Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

https://liacs.leidenuniv.nl/~takesfw/SNACS

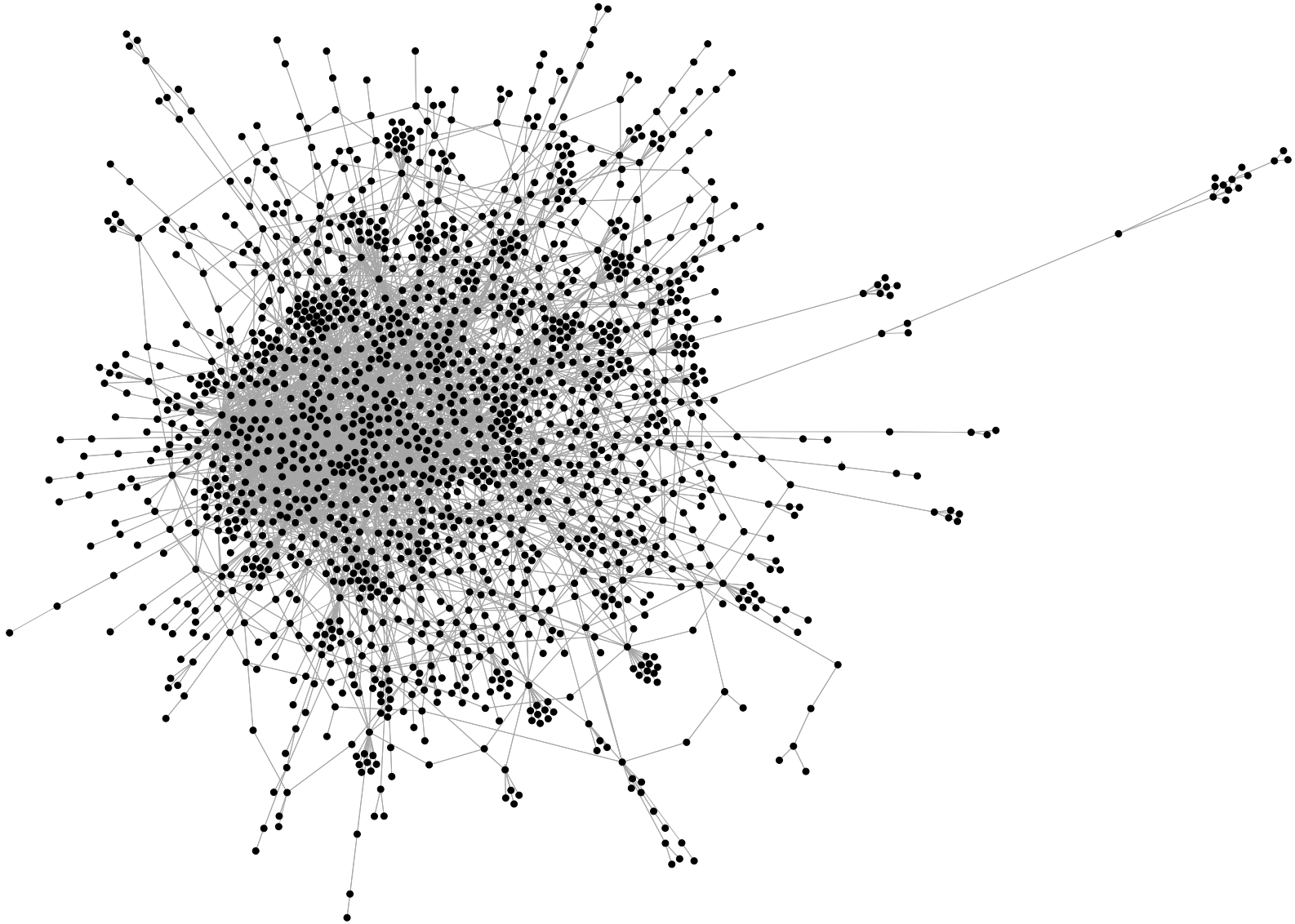
Lecture 6 — Network processes & Network science challenges

Today

- Recap
- Community detection — resolution parameter and interpretation
- Betweenness centrality — computational details
- Network science challenges
- (If time allows: Epidemic spread: SIR model)
- (If time allows: Corporate Ownership Networks and Tax Evasion)

Recap

Networks



Notation

Concept	**Symbol**
Network (graph)	$G = (V, E)$
Nodes (objects, vertices, ...)	V
Links (ties, relationships, ...)	E
Directed — $E \subseteq V \times V$ — "links"	
Undirected — "edges"	
Number of nodes — $\|V\|$	n
Number of edges — $\|E\|$	m
Degree of node u	$deg(u)$
Distance from node u to v	$d(u, v)$

Real-world networks

1	Sparse networks	density
2	Fat-tailed power-law degree distribution	degree
3	Giant component	components
4	Low pairwise node-to-node distances	distance
5	Many triangles	clustering coefficient

Real-world networks

1. Sparse networks — density
2. Fat-tailed power-law degree distribution — degree
3. Giant component — components
4. Low pairwise node-to-node distances — distance
5. Many triangles — clustering coefficient

- Many examples: communication networks, citation networks, collaboration networks (Erdös, Kevin Bacon), protein interaction networks, information networks (Wikipedia), webgraphs, financial networks (Bitcoin) ...

Advanced concepts

- Assortativity, homophily
- Reciprocity
- Power law exponent
- Planar graphs
- Complete graphs
- Subgraphs
- Trees
- Spanning trees
- Diameter, eccentricity
- Bridges
- Graph traversal: DFS, BFS

Centrality measures

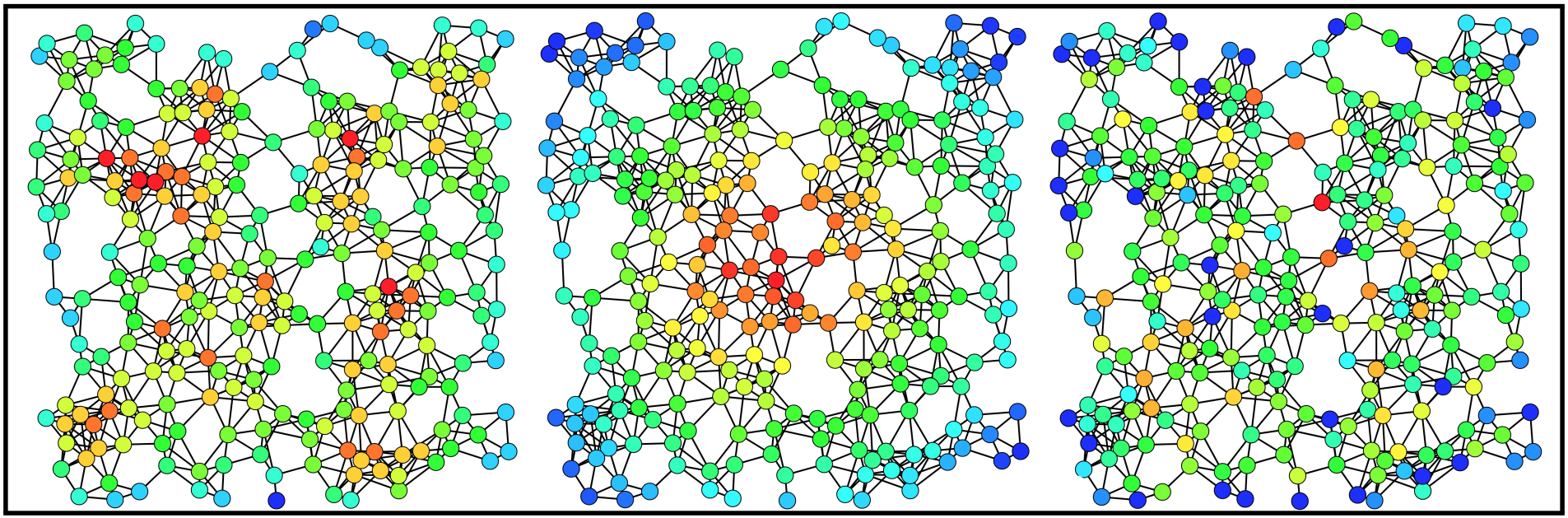
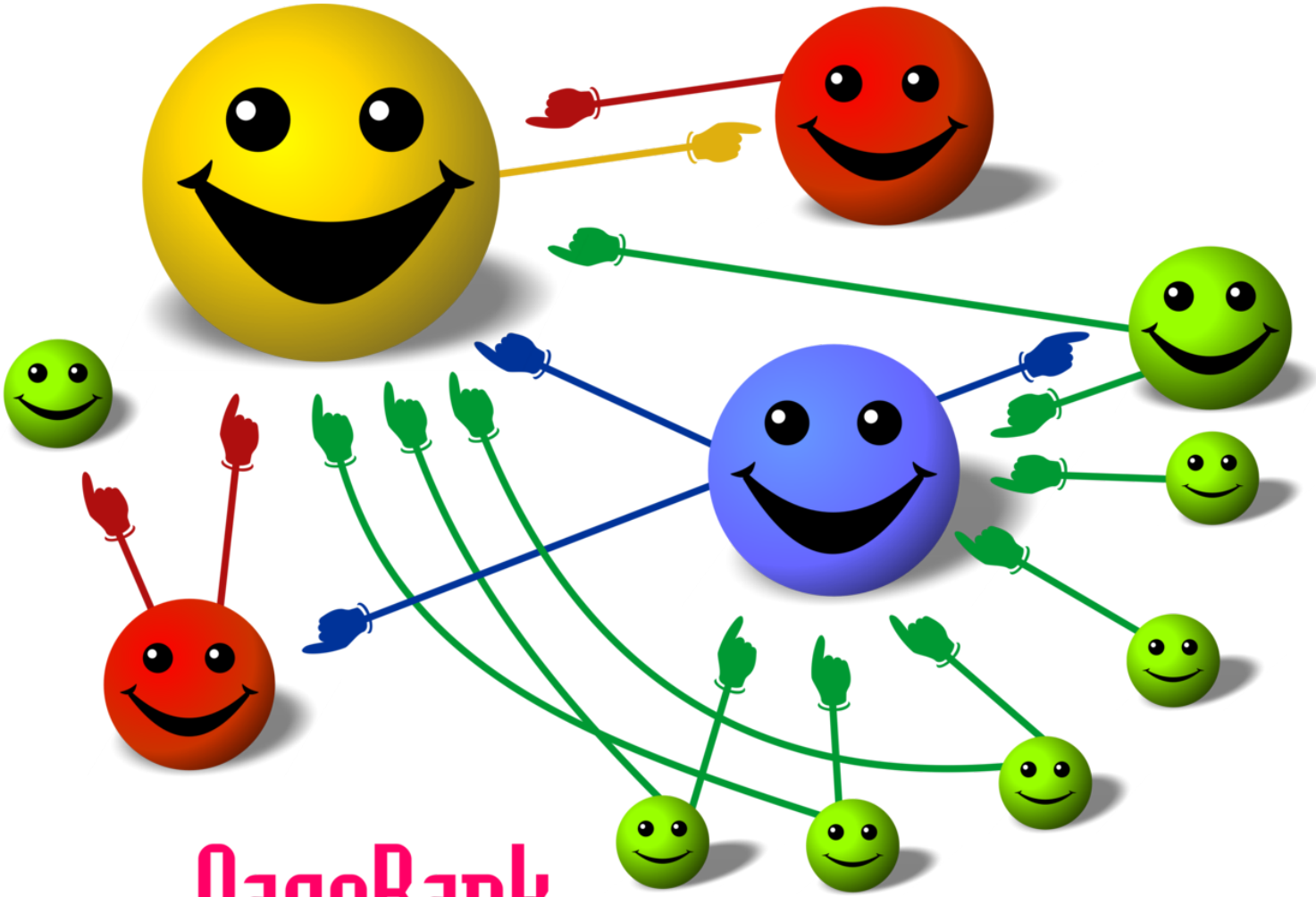


Figure: Degree, closeness and betweenness centrality

Source: "Centrality"' by Claudio Rocchini, Wikipedia `File:Centrality.svg`

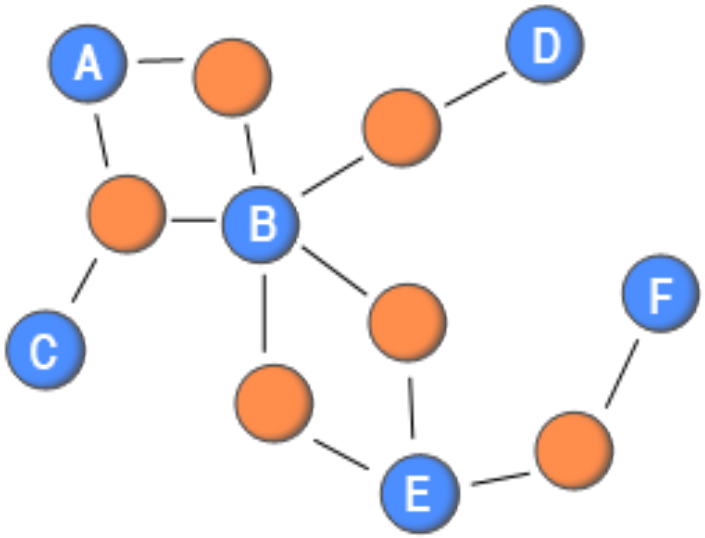
Centrality measures: PageRank



Centrality measures

- Distance/path-based measures:
  - Degree centrality $O(n)$
  - Closeness centrality $O(mn)$
  - Betweenness centrality $O(mn)$
  - Eccentricity centrality $O(mn)$
- **Propagation-based** measures:
  - Hyperlink Induced Topic Search (HITS) $O(m)$
  - PageRank $O(m)$

Network projection



$\rightarrow$

Network projection

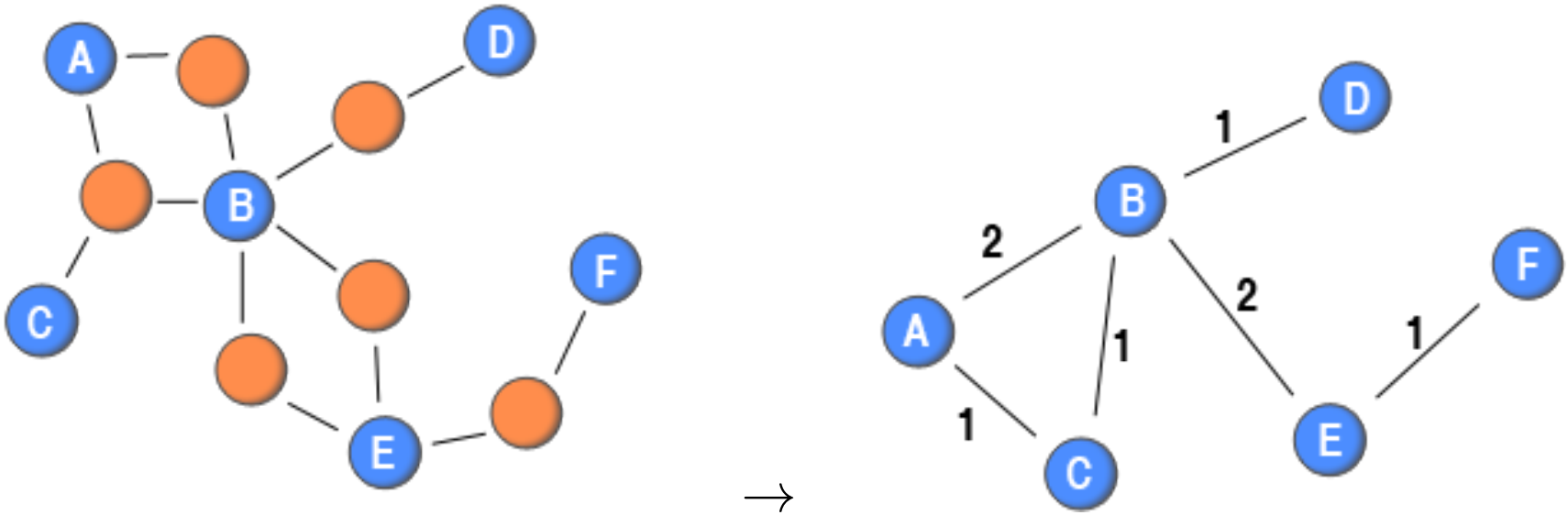


Image: `http://toreopsahl.com`

Community detection

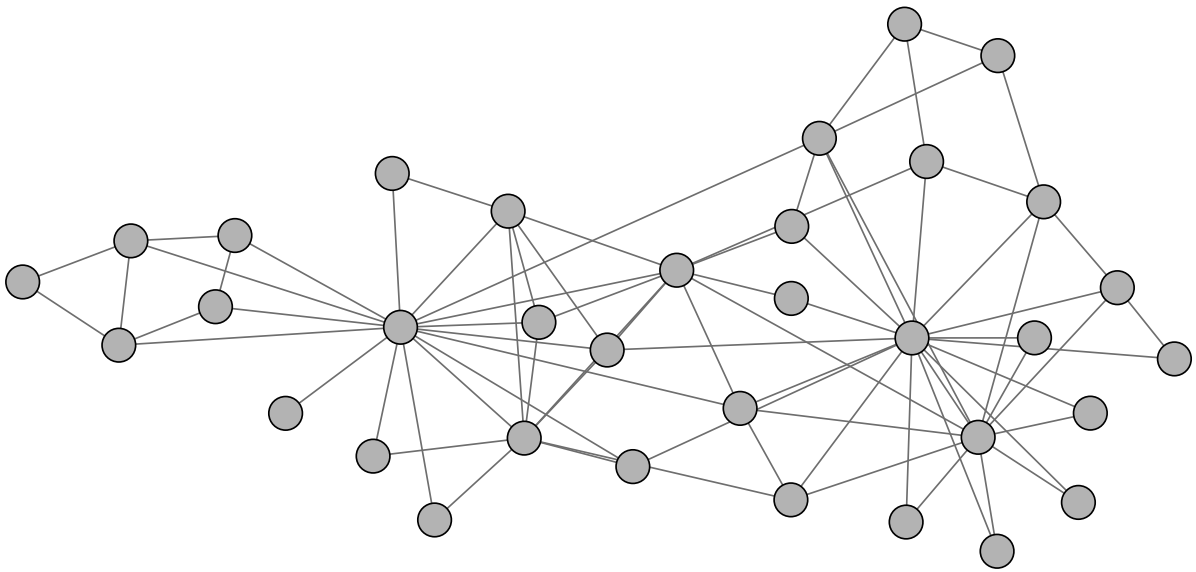


Figure: Communities: node subsets connected more strongly with each other

Community detection

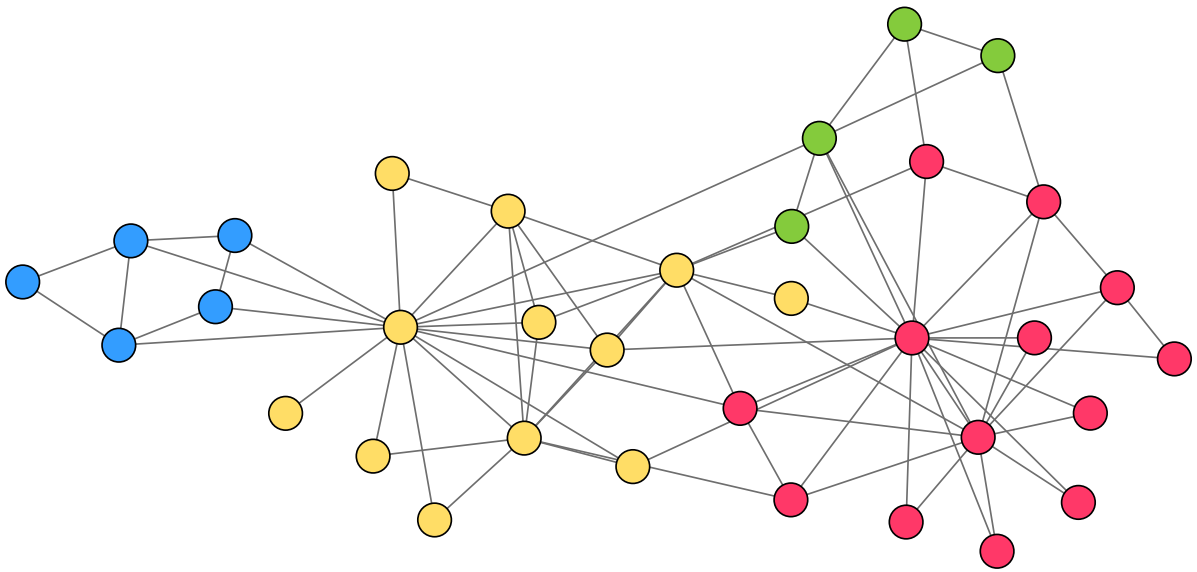
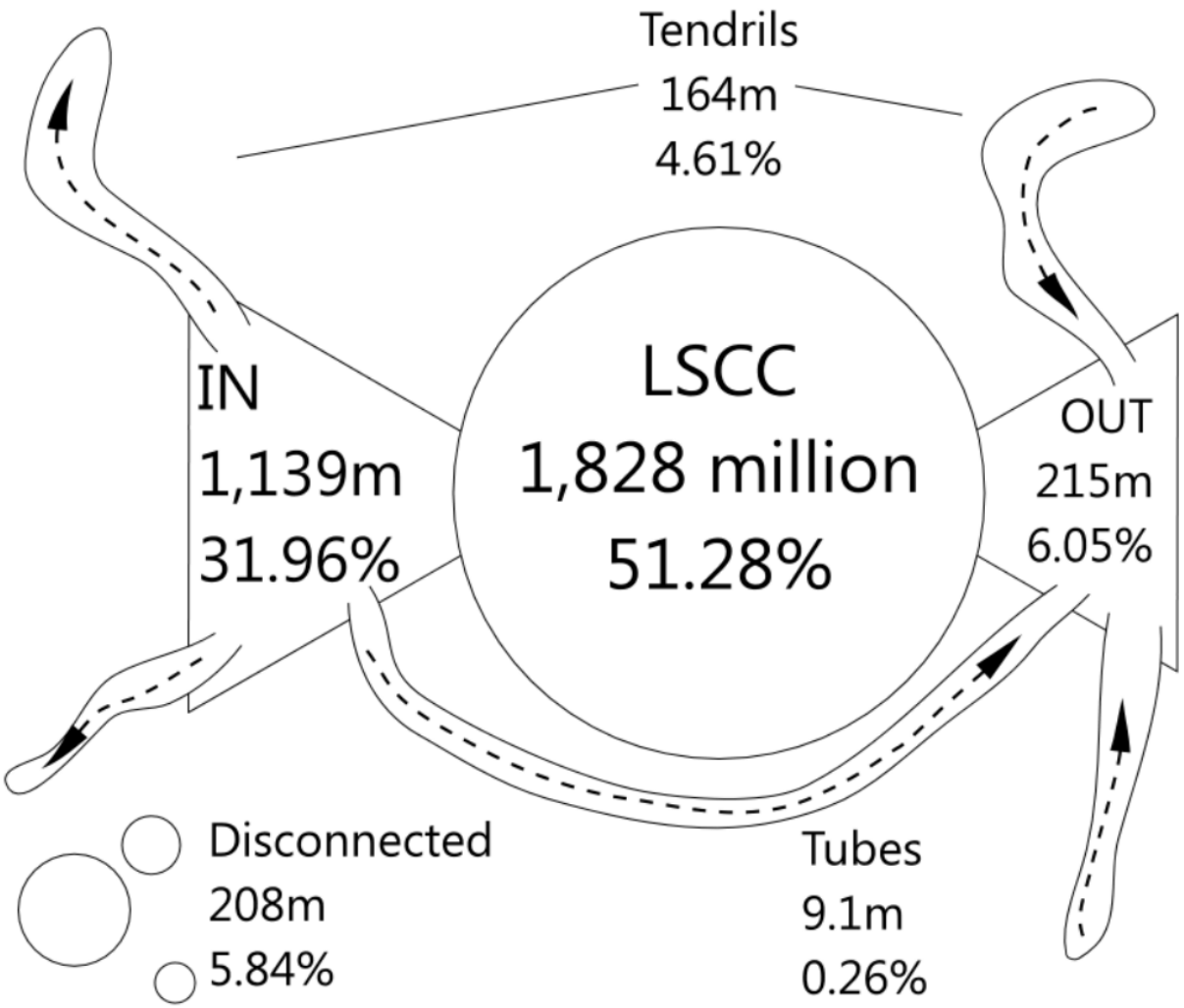


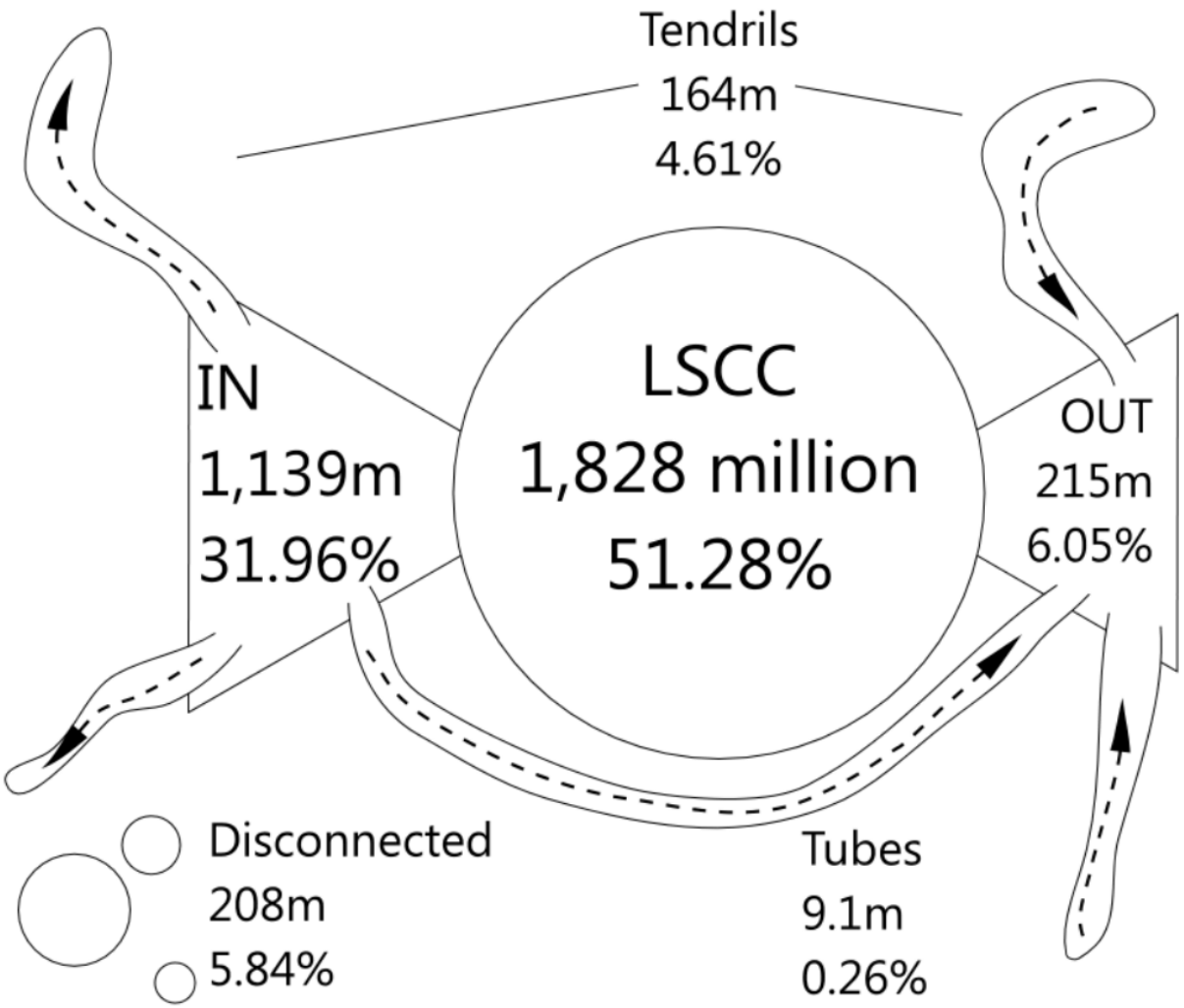
Figure: Communities: node subsets connected more strongly with each other

Bow-tie structure of the web



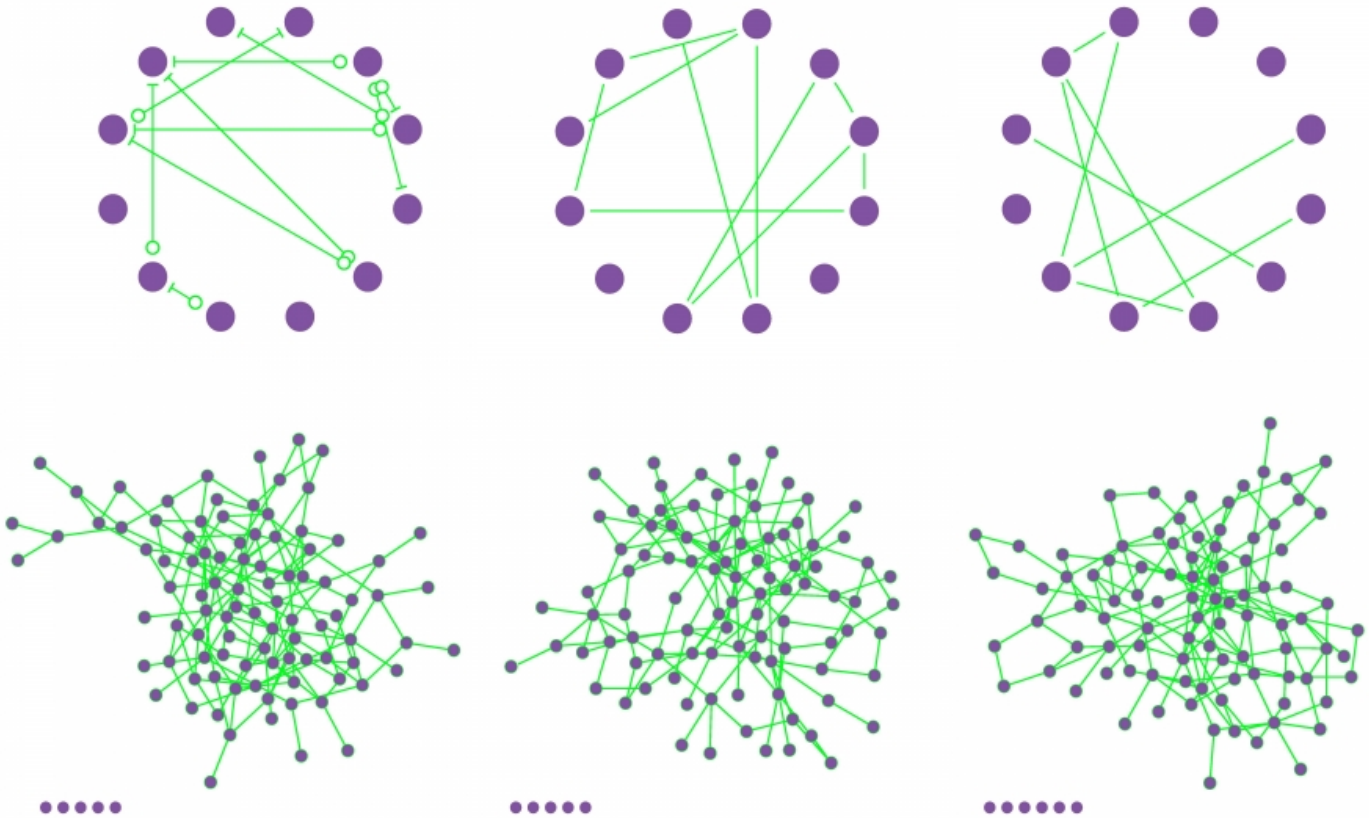
Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Bow-tie structure of the web



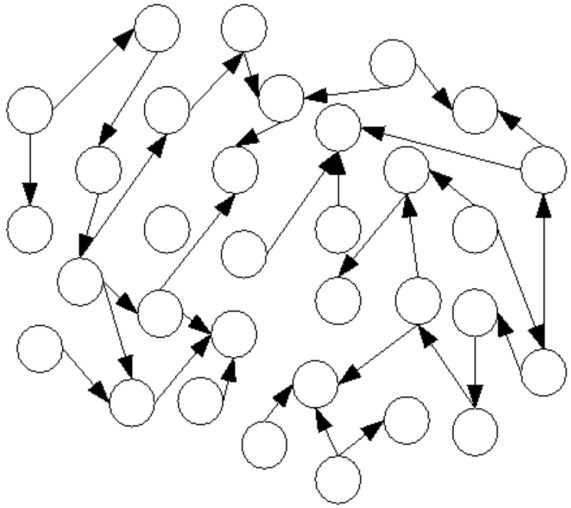
Meusel et al., Graph Structure in the Web — Revisited, WWW 2014: 427–431, 2014.

Erdös-Rényi random graphs

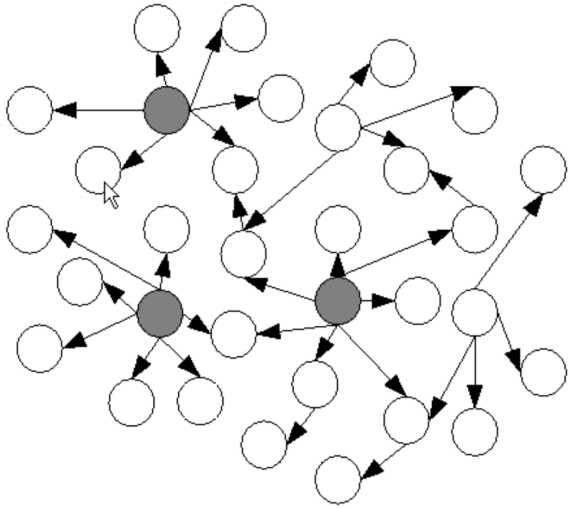


http://barabasi.com/networksciencebook/chapter/3

Scale-free / Barábasi-Albert graphs



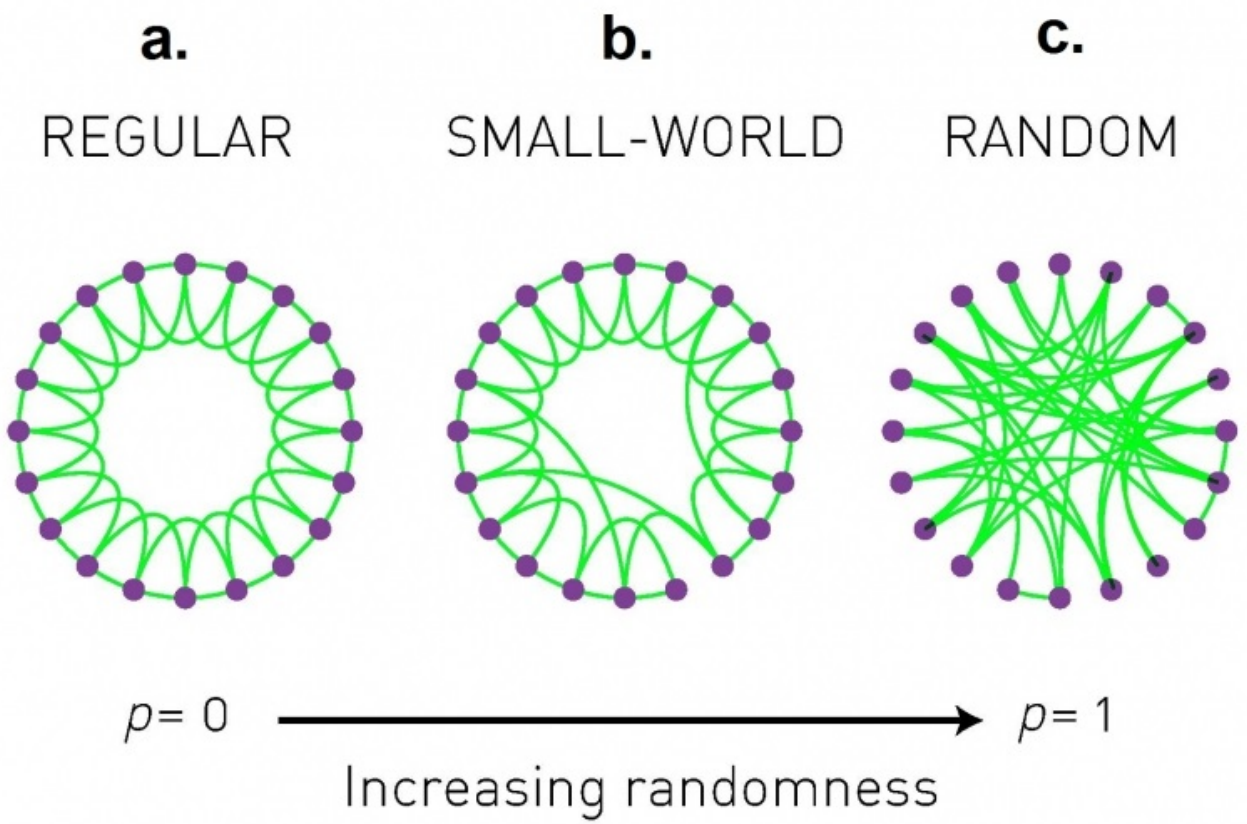
(a) Random network



(b) Scale-free network

B. Svenson, Complex networks and social network analysis in information fusion

Small-world / Watts & Strogatz graphs



http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife/bgc-sci.jpg

Community detection — very brief recap

Community detection

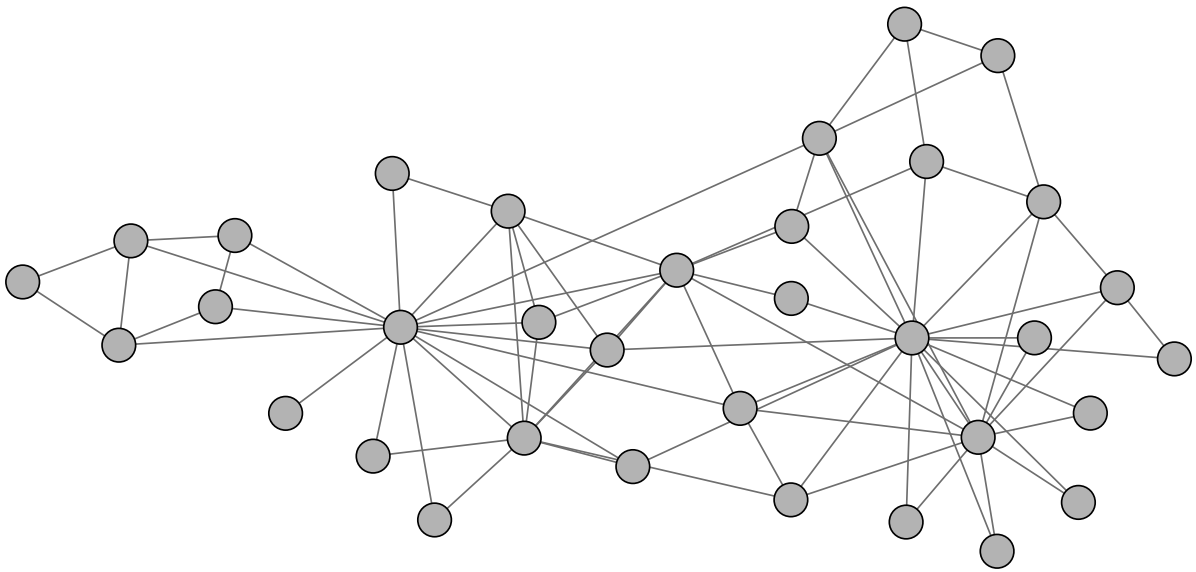


Figure: Communities: node subsets connected more strongly with each other

Community detection

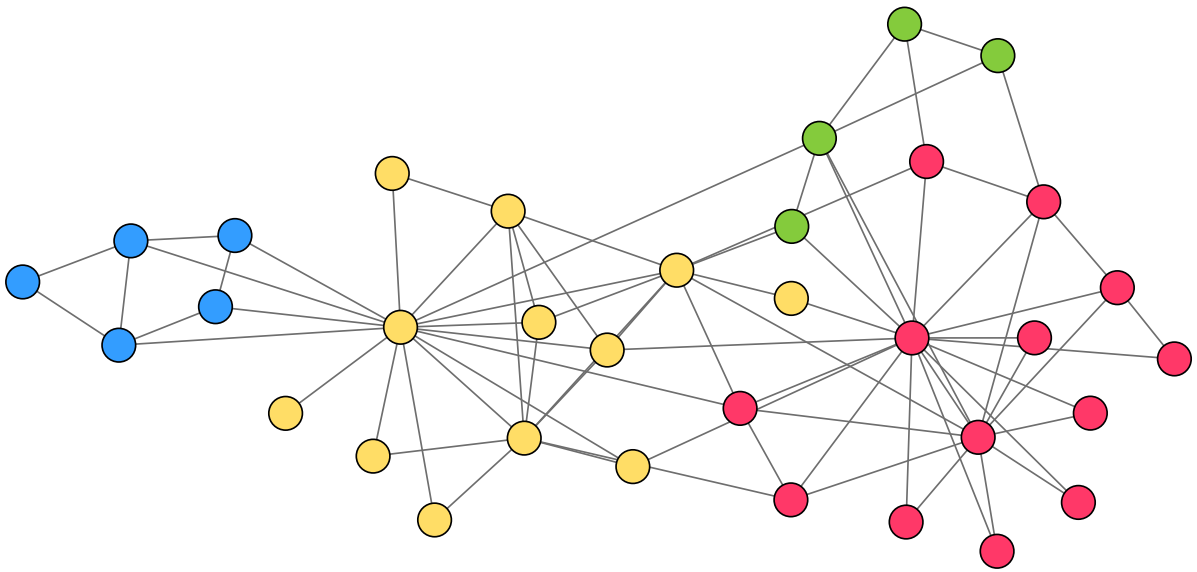


Figure: Communities: node subsets connected more strongly with each other

Modularity

- **Community** (alternative definition): subset of nodes for which the fraction of links inside the community is higher than expected
- **Modularity**: numerical value Q indicating the quality of a given division of a network into communities. Higher value of Q means more links within communities (and fewer between)
- Resolution parameter r indicating how "tough" the algorithm should look for communities
- Algorithms optimize (maximize) the modularity score Q given some r (using local search, heuristics, hill climbing, genetic algorithms or other optimization techniques)

V.D. Blondel, J-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks in *Journal of Statistical Mechanics: Theory and Experiment* 10: P10008, 2008.

Yet another network

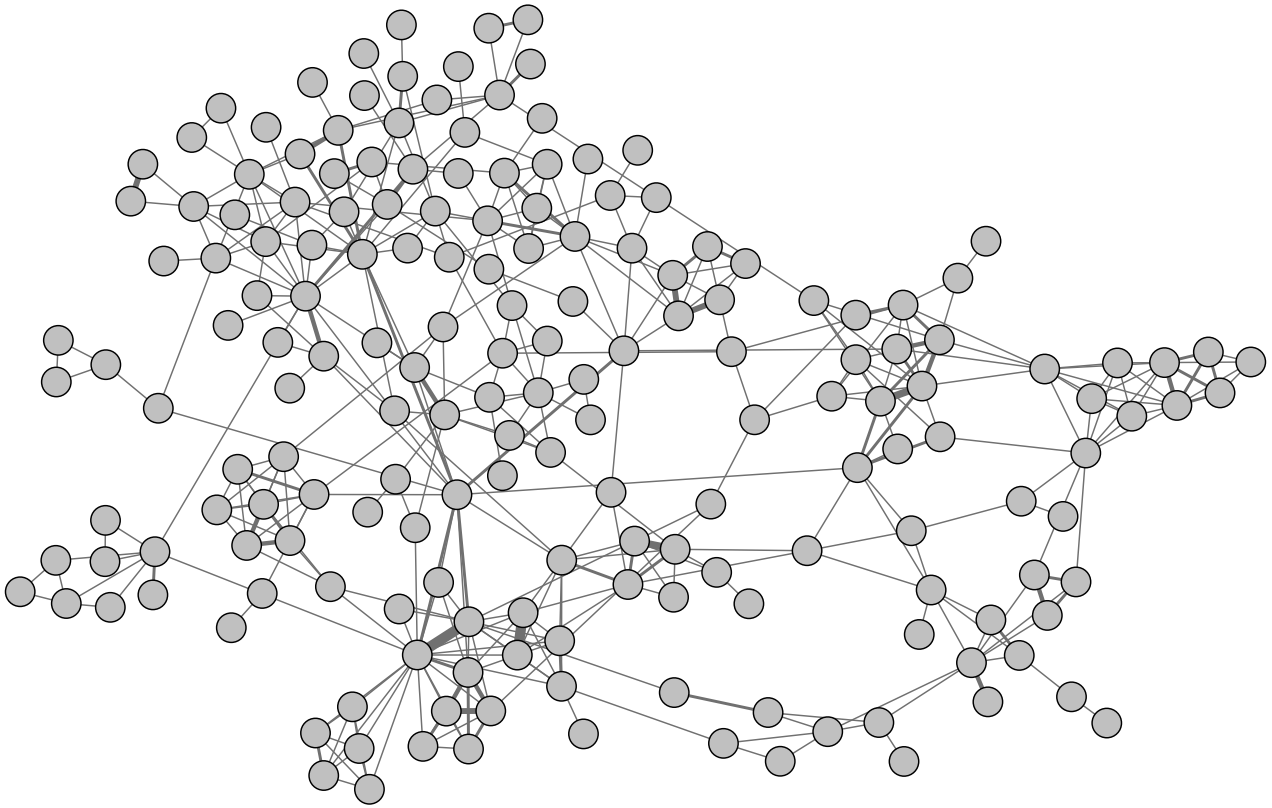


Figure: Do you see communities?

Resolution = 1.0

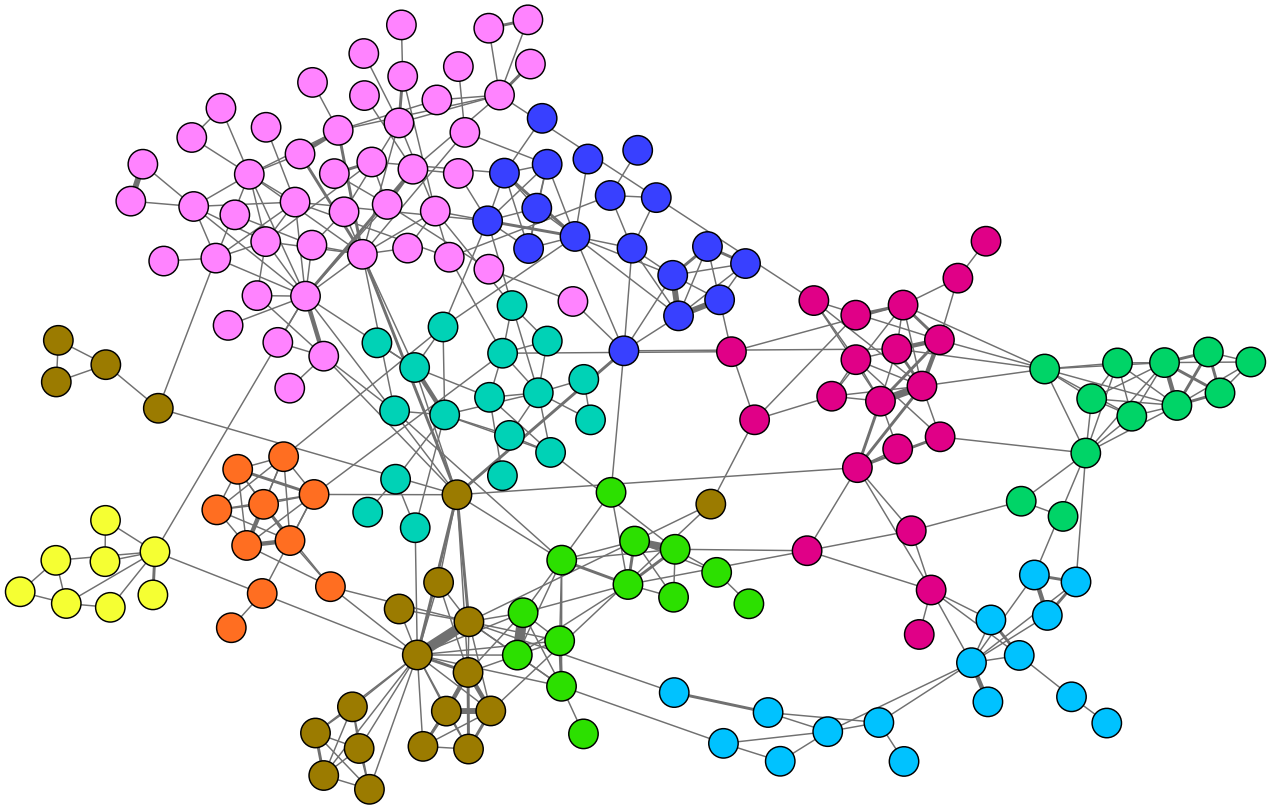


Figure: Modularity = 0.747; 10 communities

Resolution = 2.0

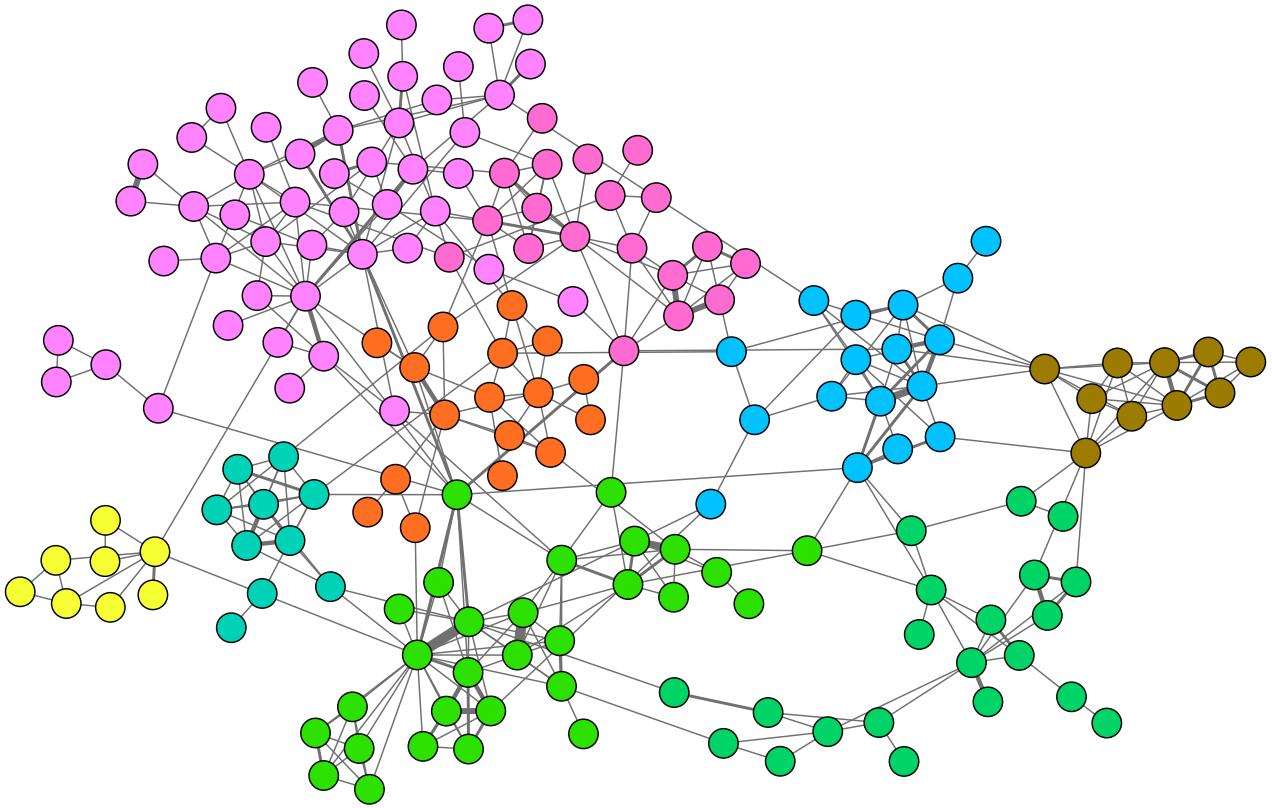


Figure: Modularity = 0.732; 8 communities

Resolution = 4.0

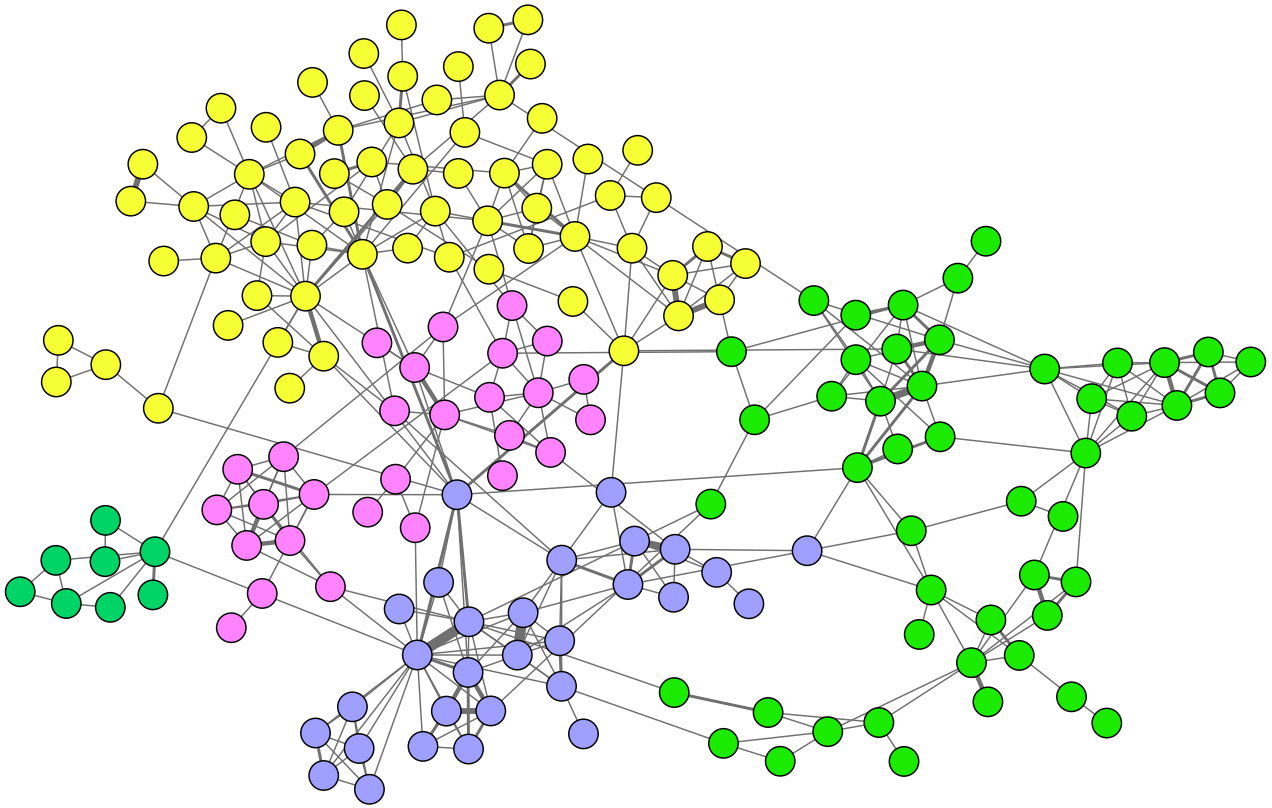


Figure: Modularity = 0.641; 5 communities

Evaluating communities and partitions

- **Communities**: groups of nodes that are more connected amongst each other than with the other nodes of the network
- **Partitions**: non-overlapping communities
- Compare with groups of nodes based on common attributes
- Human interpretation by hand can suffer from subjective bias

Communities in corporate networks

Corporate board interlocks

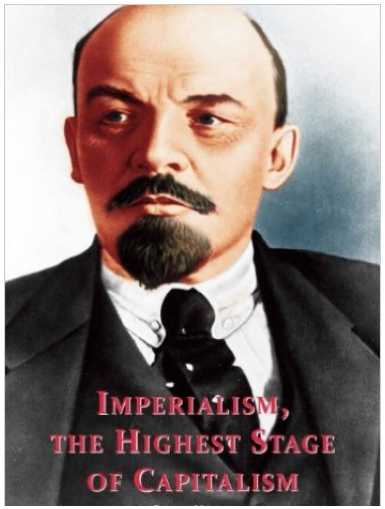
- **Nodes** are organizations/firms/companies/corporations

Corporate board interlocks

- **Nodes** are organizations/firms/companies/corporations
- **Edge** are **board interlocks**: relationships between firms because they share a board member or director

Corporate board interlocks

- **Nodes** are organizations/firms/companies/corporations
- **Edge** are **board interlocks**: relationships between firms because they share a board member or director
- Vladimir I. Lenin, *Imperialism, The Highest Stage of Capitalism*, 1916.
- "... a personal union, so to speak, is established between the banks and the biggest industrial and commercial enterprises, the merging of one with another through the acquisition of shares, through the appointment of bank directors to the Supervisory Boards (or Boards of Directors) of industrial and commercial enterprises, and vice versa."



Board interlocks

- **Causes** of interlocks:
  - Collusion
  - Cooptation and monitoring
  - Legitimacy
  - Career advancement
  - Social cohesion
- **Consequences** of interlocks:
  - Corporate control
  - Economic performance
  - Access to resources

Board interlocks

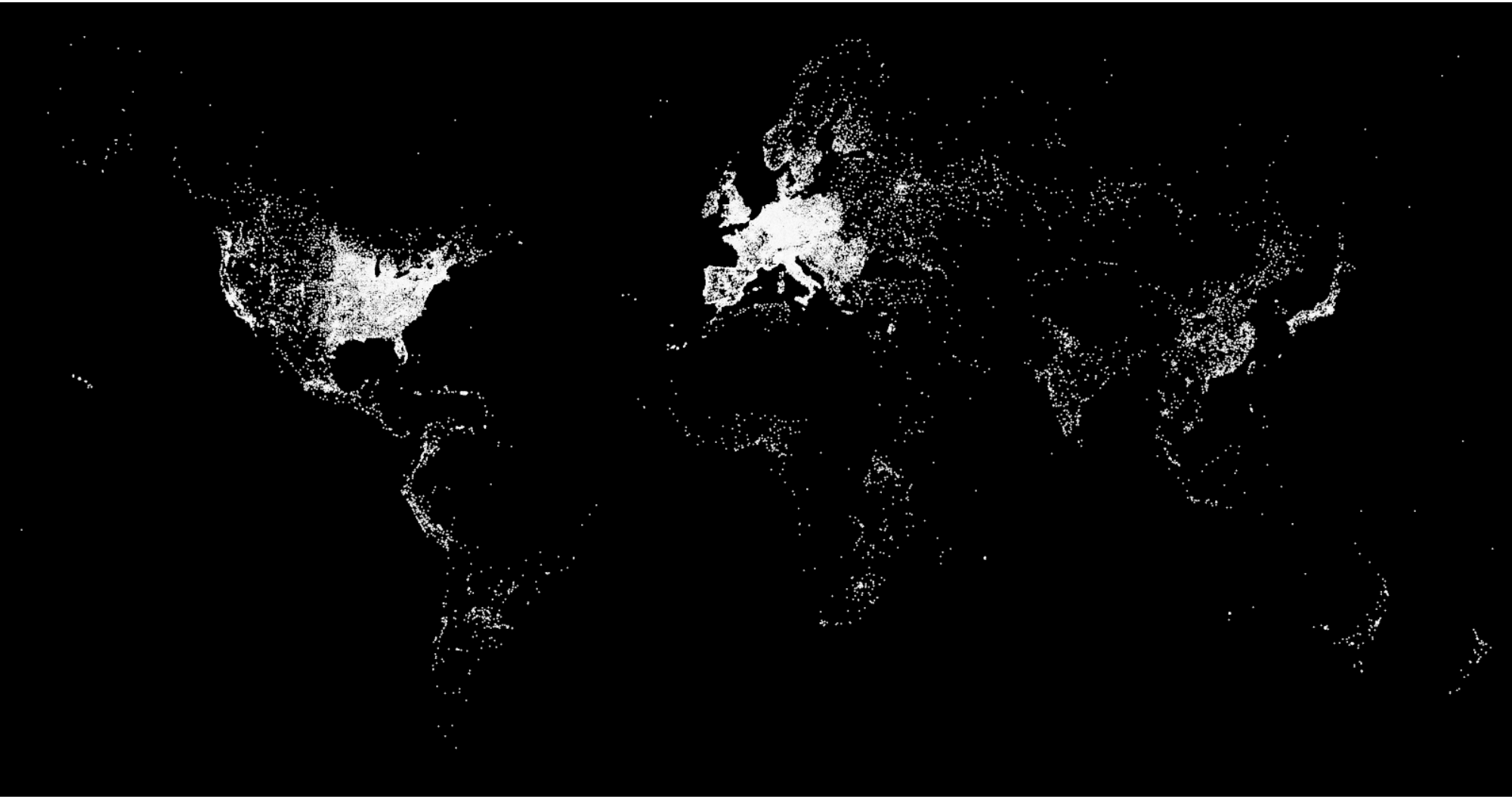
- **Causes** of interlocks:
  - Collusion
  - Cooptation and monitoring
  - Legitimacy
  - Career advancement
  - Social cohesion
- **Consequences** of interlocks:
  - Corporate control
  - Economic performance
  - Access to resources

M. Mizruchi, What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates, *Annual review of Sociology* 22: 271–298, 1996.

Corporate city networks

- Nodes are cities
- Edges between cities are based on firms sharing directors
- Weights on edges denote the number of connections
- Each city has an associated country
- Provides insight in geographical orientation of global economy

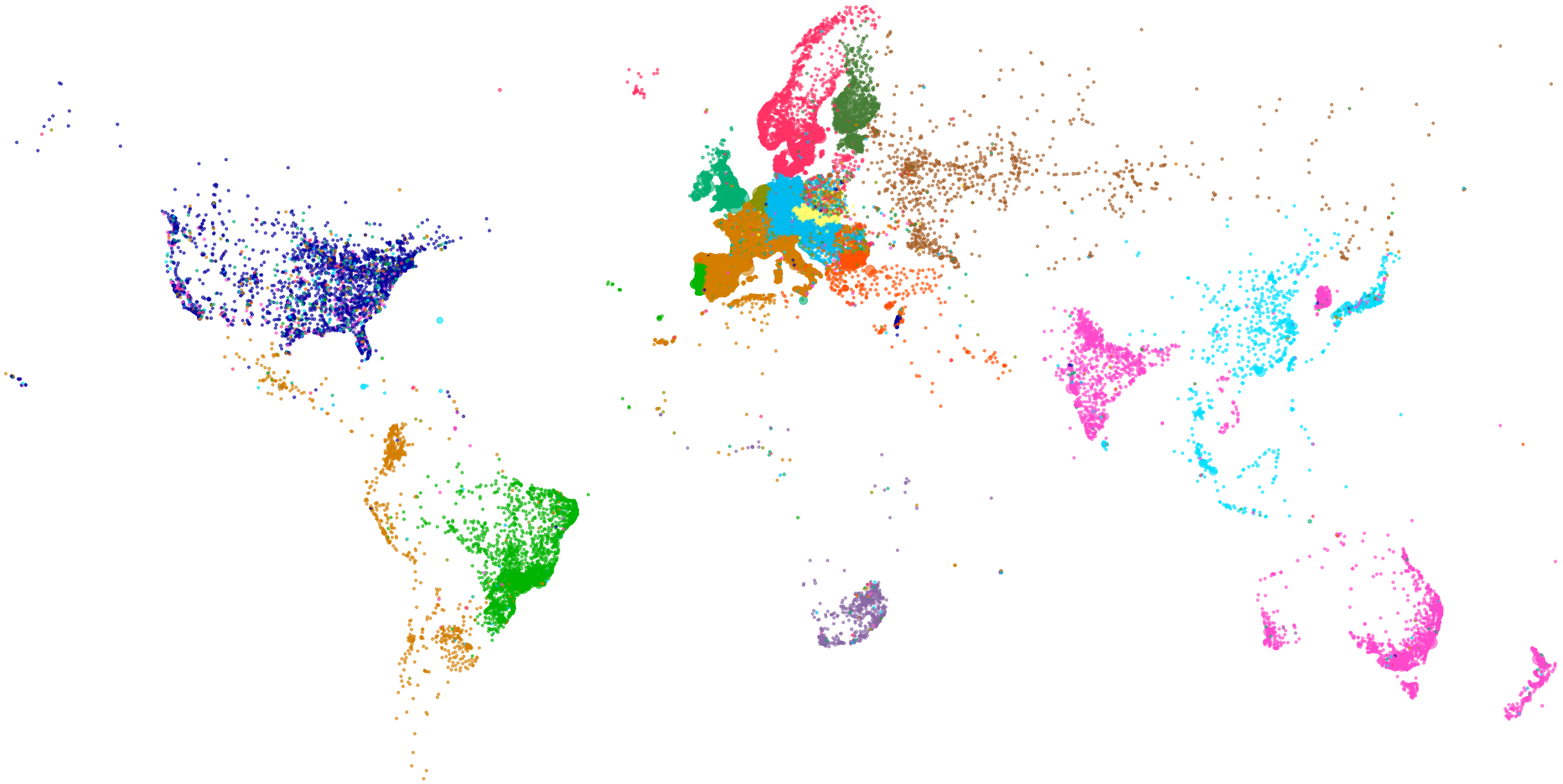
Corporations



Corporate network

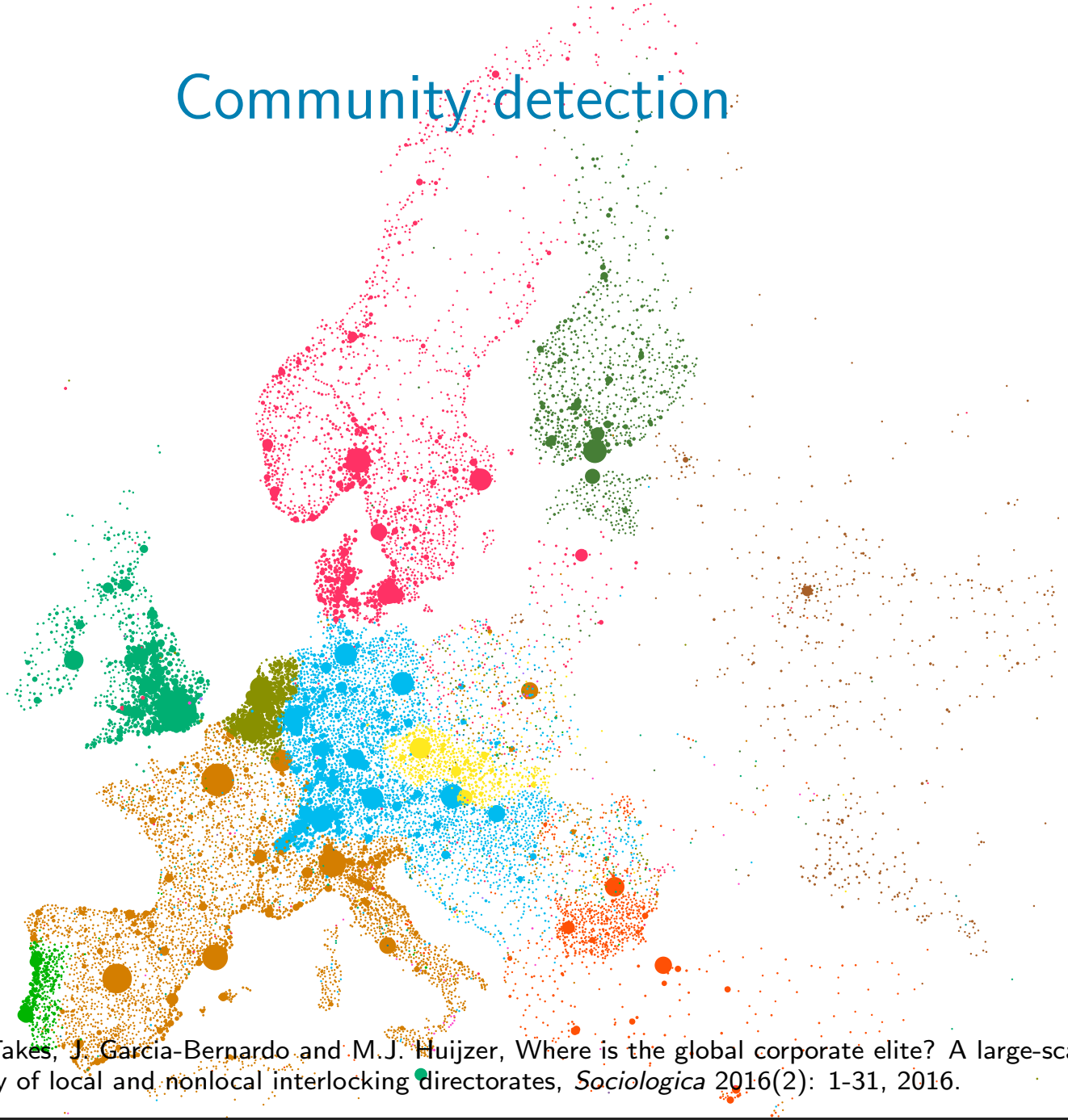
F.W. Takes and E.M. Heemskerk, Centrality in the Global Network of Corporate Control, *Social Network Analysis and Mining* 6(1): 1-18, 2016.

Community detection



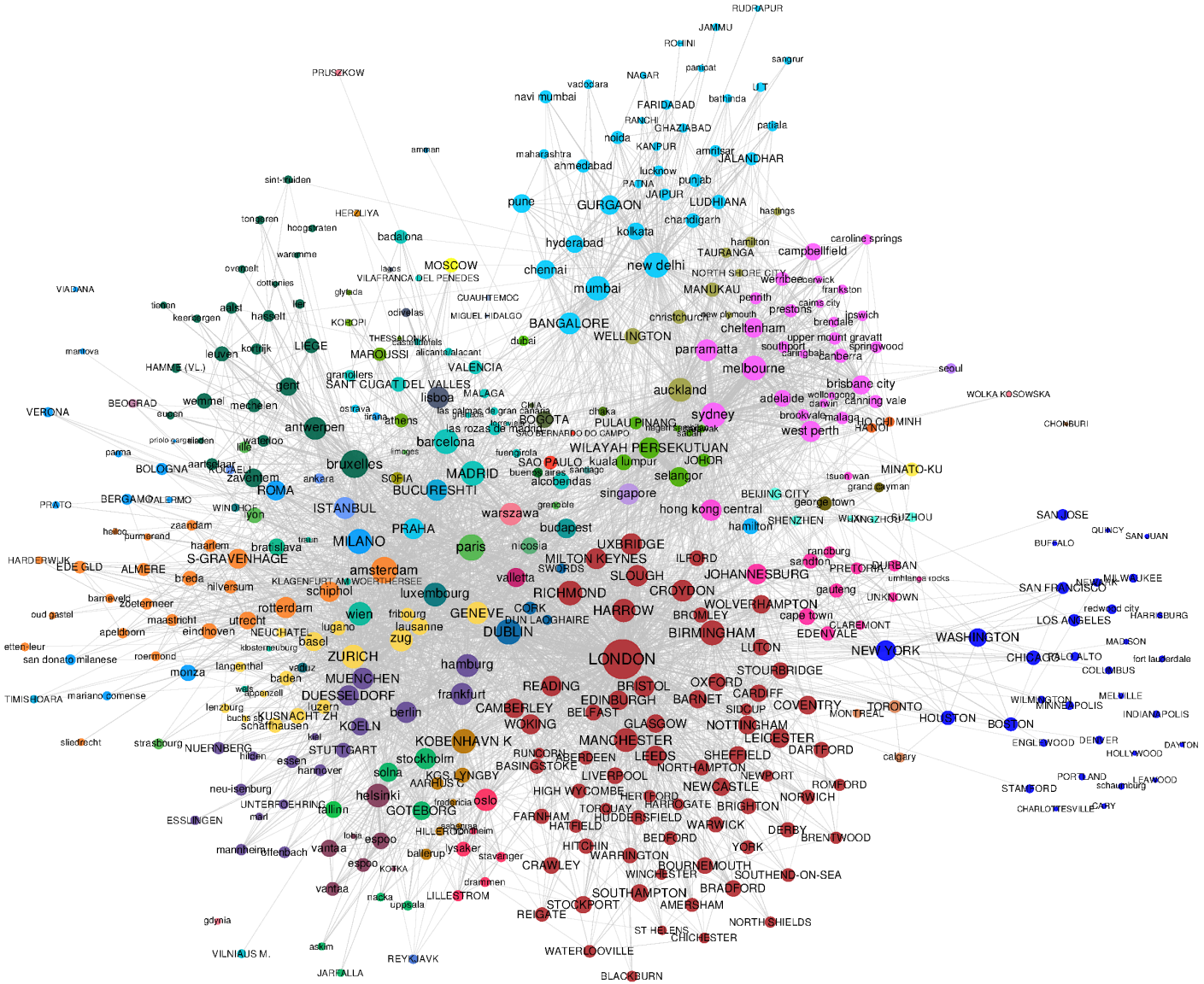
E.M. Heemskerk and F.W. Takes, The Corporate Elite Community Structure of Global Capitalism, *New Political Economy* 21(1): 90-118, 2016.

Community detection



E.M. Heemskerk, F.W. Takes, J. Garcia-Bernardo and M.J. Huijzer, Where is the global corporate elite? A large-scale network study of local and nonlocal interlocking directorates, *Sociologica* 2016(2): 1-31, 2016.

Nodes colored by country (sample)



Community composition

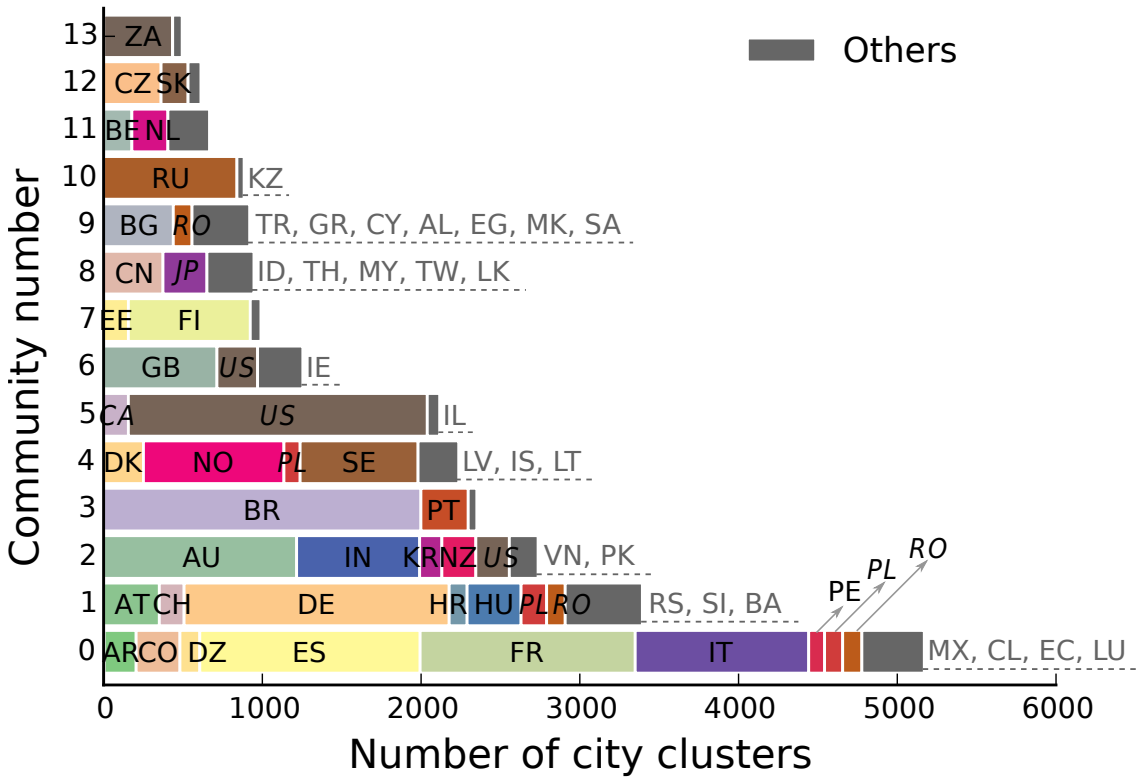


Figure: Country involved in each community

Communities in scientific co-citation networks

Co-citation network

- Nodes are scientific publications
- Edges indicate that papers cite the same previous work
- Each node has an associated scientific field
- Network provides insight in how scientific fields interact

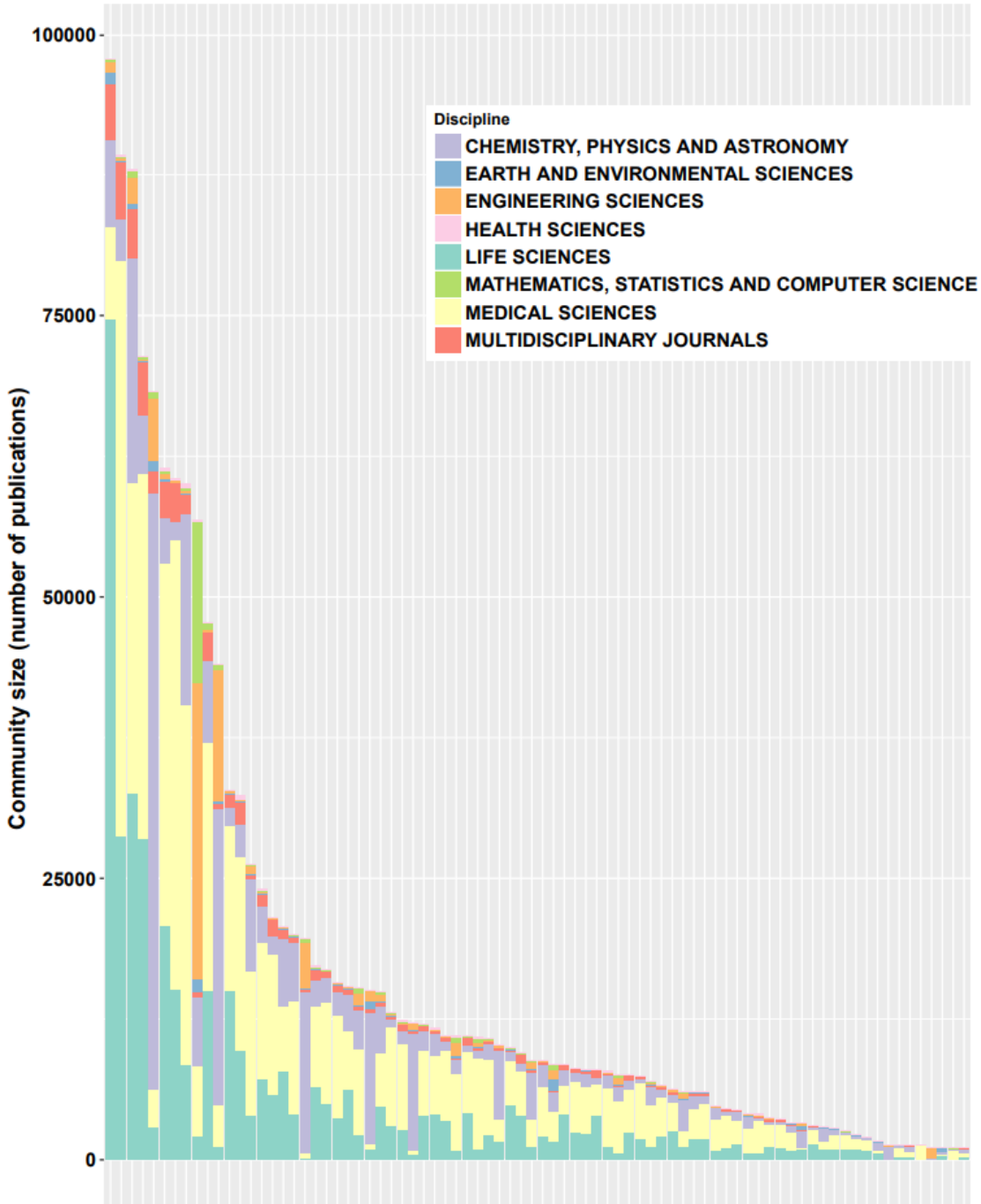
Co-citation network

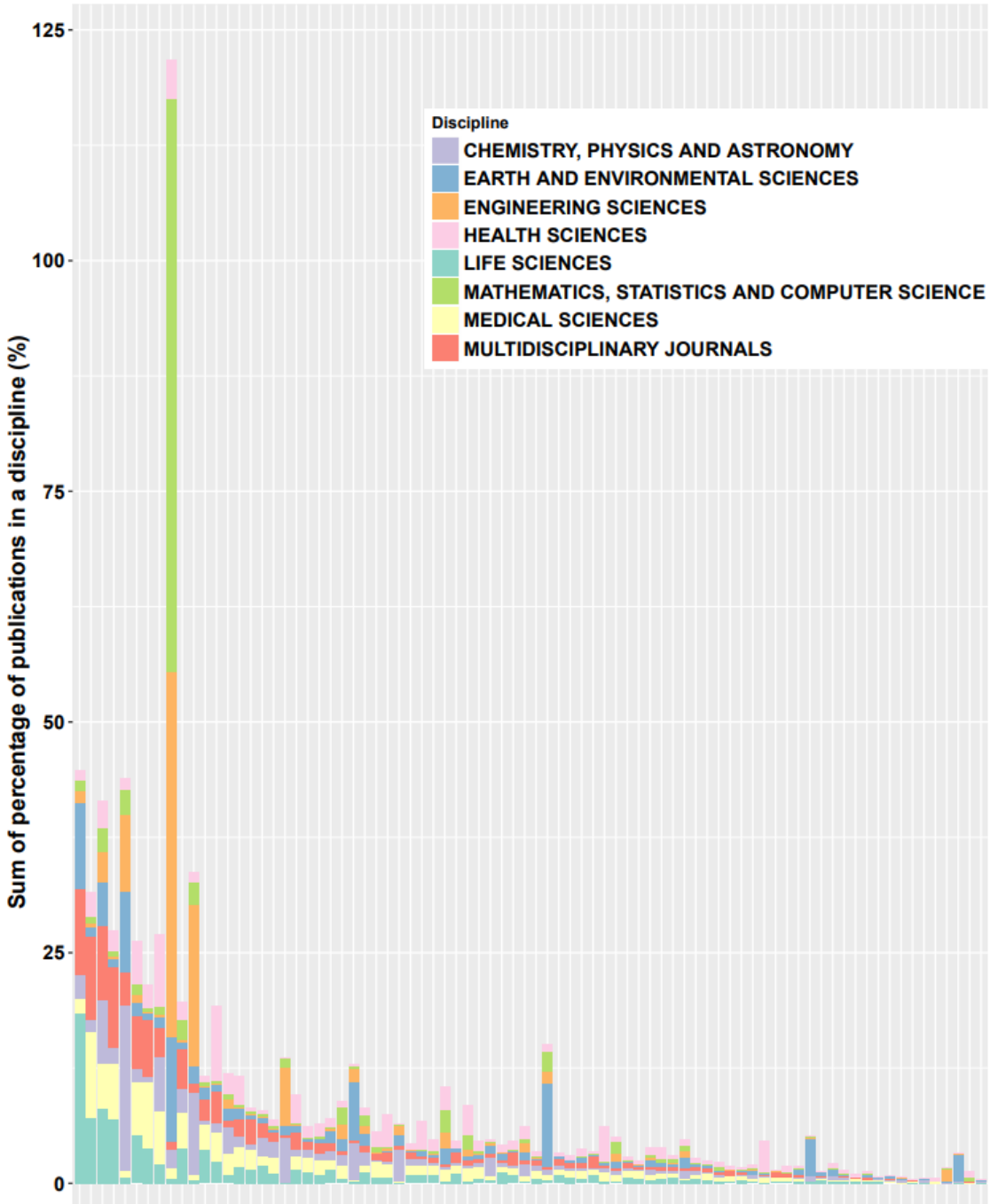
- Nodes are scientific publications
- Edges indicate that papers cite the same previous work
- Each node has an associated scientific field
- Network provides insight in how scientific fields interact
- Size: 1.6 million nodes and 44 million edges
- 99% in giant component, scale-free, small-world

Co-citation network

Discipline	Number of publications
MEDICAL SCIENCES	550672.68
LIFE SCIENCES	403633.85
CHEMISTRY, PHYSICS AND ASTRONOMY	293971.77
ENGINEERING SCIENCES	66186.33
MULTIDISCIPLINARY JOURNALS	55394.00
MATHEMATICS, STATISTICS AND COMPUTER SCIENCE	23192.52
EARTH AND ENVIRONMENTAL SCIENCES	10596.43
HEALTH SCIENCES	5043.42

Figure: Categories of publications (weighting applied if multiple apply)


100000
75000
50000
25000
0
Community size (number of publications)
Discipline
CHEMISTRY, PHYSICS AND ASTRONOMY
EARTH AND ENVIRONMENTAL SCIENCES
ENGINEERING SCIENCES
HEALTH SCIENCES
LIFE SCIENCES
MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
MEDICAL SCIENCES
MULTIDISCIPLINARY JOURNALS


Discipline
CHEMISTRY, PHYSICS AND ASTRONOMY
EARTH AND ENVIRONMENTAL SCIENCES
ENGINEERING SCIENCES
HEALTH SCIENCES
LIFE SCIENCES
MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
MEDICAL SCIENCES
MULTIDISCIPLINARY JOURNALS
Sum of percentage of publications in a discipline (%)
125
100
75
50
25
0

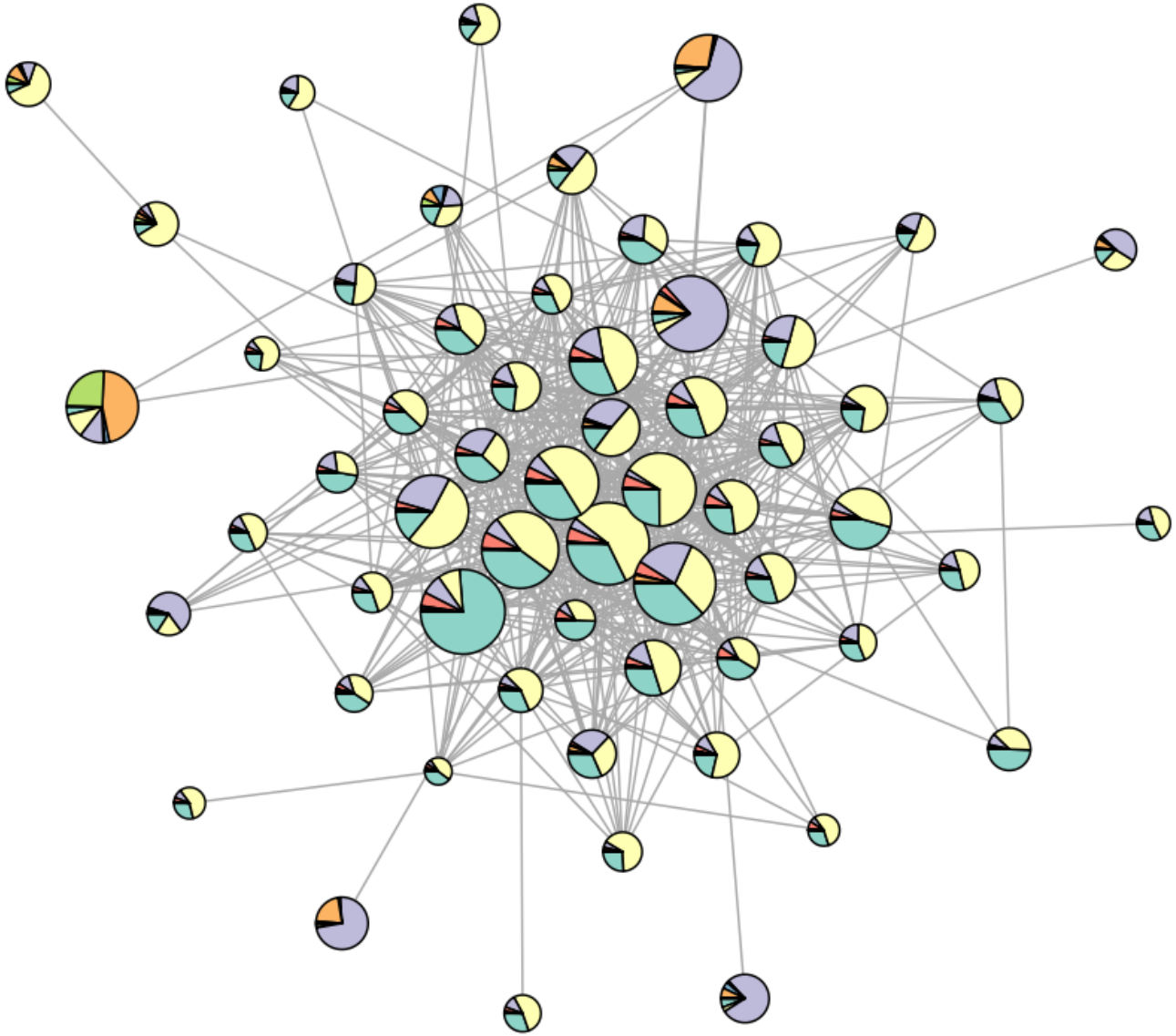


Figure: Community composition and connections

Computing betweenness centrality

Centrality measures compared

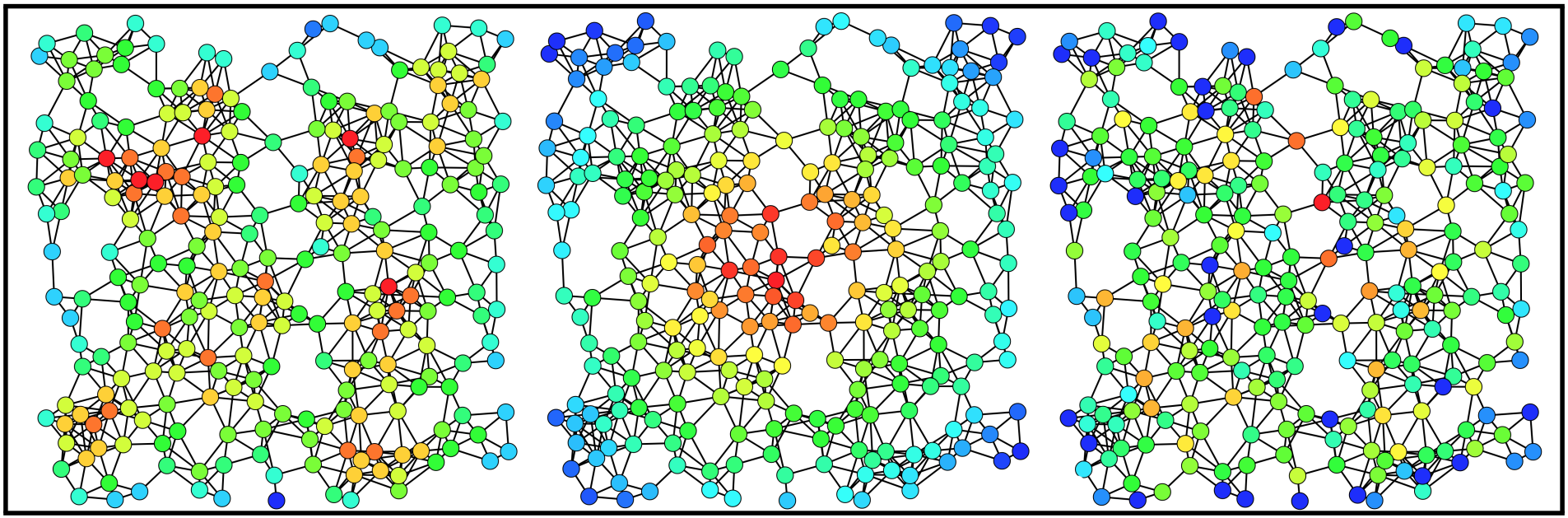


Figure: Degree, closeness and betweenness centrality

Source: "Centrality"' by Claudio Rocchini, Wikipedia `File:Centrality.svg`

Recap: Betweenness centrality

- **Betweenness centrality**: measure the number of shortest paths that run through a node

$$C_b(u) = \sum_{\substack{v,w \in V \\ v \neq w, u \neq v, u \neq w}} \frac{\sigma_u(v,w)}{\sigma(v,w)}$$

- $\sigma(v,w)$ is the number of shortest paths from v to w
- $\sigma_u(v,w)$ is the number of such shortest paths that run through u
- Divide by largest value to normalize to $[0;1]$
- Global path-based measure
- $O(2mn)$ time to compute (two "BFSes" for each node)

U. Brandes, "A faster algorithm for betweenness centrality", Journal of Mathematical Sociology 25(2): 163–177, 2001

Counting shortest paths

- **Bellman criterion**: v lies on a shortest path from u to w if $d(u,v) + d(v,w) = d(u,w)$.
- Predecessors: $P_u(w)$ is the set of predecessors of node w on a shortest path from u, formally:

$$P_u(w) = \{v \in V : (v,w) \in E, d(u,w) = d(u,v) + 1\}$$

- Counting the number of shortest paths $\sigma(u,w)$ for $u \neq w$:

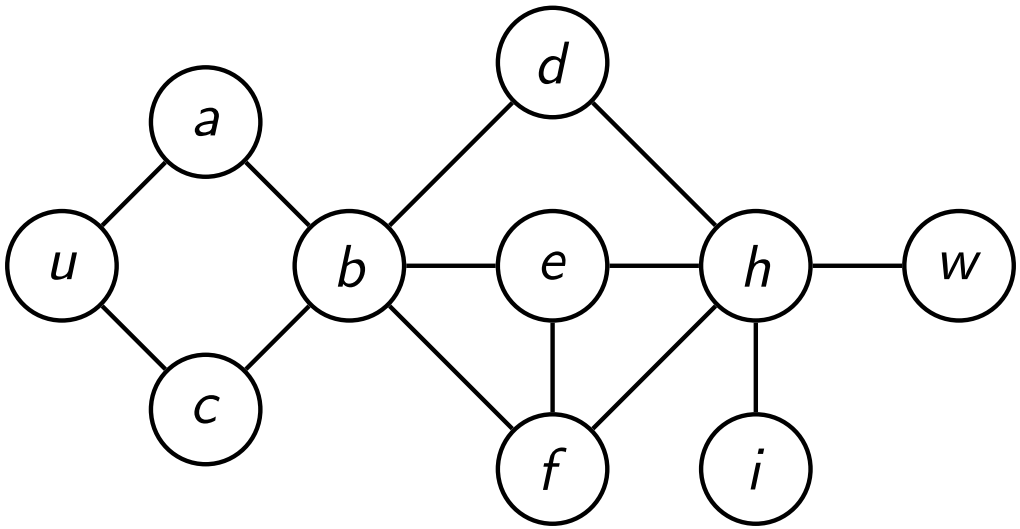
$$\sigma(u,w) = \sum_{v \in P_u(w)} \sigma(u,v)$$

$$\text{with } \sigma(u,u) = 1$$

Counting shortest paths

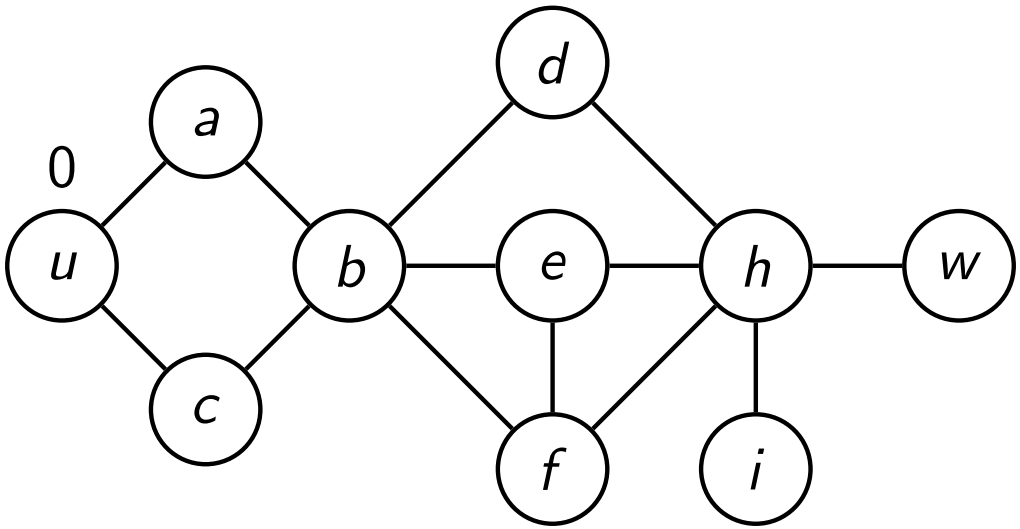
Task: compute $\sigma(u, w)$, the number of shortest paths from u to w

...



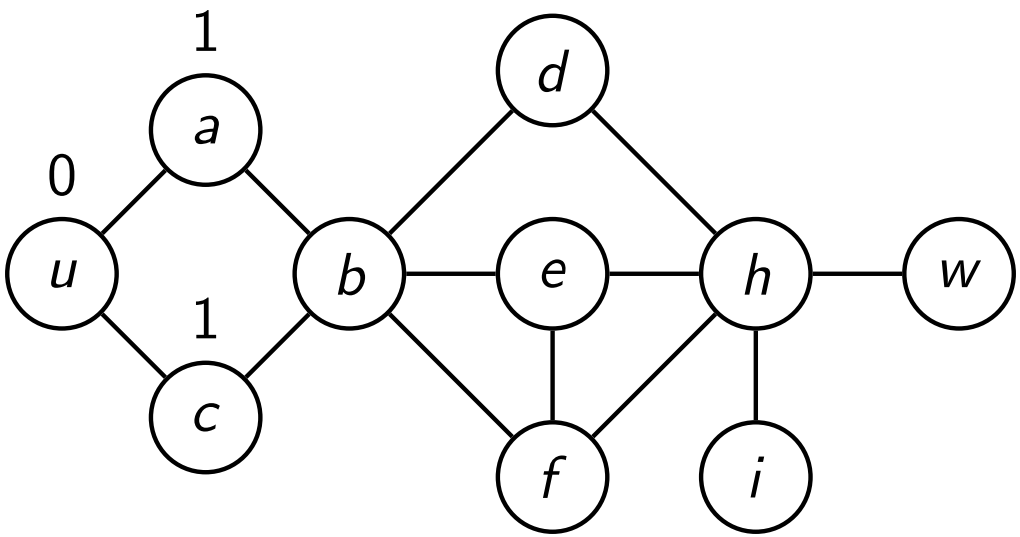
Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w
First a BFS from u to find w



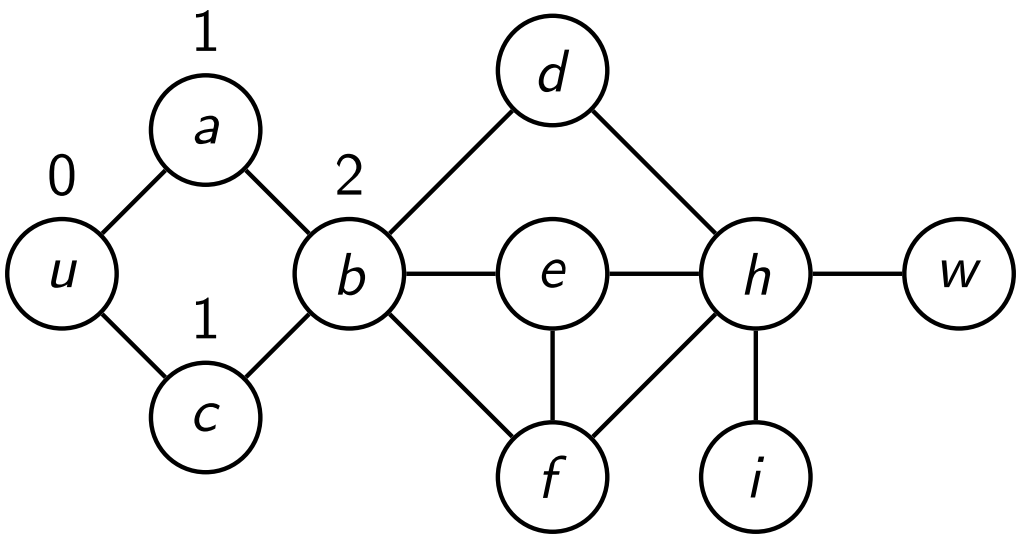
Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w
First a BFS from u to find w



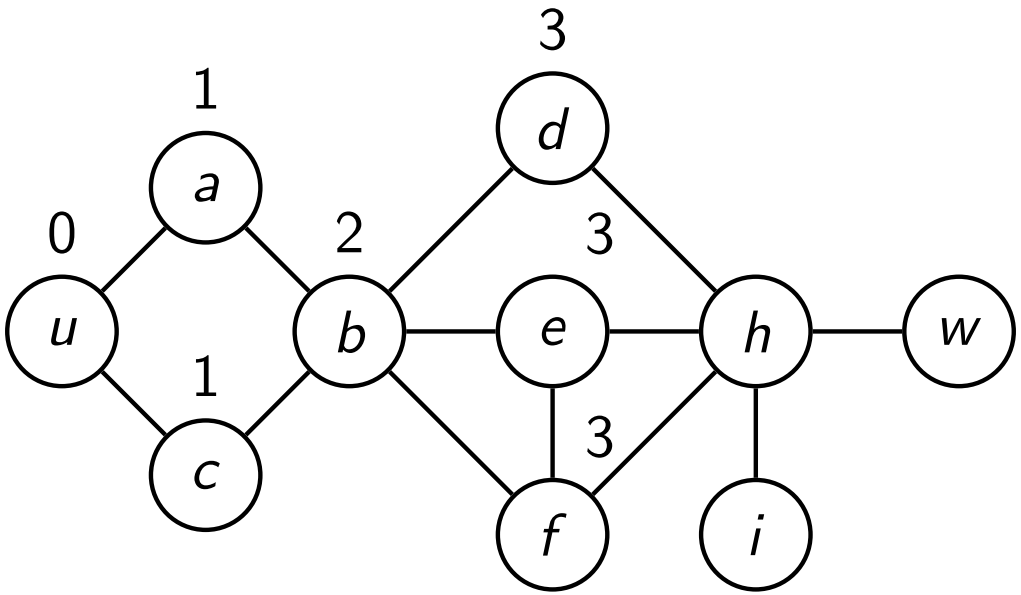
Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w
First a BFS from u to find w



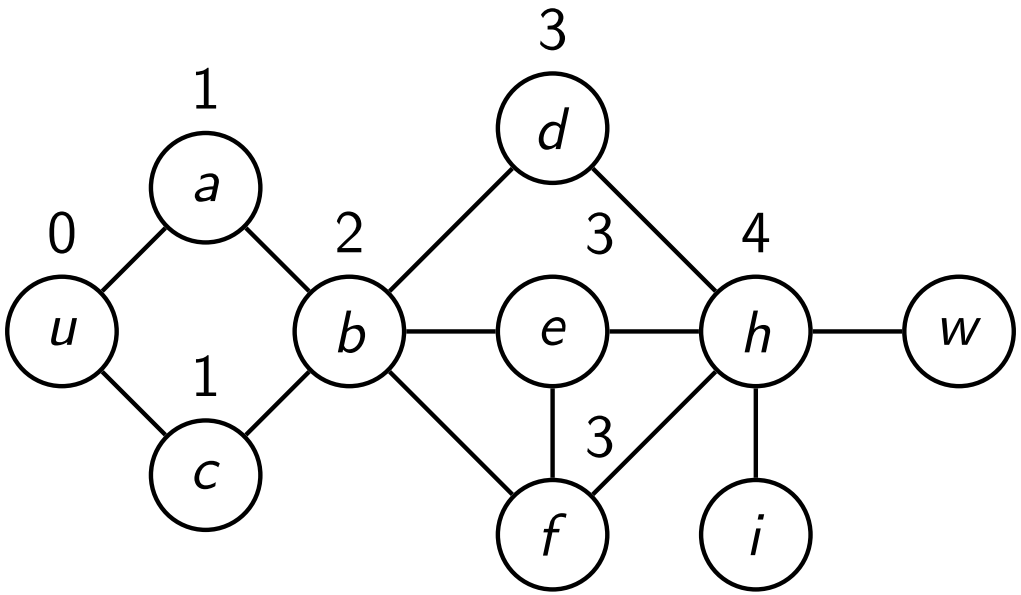
Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w
First a BFS from u to find w



Counting shortest paths

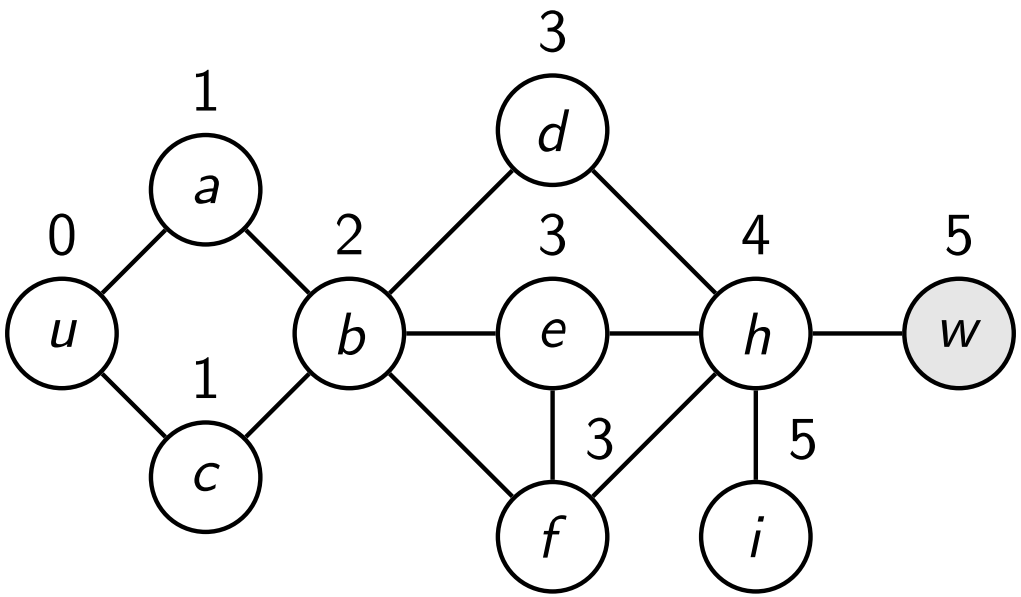
Task: compute $\sigma(u, w)$, the number of shortest paths from u to w
First a BFS from u to find w



Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w

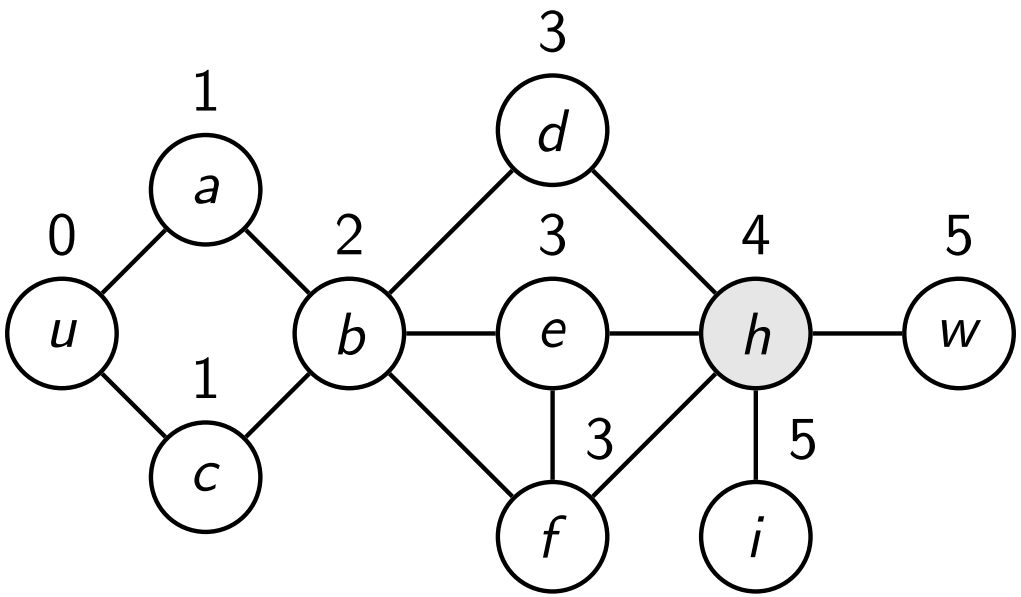
$\sigma(u, w)$? Ask predecessors $P_u(w) = \{h\}$ for its value



Counting shortest paths

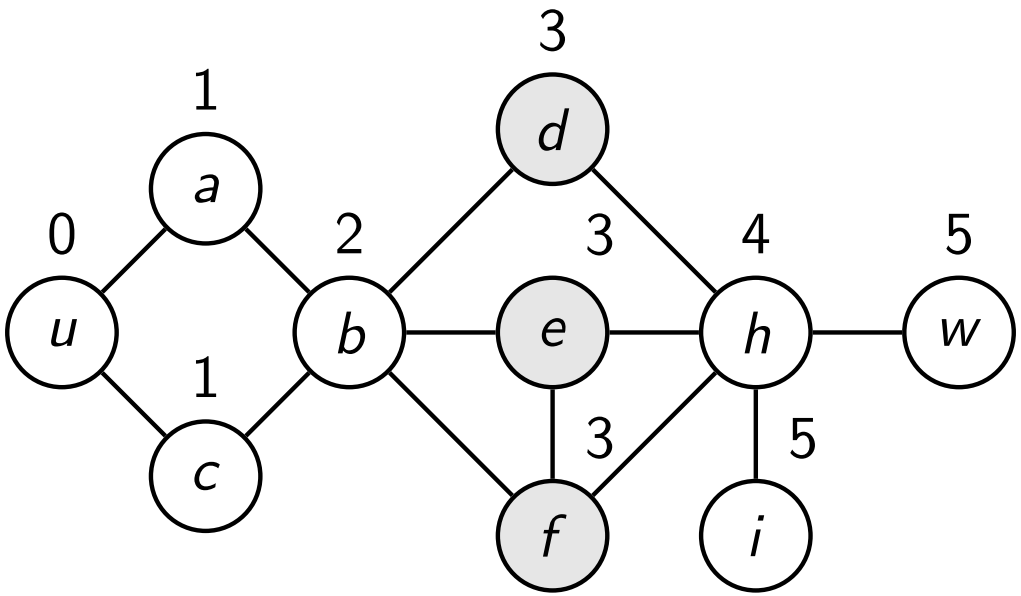
Task: compute $\sigma(u, w)$, the number of shortest paths from u to w

$\sigma(u, h)$? Ask predecessors $P_u(h) = \{d, e, f\}$ for its value



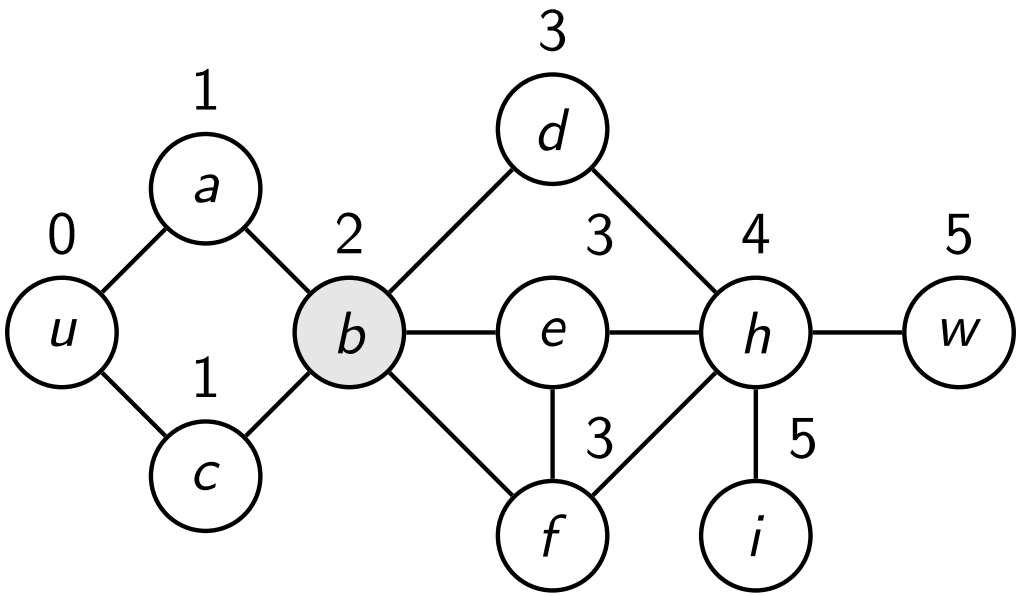
Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w
Ask predecessors ...



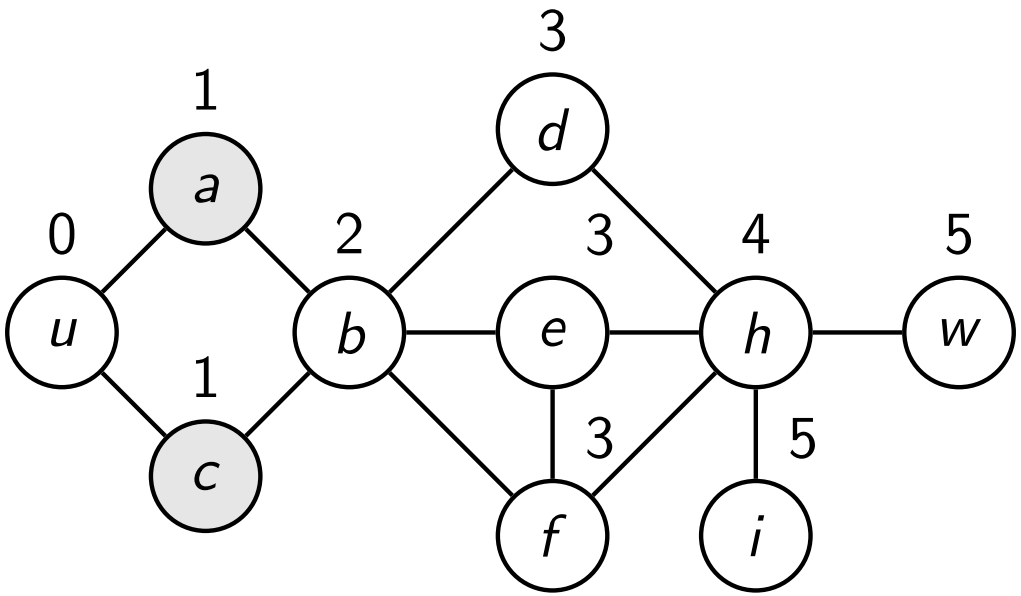
Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w
Ask predecessors ...



Counting shortest paths

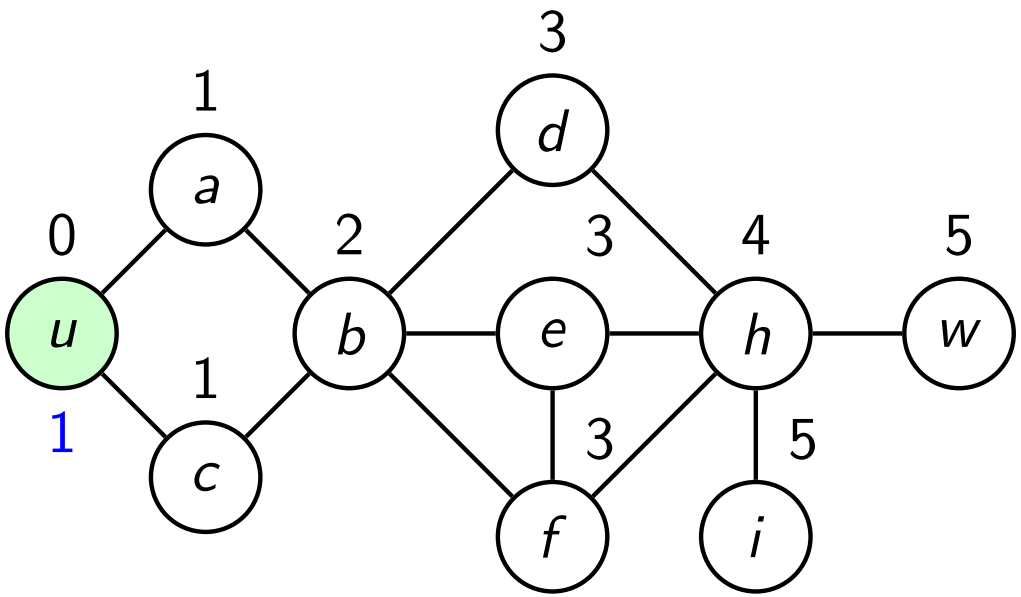
Task: compute $\sigma(u, w)$, the number of shortest paths from u to w
Ask predecessors ...



Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w

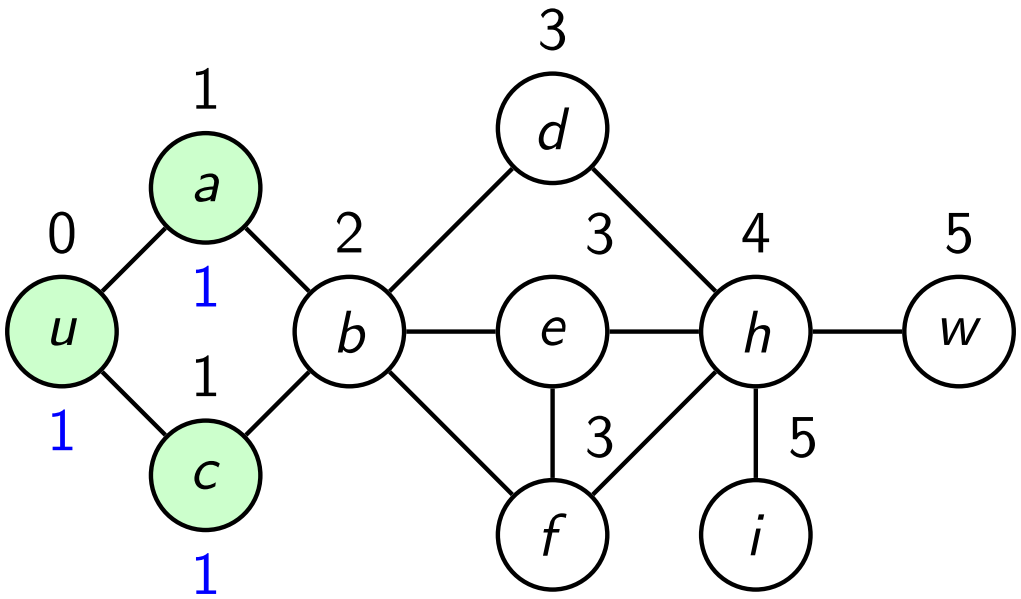
$\sigma(u, u) = 1$, now propagate back...



Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w

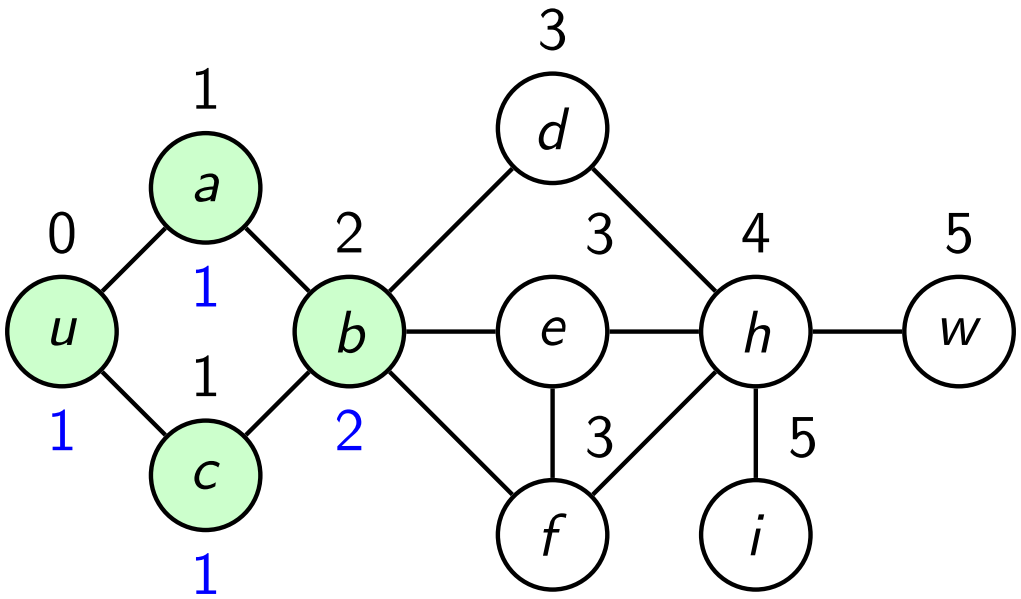
$$\sigma(u, a) = \sigma(u, c) = 1$$



Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w

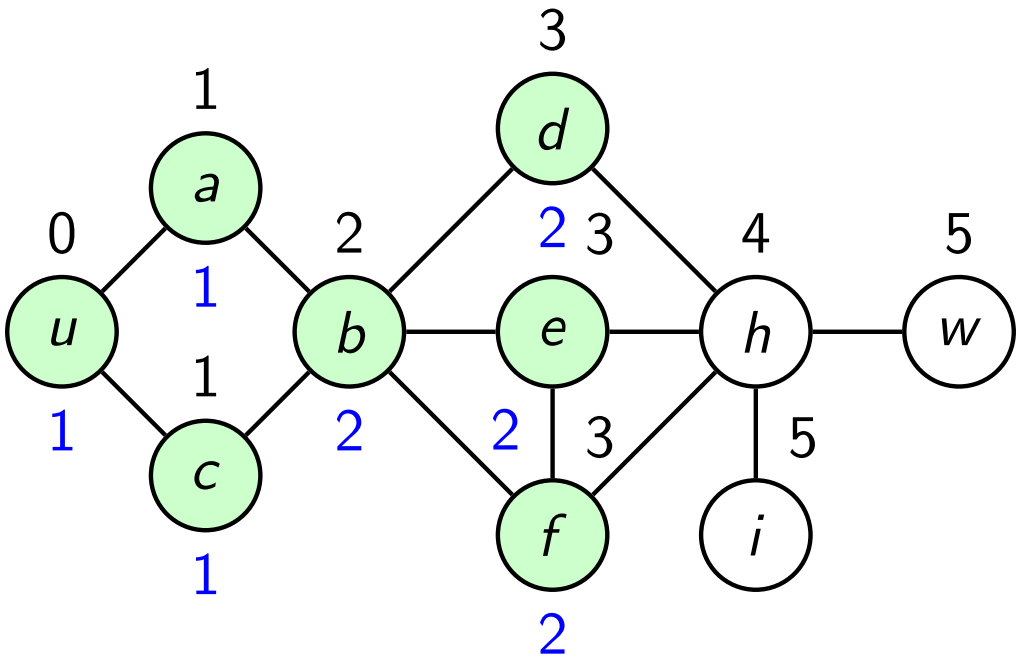
$$\sigma(u, b) = \sigma(u, a) + \sigma(u, c) = 2$$



Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w

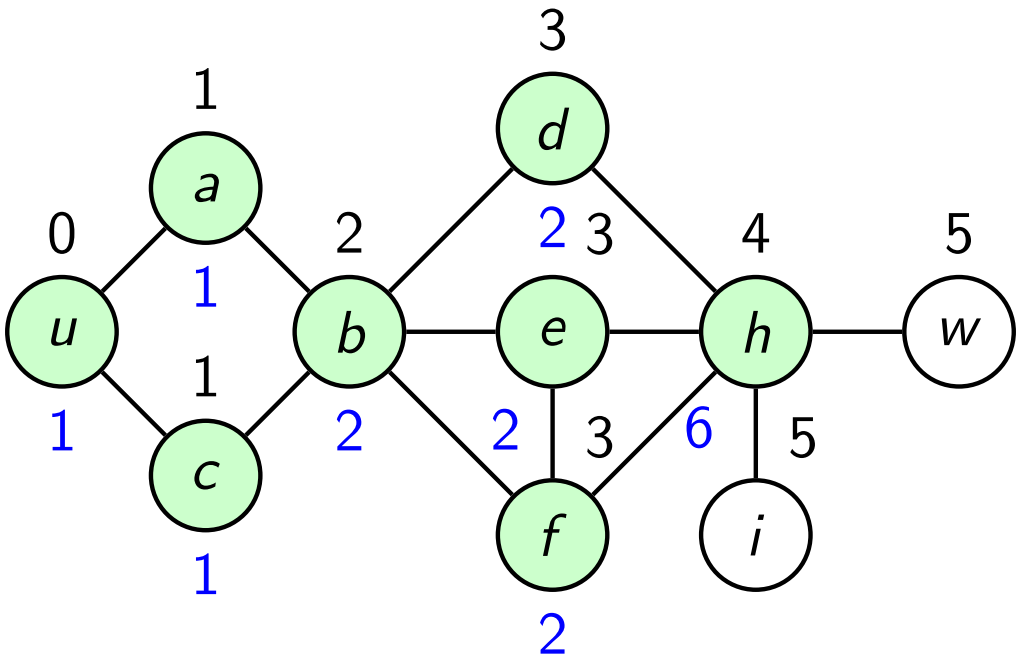
$$\sigma(u, d) = \sigma(u, e) = \sigma(u, f) = \sigma(u, c) = 2$$



Counting shortest paths

Task: compute $\sigma(u, w)$, the number of shortest paths from u to w

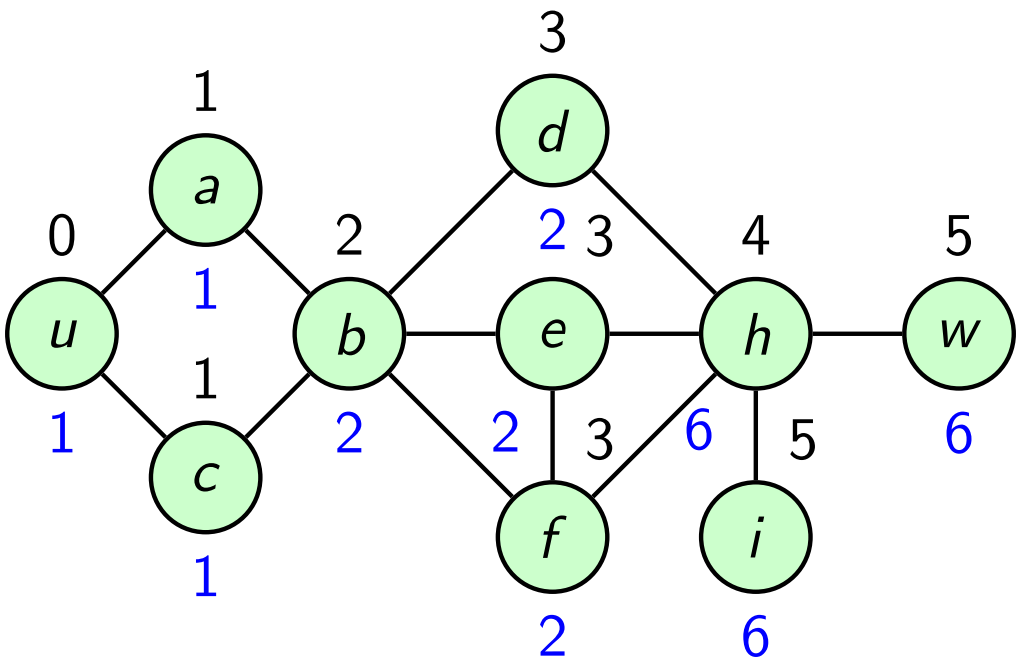
$$\sigma(u, h) = \sigma(u, d) + \sigma(u, e) + \sigma(u, f) = 6$$



Counting shortest paths

Done!

$$\sigma(u, w) = \sigma(u, h) = 6$$



Epidemic spread

Based on https://www.youtube.com/watch?v=XWXqXzAYe4E

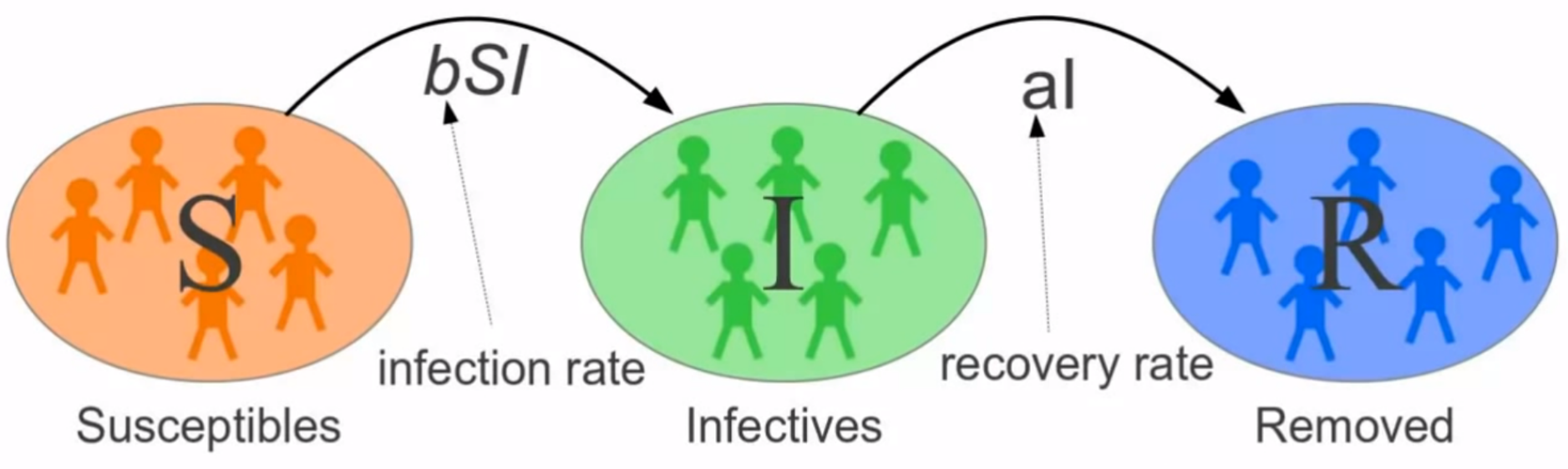
Epidemic spread

- Given a social network, how does a disease spread? (modelling)
- Related: how does a message or rumor spread? (influence maximization)
- Great societal relevance
- Applications in health, viral marketing, innovation networks, etc.

SIR

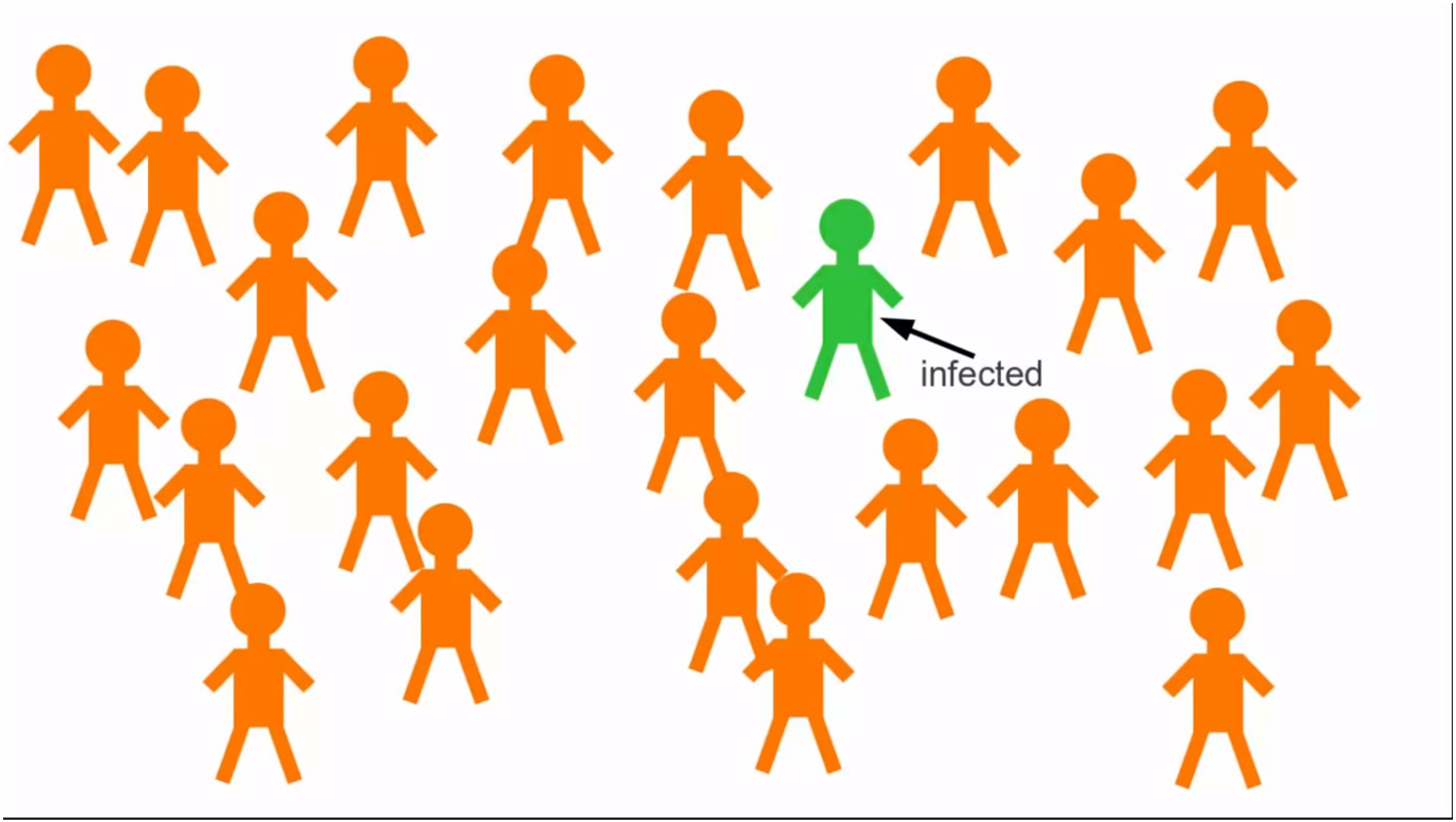
- Susceptible: individuals that can get infected S
- Infected: individuals that are infected I
- Removed: individuals who are immune R
- Infection rate (based on S and I) b
- Recovery rate (based on I) a

SIR

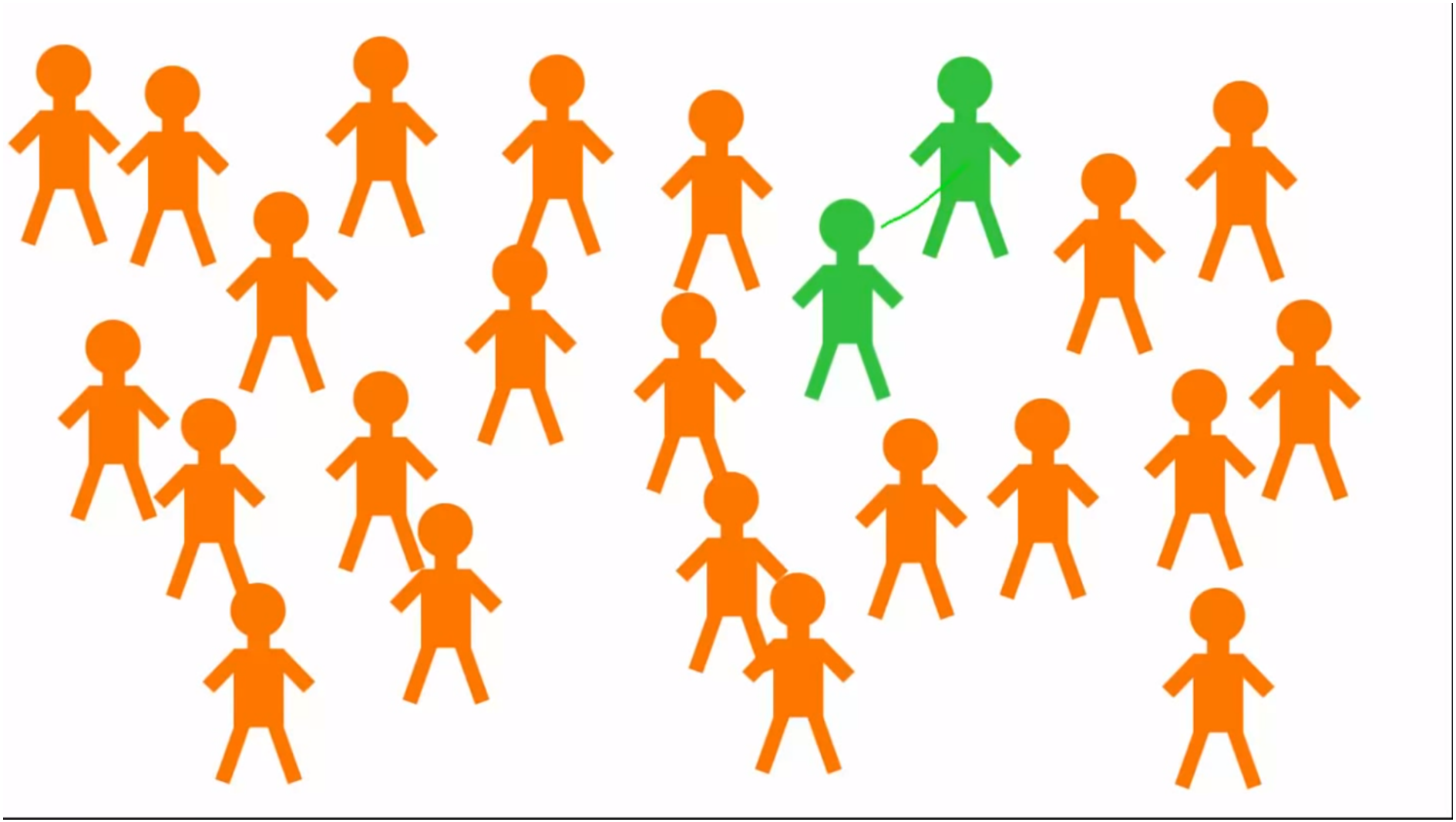

bSI
aI
S
I
R
infection rate
recovery rate
Susceptibles
Infectives
Removed

SIR

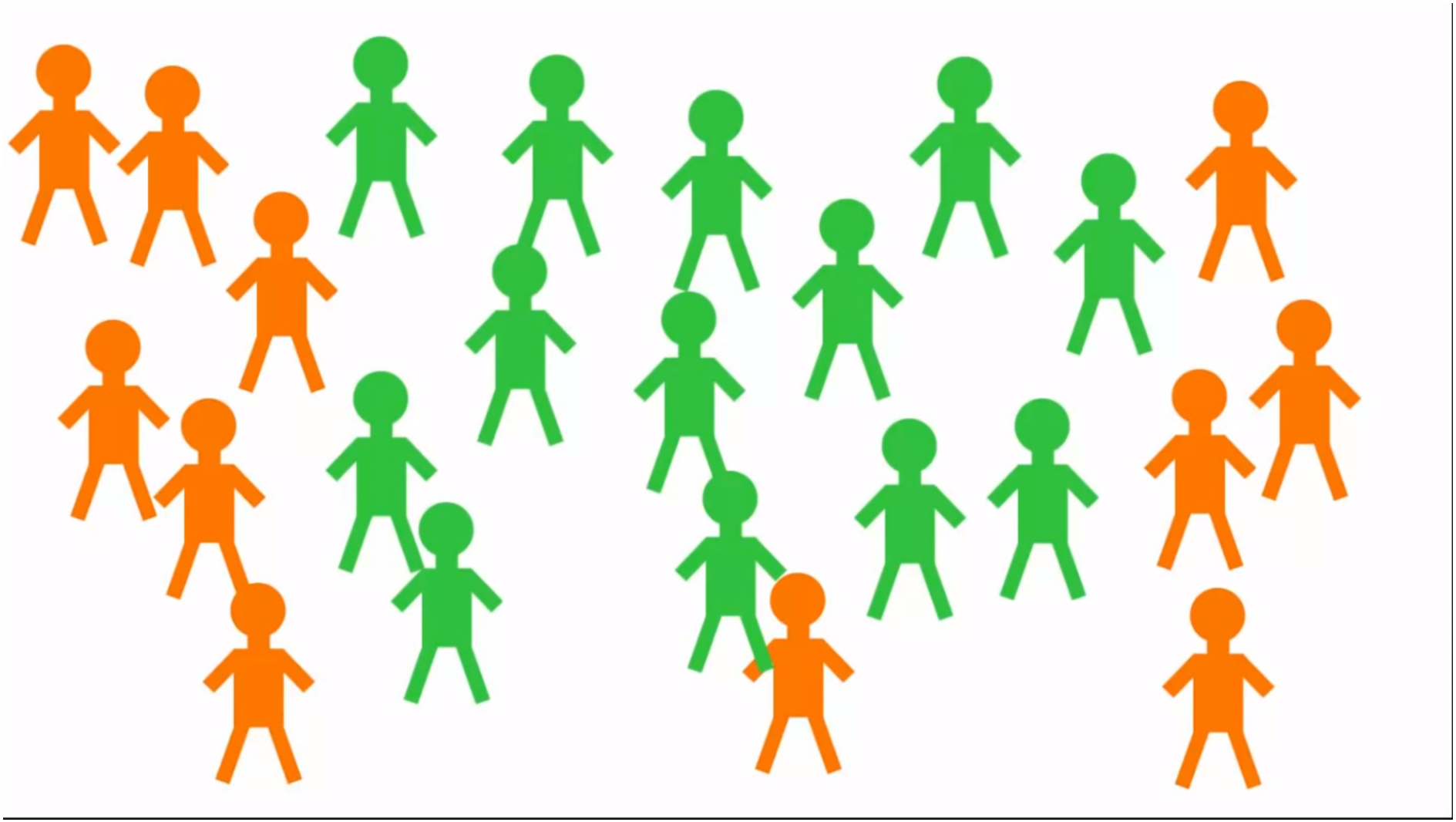
SIR



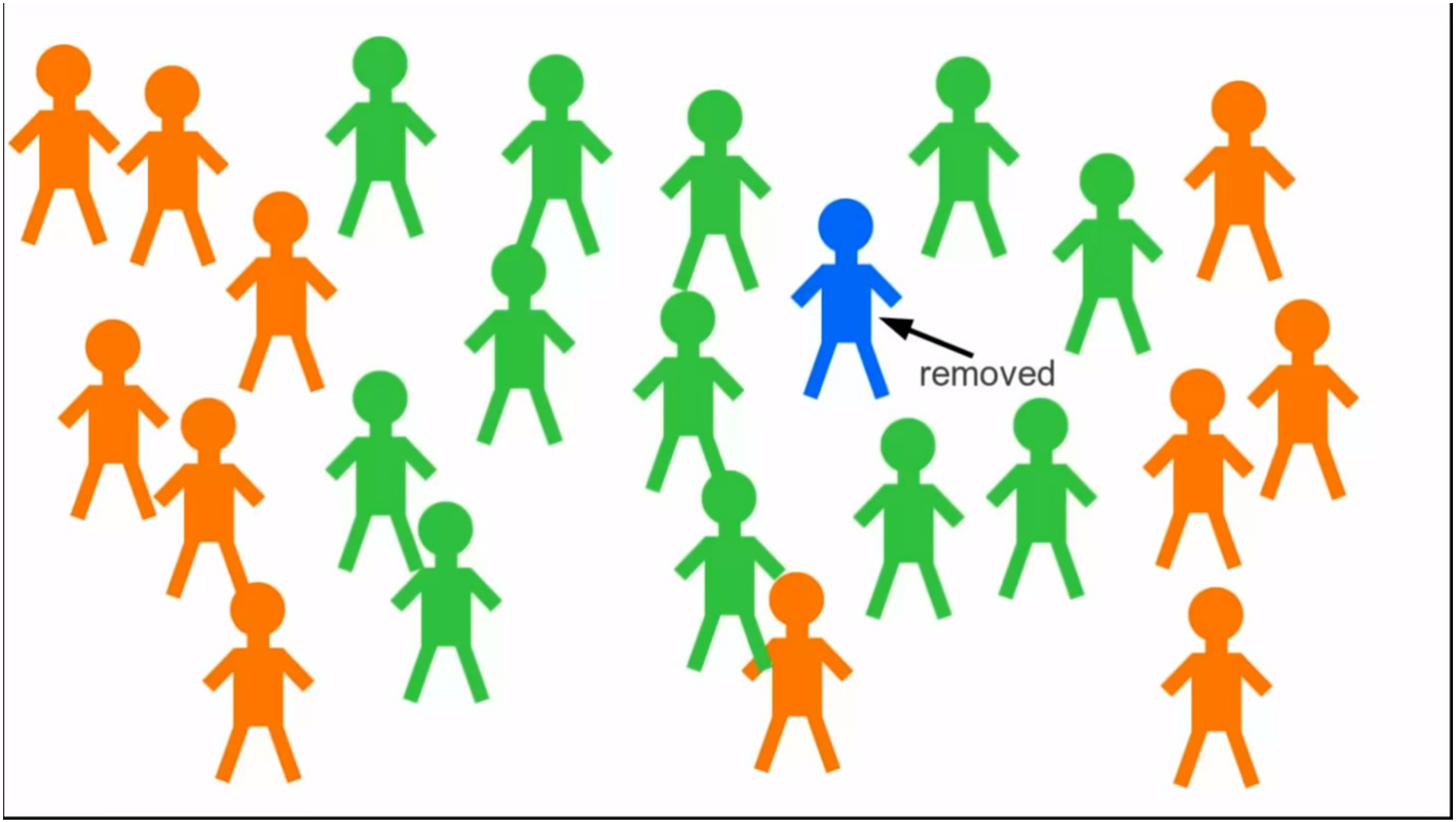
SIR



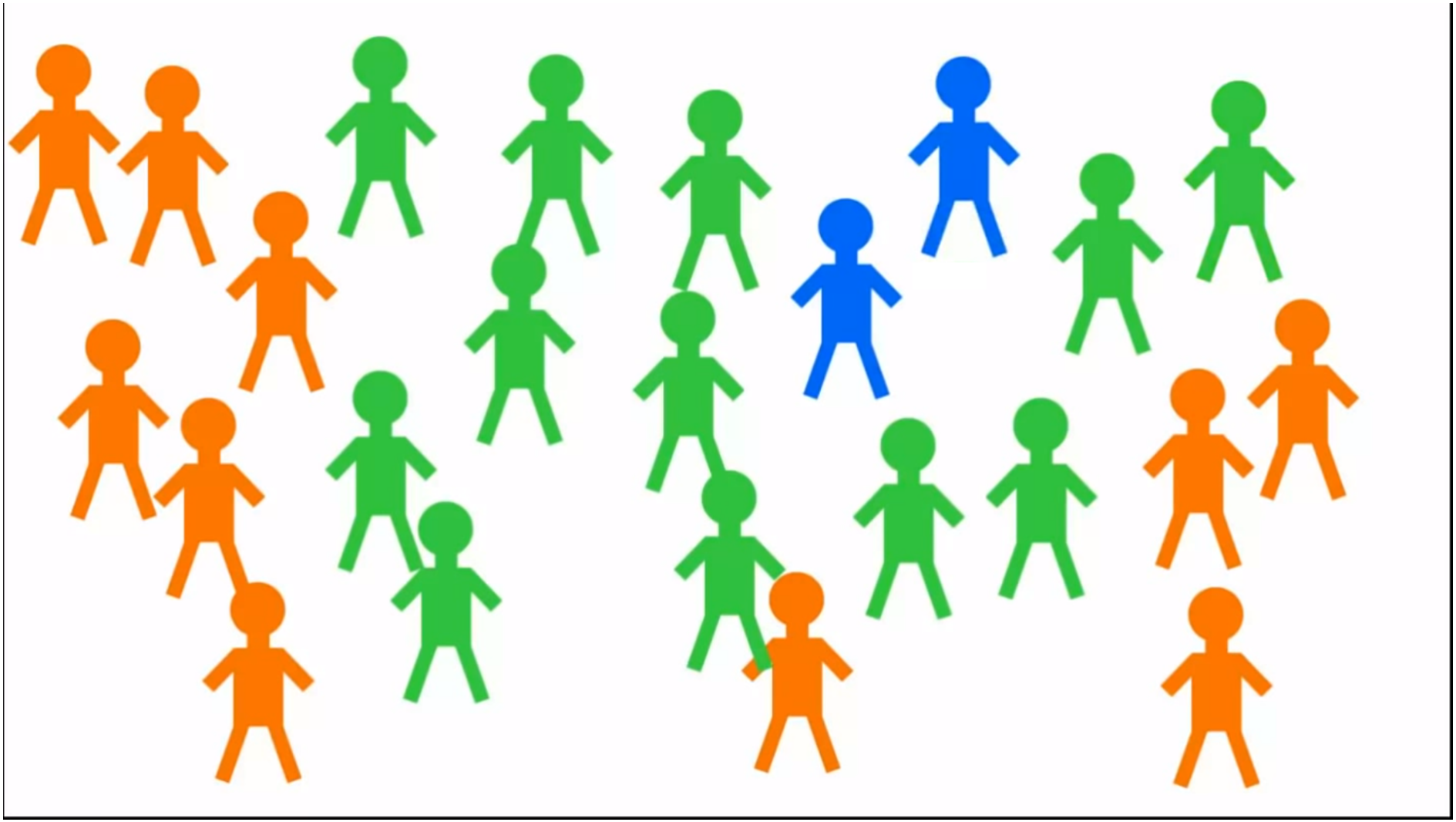
SIR



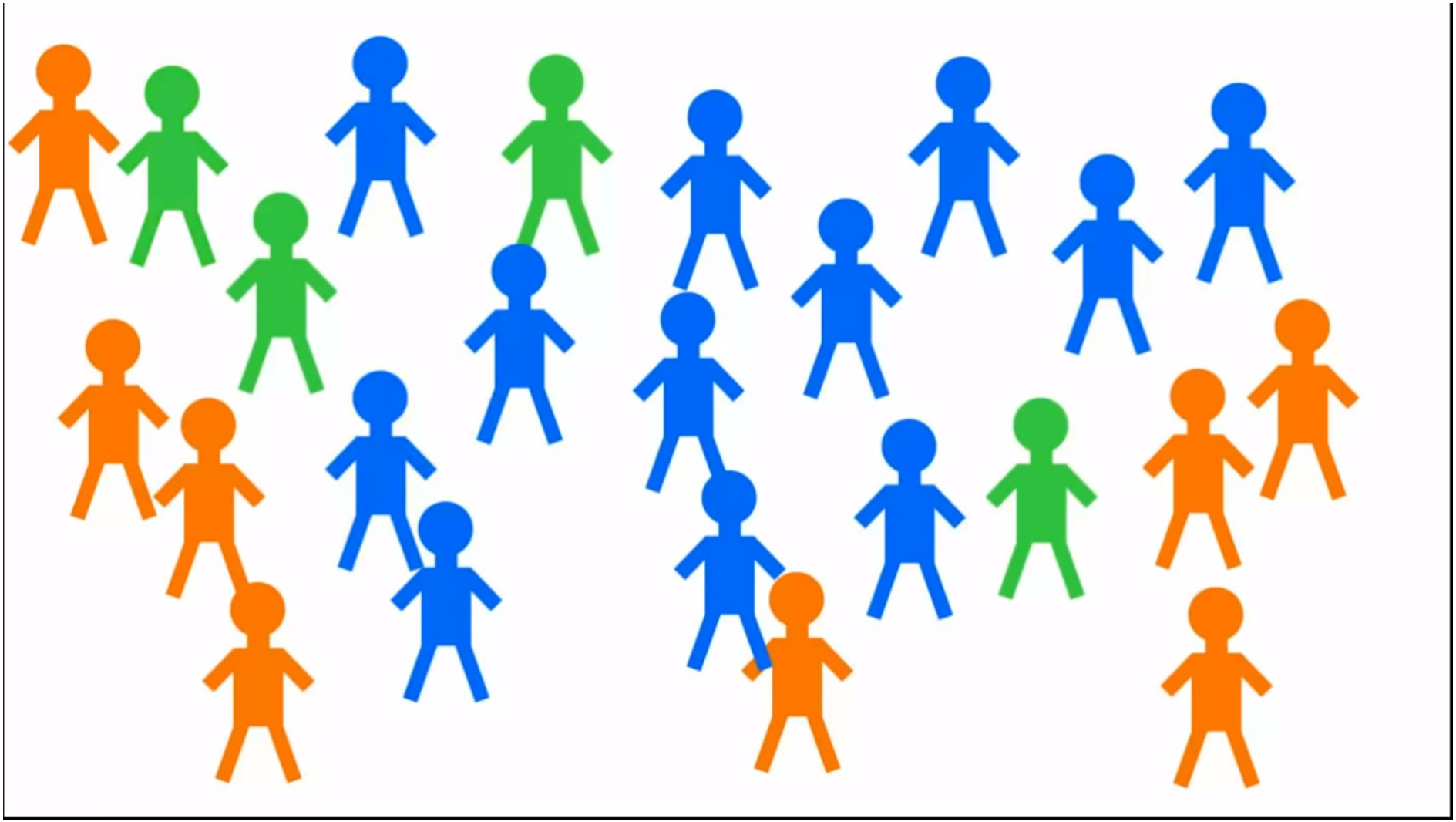
SIR



SIR



SIR



SIR in networks

- Infection rate b based on neighborhood and link weights
- Recovery rate a based on neighborhood and link weights

SIR in networks

- Infection rate b based on neighborhood and link weights
- Recovery rate a based on neighborhood and link weights
- Elegant, theoretically ...

SIR in networks

- Infection rate b based on neighborhood and link weights
- Recovery rate a based on neighborhood and link weights
- Elegant, theoretically ...
- What are a and b?
- What is the influence of link weights?
- Need (a lot of (hard to get)) empirical data

SIR in networks

- Infection rate b based on neighborhood and link weights
- Recovery rate a based on neighborhood and link weights
- Elegant, theoretically ...
- What are a and b?
- What is the influence of link weights?
- Need (a lot of (hard to get)) empirical data
- Difficult, practically ...

Network science challenges

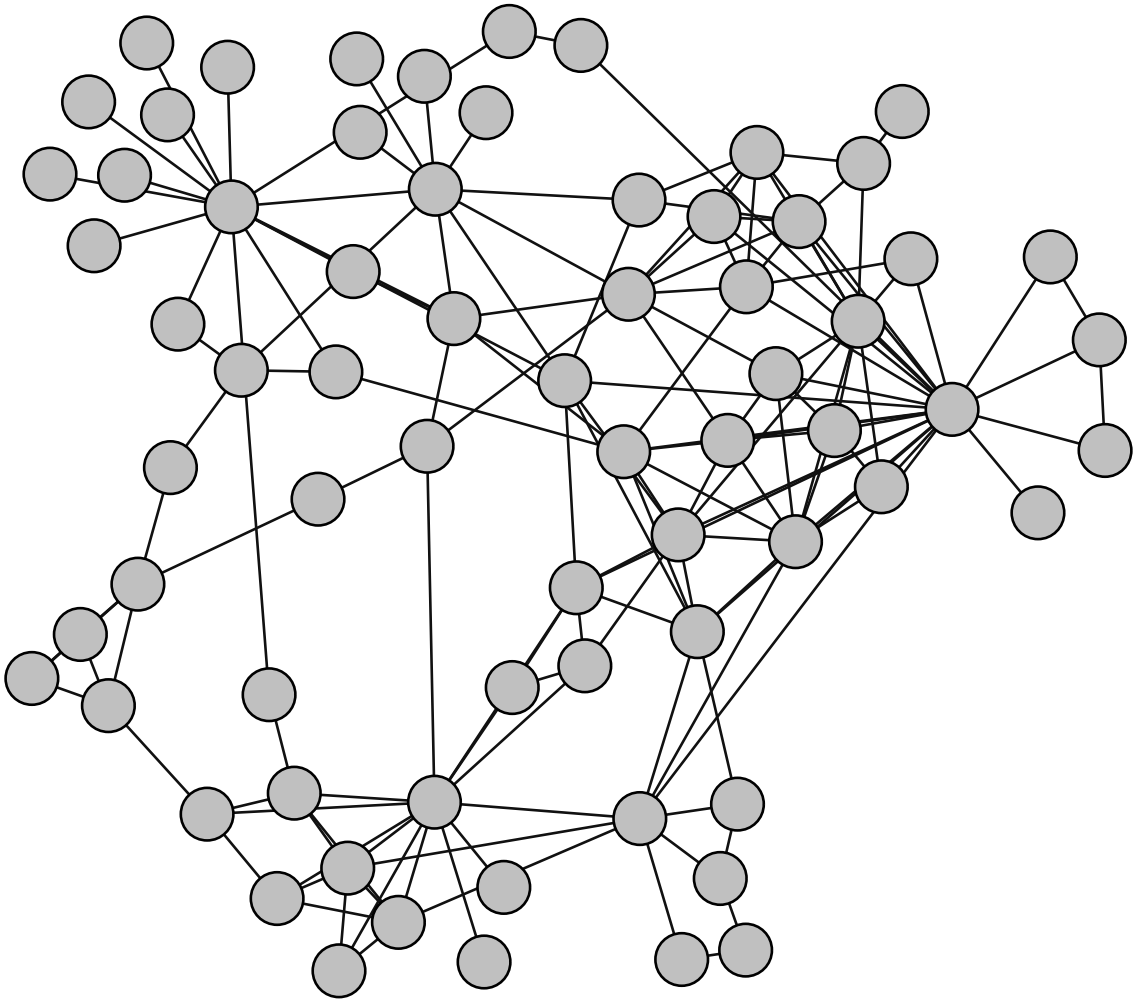
Network science

- **Network science**: understanding data by investigating interactions and relationships between individual data objects as a network
- **Networks** are the central model of computation

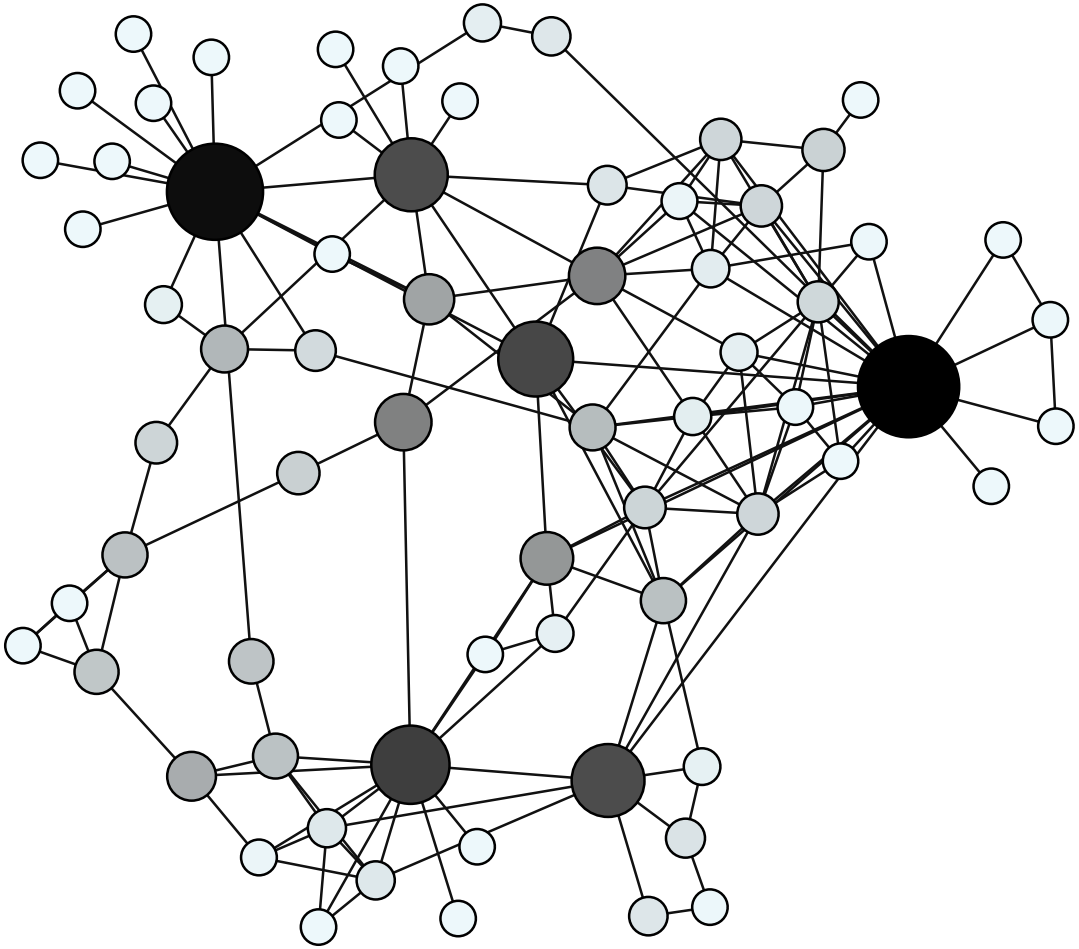
Network science

- **Network science**: understanding data by investigating interactions and relationships between individual data objects as a network
- **Networks** are the central model of computation
- Branch of data science focusing on network data
- Method in complexity research
- Complex systems approach: the behavior emerging from the network reveals patterns not visible when studying the individuals
- For now assume: network science = social network analysis

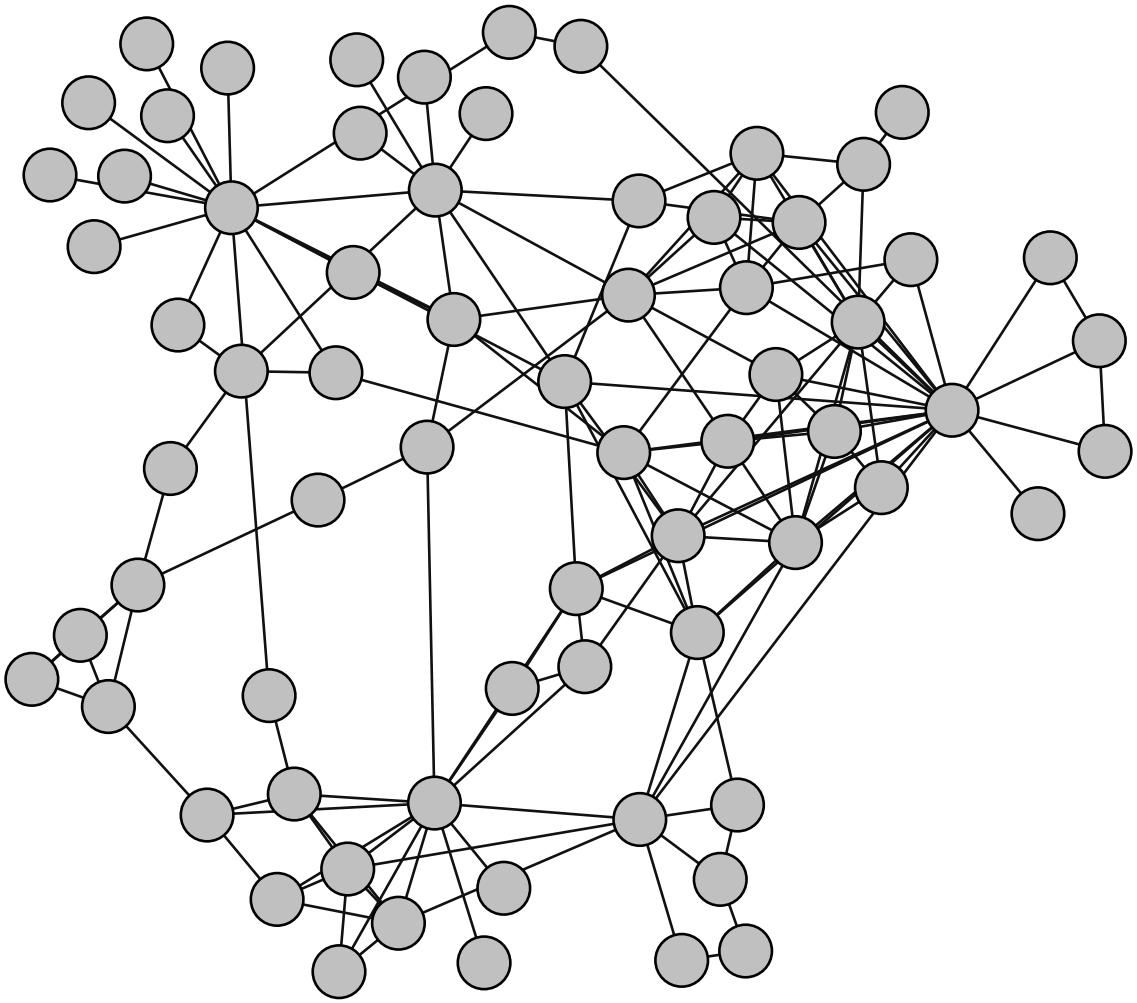
Network analysis



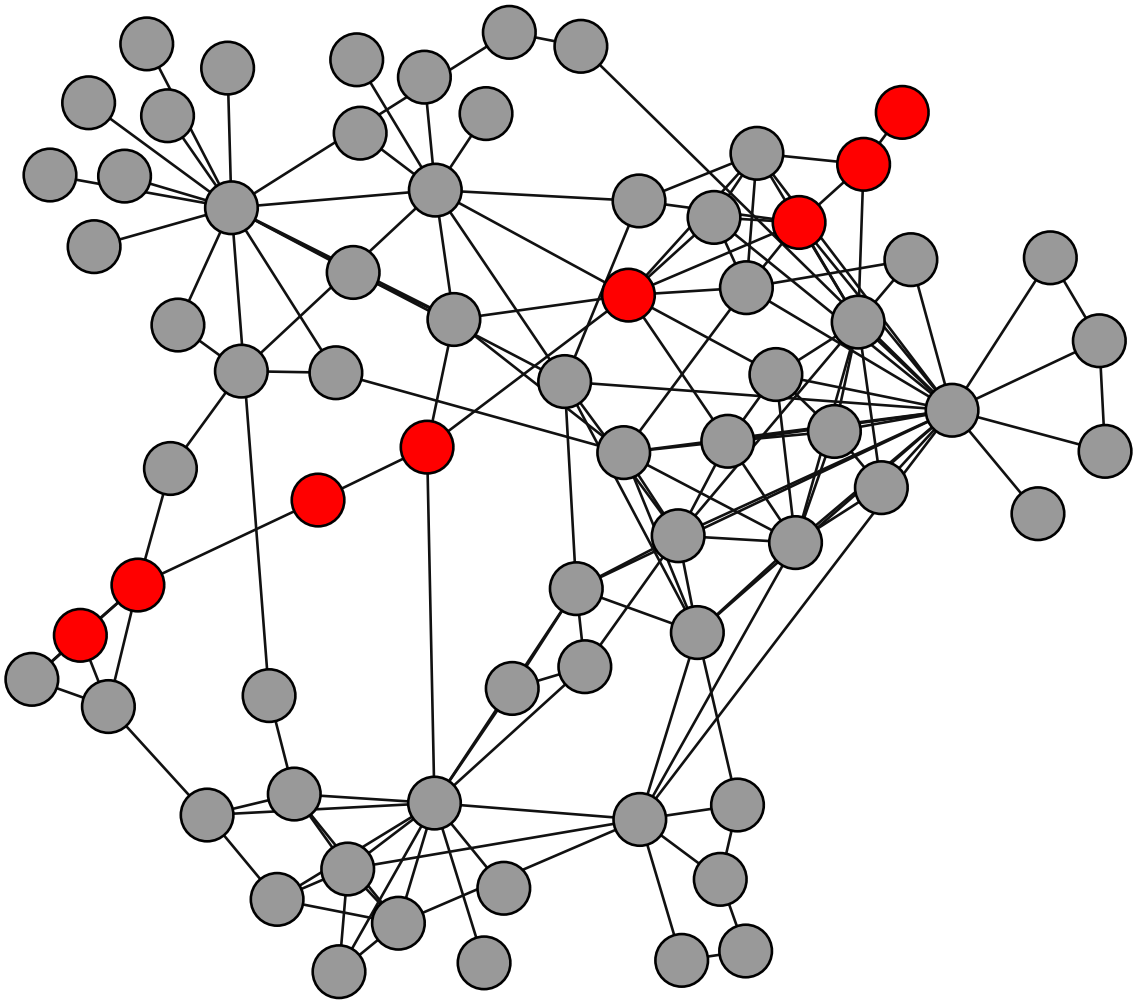
Micro scale



Macro scale



Macro scale



Network analysis

- **Micro** scale: analyzing the position of individual nodes, based on their structural position in the network (e.g., node centrality, etc.)
- **Macro** scale: analyzing the structure of the network as a whole (e.g., network diameter, small-world effect, etc.)

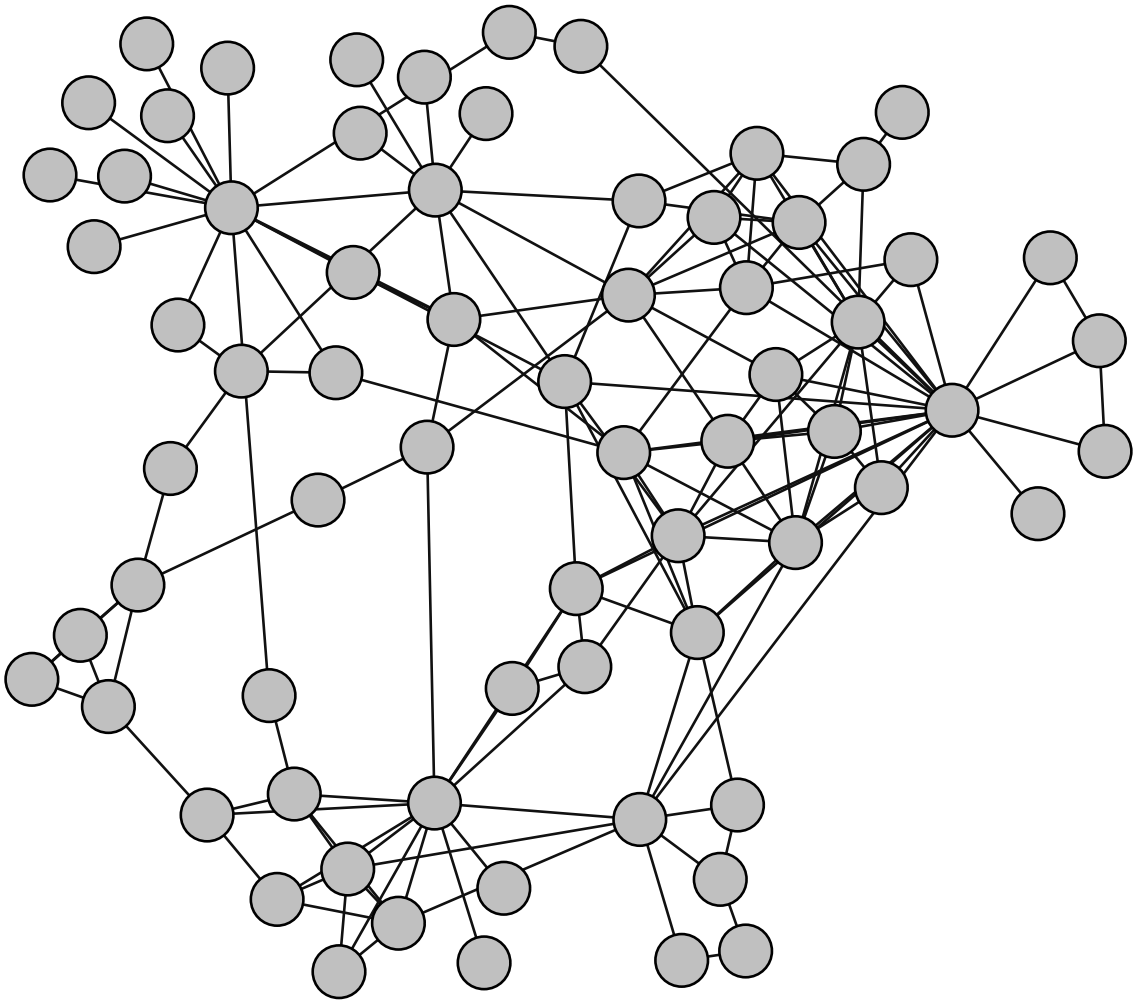
Network analysis

- **Micro** scale: analyzing the position of individual nodes, based on their structural position in the network (e.g., node centrality, etc.)
- **Macro** scale: analyzing the structure of the network as a whole (e.g., network diameter, small-world effect, etc.)
- **Meso** scale: analyzing groups of nodes occurring in a particular configuration

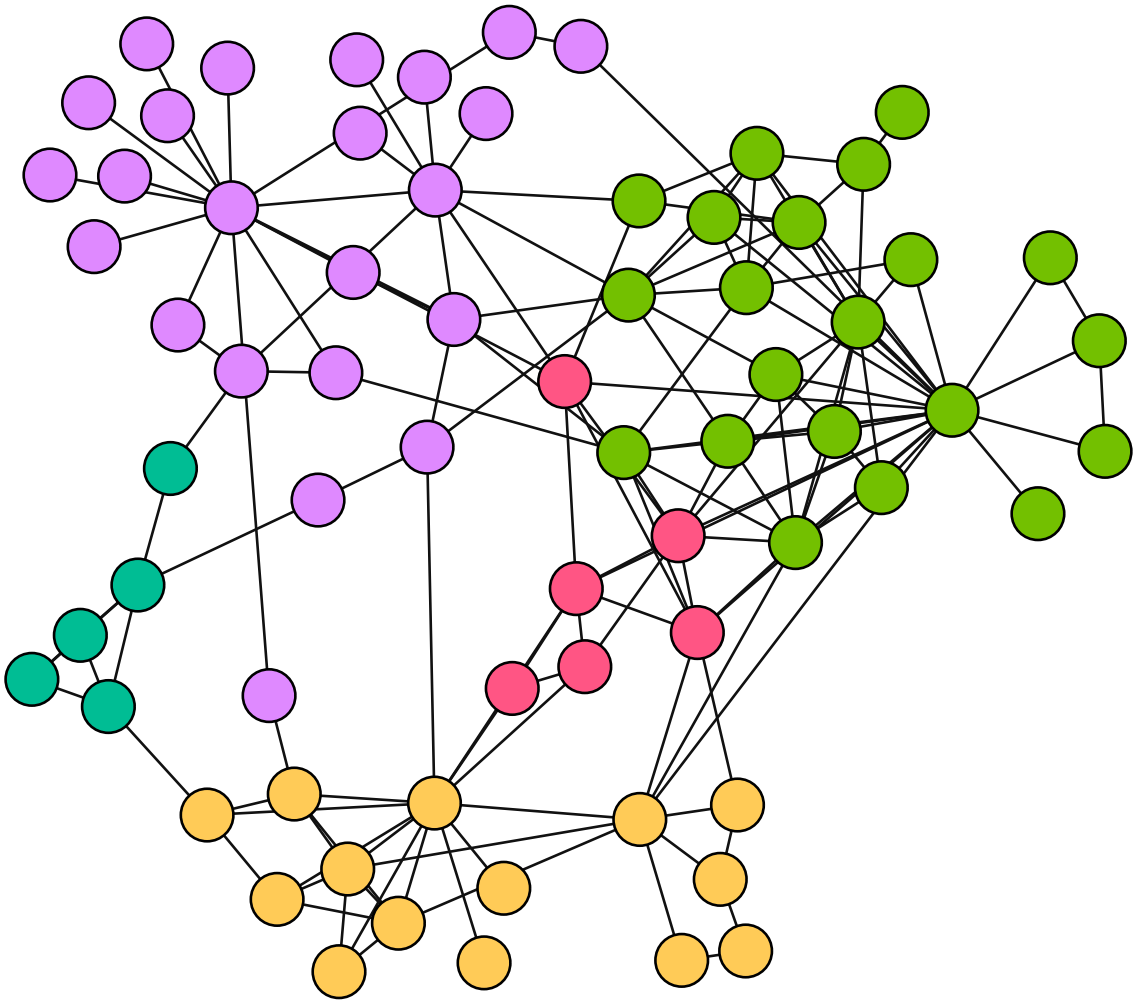
Network analysis

- **Micro** scale: analyzing the position of individual nodes, based on their structural position in the network (e.g., node centrality, etc.)
- **Macro** scale: analyzing the structure of the network as a whole (e.g., network diameter, small-world effect, etc.)
- **Meso** scale: analyzing groups of nodes occurring in a particular configuration (e.g., communities or networks motifs)

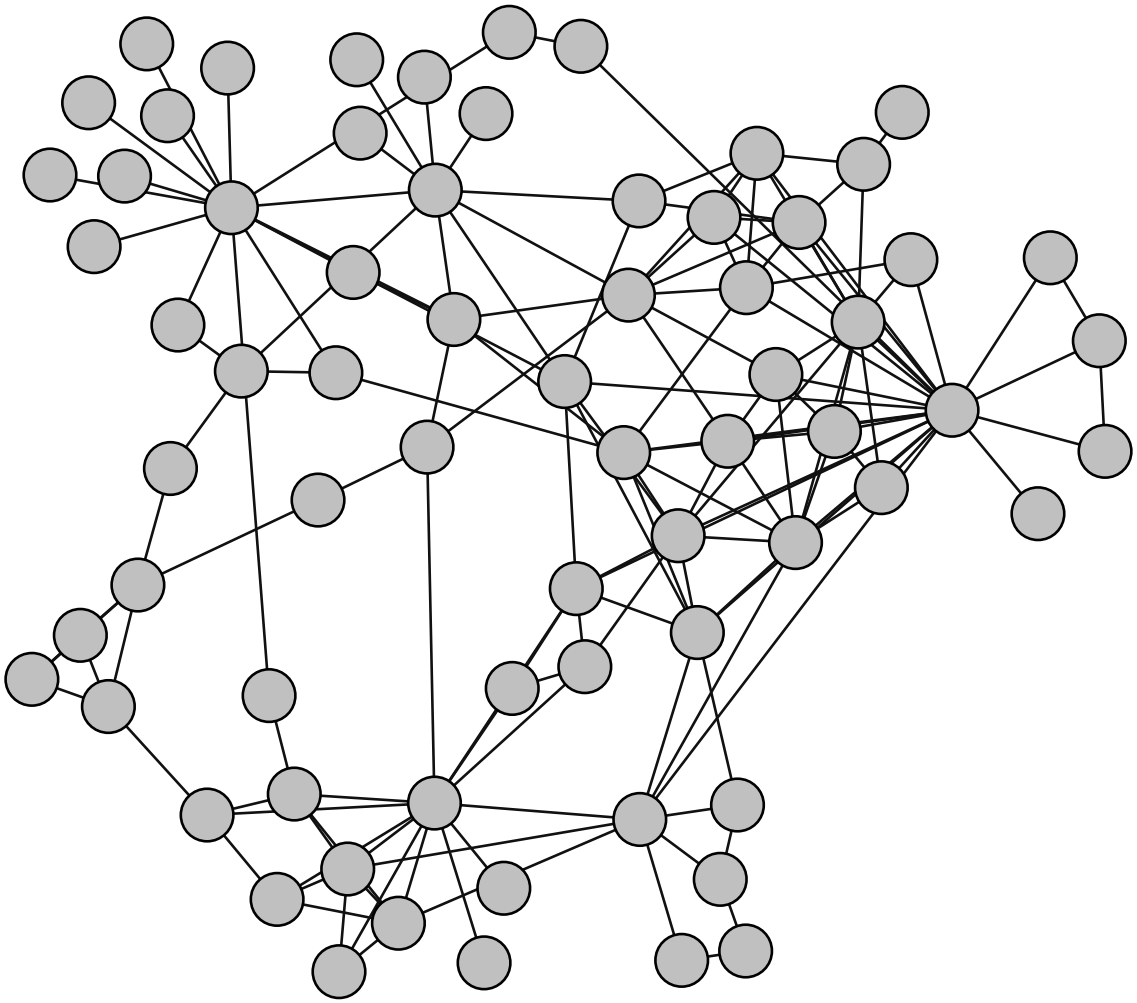
Meso scale: communities



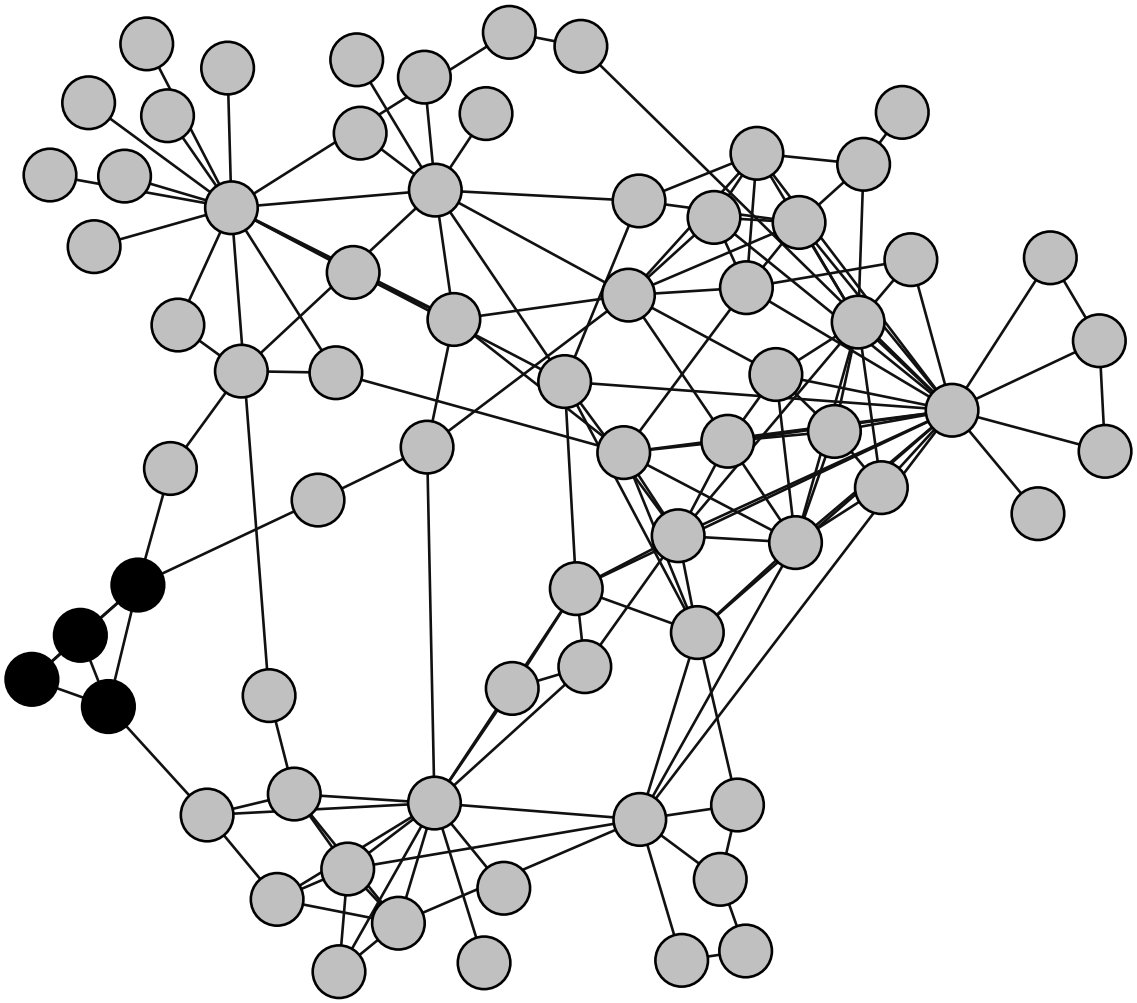
Meso scale: communities



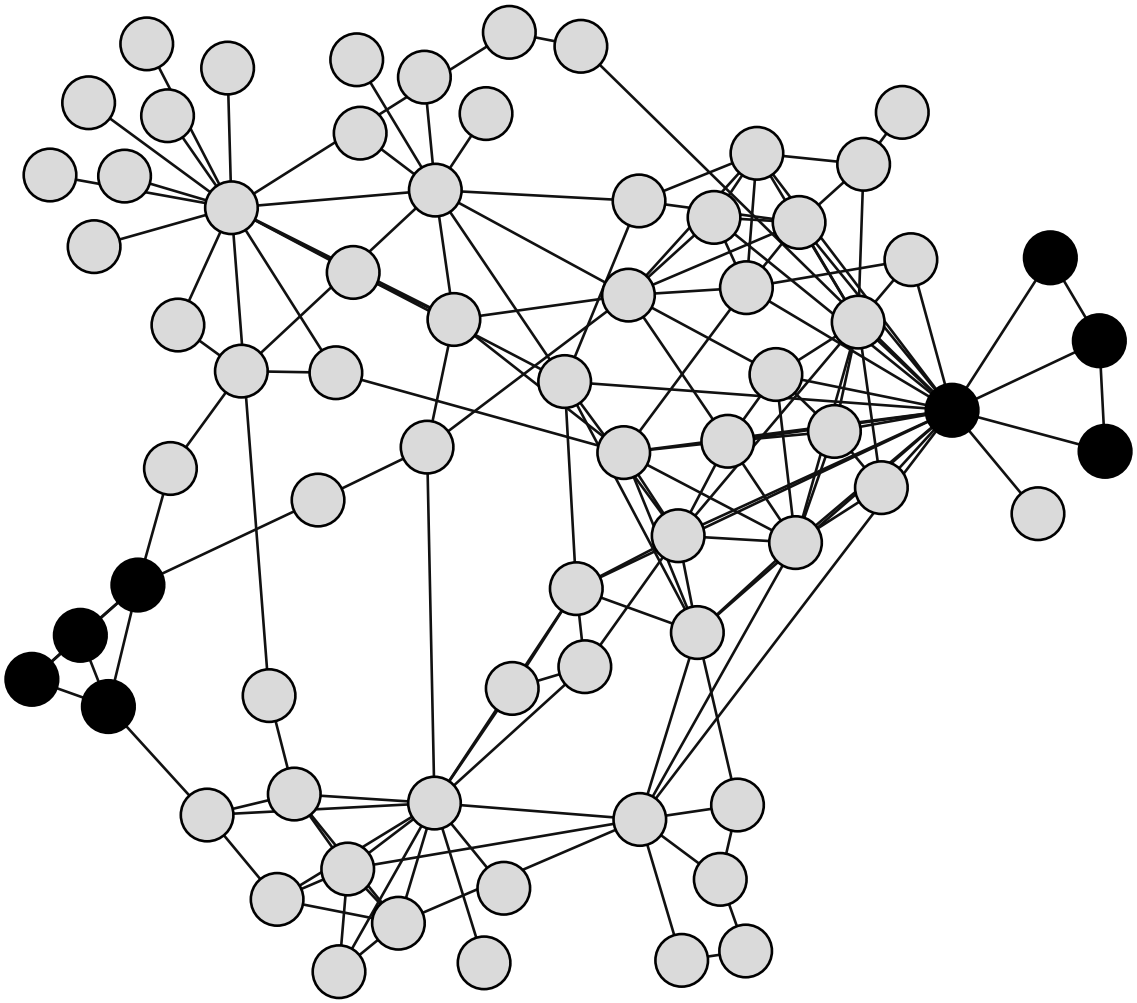
Meso scale: motifs



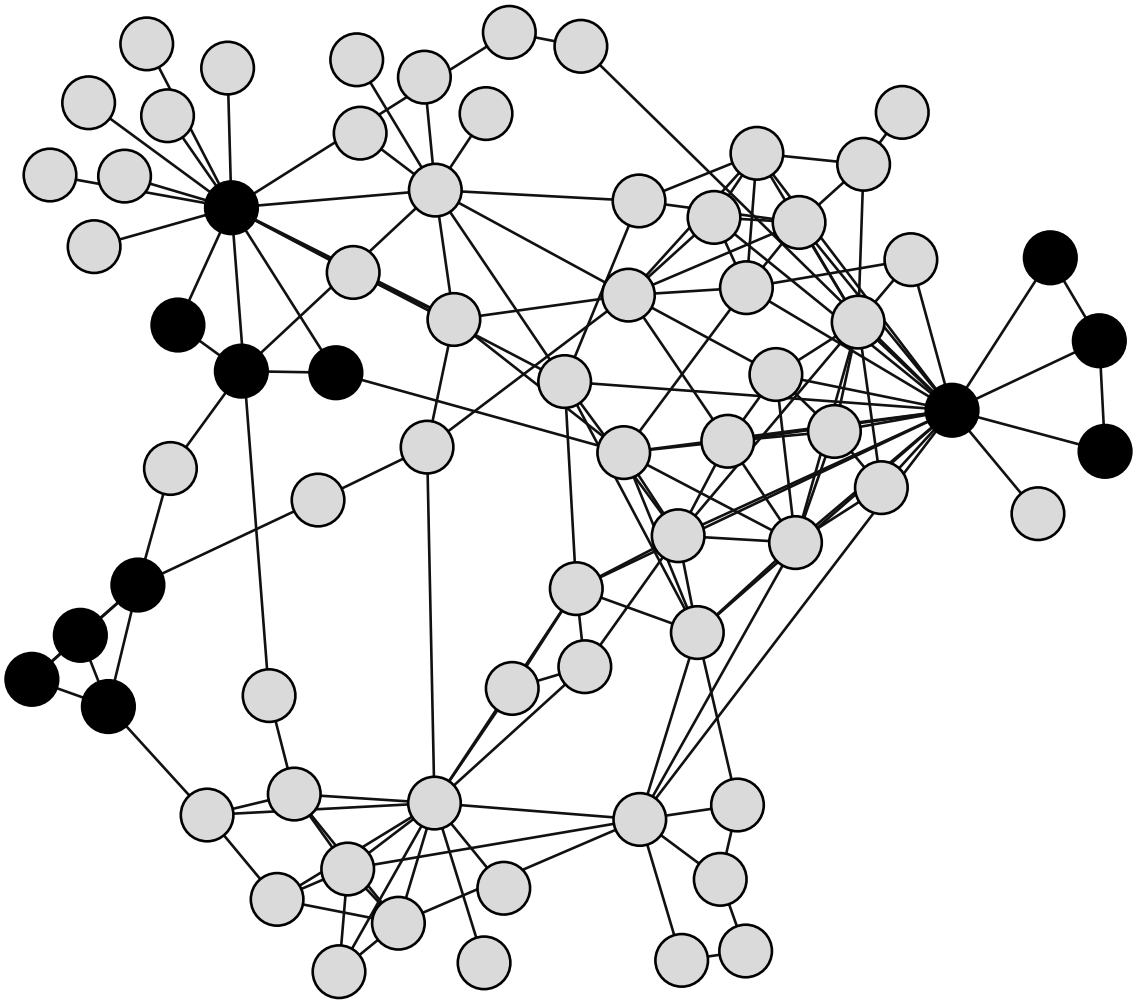
Meso scale: motifs

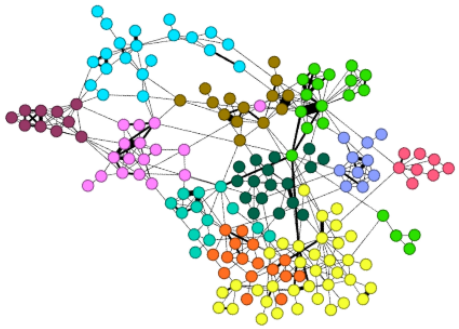


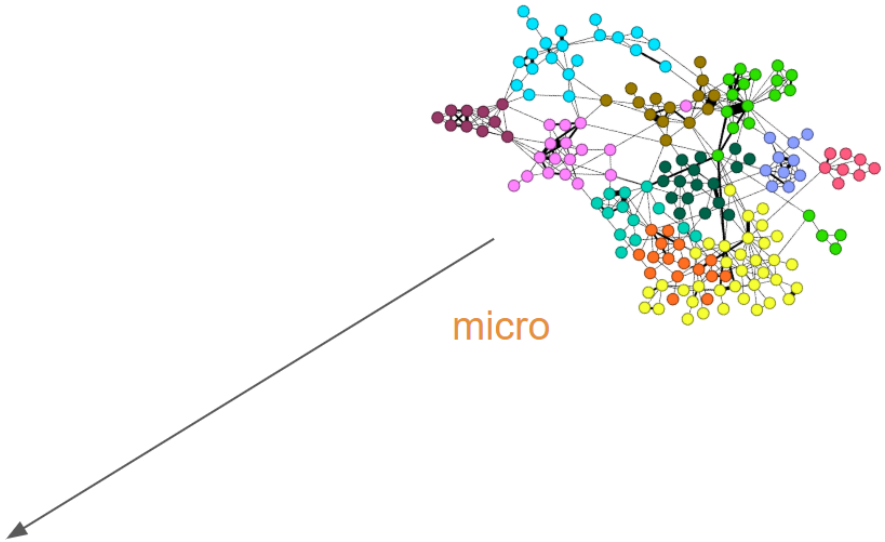
Meso scale: motifs

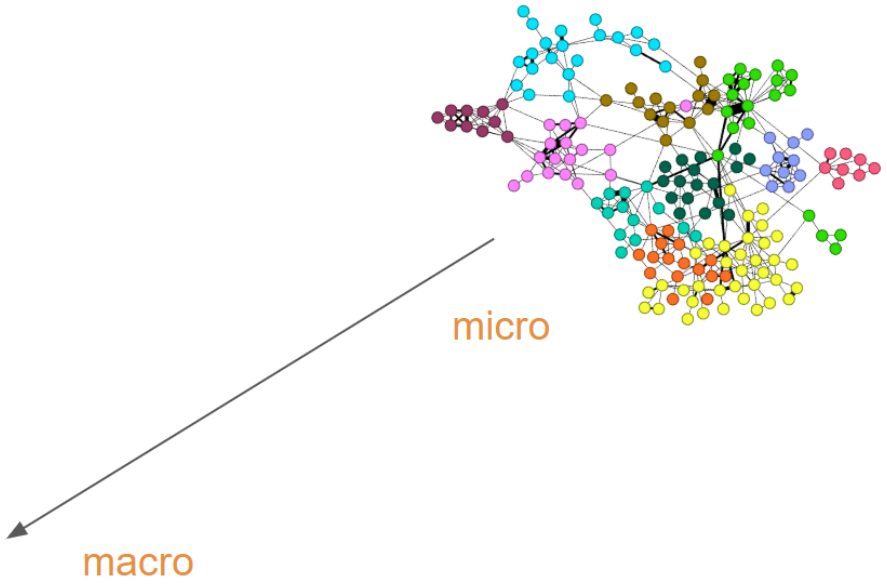


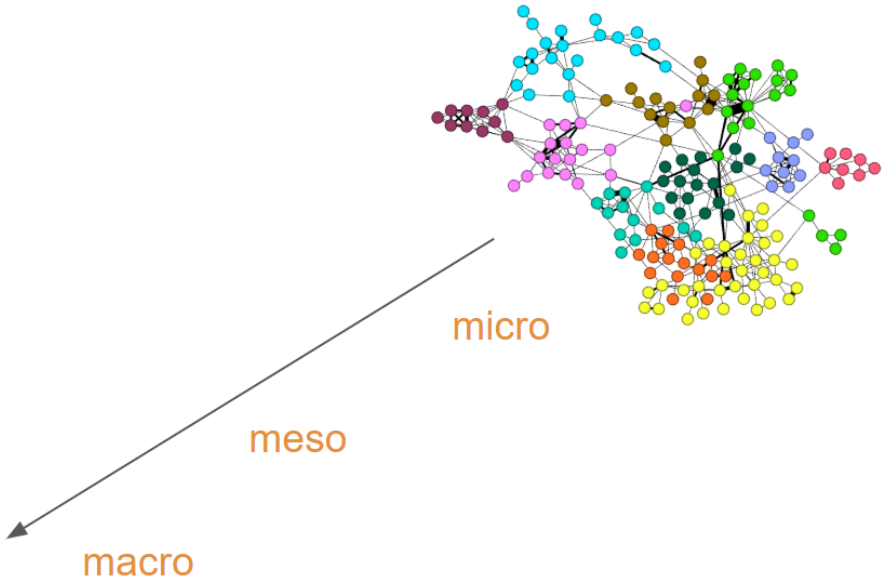
Meso scale: motifs

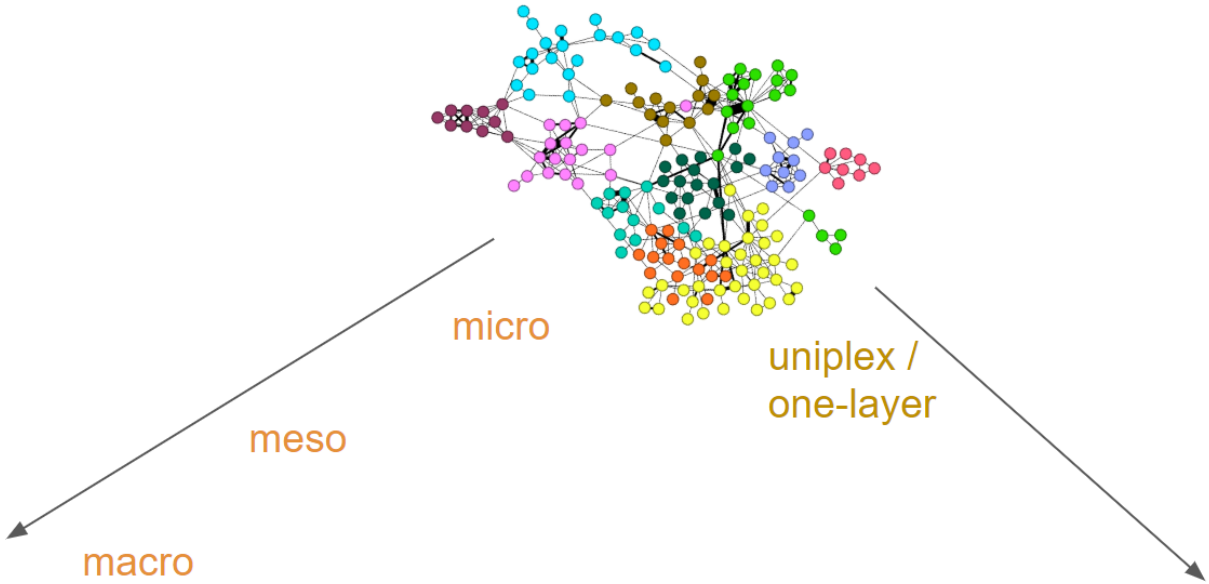


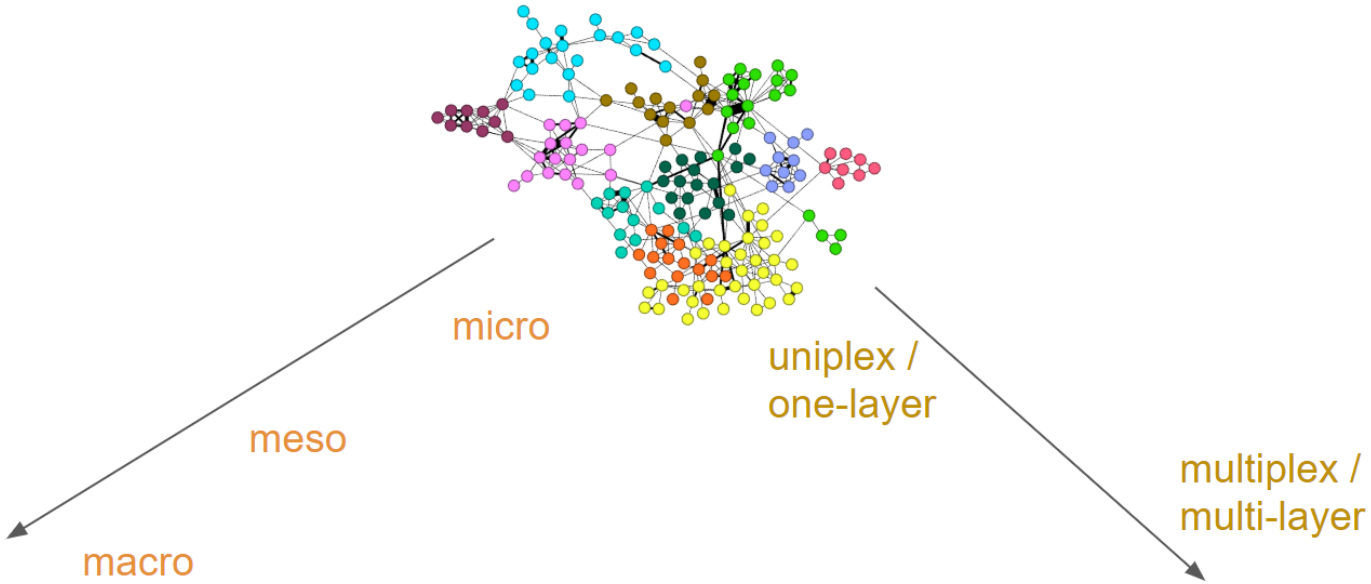


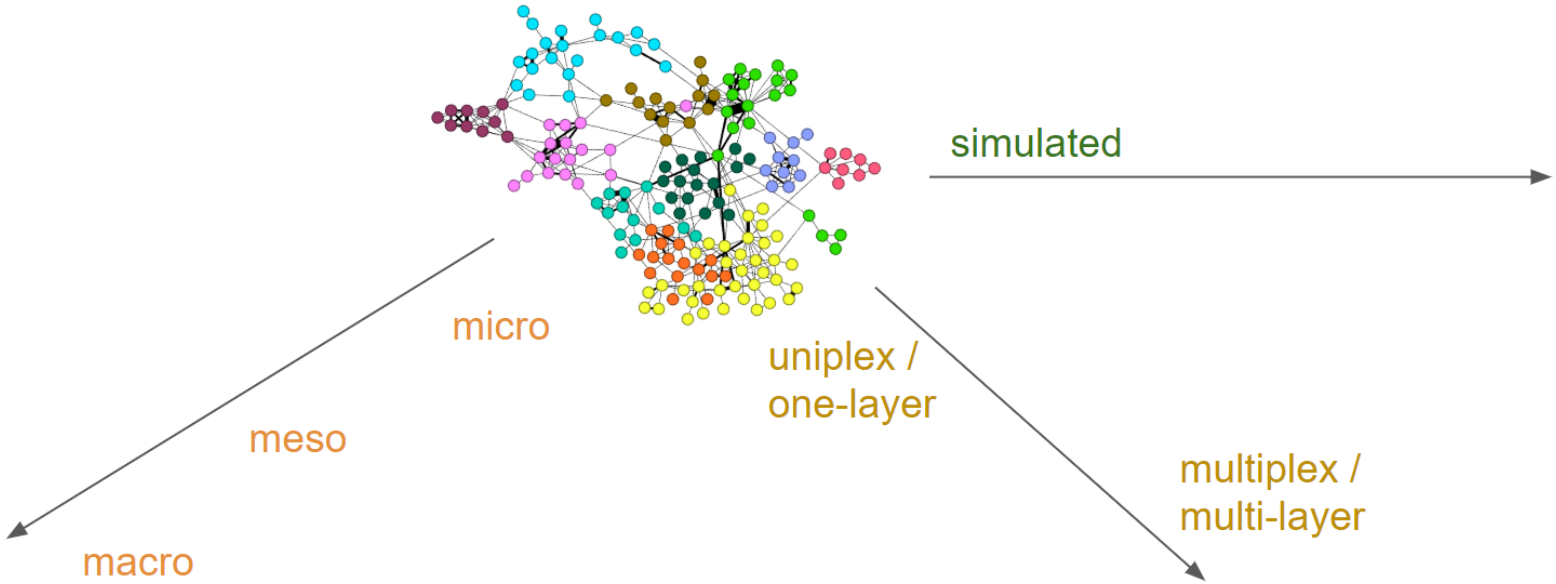

micro

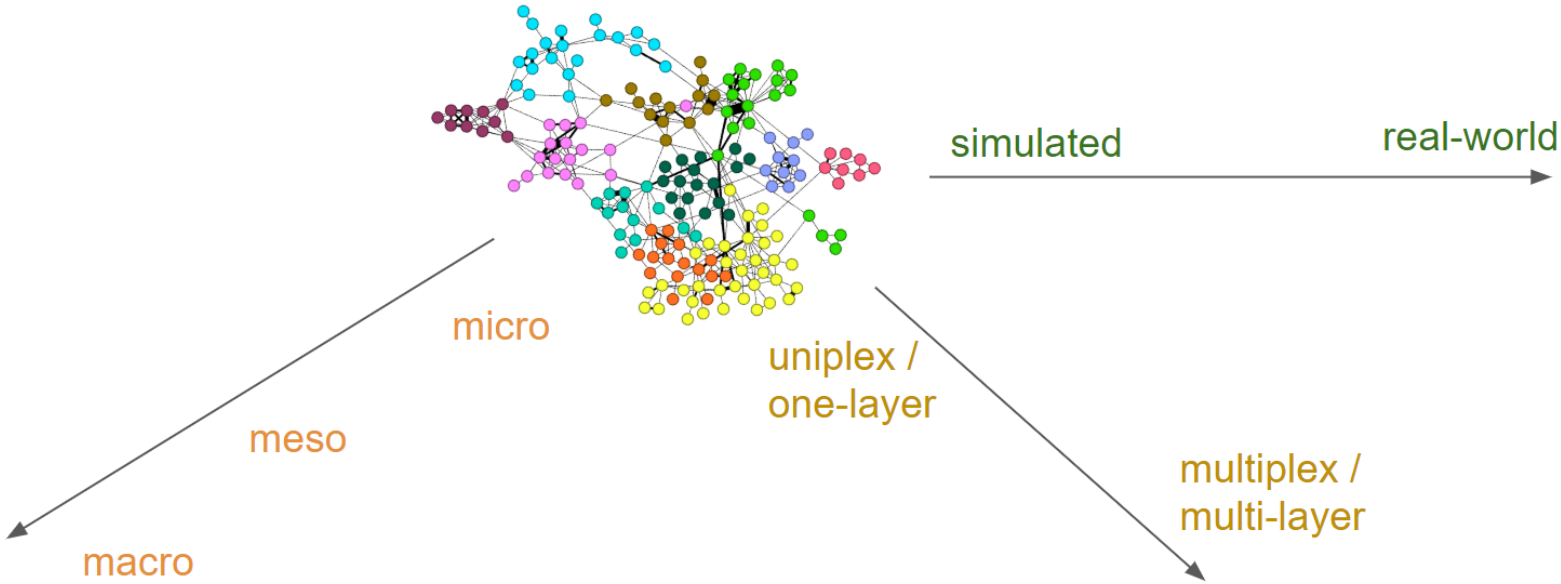

micro
macro

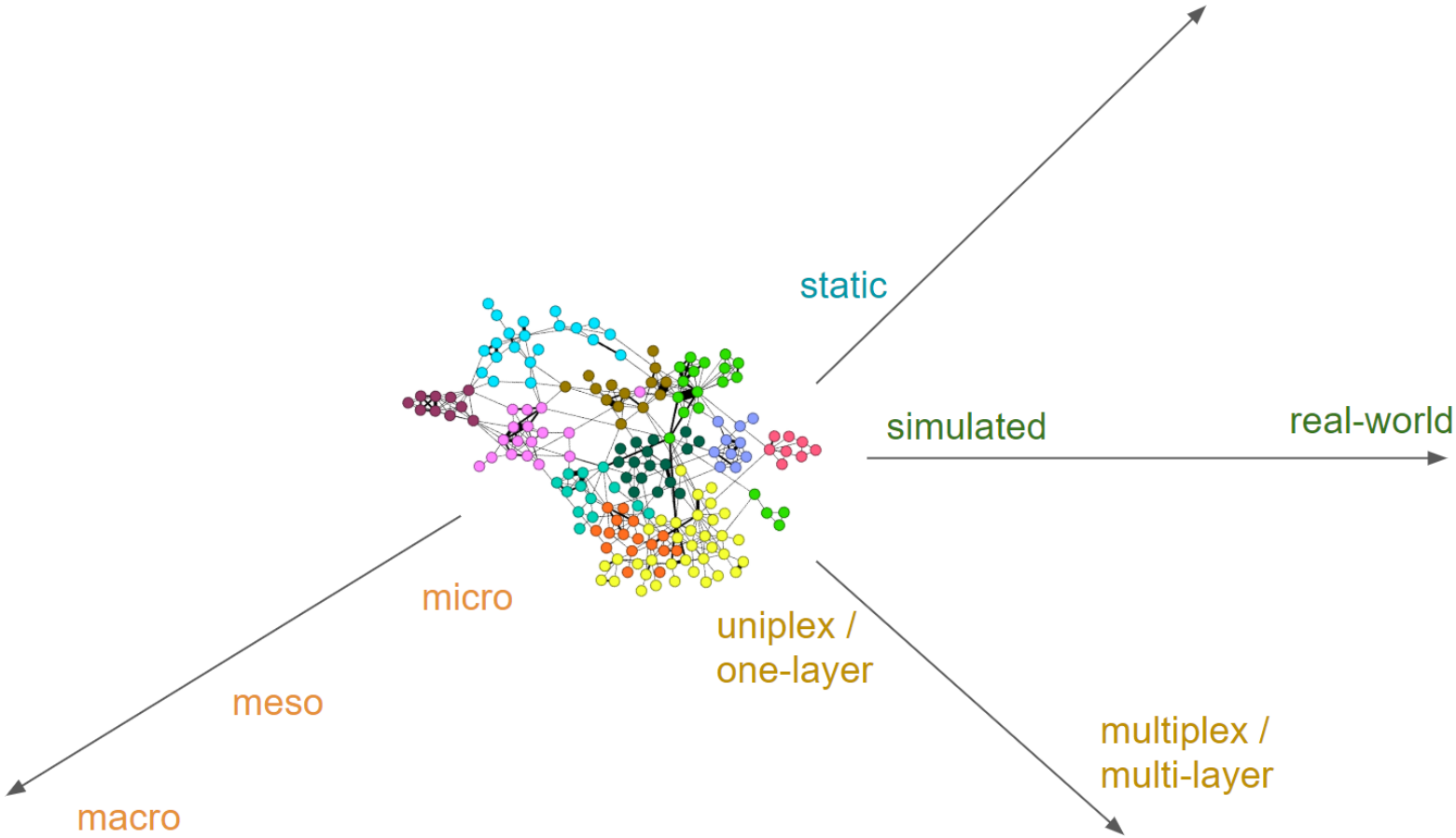

micro
meso
macro

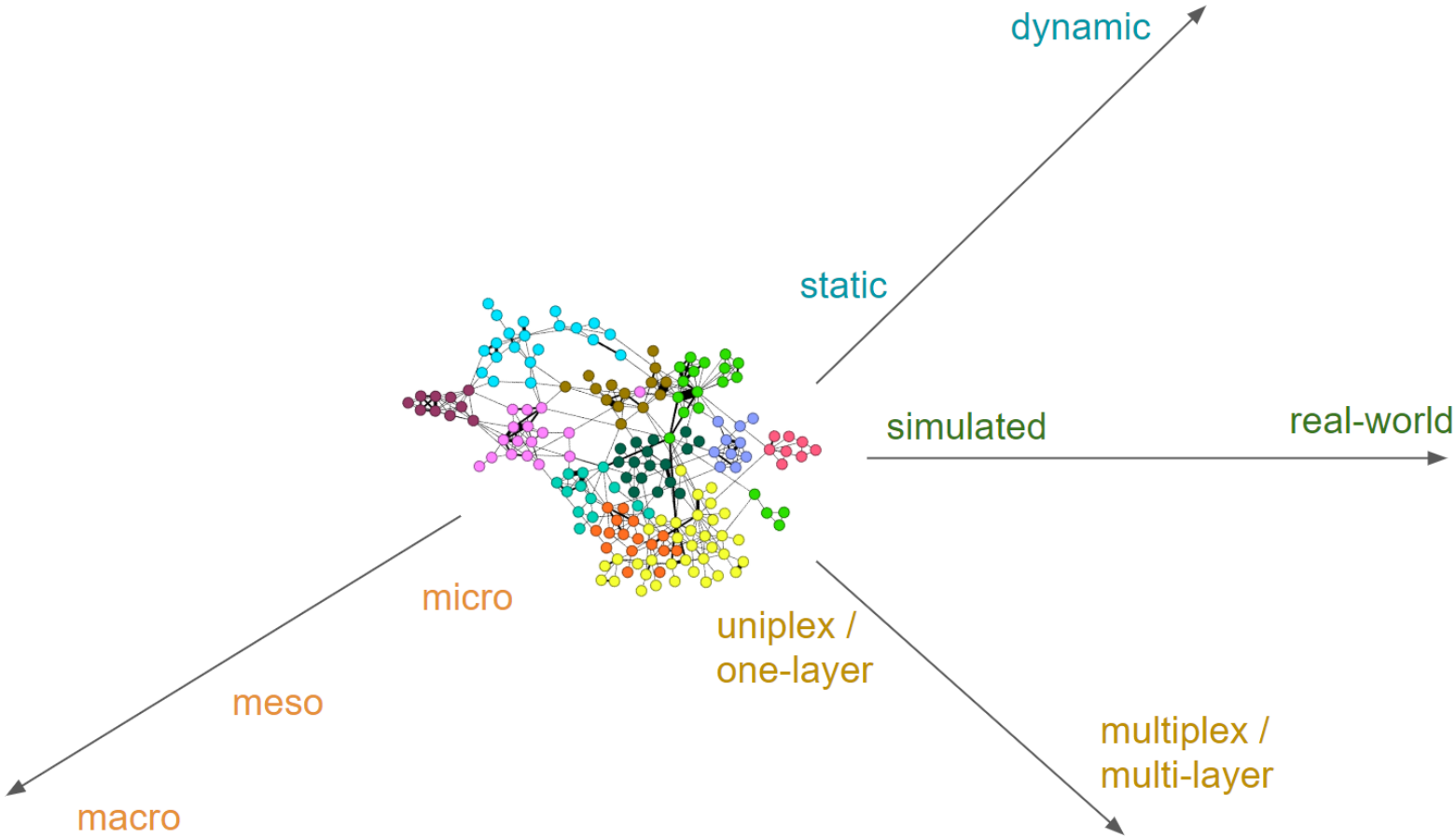

micro
uniplex /
one-layer
meso
macro

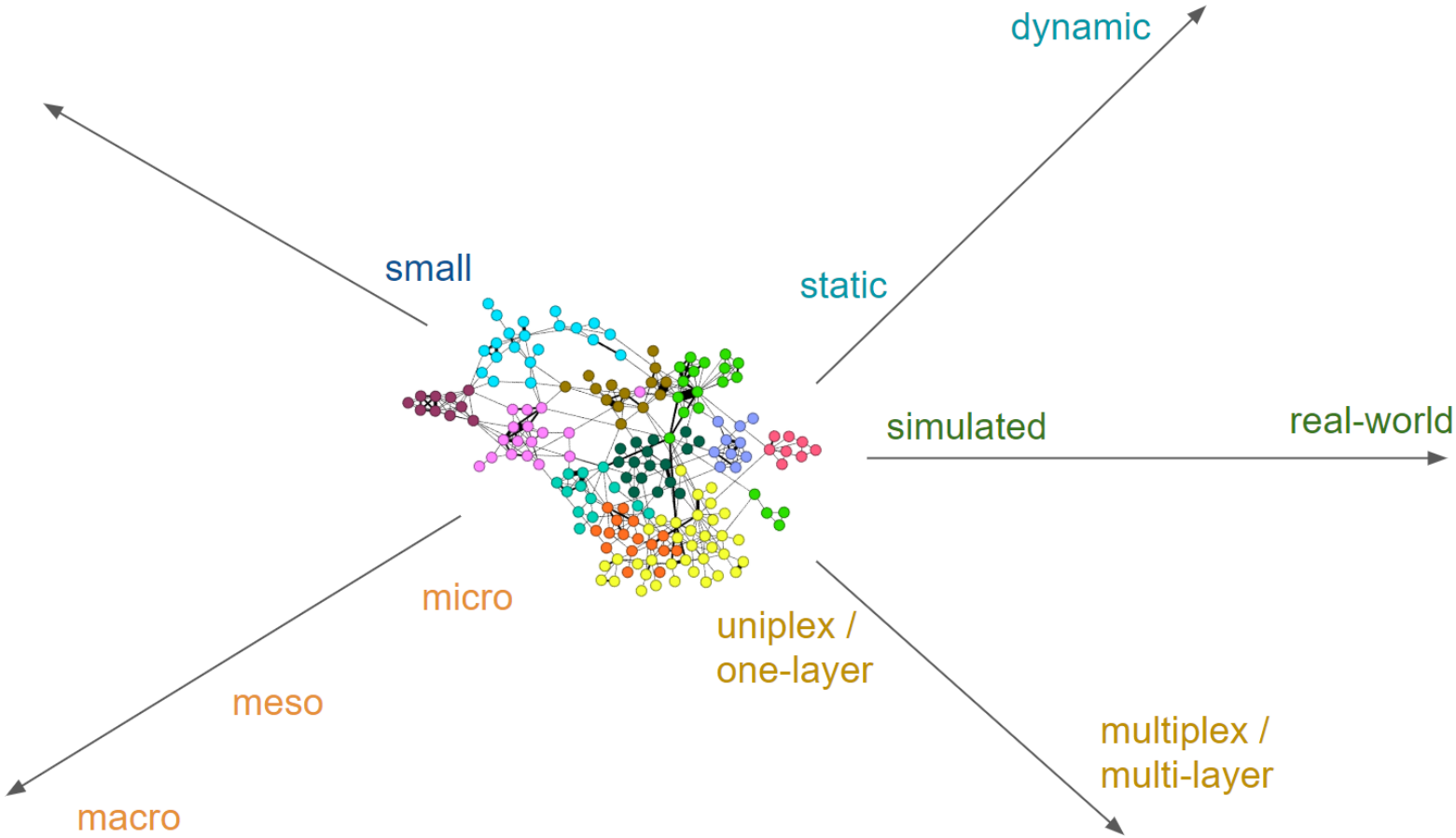

micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer

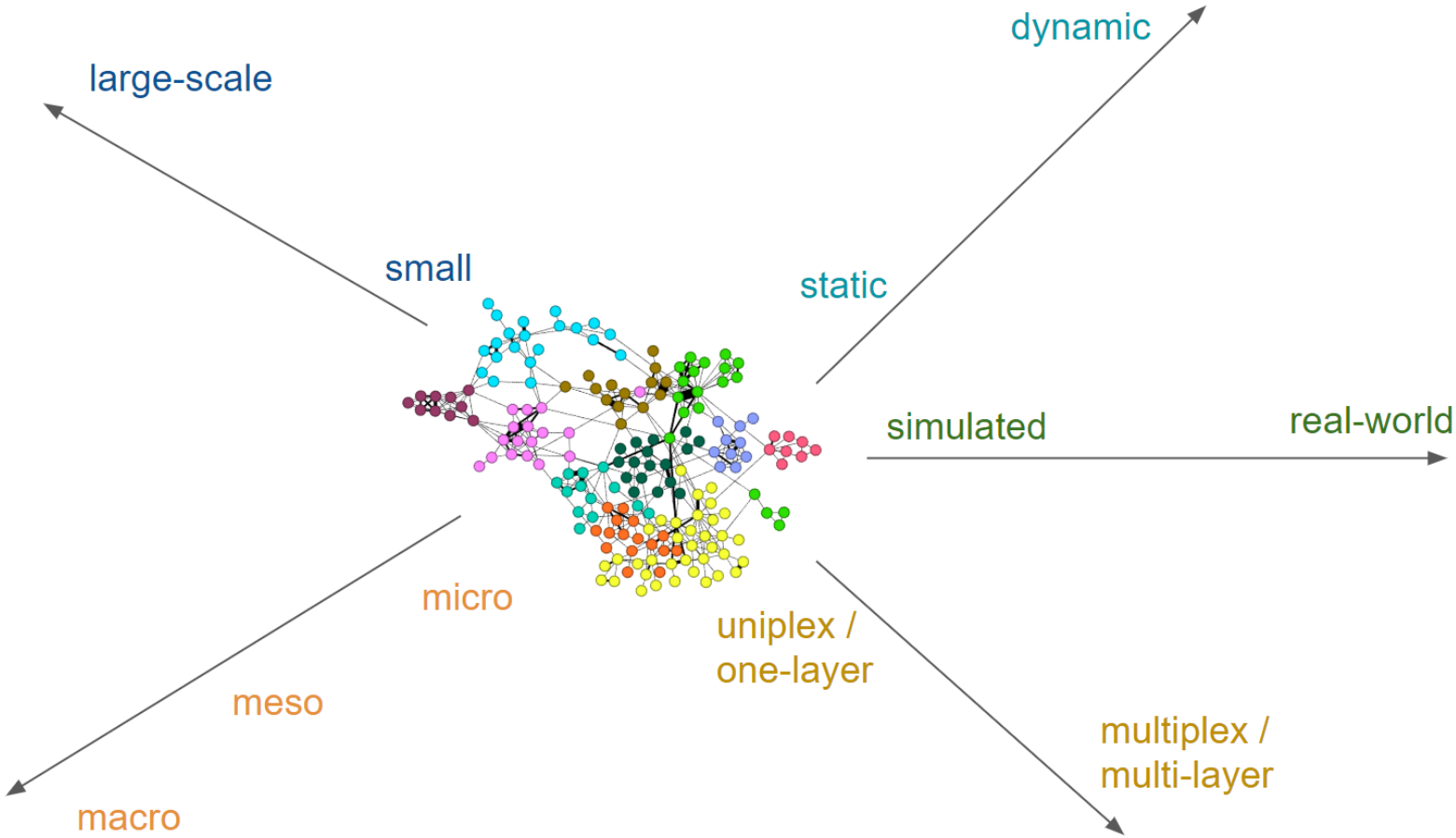

simulated
micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer

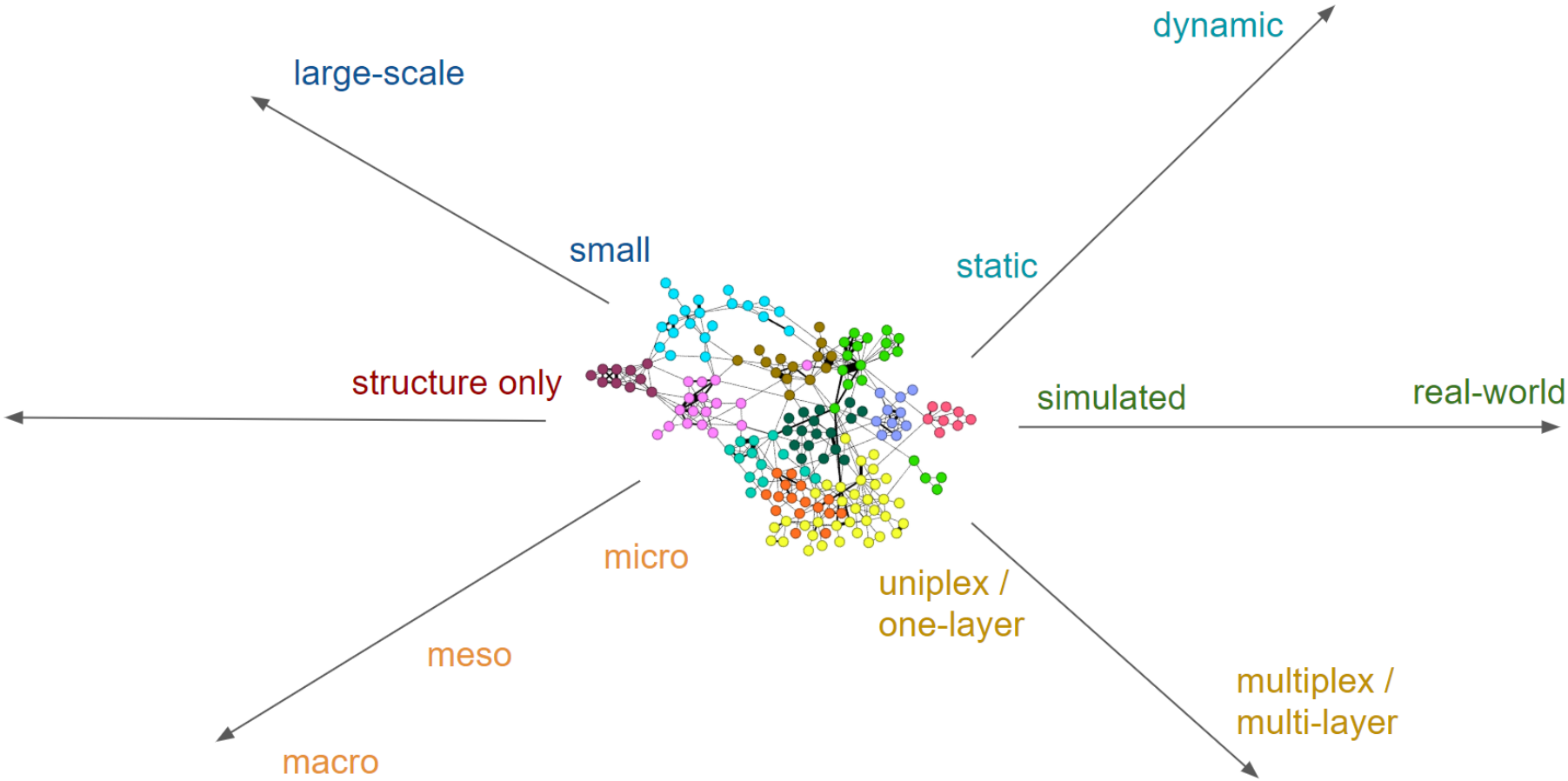

simulated
real-world
micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer

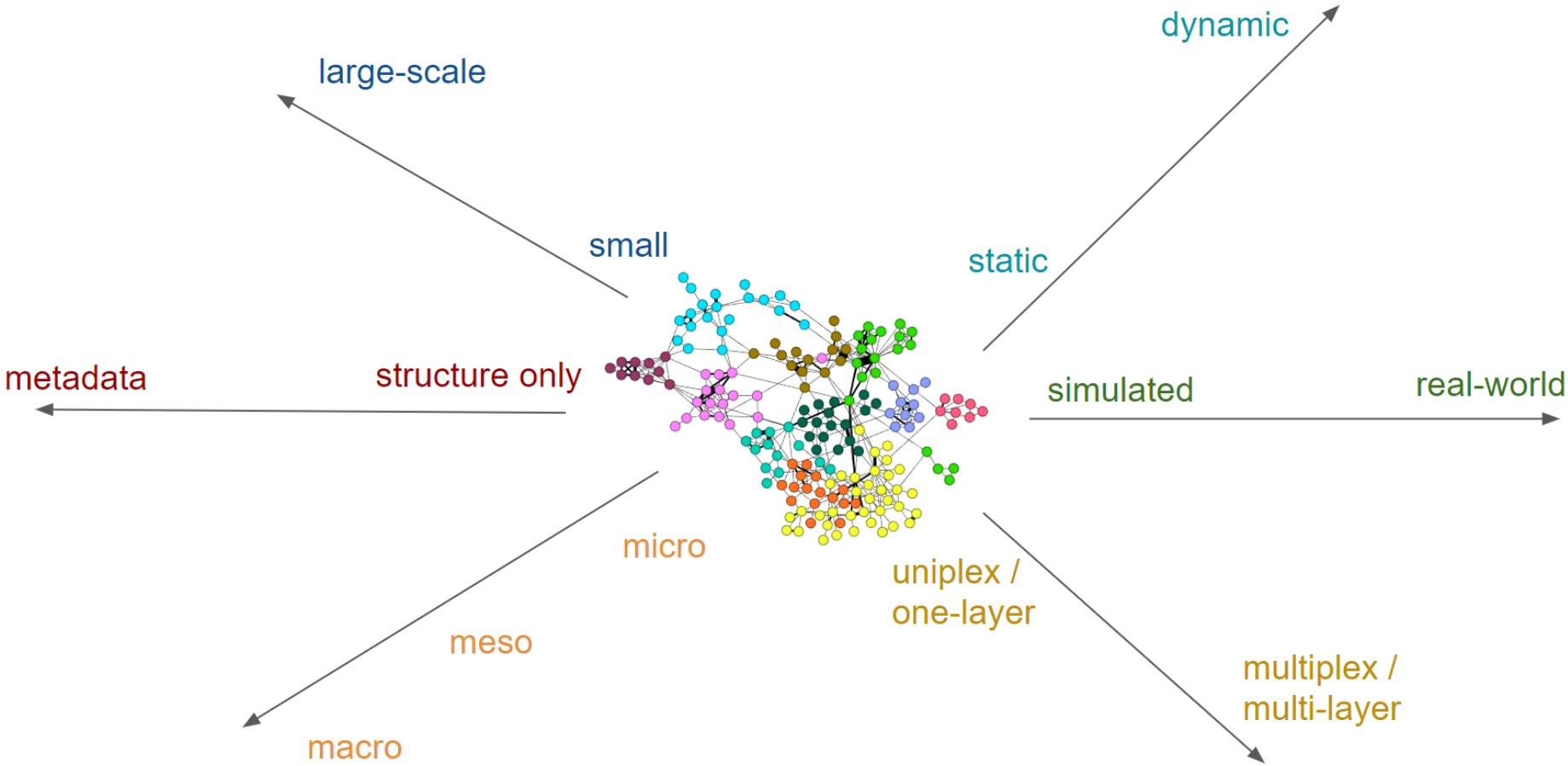

static
simulated
real-world
micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer


dynamic
static
simulated
real-world
micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer

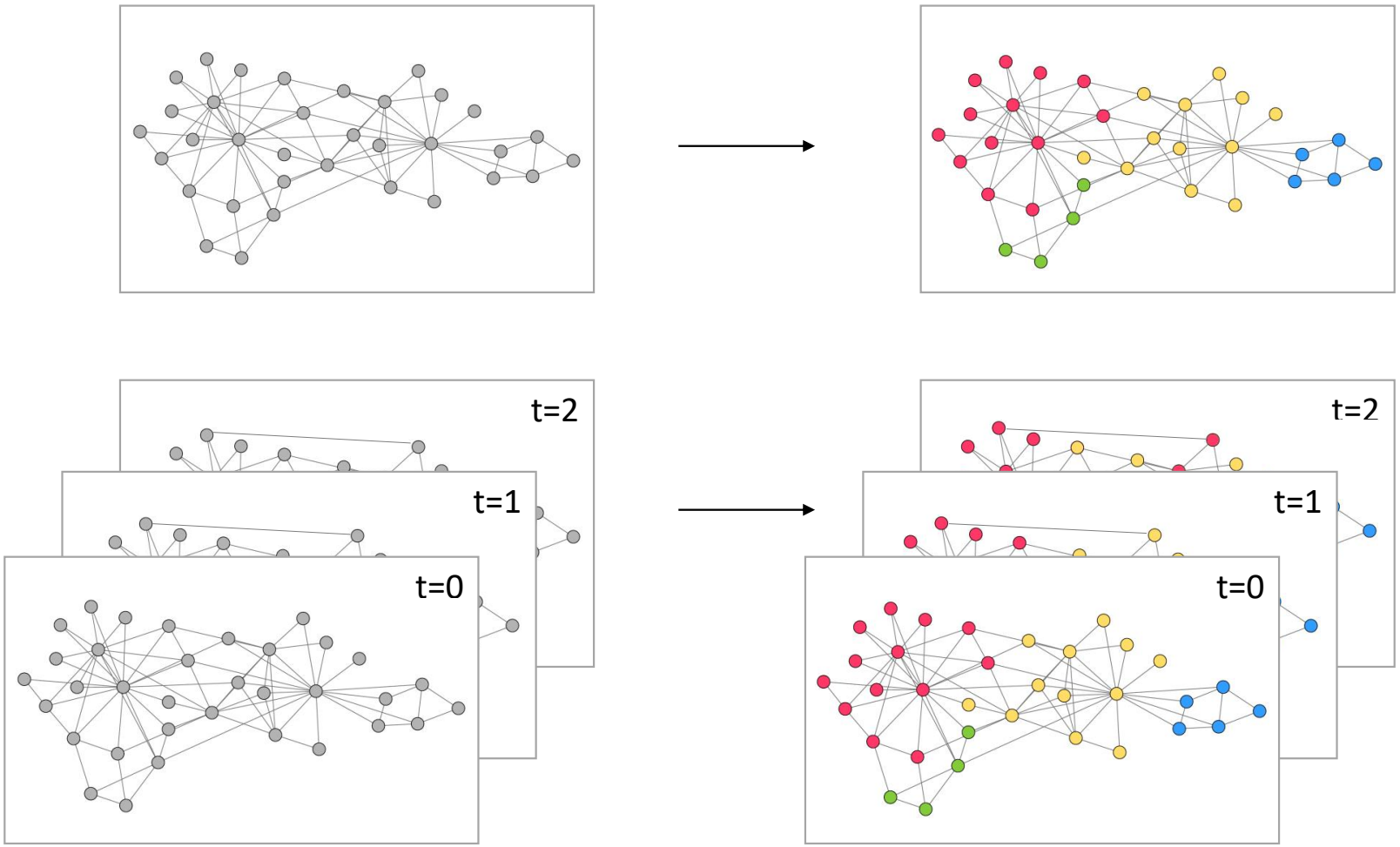

dynamic
small
static
simulated
real-world
micro
uniplex /
one-layer
meso
multiplex /
multi-layer
macro


large-scale
small
dynamic
static
simulated
real-world
micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer

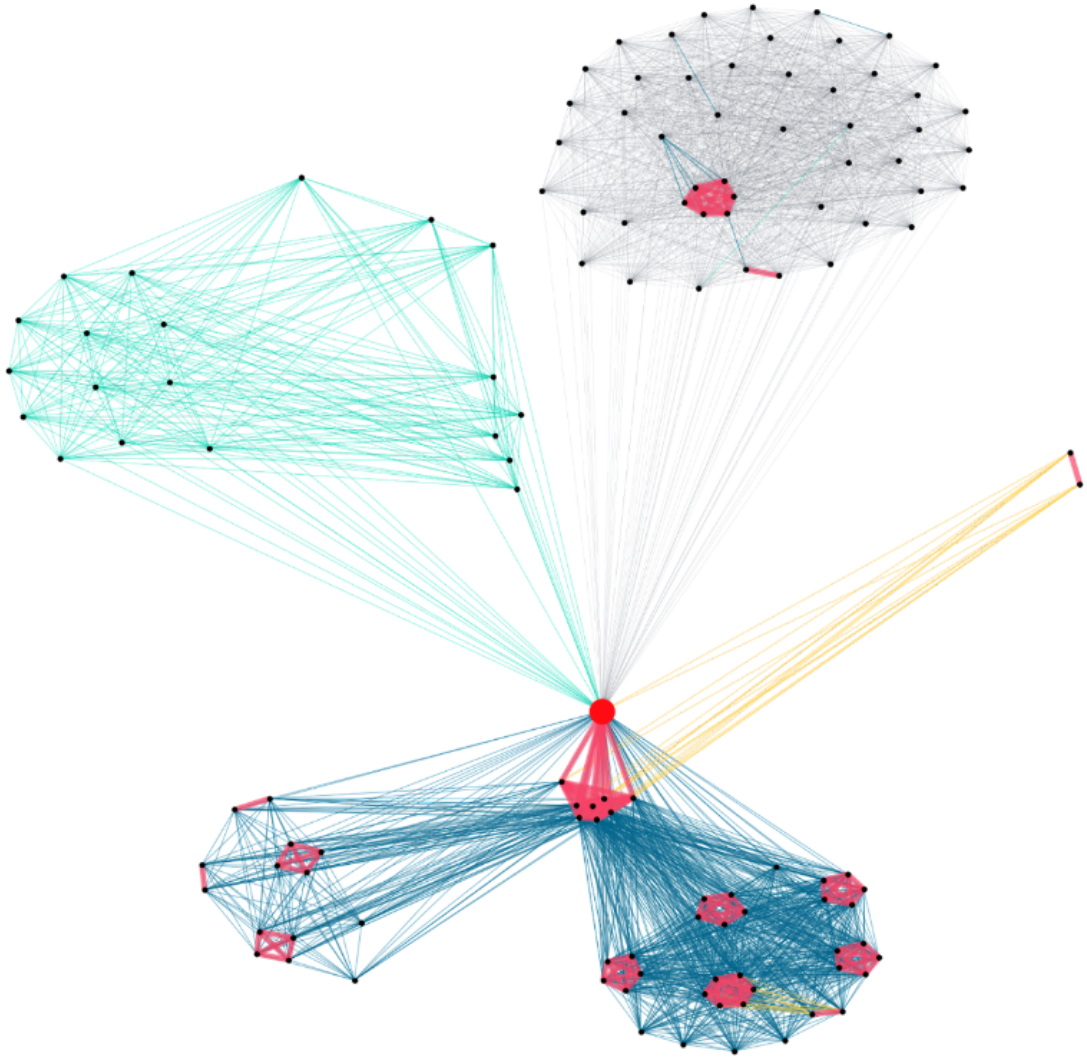

dynamic
large-scale
small
static
structure only
simulated
real-world
micro
uniplex /
one-layer
meso
multiplex /
multi-layer
macro


large-scale
dynamic
small
static
metadata
structure only
simulated
real-world
micro
uniplex /
one-layer
meso
multiplex /
multi-layer
macro

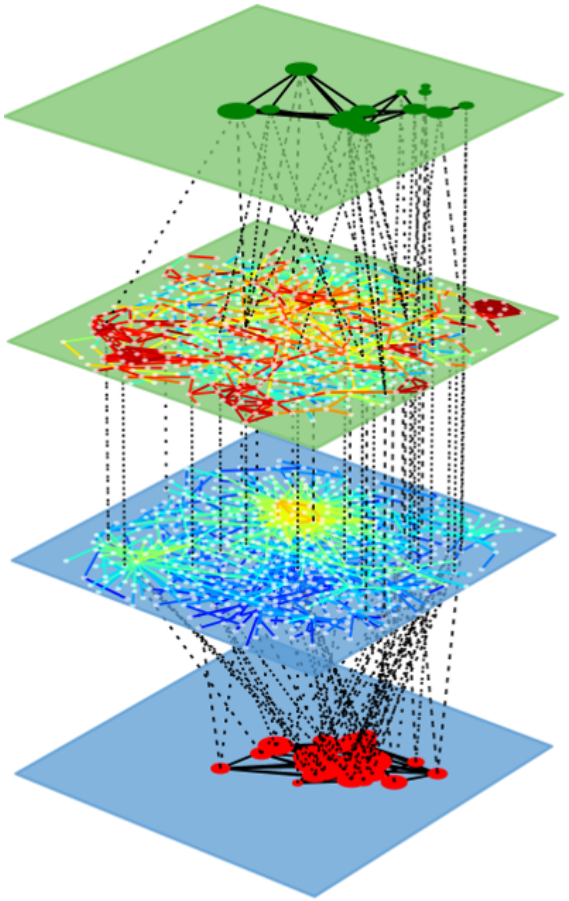
Network (community) dynamics



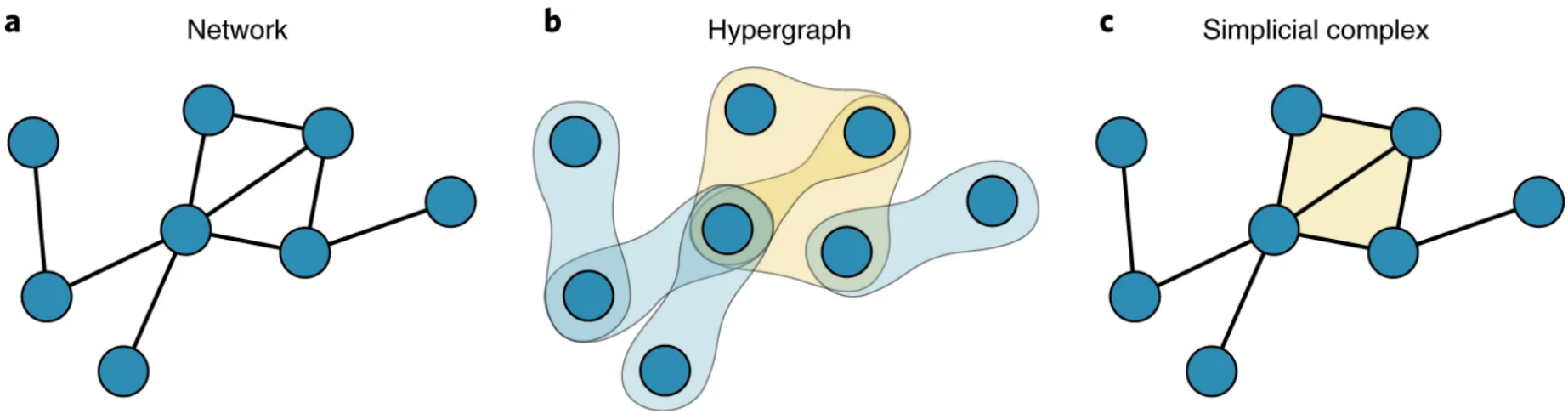
Multilayer networks



Multilevel networks



Higher-order networks / Simplicial complexes



Battiston et al. "The physics of higher-order interactions in complex systems." Nature Physics 17 (2021): 1093–1098.

Network flow

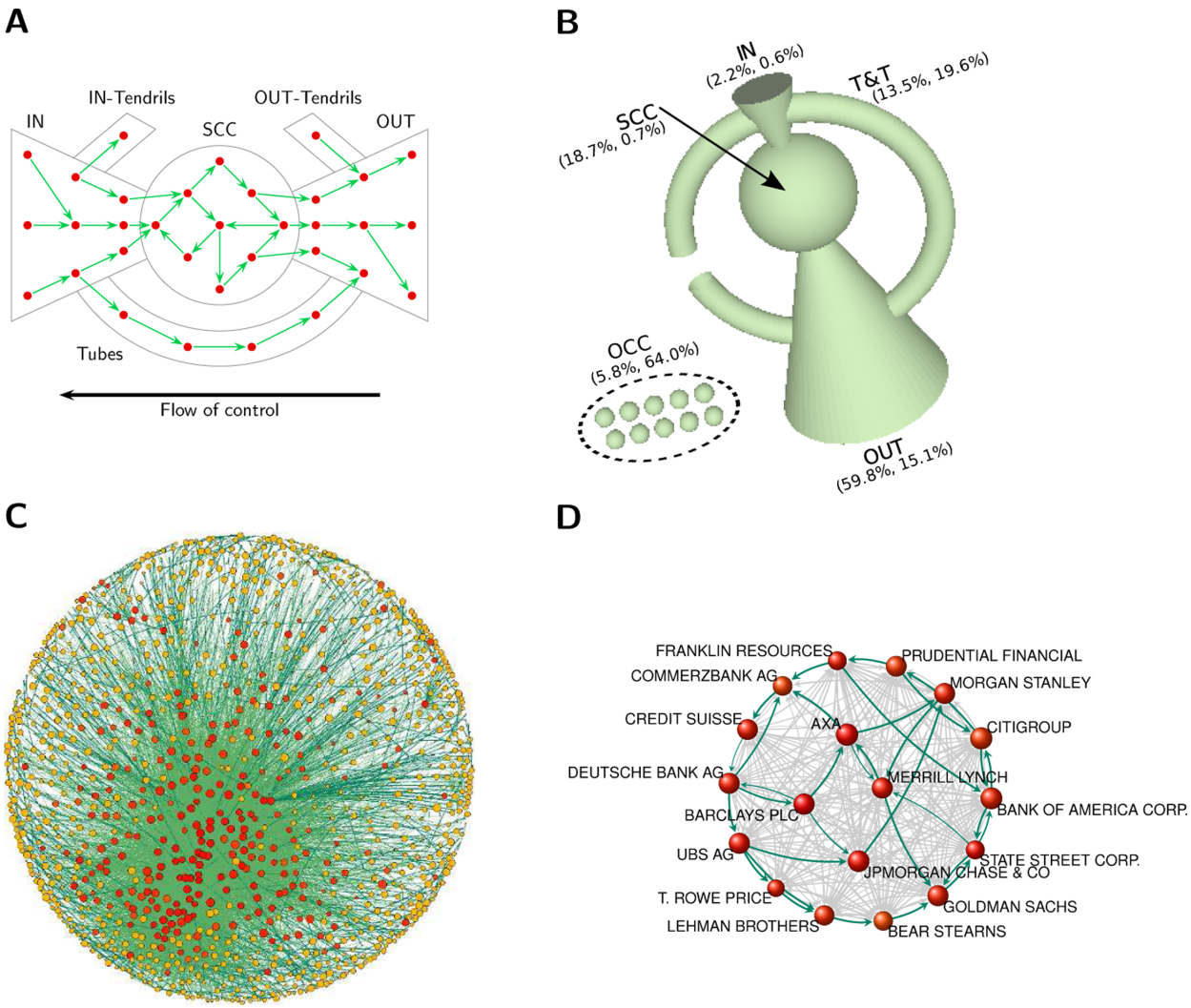
Uncovering Offshore Financial Centers in the Global Corporate Ownership Network

J. Garcia-Bernardo, J. Fichtner, F.W. Takes and E.M. Heemskerk, Uncovering Offshore Financial Centers: Conduits and Sinks in the Global Corporate Ownership Network, *Scientfic Reports* 7: 6246, Nature Publishing Group, 2017.

Corporate networks

- **Nodes** are organizations/firms/companies/corporations
- **Links** represent:
  - Trade
  - Loans
  - **Ownership**
  - Board interlocks

Ownership network



S. Vitali, J.B. Glattfelder and S. Battiston, S, The network of global corporate control, *PloS one* 6(10), e25995, 2011.

Offshore financial centers

- **Offshore Financial Center** (OFC): jurisdiction (country) that attracts financial activities from abroad through low taxation and lenient regulation.

Offshore financial centers

- **Offshore Financial Center** (OFC): jurisdiction (country) that attracts financial activities from abroad through low taxation and lenient regulation.
- OFCs "process" around $\$6\,000\,000\,000\,000$ yearly

Offshore financial centers

- **Offshore Financial Center** (OFC): jurisdiction (country) that attracts financial activities from abroad through low taxation and lenient regulation.
- OFCs "process" around $\$6\,000\,000\,000\,000$ yearly
- Problematic because:
  - Unnecessary system complexity
  - Accountability
  - Tax avoidance

Characterizing OFCs

- So, which countries are OFCs?

Characterizing OFCs

- So, which countries are OFCs?
  - Heavily debated
  - Definitions differ
  - Highly contested and politicized

Characterizing OFCs

- So, which countries are OFCs?
  - Heavily debated
  - Definitions differ
  - Highly contested and politicized
- FDI/GDP ratio approach:
  substantially more Foreign Direct Investment than expected based on GDP? $\rightarrow$ OFC

Characterizing OFCs

- So, which countries are OFCs?
  - Heavily debated
  - Definitions differ
  - Highly contested and politicized
- FDI/GDP ratio approach:
  substantially more Foreign Direct Investment than expected based on GDP? $\rightarrow$ OFC...
  Problematic because:
  - FDI systematically underreported and subject to politics
  - No exact money flows, just A to B
  - OFC homogeinity assumption

Characterizing OFCs

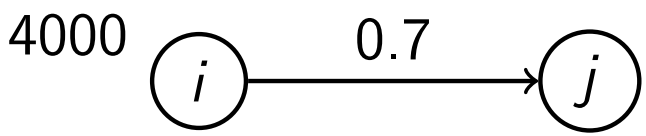
- So, which countries are OFCs?
  - Heavily debated
  - Definitions differ
  - Highly contested and politicized
- FDI/GDP ratio approach:
  substantially more Foreign Direct Investment than expected based on GDP? $\rightarrow$ OFC...
  Problematic because:
  - FDI systematically underreported and subject to politics
  - No exact money flows, just A to B
  - OFC homogeinity assumption
- Solution: **Networks!**

Data

- Orbis database
- Based on data at Chambers of Commerce
- 71,201,304 ownership links between firms
- Better coverage for high-income countries
- Poor data quality $\rightarrow$ fiscal secrecy
- Findings likely represent lower bound

Ownership network

- Ownership network $N = (V, E)$
- Nodes are firms $i \in V$
  - Have value associated to them R_i
  - Are based in a particular country $\phi(i)$
- Directed links indicate ownership relation / value flow $(i, j) \in E$
  - Have an associated weight in $[0; 1]$ $w(i, j)$
- $w(i, j) = 0.7$, $R_i = 4000$ gives:



Ownership paths and value

- Ownership path

$$p = (v_1, v_2, \ldots, v_\ell) \text{ with } (v_i, v_{i+1}) \in E \text{ for } 1 \leq i < \ell$$

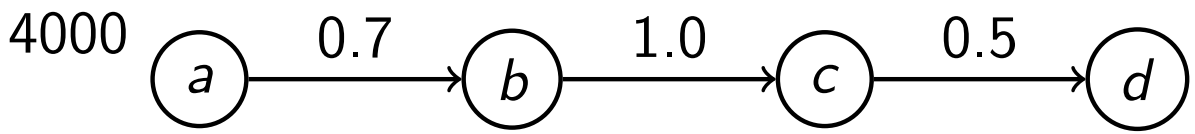
- Multiplicative ownership

$$w_p = \prod_{i=1}^{\ell-1} w(v_i, v_{i+1})$$

- Path value

$$V_p = R_{v_1} \cdot w_p$$

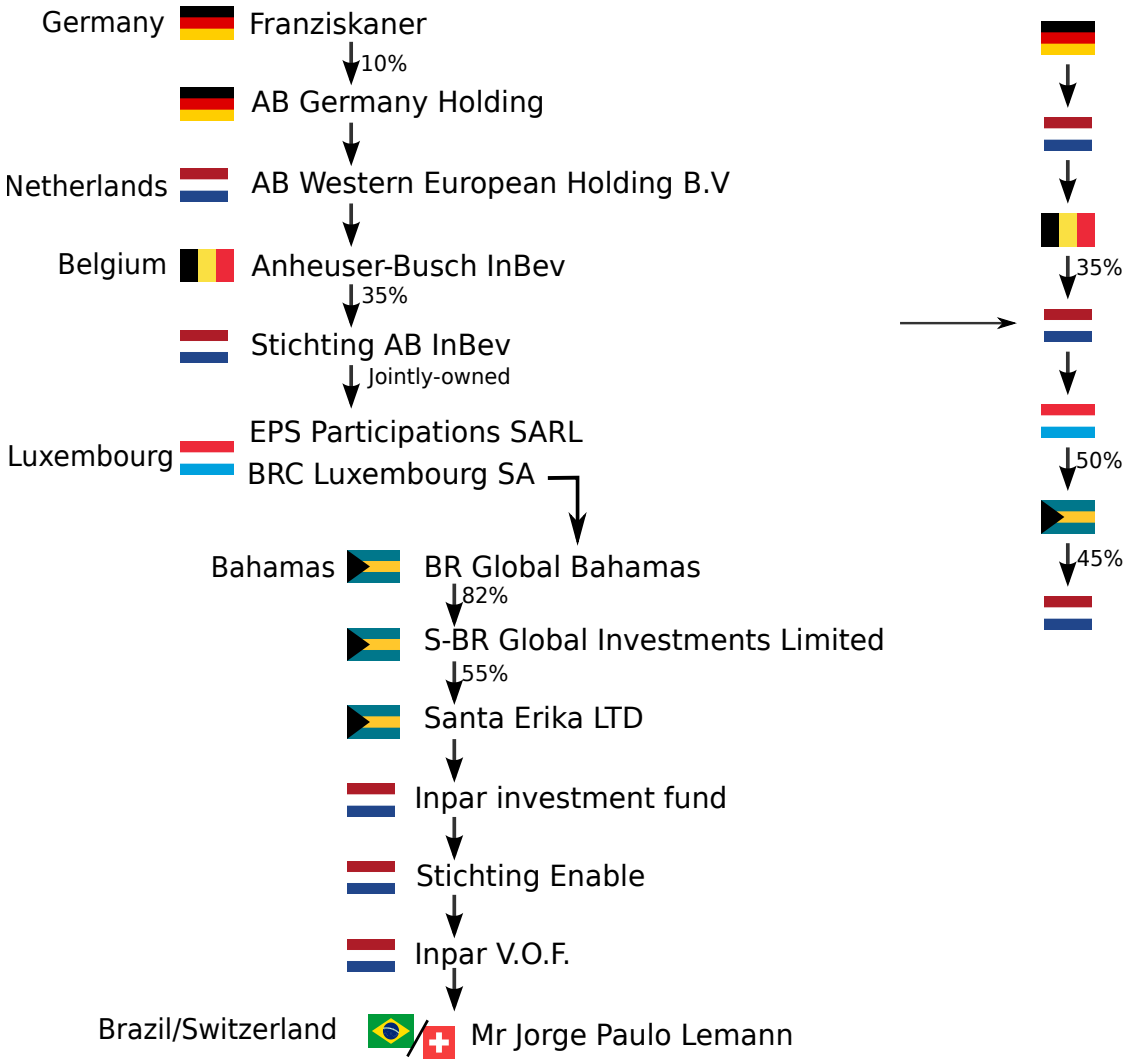
- $p = (a, b, c, d)$ and $w_p = 0.35$ and $V_p = 1400$



Ownership paths $\rightarrow$ OFCs

Ownership paths → OFCs

Ownership paths → OFCs



Ownership chains

- An **ownership chain** of a firm v is an ownership path p which:
  1. starts at node v,
  2. is a simple path (has no repeated nodes),
  3. has multiplicative ownership value greater than θ, i.e., $w_p \geq \theta$ and
  4. is maximal in length, i.e., cannot be extended by adding another node.

Ownership chains

- An **ownership chain** of a firm v is an ownership path p which:
  1. starts at node v,
  2. is a simple path (has no repeated nodes),
  3. has multiplicative ownership value greater than θ, i.e., $w_p \geq \theta$ and
  4. is maximal in length, i.e., cannot be extended by adding another node.
- Set of all ownership chains starting at v is denoted C_v
- Enumeration is possible using a DFS search for each node

Ownership chains

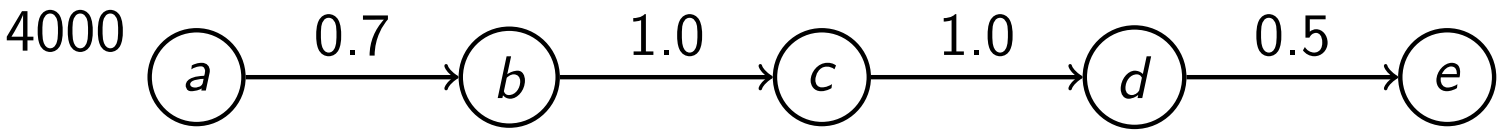
- An **ownership chain** of a firm v is an ownership path p which:
  1. starts at node v,
  2. is a simple path (has no repeated nodes),
  3. has multiplicative ownership value greater than θ, i.e., $w_p \geq \theta$ and
  4. is maximal in length, i.e., cannot be extended by adding another node.
- Set of all ownership chains starting at v is denoted C_v
- Enumeration is possible using a DFS search for each node
- Ultimately, $C = \bigcup_{v \in V} C_v$ is the set of all ownership chains

Ownership chunks

- **Ownership chunk**: subpath of length $2, 3, \ldots, \ell$ (generated from a chain of length ℓ)
- From all chains C, generate all ownership chunks, forming set H
- Ownership chunks of length x: H^x.
- Chunk $q \in H$ has an associated value $V^p(q)$ or in short V^p_q.

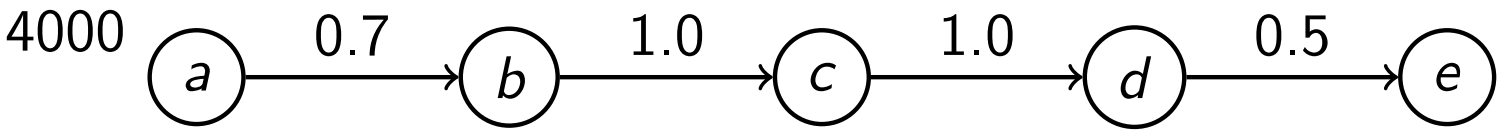
Ownership chunks

- **Ownership chunk**: subpath of length $2, 3, \ldots, \ell$ (generated from a chain of length ℓ)
- From all chains C, generate all ownership chunks, forming set H
- Ownership chunks of length x: H^x.
- Chunk $q \in H$ has an associated value $V^p(q)$ or in short V^p_q.



Ownership chunks

- **Ownership chunk**: subpath of length $2, 3, \ldots, \ell$ (generated from a chain of length ℓ)
- From all chains C, generate all ownership chunks, forming set H
- Ownership chunks of length x: H^x.
- Chunk $q \in H$ has an associated value $V^p(q)$ or in short V^p_q.



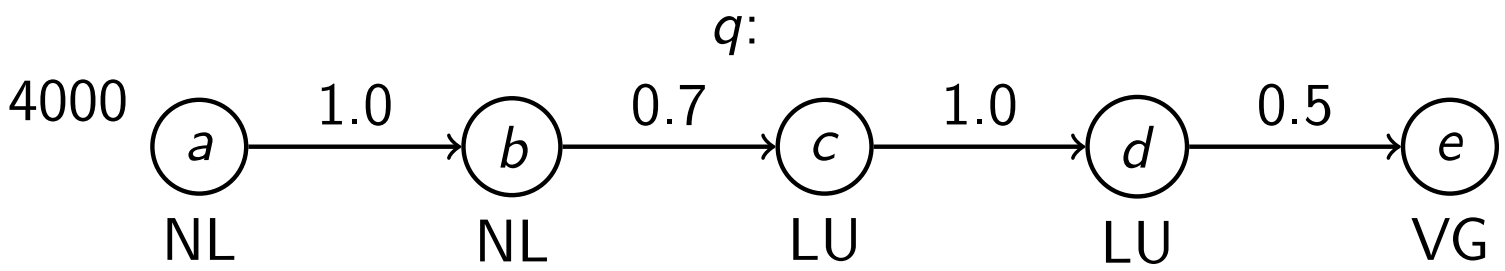
- Chunks of length 3 of $p = (a, b, c, d, e)$:
  - $q_1 = (a, b, c)$ with $V^p_{q_1} = 2800$
  - $q_2 = (b, c, d)$ with $V^p_{q_2} = 2800$
  - $q_2 = (c, d, e)$ with $V^p_{q_3} = 1400$

Country chains

- Map each node v in a chunk q to its country $\phi(v)$
- Merge subsequent same country nodes into **country chain** g
- New valuation function $V^{\phi}(g)$ for country chains, aggregating value through chunks with different value origins to the country level

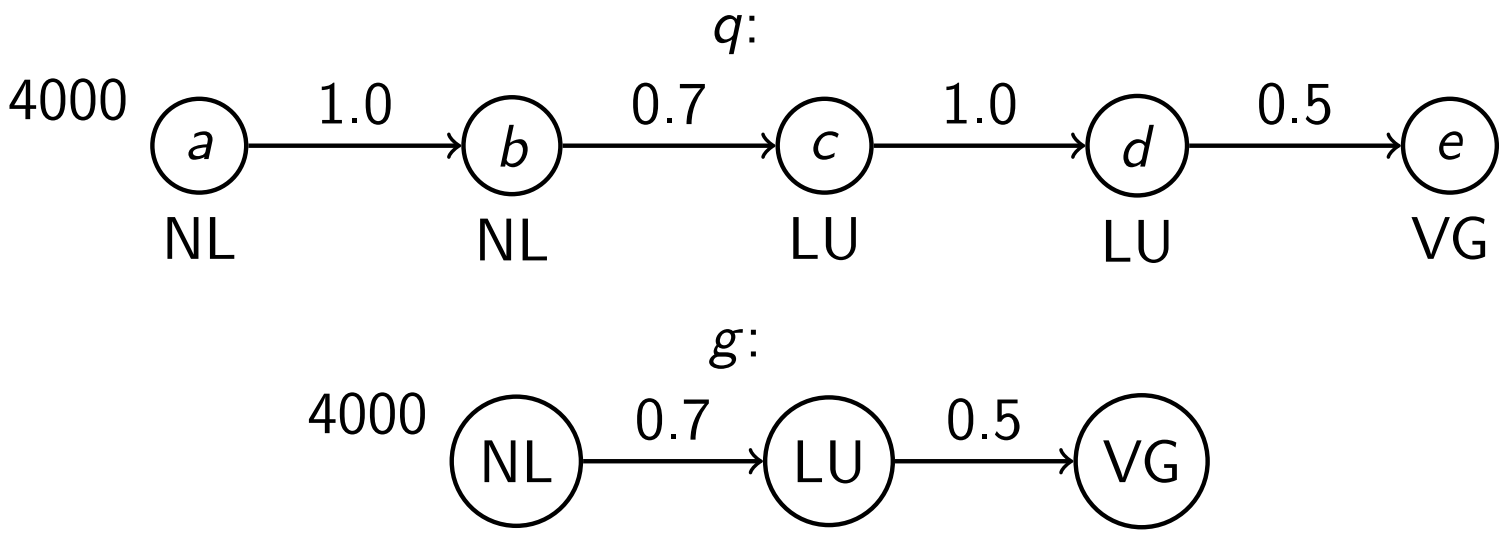
Country chains

- Map each node v in a chunk q to its country $\phi(v)$
- Merge subsequent same country nodes into **country chain** g
- New valuation function $V^{\phi}(g)$ for country chains, aggregating value through chunks with different value origins to the country level

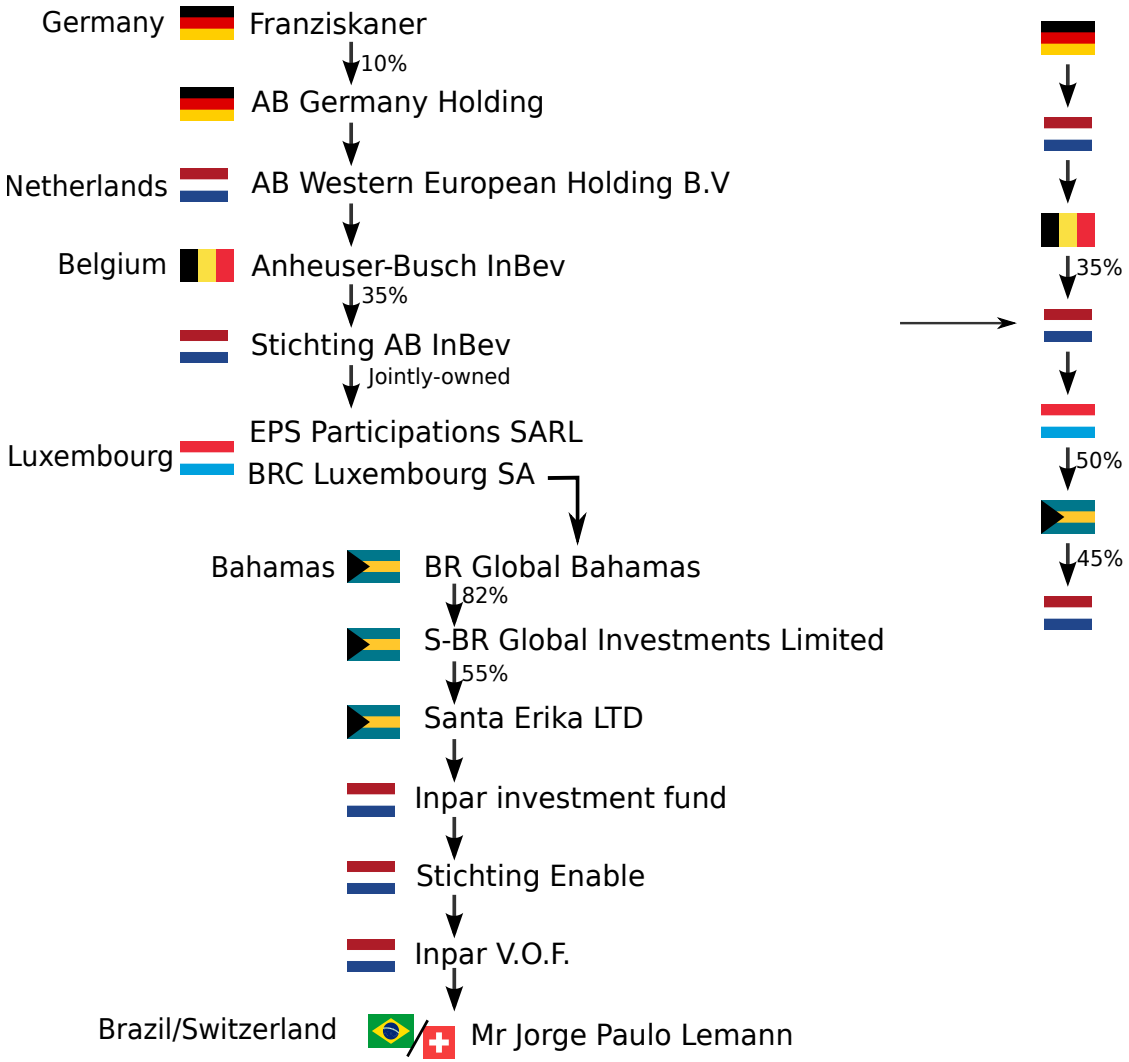


Country chains

- Map each node v in a chunk q to its country $\phi(v)$
- Merge subsequent same country nodes into **country chain** g
- New valuation function $V^{\phi}(g)$ for country chains, aggregating value through chunks with different value origins to the country level



Country chains



Results on our data

- 71,201,304 ownership links between firms
- 11,404,819 transnational ownership chains
  - threshold $\theta = 10^{-3}$
  - chain value > 1
- 108,159,506 chunks
- 377,098 country chains

Results on our data

- 71,201,304 ownership links between firms
- 11,404,819 transnational ownership chains
  - threshold $\theta = 10^{-3}$
  - chain value > 1
- 108,159,506 chunks
- 377,098 country chains
  - 7,172 chains of size two
  - 52,655 chains of size three

Sink countries

- A country is a **sink** if less value leaves the country than enters the country. Measured using G^2: country chains of length 2.
- **Sink-OFC centrality** of a country c is defined as

$$C_{sink}(c) = \frac{\sum_{g \in G^2 \wedge g[1]=c} V^{\phi}(g) - \sum_{g \in G^2 \wedge g[0]=c} V^{\phi}(g)}{\sum_{g \in G^2} V^{\phi}(g)}$$

Sink countries

- A country is a **sink** if less value leaves the country than enters the country. Measured using G^2: country chains of length 2.
- **Sink-OFC centrality** of a country c is defined as

$$C_{sink}(c) = \frac{\sum_{g \in G^2 \wedge g[1]=c} V^{\phi}(g) - \sum_{g \in G^2 \wedge g[0]=c} V^{\phi}(g)}{\sum_{g \in G^2} V^{\phi}(g)}$$

- Problem: metric depends on size of country.
- Solution: normalize based on value GDP_c of the country, obtaining:

$$C_{sink-normalized}(c) = C_{sink}(c) \cdot \frac{\sum_{i \in I} GDP_i}{GDP_c}$$

Sink countries

ISO	Country name	C_s	ISO	Country name	C_s	ISO	Country name	C_s
VG	British Virgin Islands	5235	MH	Marshall Islands	100	BZ	Belize	38
TW	Taiwan	2278	MT	Malta	100	GI	Gibraltar	34
JE	Jersey	397	MU	Mauritius	75	AI	Anguilla	27
BM	Bermuda	374	LU	Luxembourg	71	LR	Liberia	17
KY	Cayman Islands	331	NR	Nauru	67	VC	St. Vincent & Gran.	14
WS	Samoa	277	CY	Cyprus	62	GY	Guyana	14
LI	Lichtenstein	225	SC	Seychelles	60	HK	Hong Kong	14
CW	Curacao	115	BS	Bahamas	40	MC	Monaco	11

Table: List of sink-OFCs, ordered by sink centrality value C_s.

Conduits

- The extent to which a country is **conduit** is measured using G^3: country chains of length 3.
- **Inward conduit-OFC centrality** measures the value flowing from a sink-OFC, into the conduit country c, out to any country:

$$C_{conduit_{in}}(c) = \frac{\sum_{g \in G^3 \wedge g[1]=sink \wedge g[2]=c} V^{\phi}(g)}{\sum_{g \in G^3} V^{\phi}(g)}$$

  Here, $g[i] = sink$ denotes that the i-th node in the chain is a sink.
- **Outward conduit-OFC centrality** measures the value flowing from any country, into conduit c, out to a sink-OFC:

$$C_{conduit_{out}}(c) = \frac{\sum_{g \in G^3 \wedge g[3]=sink \wedge g[2]=c} V^{\phi}(g)}{\sum_{g \in G^3} V^{\phi}(g)}$$

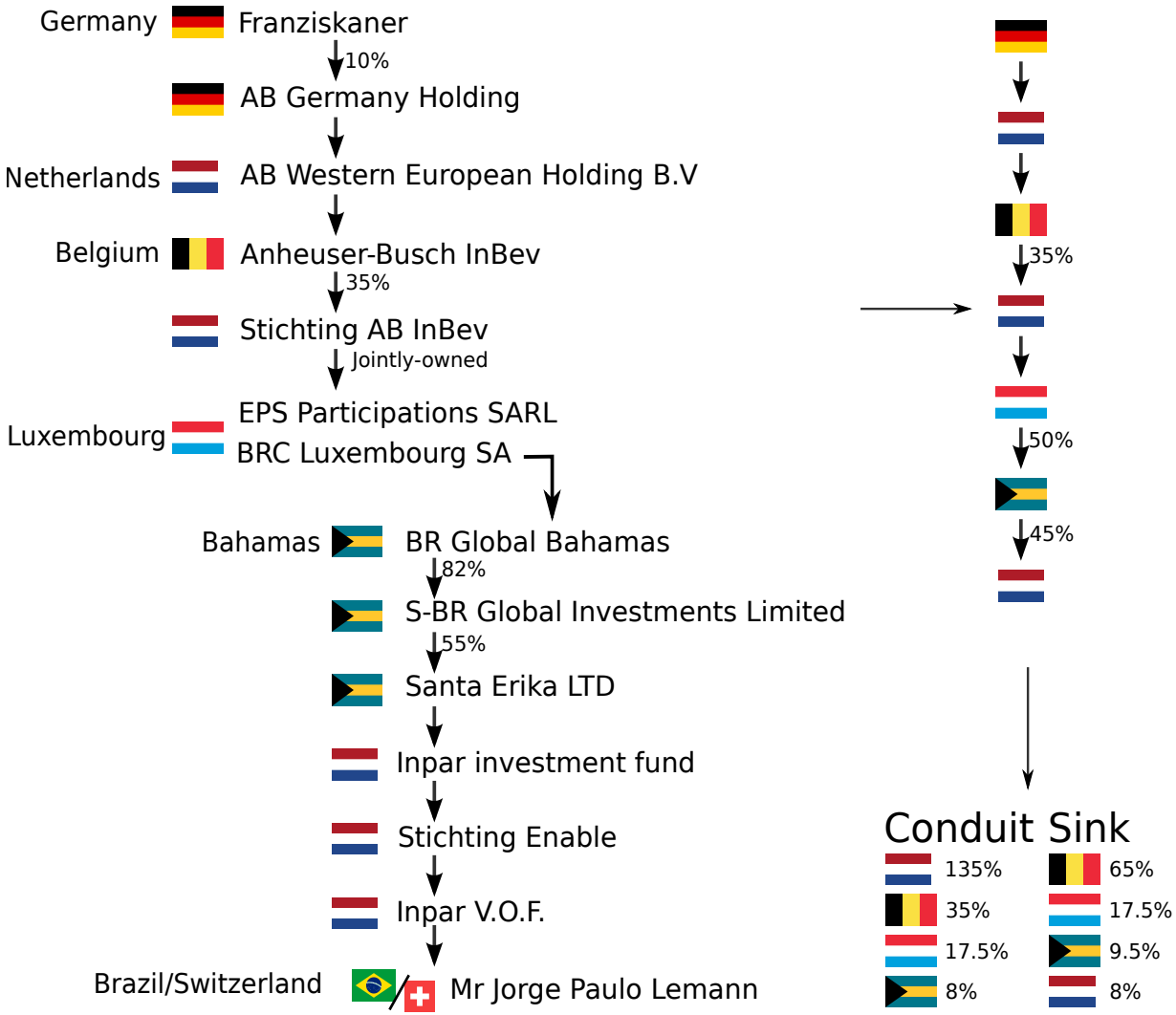
- Normalize by GDP

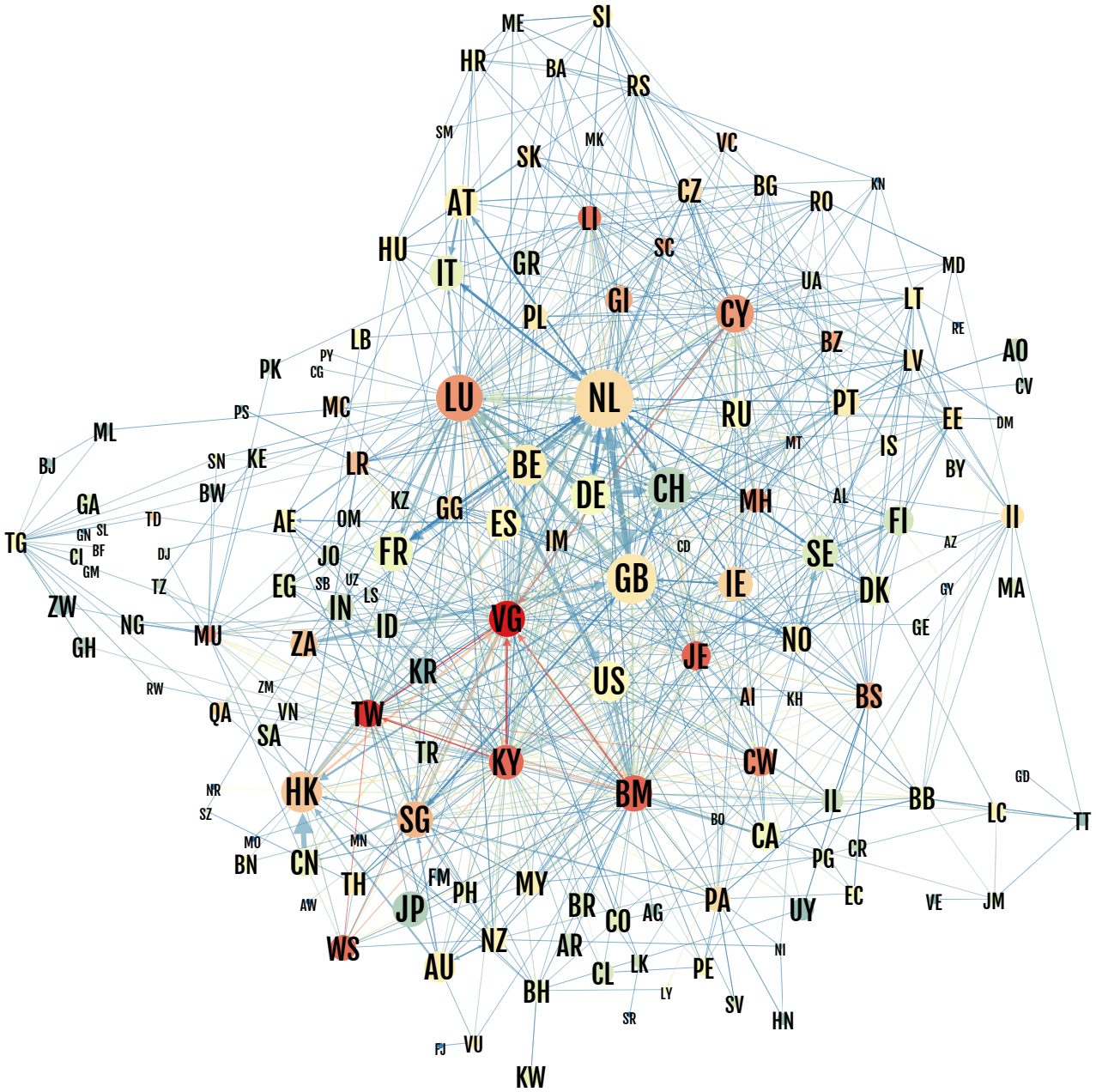
Conduits

ISO	Country name	Non-normalized $C_{c_{out}}$	Non-normalized $C_{c_{in}}$	$C_{c_{out}}$	$C_{c_{in}}$
NL	The Netherlands	$7.4 \cdot 10^{11}$	$3.8 \cdot 10^{11}$	18.6	22.5
GB	United Kingdom	$3.8 \cdot 10^{11}$	$1.3 \cdot 10^{11}$	3.1	2.4
CH	Switzerland	$2.2 \cdot 10^{11}$	$2.7 \cdot 10^{10}$	6.9	2.0
SG	Singapore	$7.2 \cdot 10^{10}$	$2.2 \cdot 10^{10}$	5.1	3.8
IE	Ireland	$6.4 \cdot 10^{10}$	$3.3 \cdot 10^{10}$	5.9	7.2

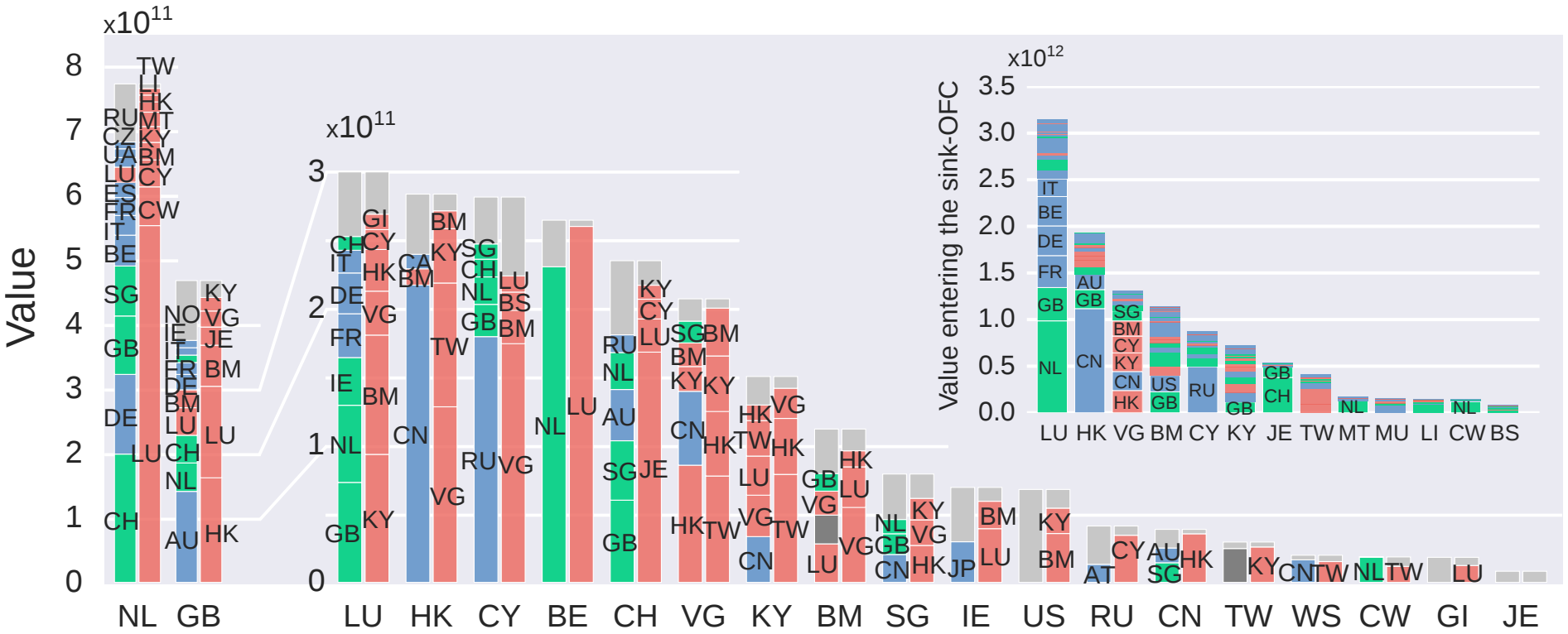
Table: List of conduit-OFCs, ordered by non-normalized $C_{c_{out}}$.

Country chains

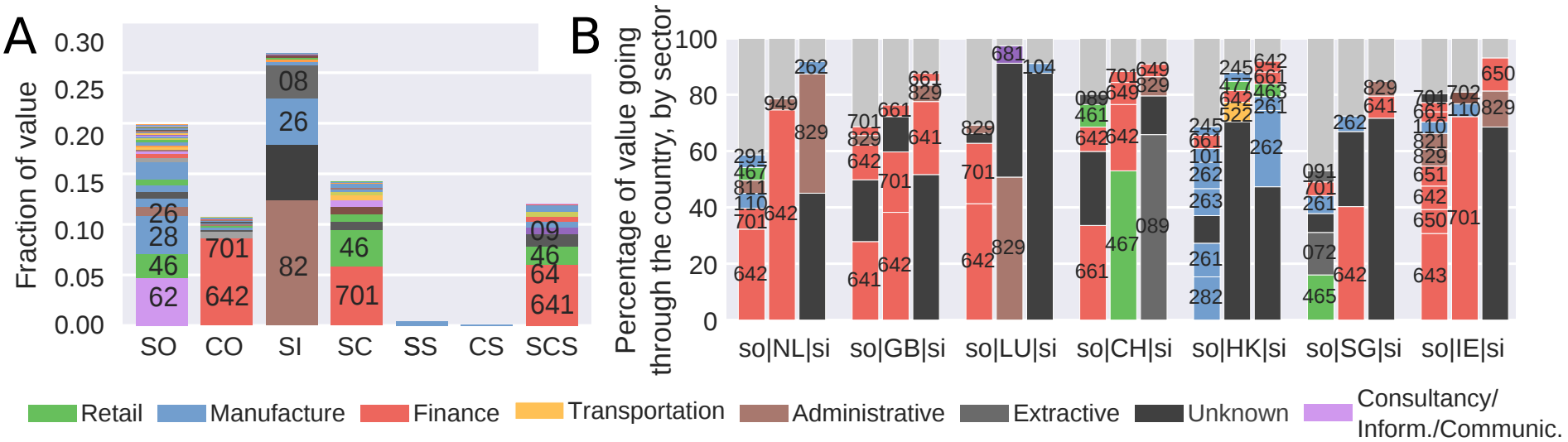



SI
ME
HR
BA
RS
SM
MK
VC
SK
AT
CZ
BG
RO
KN
LI
HU
SC
IT
GR
MD
UA
GI
CY
LT
PL
RE
LB
BZ
AO
PK
PY
CG
LV
CV
LU
NL
MC
RU
PT
PS
EE
DM
ML
MT
IS
SN
KE
LR
BE
BY
BJ
DE
CH
MH
AL
BW
KZ
GG
ES
FI
II
GA
TD
AE
OM
IM
CD
AZ
TG
GN
SL
FR
SE
CI
BF
JO
DJ
GB
IE
GM
EG
SB
UZ
LS
DK
GY
MA
TZ
IN
ZW
VG
NG
ID
MU
ZA
NO
GE
GH
JE
KR
US
RW
ZM
AI
KH
BS
QA
VN
TW
SA
TR
KY
CW
GD
NR
HK
BM
IL
BB
LC
TT
SZ
SG
BO
CA
MO
MN
PG
CR
BN
CN
TH
FM
PH
MY
EC
VE
JM
AW
JP
BR
CO
AG
PA
UY
NZ
AR
NI
WS
AU
CL
LK
PE
BH
LY
SV
SR
HN
FJ
VU
KW

Conduit country specialization



Conduit sector specialization



Findings

- The Netherlands is the conduit between European companies and Luxembourg
- Hong Kong (for China) and Luxembourg (for EU countries) serve as the main countries in the route to typical tax havens
- The United Kingdom is the conduit between European countries and former members of the British Empire, such as Hong Kong, Jersey, Guernsey and Bermuda
- Cyprus is the conduit for Russian firms
- Ireland is prominent in financial leasing and head offices
- Luxembourg specializes in support activities
- Hong Kong and Switzerland are dominant in financial intermediation and derivatives dealing

Questions in parliament . . .

Tweede Kamer der Staten-Generaal **2**

Vergaderjaar 2016–2017 **Vragen gesteld door de leden der Kamer**

2017Z10585

Vragen van het lid **Leijten** (SP) aan de Staatssecretaris van Financiën over *het bericht dat Nederland wereldwijd het belangrijkste knooppunt voor belastingconstructies is* (ingezonden 26 juli 2017).

Vraag 1
Bent u verrast door het bericht «Nederland wereldwijd belangrijkste knooppunt voor belastingconstructies»? Zo ja, waarom? Zo nee, waarom niet? [1] [2]

Vraag 2
Is het de bedoeling geweest van het «financieel gunstige vestigingsbeleid» dat Nederland in 23% van de onderzochte bedrijfsstructuren met belastingparadijzen een rol speelt als tussenschakel? Kunt u uw antwoord onderbouwen?

Vraag 3
Erkent u dat door het op vele manieren en op grote schaal mogelijk maken om belasting te ontwijken, Nederland andere landen ernstig benadeelt omdat zij niet de benodigde inkomsten genereren voor diensten voor hun bevolking? Kunt u uw antwoord toelichten?

Vraag 4
Welke landen worden het meeste benadeeld door de belastingontwijking via Nederland? Bent u bereid dit in kaart te (laten) brengen? Zo nee, waarom niet? Zo ja, wanneer verwacht u daarvan resultaat?

Vraag 5
Erkent u tevens dat het ontwijken van belastingen door grote internationale bedrijven, maakt dat inwoners van een land een steeds groter gedeelte van de belastingen moeten opbrengen en dat dit de ongelijkheid vergroot?

Vraag 6
Was het een verrassing dat onderzoeksgroep Corpnet concludeert dat de erosie van belastinggrondslag over winst mogelijk maakt, onder andere door de bepalingen over informeerkapitaal en de innovatiebox, waardoor

[1] https://fd.nl/economie-politiek/1211312/nederland-belangrijkste-schakel-in-internationale-offshorestructuren
[2] Het onderzoek zelf: https://www.nature.com/articles/s41598-017-06322-9

kv-tk-2017Z10585
's-Gravenhage 2017

Answer by Wiebes

Tweede Kamer der Staten-Generaal **2**

Vergaderjaar 2016–2017 **Aanhangsel van de Handelingen**

Vragen gesteld door de leden der Kamer, met de daarop door de regering gegeven antwoorden

2540

Vragen van het lid **Leijten** (SP) aan de Staatssecretaris van Financiën over *het bericht dat Nederland wereldwijd het belangrijkste knooppunt voor belastingconstructies is* (ingezonden 26 juli 2017).

Antwoord van Staatssecretaris **Wiebes** (Financiën) (ontvangen 25 augustus 2017).

Vraag 1, 2
Bent u verrast door het bericht «Nederland wereldwijd belangrijkste knooppunt voor belastingconstructies»? Zo ja, waarom? Zo nee, waarom niet? [1] [2]

become a supporter subscribe find a job

theguardian

news / opinion / sport / arts / life

world / europe / US / americas / asia / australia / middle east / africa / more

Tax havens

Netherlands and UK are biggest channels for corporate tax avoidance

The two countries are conduits for 37% of money heading to tax havens, most of which have strong links to Britain

Conclusions

- **Corporate networks** explicitly and implicitly contain rich information on corporations and jurisdictions
- Offshore financial centers (OFCs) come in two flavors: **conduits** and **sinks**
- OFCs can be detected using **corporate ownership chains**
- Conduits are highly **specialized** in serving particular countries and industrial sectors

Conclusions

- **Corporate networks** explicitly and implicitly contain rich information on corporations and jurisdictions
- Offshore financial centers (OFCs) come in two flavors: **conduits** and **sinks**
- OFCs can be detected using **corporate ownership chains**
- Conduits are highly **specialized** in serving particular countries and industrial sectors
- The Netherlands is the largest value conduit in the world

Upcoming week

- Next week student presentations start
- Be sure you know the following:
  - Your track letter
    - Track A: room BW.0.17 (Frank)
    - Track B: room DM.1.15 (Rachel)
    - Track C: room DM.1.19 (Gamal); starts at 10:00 in the first week
- Tracks and rooms are fixed for the semester
- In the lab session in the week before you are presenting, you can (make an appointment with the lecturer assigned to your room to) receive input on your draft slides
- Somewhat finalize Assignment 2