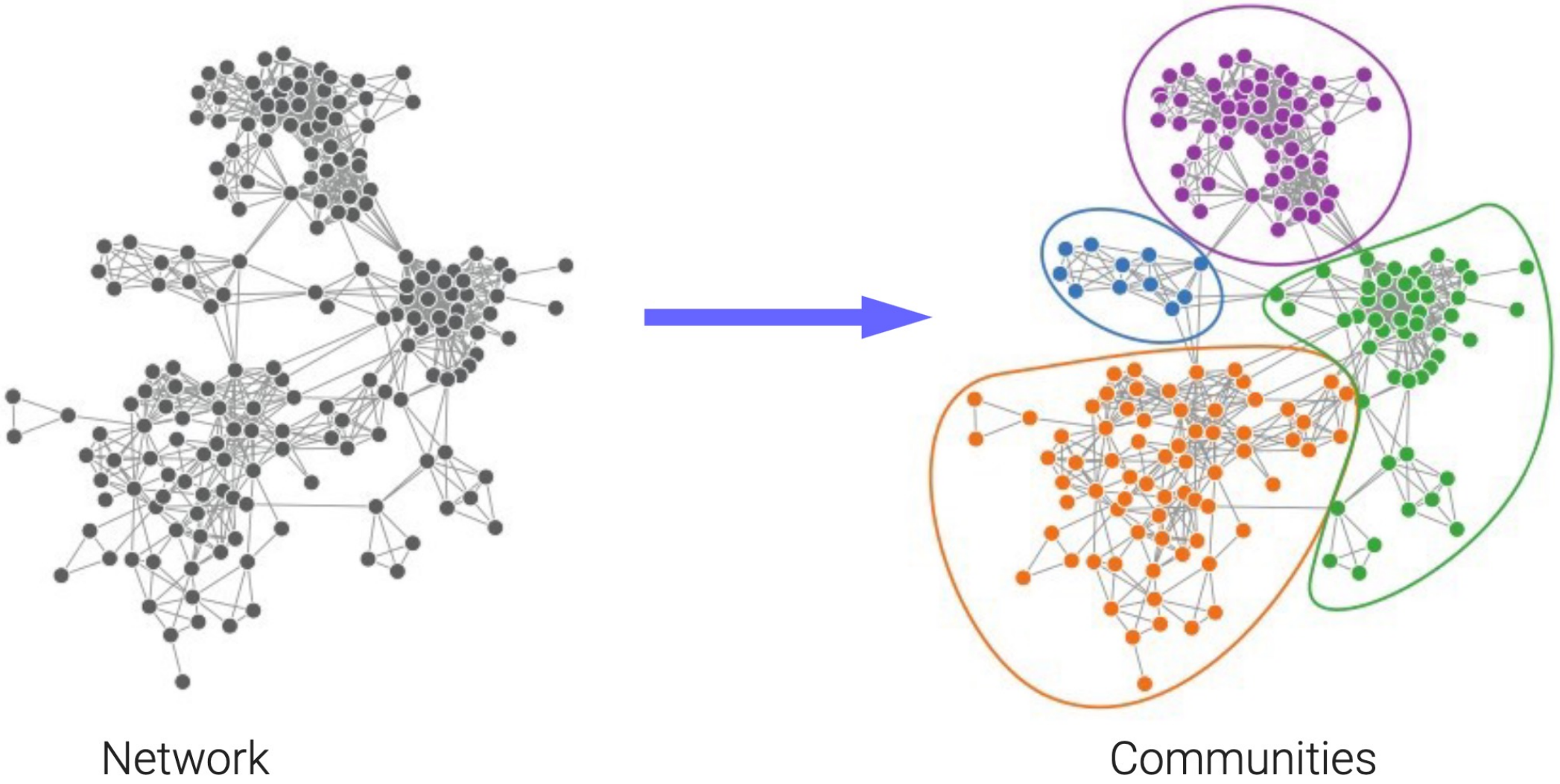


Image source: Leto Peel

How to partition the network?

There are many algorithms to partition the network into communities

Often: the algorithms detect communities so that there are **many links within communities and few links across communities**

Main example: communities that maximize modularity (Q)

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

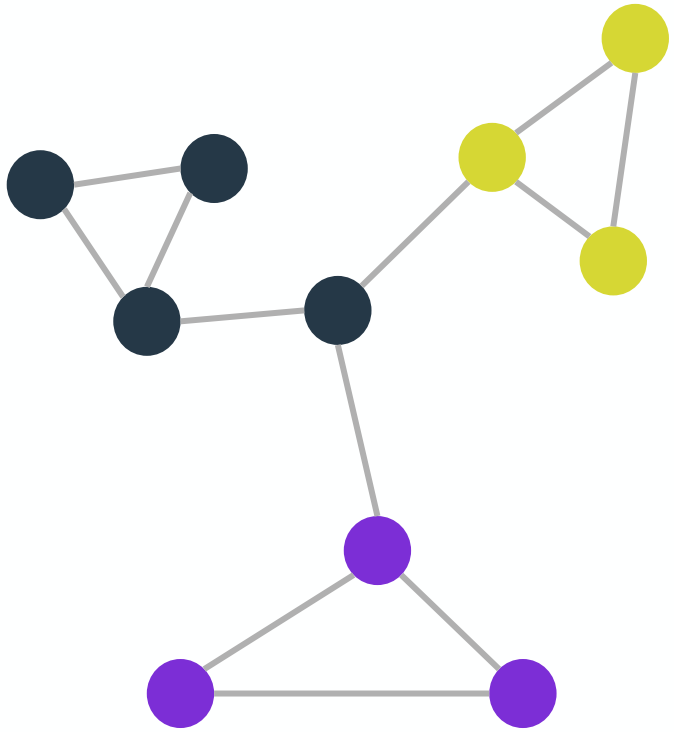
Fraction of links inside community c

Expected fraction of links within a community in a random network

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

$k_i = degree\ of\ i.$
m = number of edges

How to partition the network?



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

Fraction of links inside community c

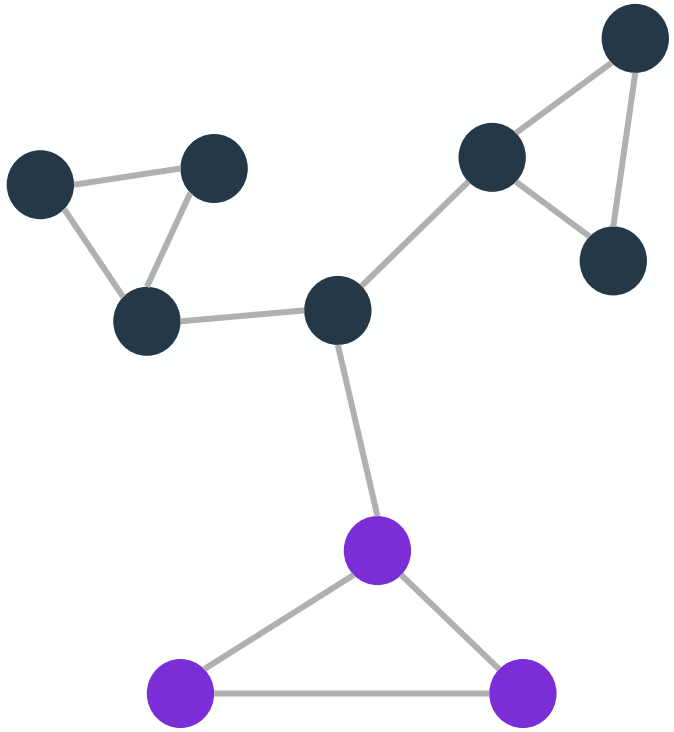
Expected fraction of links within a community in a random network

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

$k_i = degree\ of\ i.$
m = number of edges

	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

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

$k_i = degree\ of\ i.$
m = number of edges

	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**

How to partition the network?

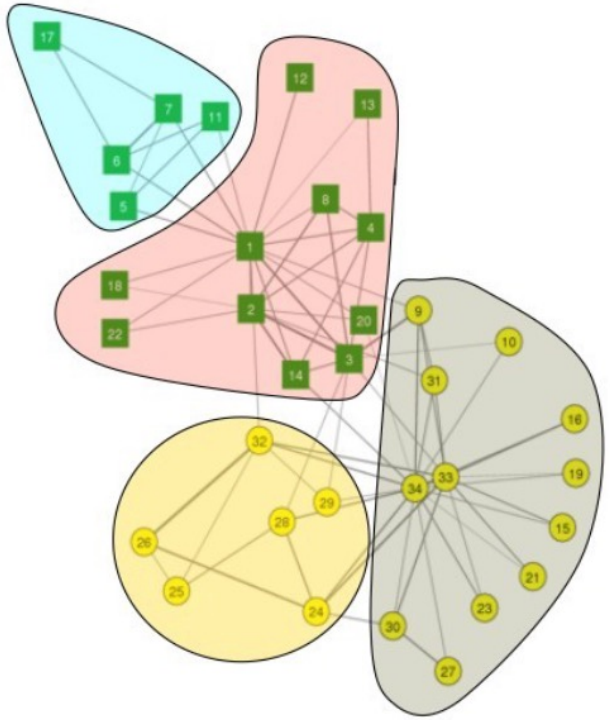
Algorithms for modularity maximization (and related methods)

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

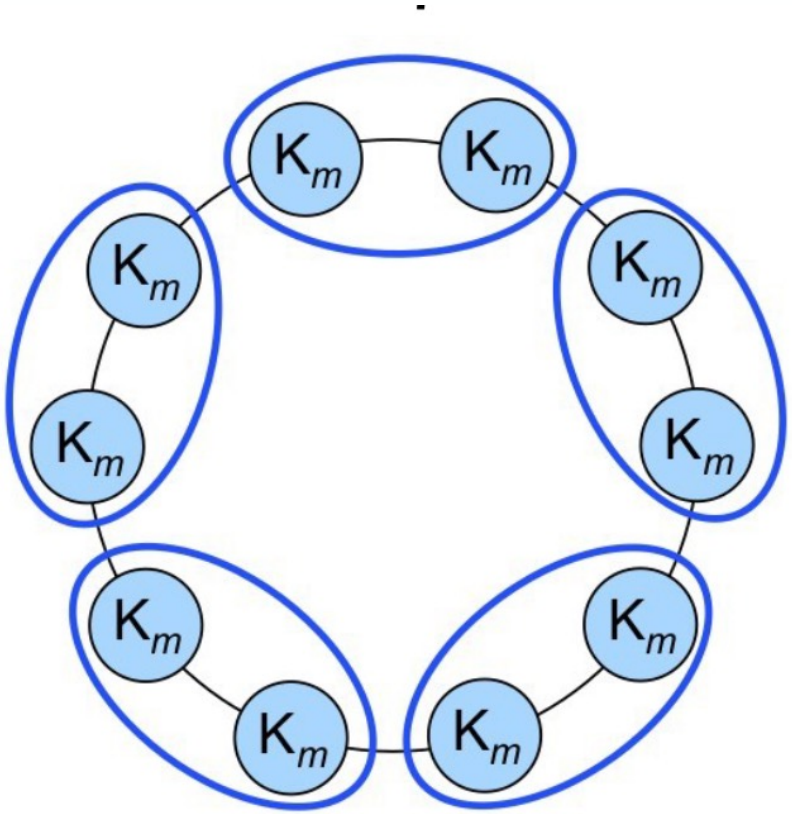
Other algorithms

- 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 maximization



Finds spurious communities
(overfitting)



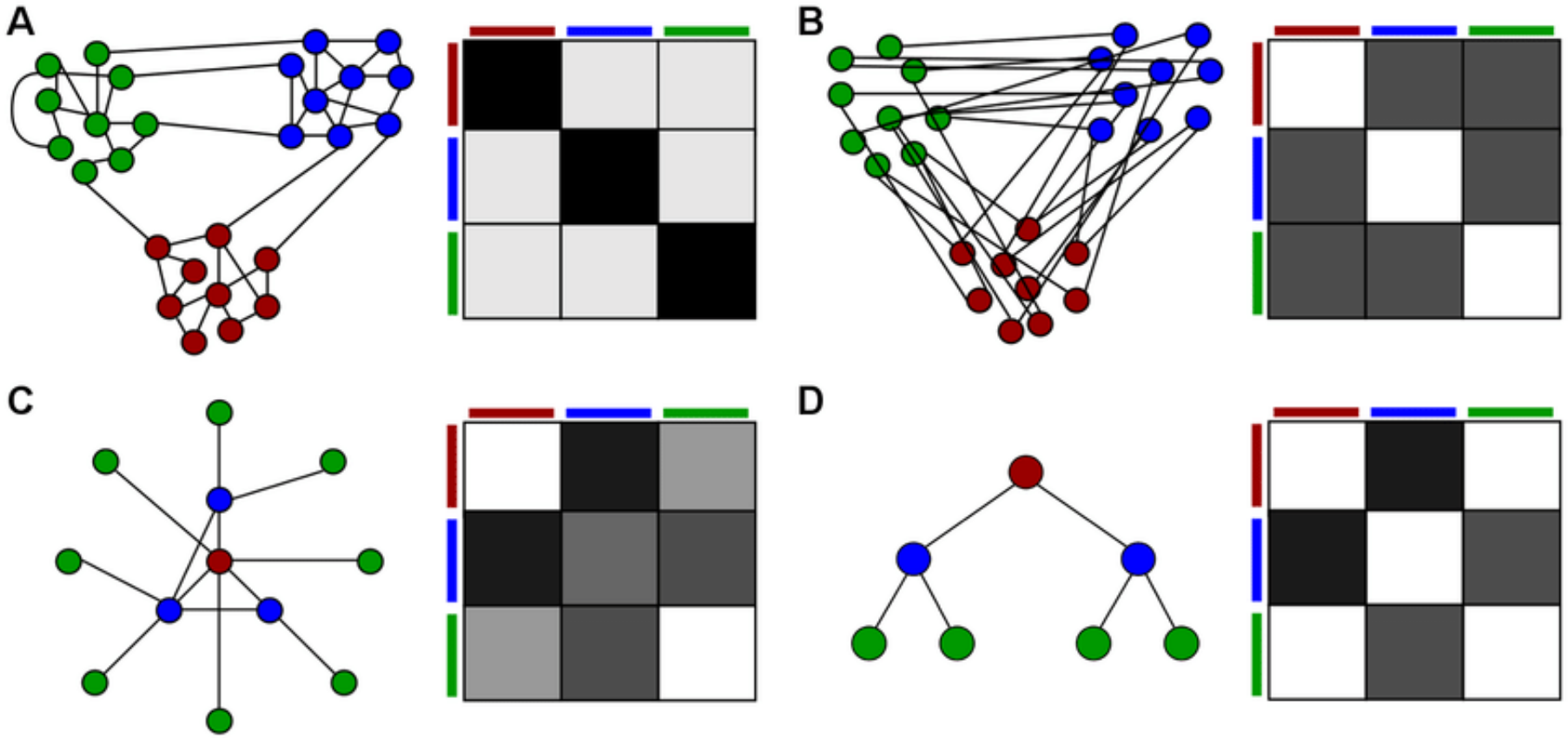
Resolution limit
(underfitting)

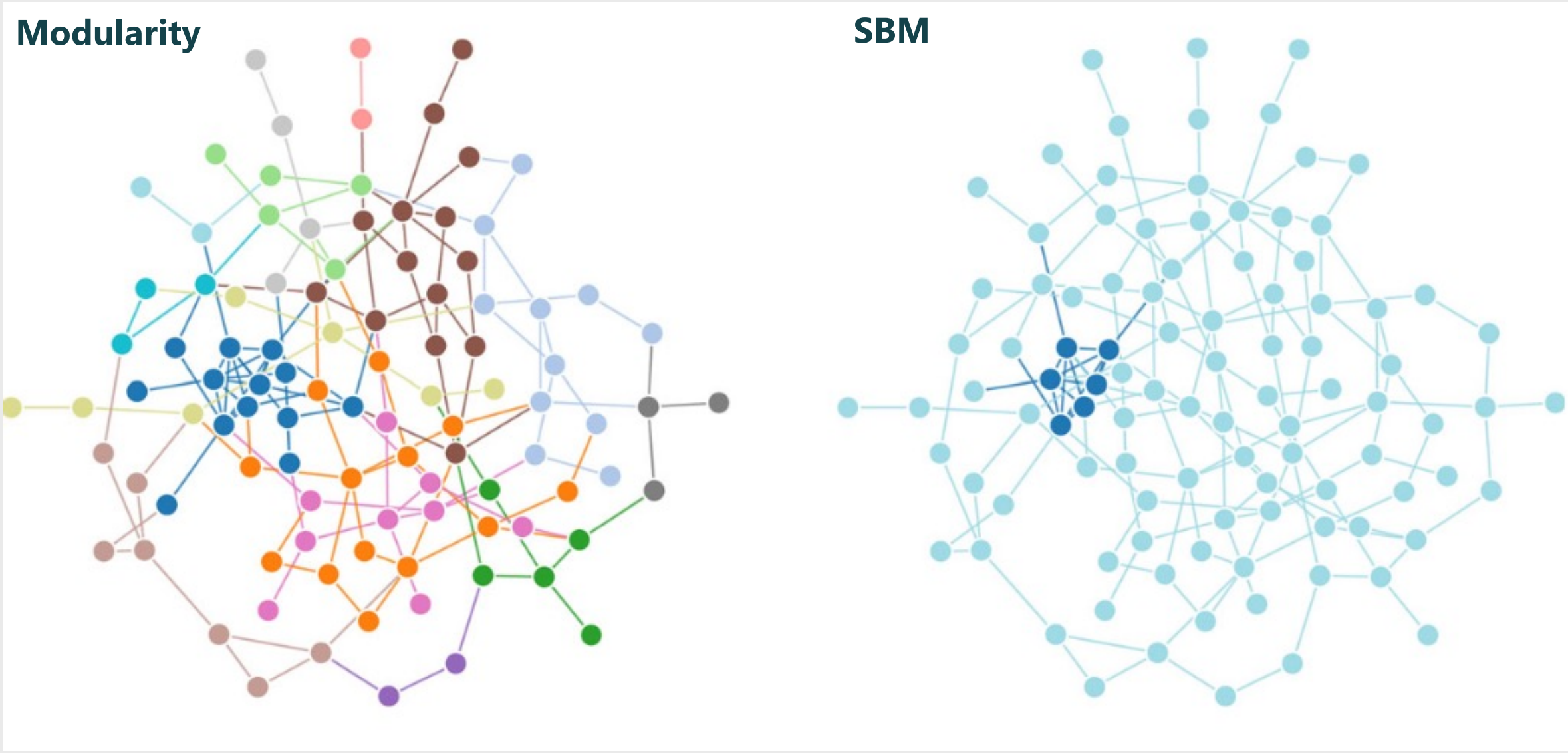
Stochastic Block Model (SBM)

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

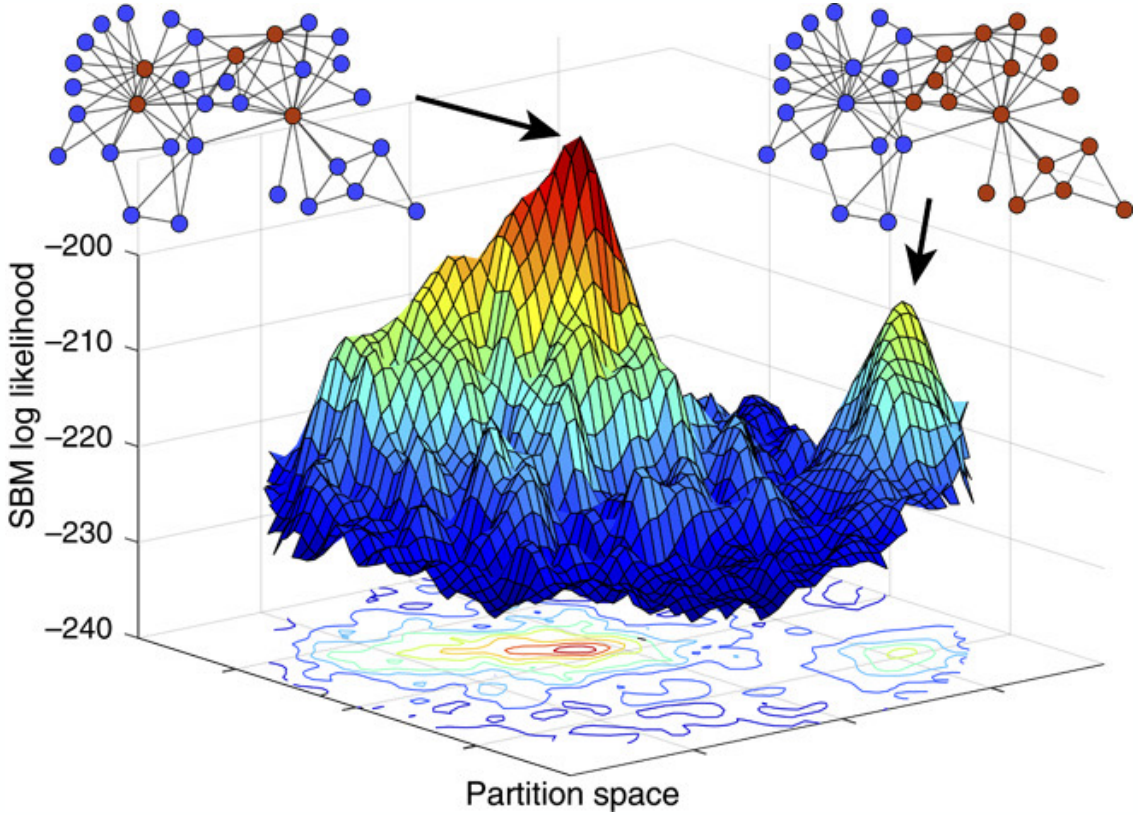
Can find other connectivity patterns apart from the ones defined by modularity maximization (“many connections within communities, few between communities”)

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.



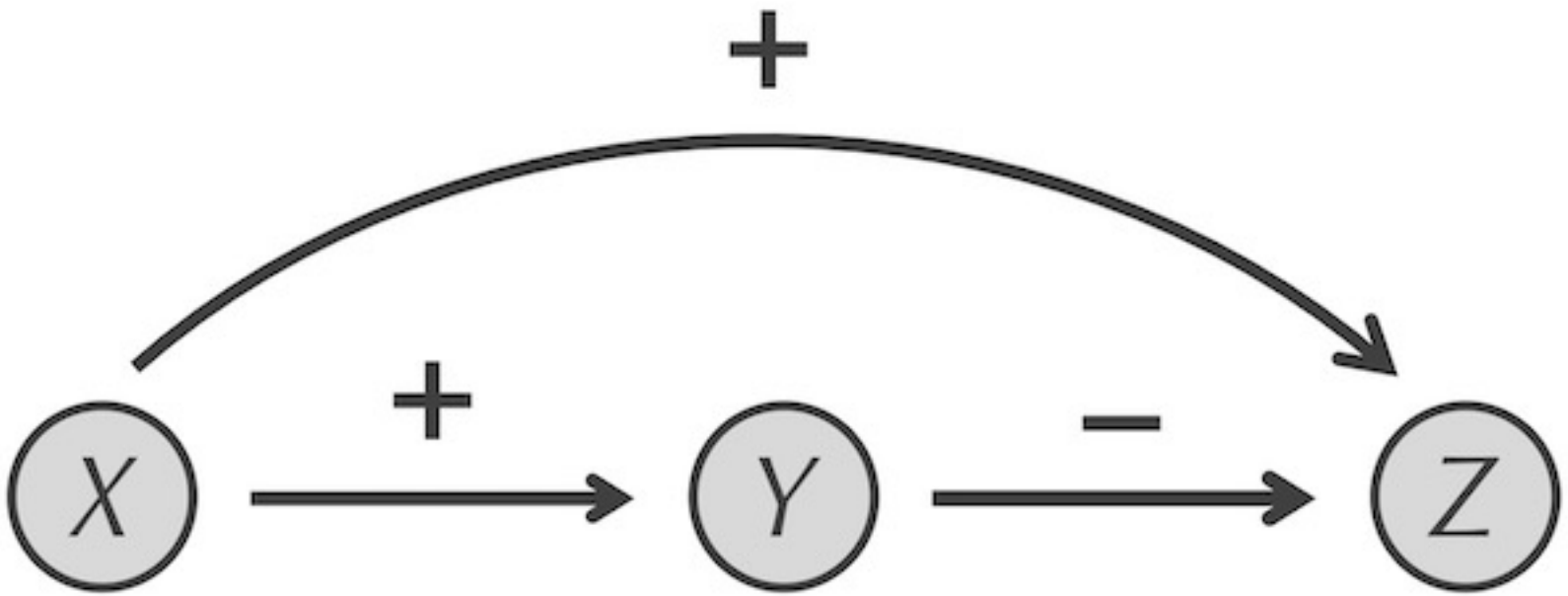
Many good partitions exist!

"It is standard practice to treat some observed discrete-valued node attributes, or metadata, as ground truth. **We show that metadata are not the same as ground truth and that treating them as such induces severe theoretical and practical problems**. We prove that no algorithm can uniquely solve community detection, and we prove a general No Free Lunch theorem for community detection, which implies that there can be no algorithm that is optimal for all possible community detection tasks" *(Peel, Larremore, Clauset, 2017)*

Some other type of analysis

Types of analysis: Motif detection

Find overrepresented patterns

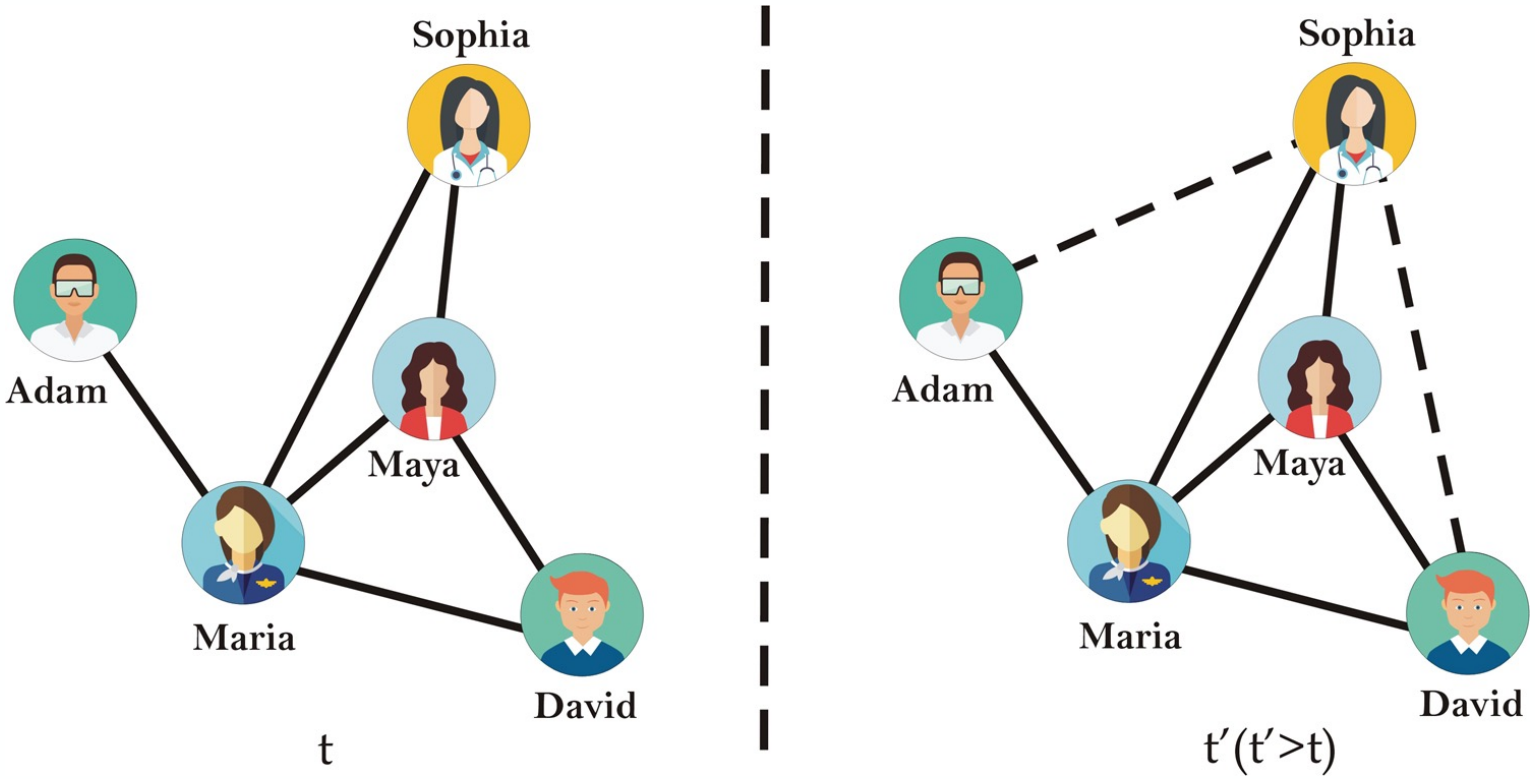


Feed-forward loop (https://biologicalmodeling.org/motifs/feedforward)

Types of analysis: Link/metadata prediction

Networks are rarely complete

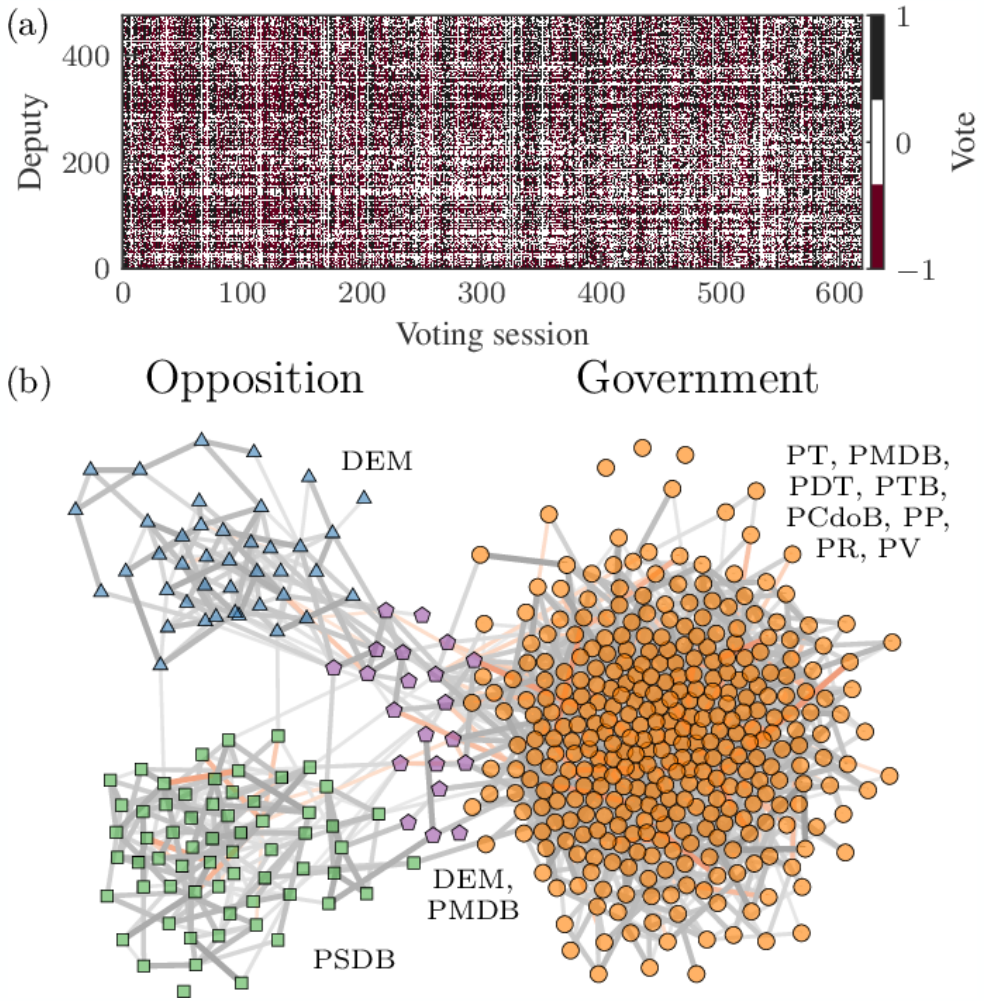
Link prediction approach: Approaches such as triangle closure



Ahmad et al 2020

Types of analysis: Network reconstruction

Network from co-occurrences



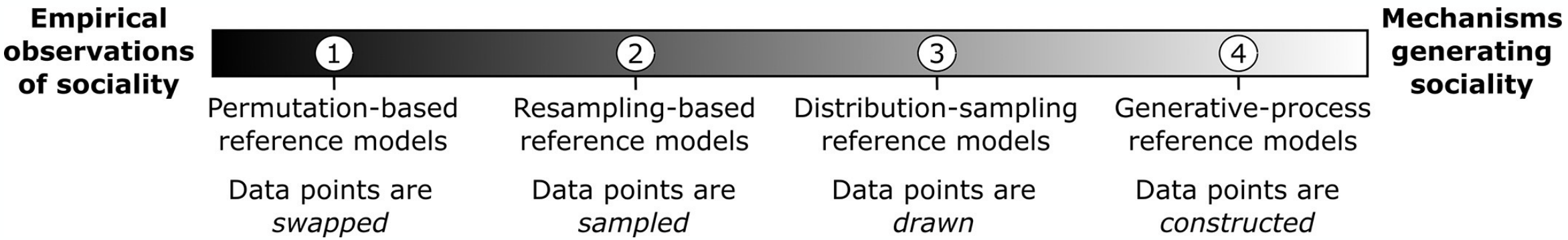
Network Reconstruction and Community Detection from Dynamics, Peixoto 2019

Types of analysis: Testing hypothesis

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

Approach: Create a reference model (see *Hobson 2021* for a great guide) to compare with it

- Configuration model (permuting edges)
- Generative models (e.g. rich get richer model)
- ERGM (which features of dyads affect the presence or strength of edges.)
- ABM

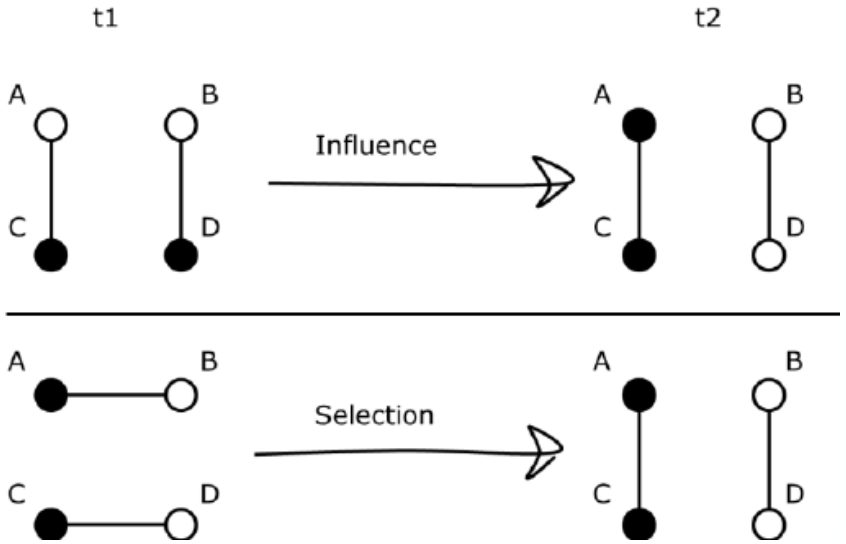


Hobson 2021

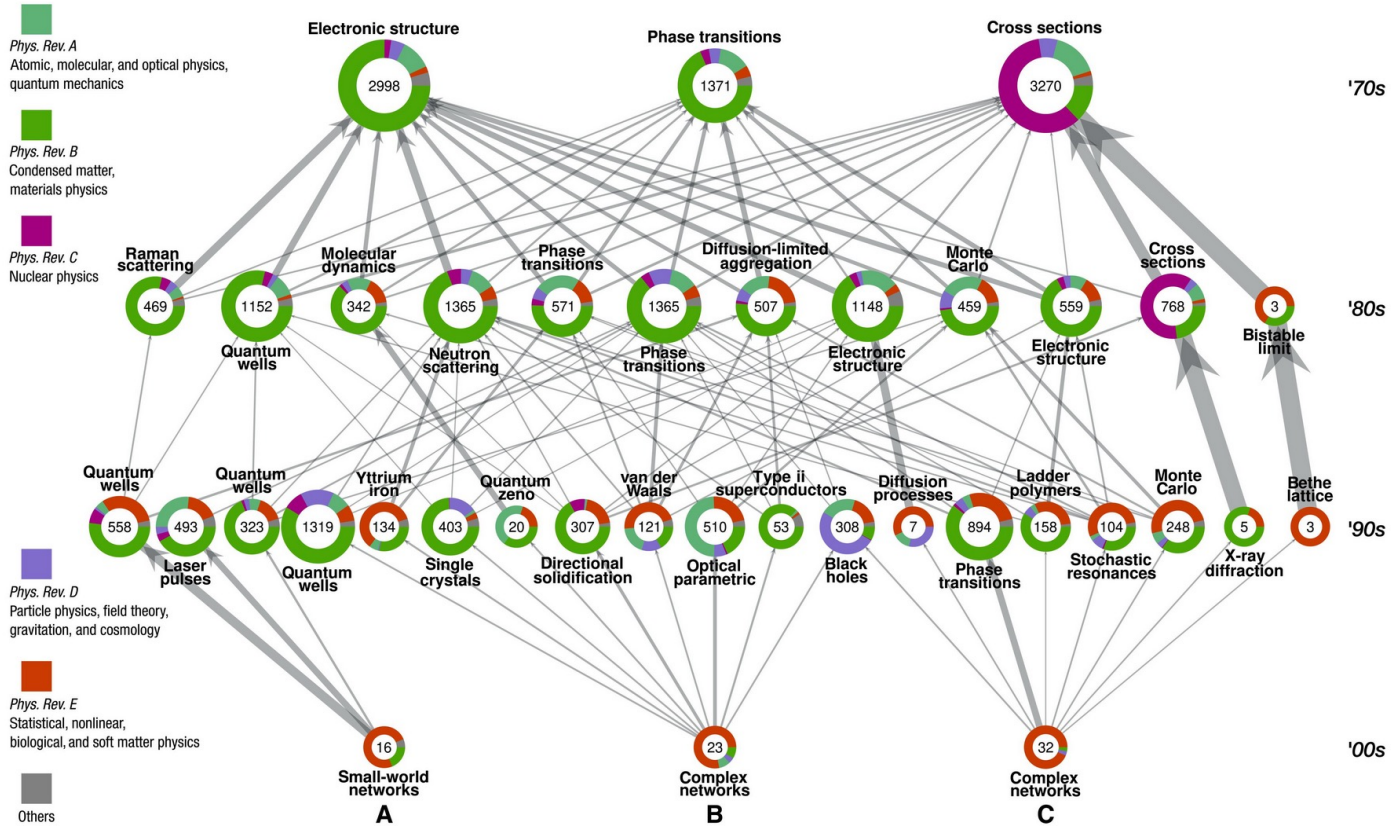
Types of analysis: Dynamics

How does behavior/diseases/information spread?

- Allow to test selection vs influence
- Run simulations on networks
  - Game theory
  - Epidemic spreading
  - Gene expression



Friemel, 2015



Bovet et al, 2022

Want to learn more about networks?

Network Science

2024 Utrecht University Summer School

Practical information

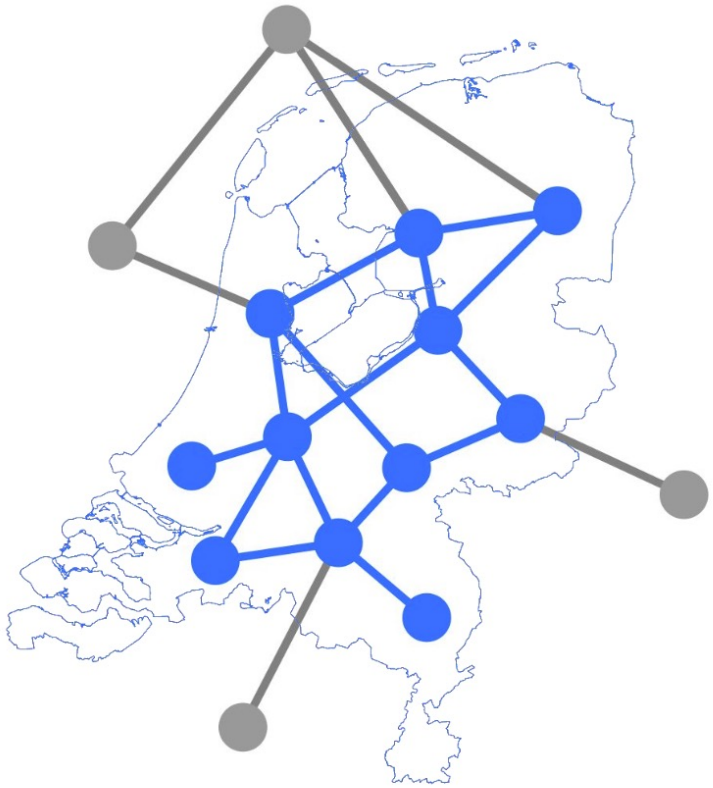
- **Dates**: July 15th — July 19nd, 2024
- **Location**: Utrecht University, Science Park
- **Instructors**: Javier Garcia-Bernardo, Leto Peel, Mahdi Shafiee Kamalabad, Elena Candellone, Jiamin Ou, Vincent Buskens
- **Preparation**: Install R, RStudio and Anacondas

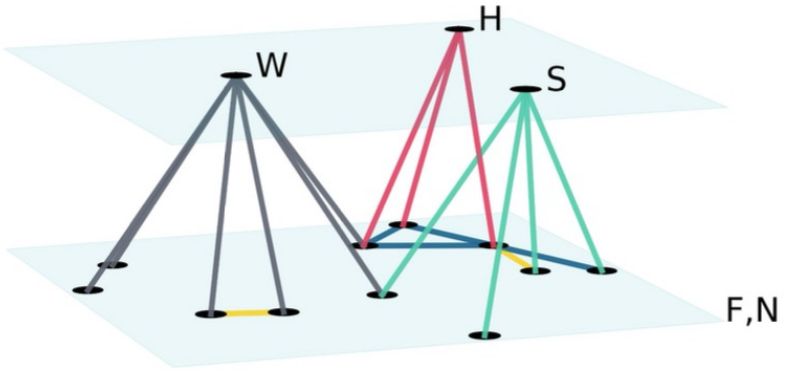
Github repository

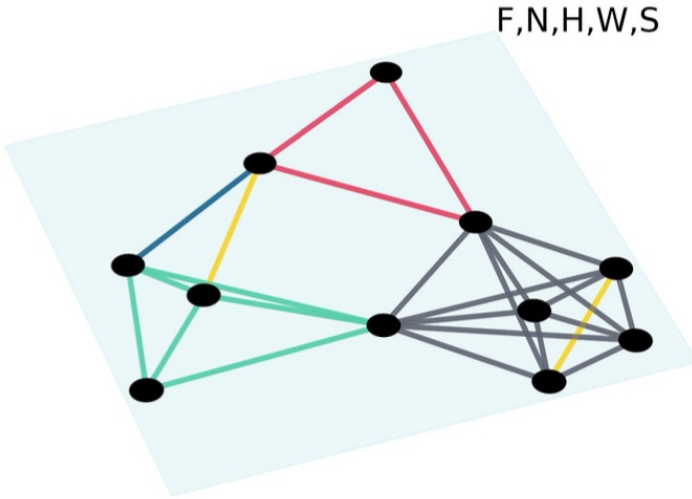
All slides, code and data can be found here. The lectures and code can also be explored using the links below.

Networks at CBS (POPNET data)

Every person in the Netherlands







Images: Eszter Bokányi (POPNET)

Data

Deel deze pagina

Familienetwerktab: familierelaties

Deze netwerklaag bevat alle familierelaties van individuen die op 1 januari JJJJ staan ingeschreven in de Basisregistratie Personen (BRP). Bestand per jaar beschikbaar over de periode 2009 t/m 2021.

Downloads

- familienetwerktab
- Overzicht verschillen oude nieuwe persoonsnetwerk

```python
import polars as pl
pl.read_csv("cbsdata/Bevolking/FAMILIENETWERKTAB/FAMILIENETWERK2021TABV1.csv",
            n_rows=10, separator=";", dtypes={"RINPERSOON": str})
```

✓ 0.0s Python

shape: (10, 5)

RINPERSOON	RINPERSOONS	RINPERSOONRELATIE	RINPERSOONSRELATIE	RELATIE
str	str	i64	str	i64
"100439107"	"R"	100342679	"R"	310
"100086293"	"R"	100108614	"R"	310
"100150055"	"R"	100058381	"R"	317
"100453751"	"R"	100448572	"R"	313
"100325504"	"R"	100041245	"R"	315
"100211113"	"R"	100402661	"R"	314
"100312768"	"R"	100415389	"R"	317
"100302277"	"R"	100091753	"R"	315
"100024777"	"R"	100126339	"R"	313
"100387455"	"R"	100439892	"R"	316

317: 'Stepparent',

Child

Stepparent

"Neighbors": "BURENNETWERKTAB"

101: 'Neighbor - 10 closest addresses',

102: 'Neighborhood acquaintance - 20 random neighbors within 200 meters’,

"Colleagues": "COLLEGANETWERKTAB"

201: 'Colleague’,

"Housemates": "HUISGENOTENNETWERKTAB"

401: 'Housemate',

402: 'Housemate - institution’,

"Schoolmates": "KLASGENOTENNETWERKTAB"

501: 'Classmate primary education',

502: 'Classmate special education',

503: 'Classmate secondary education',

504: 'Classmate vocational education',

505: 'Classmate higher professional education',

506: 'Classmate university education'

"Family": "FAMILIENETWERKTAB"

301: 'Parent',

302: 'Co-parent',

303: 'Grandparent',

304: 'Child',

305: 'Grandchild',

306: 'Full sibling',

307: 'Half sibling',

308: 'Unknown sibling',

309: 'Full cousin',

310: 'Cousin',

311: 'Aunt/Uncle',

312: 'Partner - married',

313: 'Partner - not married',

314: 'Parent-in-law',

315: 'Child-in-law',

316: 'Sibling-in-law',

317: 'Stepparent',

318: 'Stepchild',

319: 'Stepsibling',

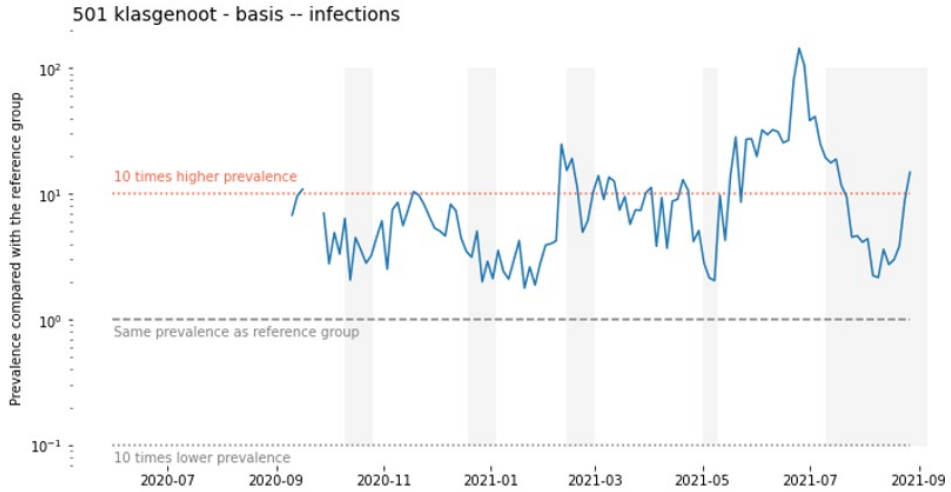
320: 'Married full cousin',

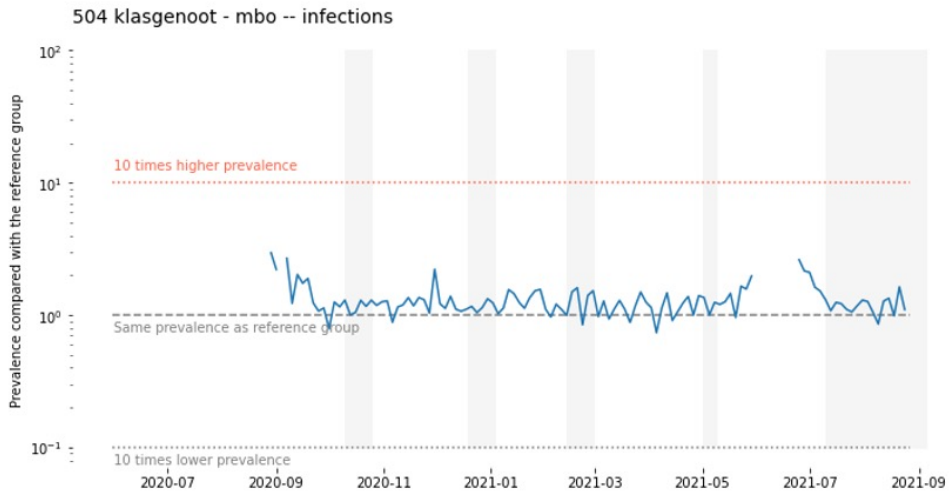
321: 'Married cousin',

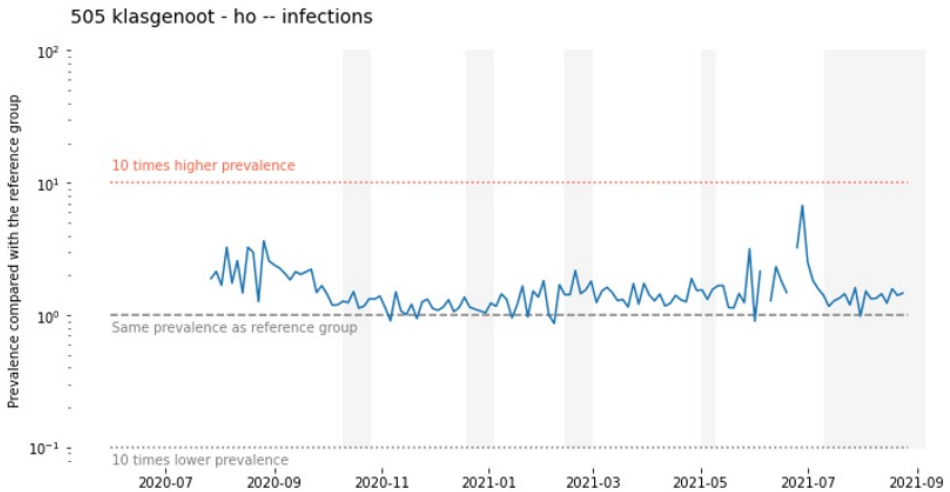
322: 'Married aunt/uncle',

Not all network files are equally relevant

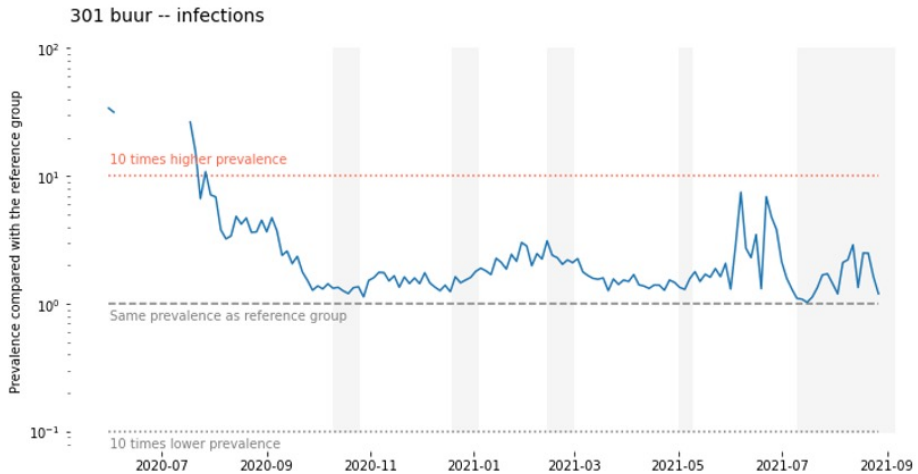
School Networks







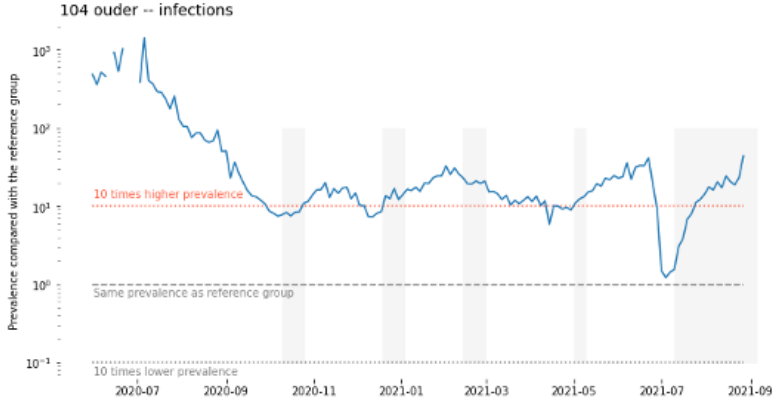
Colleagues and neighbors



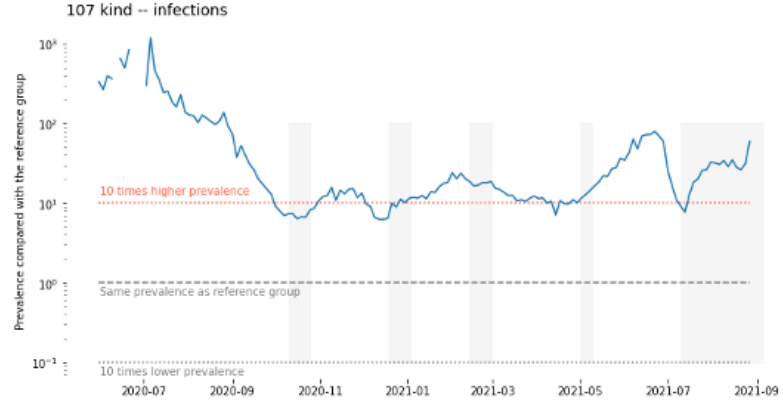
102 co-ouder -- infections

103 zus/broer - vol -- infections

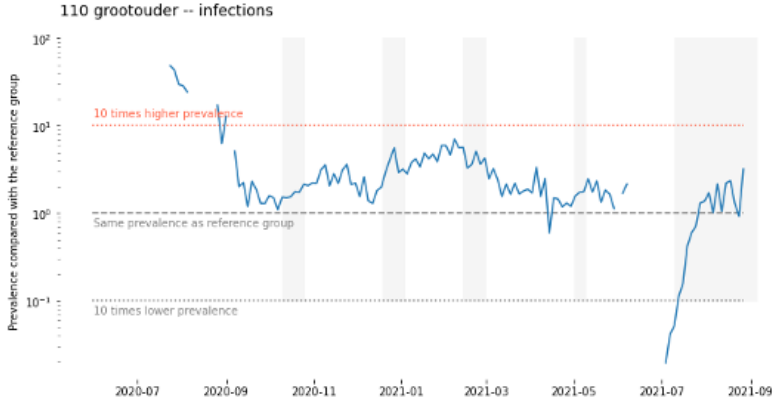
104 ouder -- infections



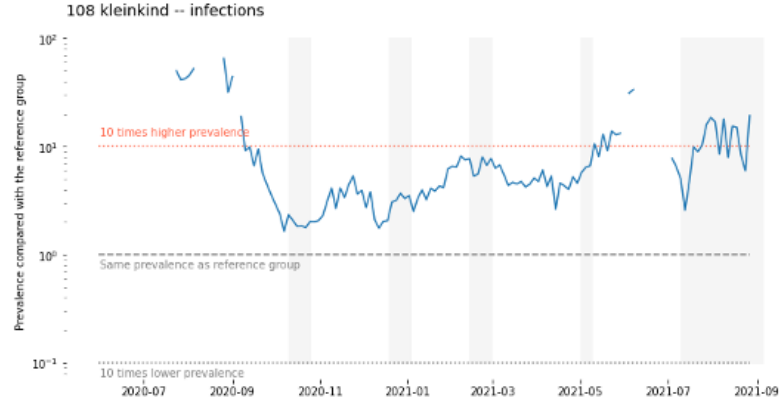
107 kind -- infections



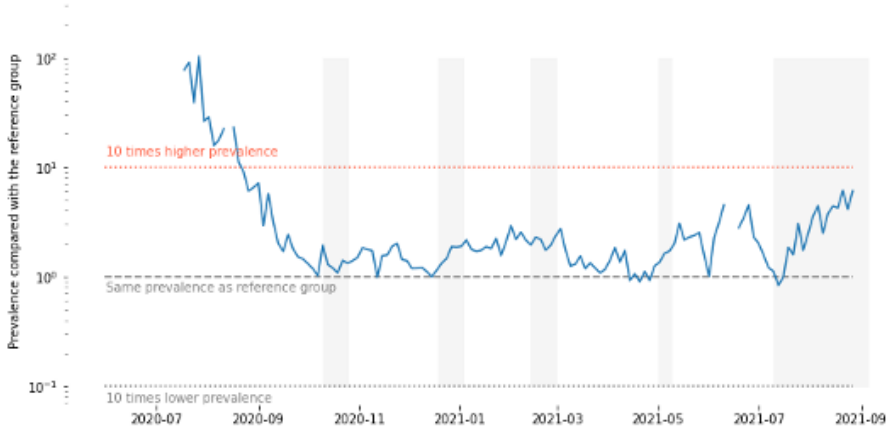
110 grootouder -- infections



108 kleinkind -- infections



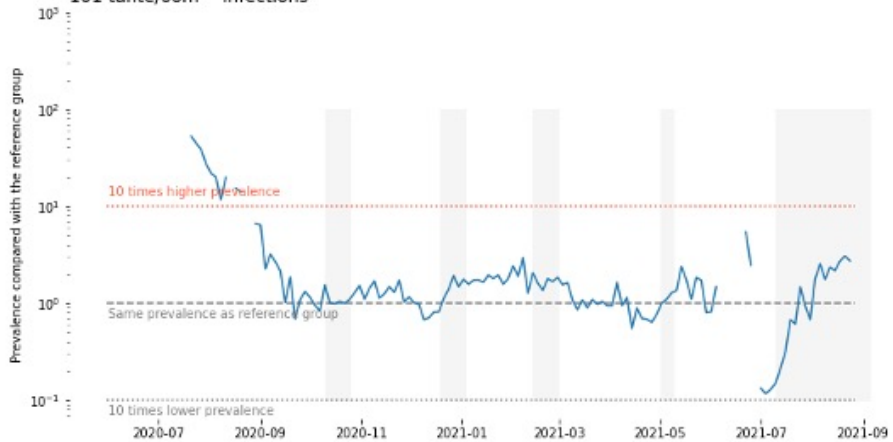
105 volle nicht/neef -- infections



106 nicht/neef -- infections

109 zus/broer - onbekend -- infections

101 tante/oom -- infections



netCBS library (Python)

Allows you to create “network” queries, such as calculate the number of children of the siblings of the people in the sample

Library available at H:/netcbs

You can import it by adding the path to Python.

Supposing you are working in H:/group_x/pythonscript.ipynb

```
import sys
sys.path.insert(0, "../netcbs")

import netcbs as net
```

```
## How to construct the query
# 1. Start with the variables that you want to aggregate, e.g. "[Income, Age] ->"
# 2. Then add the relationships between the tables, e.g., "[Income, Age] -> Family[301]".
# In square brackets you can specify the type of the relationships:
# write [all] for all, or [301,302] for parents and co-parents
# 3. You can add several tables: "[Income, Age] -> Family[301] -> Schoolmates[all]"
# 4. Finally, you must write "-> Sample"

## Other parameters
# df_sample: A polars dataframe containing the sample (with the people you want to have information
on)
# df_agg: A polars dataframe with the information you want to aggregate. For example, the income of
all people in the country
# year: the year of the data you want to use
# agg_func: the aggregation functions you want to use. For example, [pl.mean or pl.sum]
# lazy: if True, the operations are concatenated lazily and computed at the end. If False, the
operations are computed immediately

# Example
query = "[Income, Age] -> Family[301,302] -> Schoolmates[all] -> Sample"

df = netcbs.transform(query,
                      df_sample = df_sample,
                      df_agg = df_agg,
                      year=2021,
                      agg_func=[pl.mean, pl.sum, pl.max],
                      lazy=True
)
```

In general:

- These are huge files!
  - There are 700 M colleagues, 415 M neighbors, 440 M housemates, 88 M cousins
- Tips:
  - Start with small samples until you know everything works
  - Be efficient
    - Use *polars* instead of *pandas*
    - Use lazy operations
  - Clean up regularly using magic commands (remove data that you are not using from memory)
    - %reset_selective –f df_x3
  - The RA will crash
    - Save your results to the hard drive using efficient data structures (pl.write_parquet)

H:/network_examples/exercice.ipybn

- Copy the file to your group folder
- Please only one person per group! (heavy server)
- Keep the queries simple, otherwise they will take too long