Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

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

Lecture 5 — Network evolution and model extensions

Assignment feedback

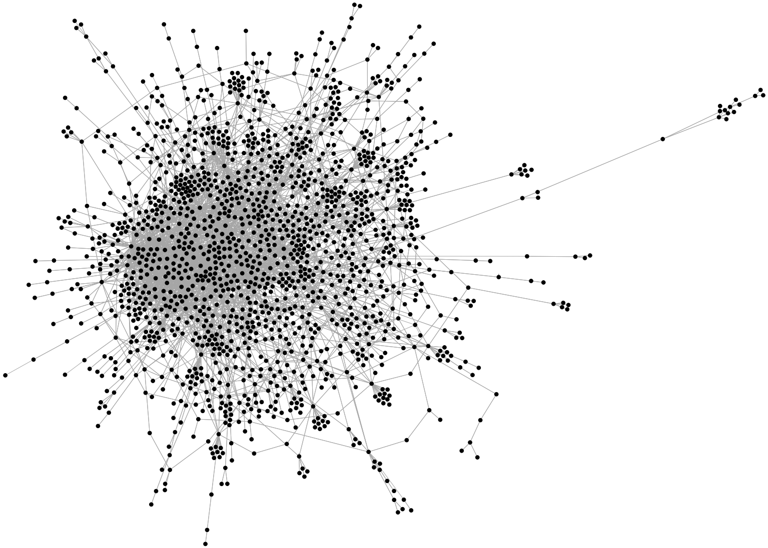
- Please study and utilize detailed assignment feedback
- Grade < 5.0: insufficient; compensate with extra assignment
- $5.0 \leq$ Grade < 5.5: insufficient, unless compensated with Assignment 2 to average of two assignments ≥ 5.5
- Grade ≥ 5.5: sufficient
- Questions? Ask your grader during the upcoming lab session (initials on work)

Today

- Recap
- Temporal networks
- Network models
- Network dynamics and evolution
- Challenges in network science

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

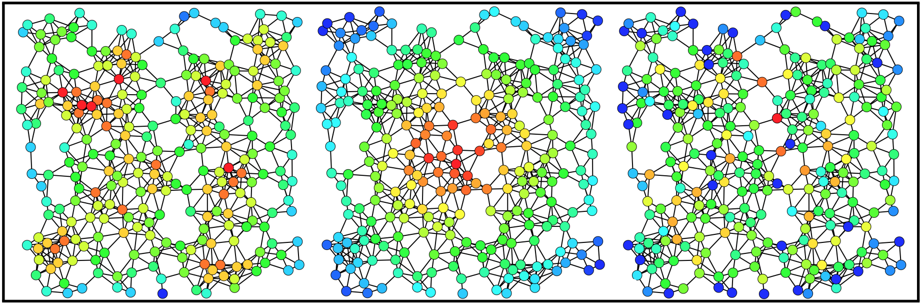
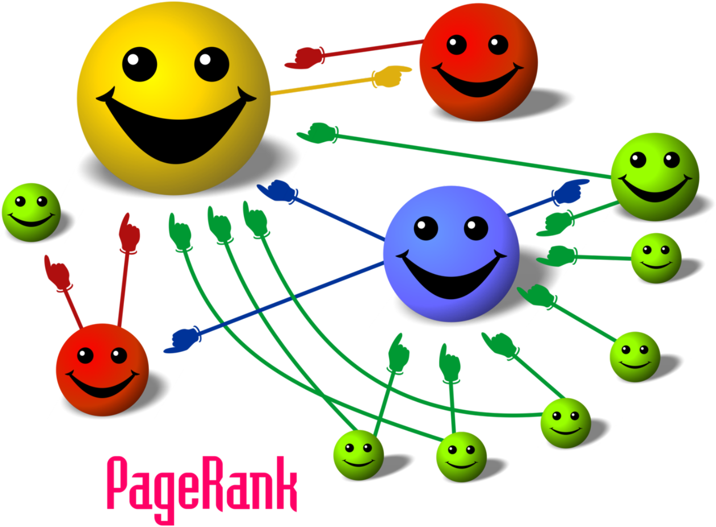


Figure: Degree, closeness and betweenness centrality

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 $O(m)$

Community detection

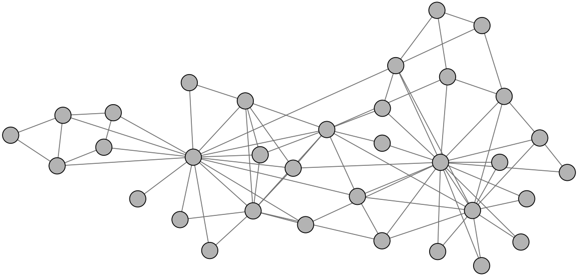


Figure: Communities: node subsets connected more strongly with each other

Community detection

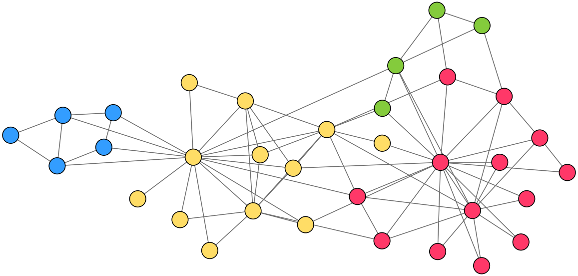
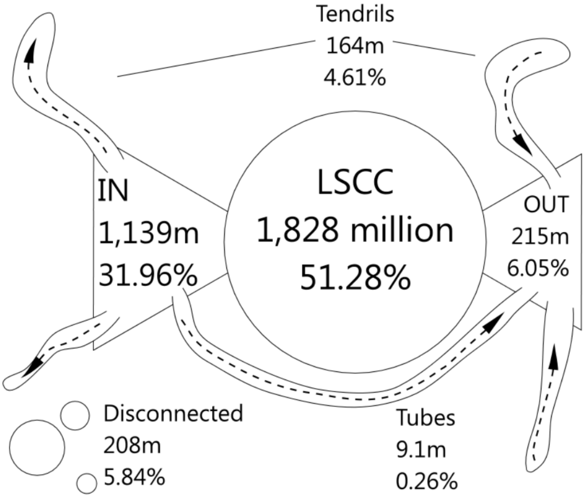


Figure: Communities: node subsets connected more strongly with each other

Bow-tie structure of the web



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

Temporal networks

Temporal network analysis

- Graphs **evolve** over time
  - Social networks: users join the network and create new friendships
  - Webgraphs: new pages and links to pages appear on the internet
  - Scientific networks: new papers are being co-authored and new citations are made in these papers
- Interesting: small world properties emerge and are preserved during evolution!

Temporal networks

- Graph $G^t = (V^t, E^t)$
- Time window $0 \leq t \leq T$
- Usually at $t = 0$, either
  - $V^0 = \emptyset$ and a new edge may bring new nodes, or
  - $V^0 = V^T$ and only edges are added at each timestamp
- Timestamp on node $v \in V$:
  $\tau(v) \in [0; T]$
- Timestamp on edge $e \in E$:
  $\tau(e) \in [0; T]$, or as common input format:
  $e = (u, v, t)$ with $u, v \in V$ and $t \in [0, T]$
  `u v t` as line contents of an edge list file

Two schools

- **Synthetic graphs** model-driven
  - Model or algorithm to generate graphs from scratch
  - Tune parameters to obtain a graph similar to an observed network
  - Statistical analysis
- **Real-world graphs** data-driven
  - Obtain data from an actual network
  - Compute and derive properties and determine similarity with other networks
  - Computational analysis

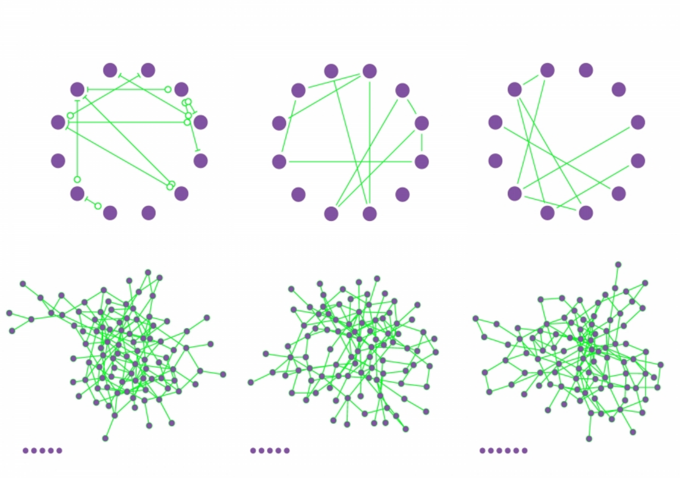
Three models

- Random graphs (Erdös-Rényi)
- Barábasi-Albert model
- Watts-Strogatz model

Random graphs (1959)

- Initially, n nodes and 0 edges
- Add edges at random
- **Edgar Gilbert** / **Erdös-Rényi**: a random graph $G(n, p)$ has n nodes and each undirected edge exists with probability $0 < p < 1$. Expected $m = p \cdot \frac{1}{2}n(n-1)$ edges
- **Erdös-Rényi**: a random graph $G(n, m)$ has n nodes and m edges, and this graph is chosen uniformly random from all possible graphs with n nodes and m edges
- Result does not really resemble real-world graphs

Erdös-Rényi



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

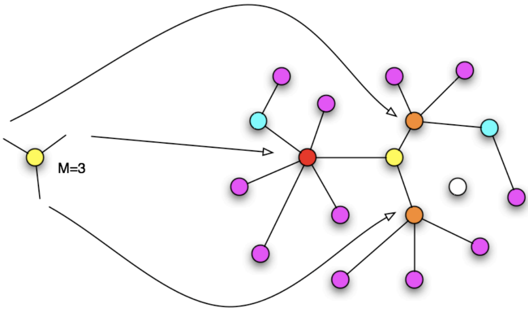
Barábasi-Albert model (1999)

- "Rich get richer"
- **Preferential attachment**: nodes with a high degree more strongly attract new links
- An edge (u, v) is added between a new node u and a non-random node v with probability:

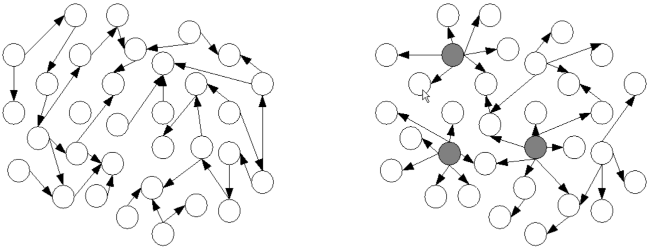
$$p(v) = \frac{deg(v)}{\sum_{w \in V} deg(w)}$$

- (Plus some dampening based on the age of the node and correction for links between high-degree nodes)
- Result: giant component and power-law degree distribution: the **scale-free** property

Barábasi-Albert model (1999)



Random vs. scale-free



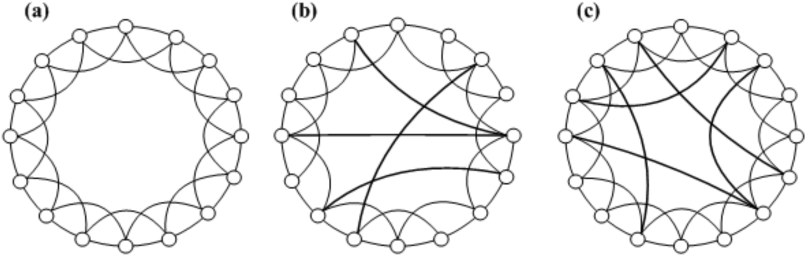
(a) Random network (b) Scale-free network

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

Watts-Strogatz model (1998)

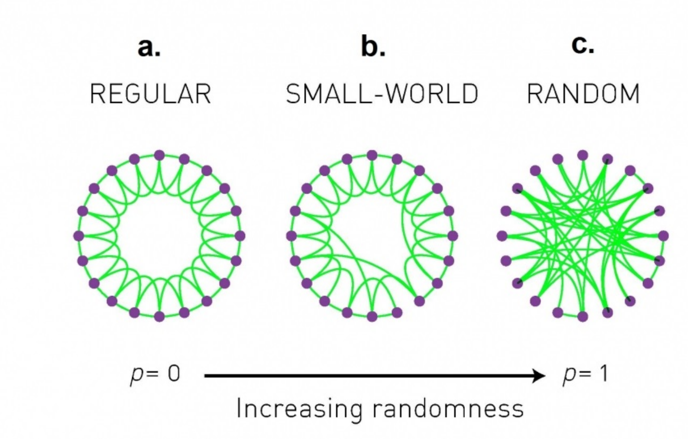
- Input number of nodes n, average degree k and parameter p
- Constructs undirected graph with n nodes and $\frac{1}{2} \cdot n \cdot k$ edges
- Start with "circle-shaped" graph connecting each node to its k nearest neighbors
- Until each edge has been considered, in clock-wise order, **Rewire** each node's edge to a closest neighbor, to a random node with probability p
- Result: low distances, giant component, high clustering

Watts-Strogatz



Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393(6684), 440-442.

Discussion of models



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

Discussion of models

- Many generative more models exist: configuration model, stub-matching model, . . .
- ERGM, SAOM, REM, stochastic block models, . . .

Discussion of models

- Many generative more models exist: configuration model, stub-matching model, . . .
- ERGM, SAOM, REM, stochastic block models, . . .
- Better understanding of system's evolution
- Compare real-world structure with model structure
- Investigate system's complexity

Discussion of models

- Many generative more models exist: configuration model, stub-matching model, . . .
- ERGM, SAOM, REM, stochastic block models, . . .
- Better understanding of system's evolution
- Compare real-world structure with model structure
- Investigate system's complexity
- Model is never perfect
- Not all small-world properties are captured

Network evolution

Levels of evolution

- **Microscopic (local)**
- Macroscopic (global)

Microscopic evolution

- Node-based investigation of evolution
- Analysis of four online social networks: DELICIOUS, FLICKR, LINKEDIN and YAHOO! ANSWERS
- Other than degree, preferential attachment (assortativity) can also be based on node **age** and the number of **hops** (distance before link is created)
- Derive model based on these properties

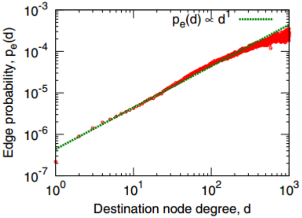
Leskovec et al., Microscopic Evolution of Social Networks, in Proceedings of KDD, pp. 462-470, 2008.

Datasets

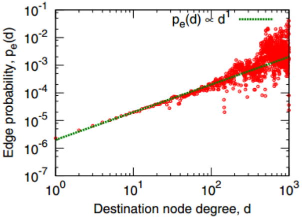
Network	T	N	E	E_b	E_u	E_Δ	%	ρ	κ
FLICKR (03/2003–09/2005)	621	584,207	3,554,130	2,594,078	2,257,211	1,475,345	65.63	1.32	1.44
DELICIOUS (05/2006–02/2007)	292	203,234	430,707	348,437	348,437	96,387	27.66	1.15	0.81
ANSWERS (03/2007–06/2007)	121	598,314	1,834,217	1,067,021	1,300,698	303,858	23.36	1.25	0.92
LINKEDIN (05/2003–10/2006)	1294	7,550,955	30,682,028	30,682,028	30,682,028	15,201,596	49.55	1.14	1.04

Table 1: Network dataset statistics. E_b is the number of bidirectional edges, E_u is the number of edges in undirected network, E_Δ is the number of edges that close triangles, $\%$ is the fraction of triangle-closing edges, ρ is the densification exponent ($E(t) \propto N(t)^\rho$), and κ is the decay exponent ($E_h \propto \exp(-\kappa h)$) of the number of edges E_h closing h hop paths

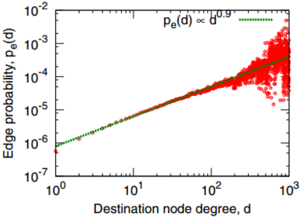
Preferential attachment: degree



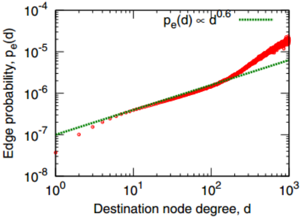
(c) FLICKR



(d) DELICIOUS

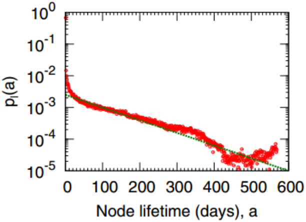


(e) ANSWERS

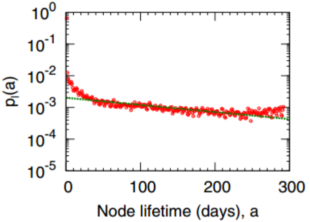


(f) LINKEDIN

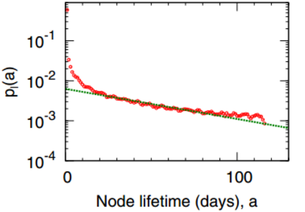
Preferential attachment: age



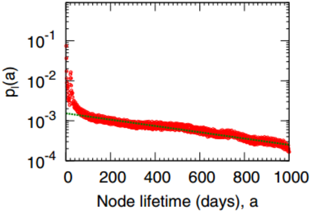
(a) FLICKR



(b) DELICIOUS

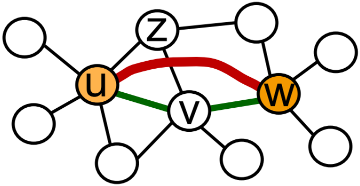


(c) ANSWERS

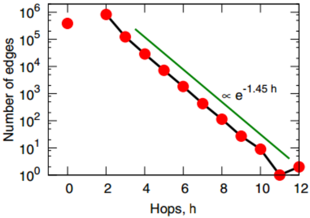


(d) LINKEDIN

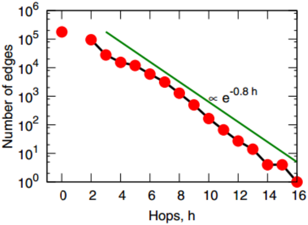
Triadic closure



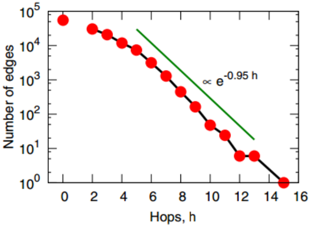
Preferential attachment: hops



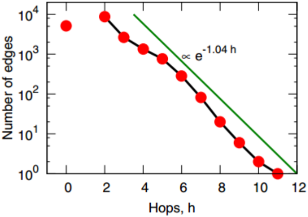
(c) FLICKR



(d) DELICIOUS



(e) ANSWERS



(f) LINKEDIN

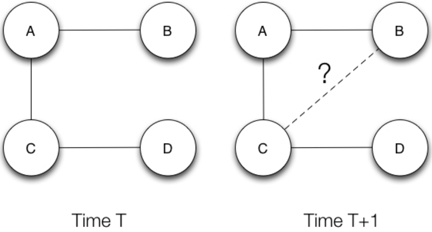
Microscopic evolution model

- Node arrival and lifetime determined using function (based on derived exponential distribution)
- Node goes to sleep for a time gap, length again sampled from a derived distribution
- Node wakes up to create an edge using (adjusted) triangle closing model and goes to sleep
- Sleep time gets shorter as the degree of a node increases
- Node dies after lifetime is reached

Leskovec et al., Microscopic Evolution of Social Networks, in Proceedings of KDD, pp. 462-470, 2008.

Link prediction

- Predict "next friendship" to be formed



Liben-Nowell et al., The Link Prediction Problem for Social Networks, in Proceedings of CIKM, pp. 556-559, 2003.

Levels of evolution

- Microscopic (local)
- **Macroscopic (global)**

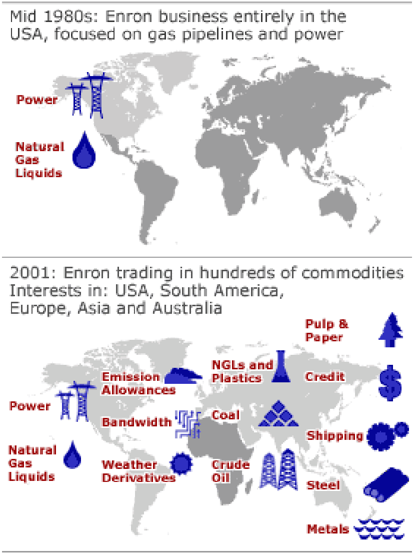
Macroscopic evolution

- Look at evolution of network as a whole
- Observe different characteristic graph properties
- Devise model that incorporates these properties

Dataset	Nodes	Edges	Time	DPL exponent
Arxiv HEP–PH	30,501	347,268	124 months	1.56
Arxiv HEP–TH	29,555	352,807	124 months	1.68
Patents	3,923,922	16,522,438	37 years	1.66
AS	6,474	26,467	785 days	1.18
Affiliation ASTRO–PH	57,381	133,179	10 years	1.15
Affiliation COND–MAT	62,085	108,182	10 years	1.10
Affiliation GR-QC	19,309	26,169	10 years	1.08
Affiliation HEP–PH	51,037	89,163	10 years	1.08
Affiliation HEP–TH	45,280	68,695	10 years	1.08
Email	35,756	123,254	18 months	1.12
IMDB	1,230,276	3,790,667	114 years	1.11
Recommendations	3,943,084	15,656,121	710 days	1.26

Leskovec et al., Graph Evolution: Densification and Shrinking Diameters, in TKDD 1(1): 2, 2007

Enron

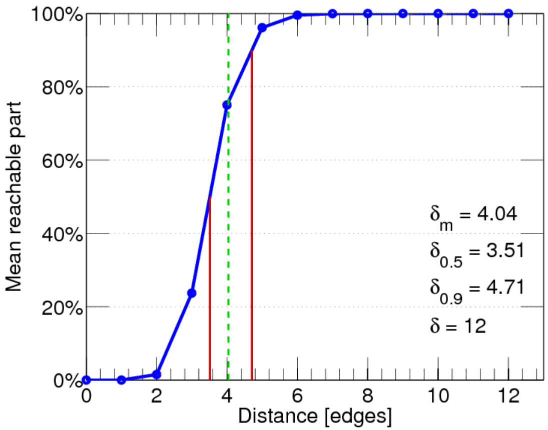


Mid 1980s: Enron business entirely in the USA, focused on gas pipelines and power
Power
Natural Gas Liquids
2001: Enron trading in hundreds of commodities
Interests in: USA, South America, Europe, Asia and Australia
Pulp & Paper
Emission Allowances
NGLs and Plastics
Credit
Power
Bandwidth
Coal
Shipping
Natural Gas Liquids
Weather Derivatives
Crude Oil
Steel
Metals

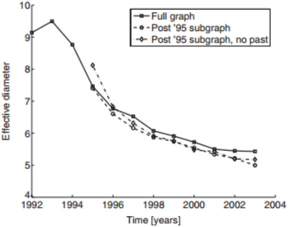
Macroscopic patterns

- Densification: density increases over time
- Giant component grows asymptotically
- Shrinking average distance: $d \sim log(n)$ does not hold over time
- Shrinking effective diameter
  - Effective diameter $\delta_{0.9}$: path length such that 90% of all node pairs are at distance $\delta_{0.9}$ or less
  - Diameter: longest shortest path length

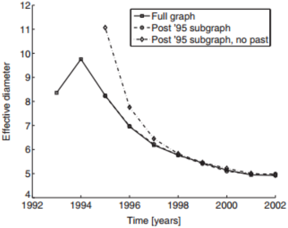
Effective diameter



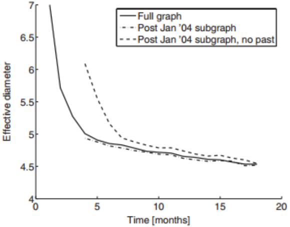
Effective diameter



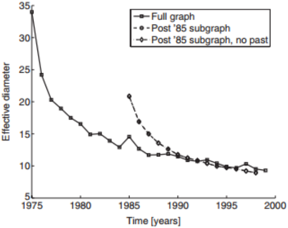
(a) arXiv citation graph



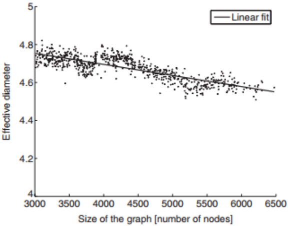
(b) Affiliation network



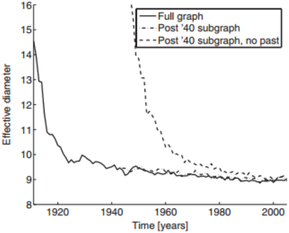
(e) Email network



(c) Patents citation graph

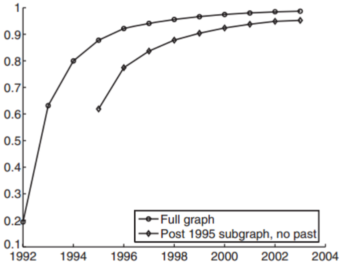


(d) Autonomous Systems



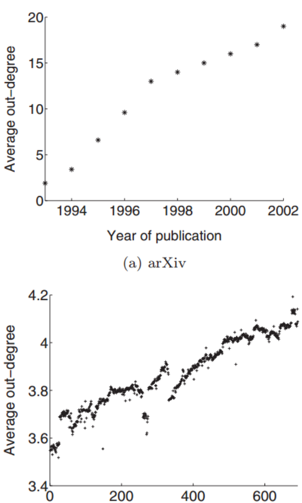
(f) IMDB actors to movies network

Giant component



(a) arXiv citation graph (b) Affiliation network

Densification



(a) arXiv

(b) Patents

(c) Autonomous Systems

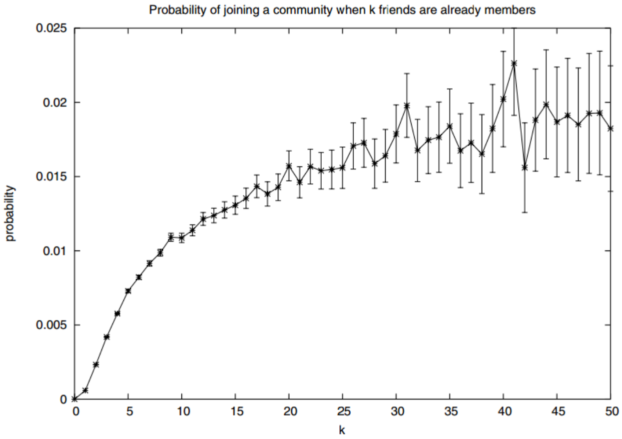
(d) Affiliation network

Community evolution

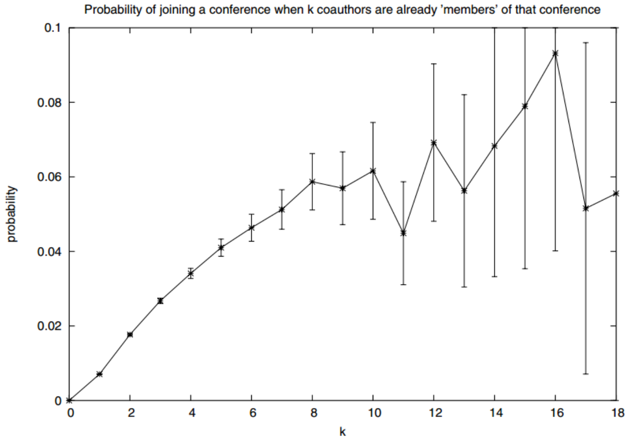
- Slightly different: user-defined communities
- DBLP: scientific collaboration network where communities are conferences that authors visit
- LiveJournal: online social network with explicit groups based on common interest
- What motivates nodes to join a community?
- What causes nodes to switch between communities?
- When do communities grow?

Backstrom et al., "Group formation in large social networks: membership, growth, and evolution", in Proceedings of KDD, pp. 44–54, 2006.

Community evolution (LiveJournal)



Community evolution (DBLP)



Features

Table 1: Features.

Feature Set	Feature
Features related to the community, C. (Edges between only members of the community are $E_C \subseteq E$.)	Number of members ($\lvert C \rvert$). Number of individuals with a friend in C (the *fringe* of C) . Number of edges with one end in the community and the other in the fringe. Number of edges with both ends in the community, $\lvert E_C \rvert$. The number of open triads: $\lvert\{(u,v,w)\lvert(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \notin E_C \wedge u \neq w\}\rvert$. The number of closed triads: $\lvert\{(u,v,w)\lvert(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \in E_C\}\rvert$. The ratio of closed to open triads. The fraction of individuals in the fringe with at least k friends in the community for $2 \leq k \leq 19$. The number of posts and responses made by members of the community. The number of members of the community with at least one post or response. The number of responses per post.
Features related to an individual u and her set S of friends in community C.	Number of friends in community ($\lvert S \rvert$). Number of adjacent pairs in S ($\lvert\{(u,v)\lvert u,v \in S \wedge (u,v) \in E_C\}\rvert$). Number of pairs in S connected via a path in E_C. Average distance between friends connected via a path in E_C. Number of community members reachable from S using edges in E_C. Average distance from S to reachable community members using edges in E_C. The number of posts and response made by individuals in S. The number of individuals in S with at least 1 post or response.

Decision tree (LiveJournal)

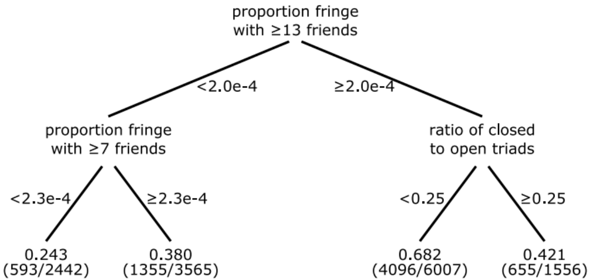


Figure 5: The top two levels of decision tree splits for predicting community growth in LiveJournal.

Decision tree (LiveJournal)

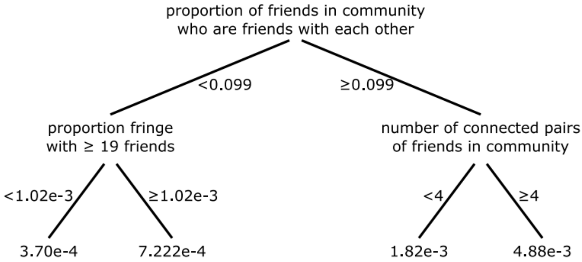
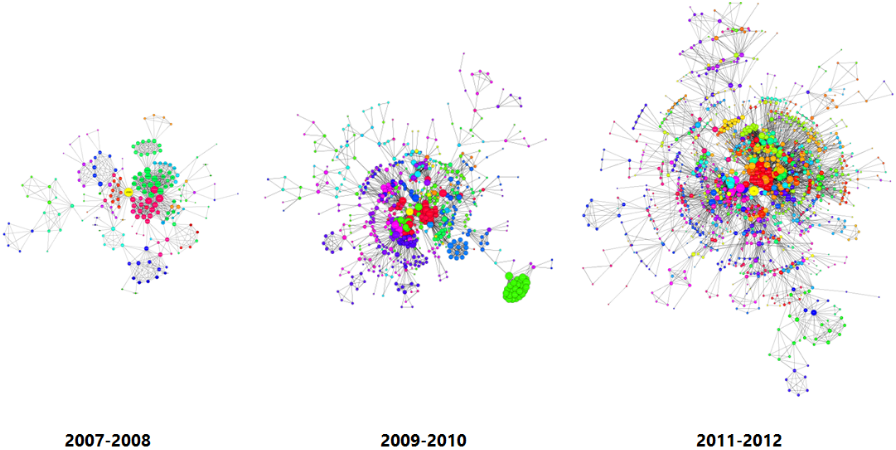


Figure 3: The top two levels of decision tree splits for predicting single individuals joining communities in LiveJournal. The overall rate of joining is 8.48e-4.

Community evolution patterns

- Number of friends already in a community correlates with decision to join a community
- Using various features, decision trees can predict community behavior
- In most models, parameters are specific for considered network
- Challenge: do not flatten data, but use actual network and community structure, perhaps even parameter-free?

Apple collaboration network



http://www.kenedict.com/apples-internal-innovation-network-unraveled/

Network contraction

- Example: social network losing members to competitor
- Deletion of nodes (and its edges)
- Deletion of edges (and ultimately nodes)
- Merging nodes (a corporate network in which companies merge)
- What happens when you remove a hub?
- How about reversing existing models?

Network science challenges

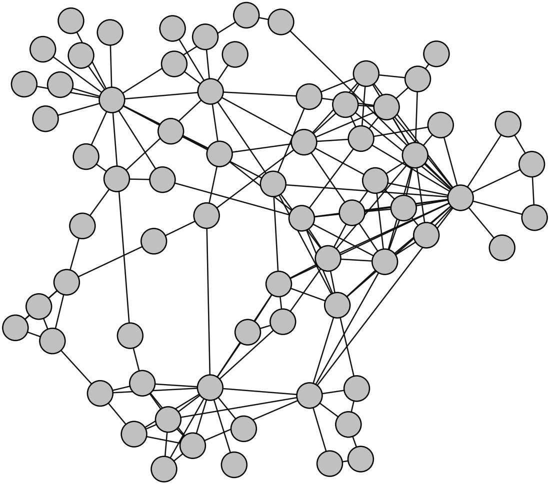
Network science

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

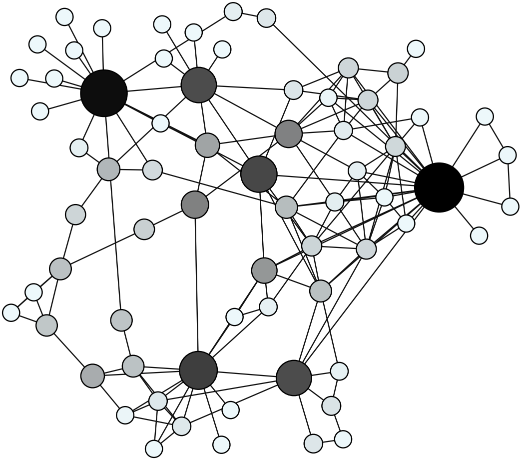
Network science

- **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

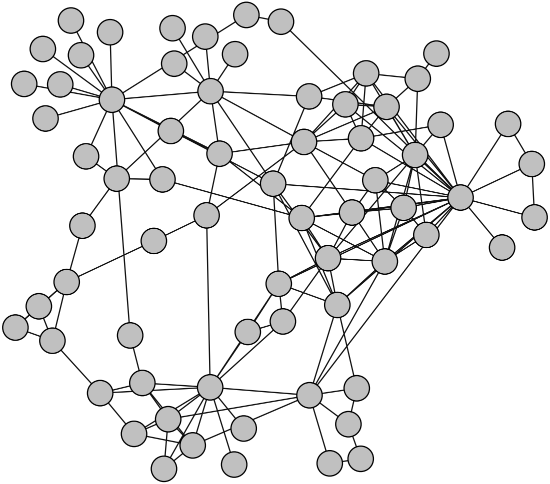
Network analysis



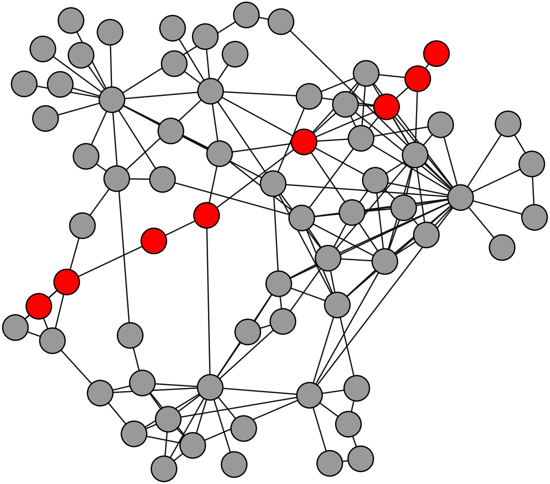
Micro scale



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.)

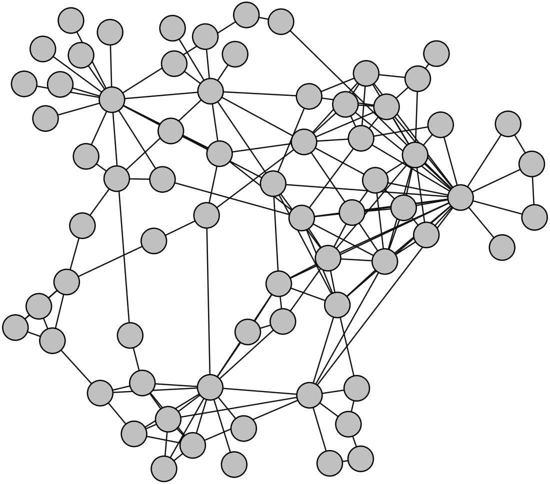
Network analysis

- **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

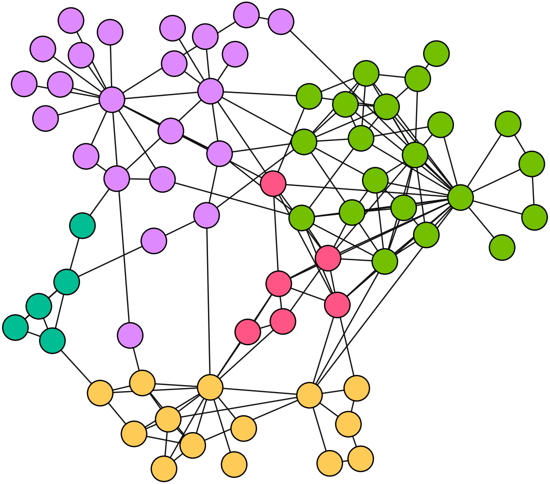
Network analysis

- **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)

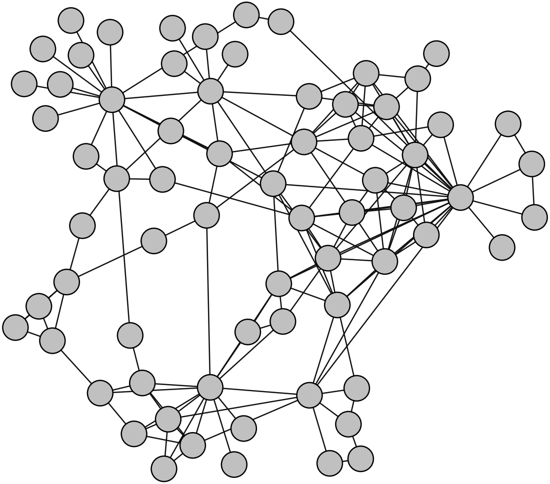
Meso scale: communities



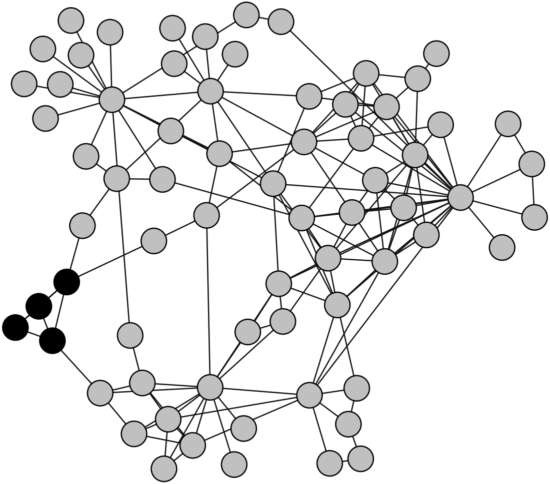
Meso scale: communities



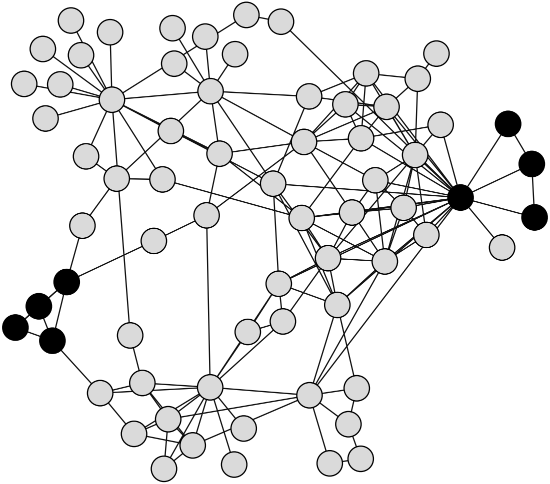
Meso scale: motifs



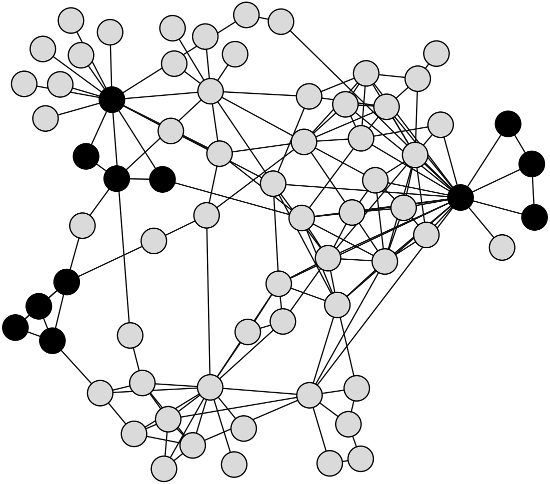
Meso scale: motifs

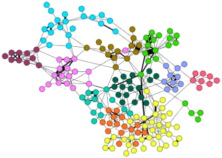


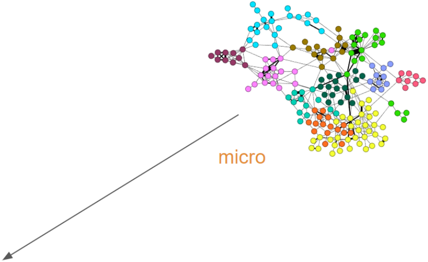
Meso scale: motifs

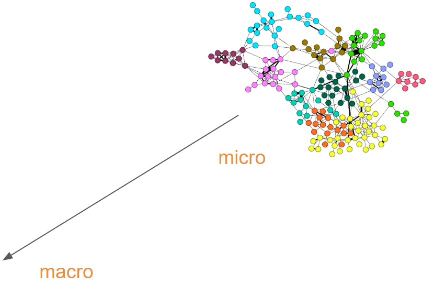


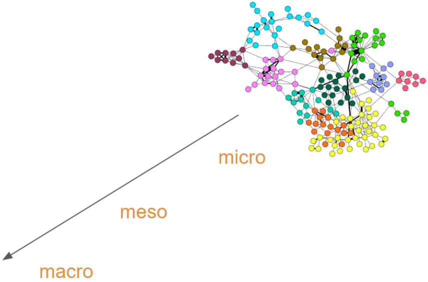
Meso scale: motifs

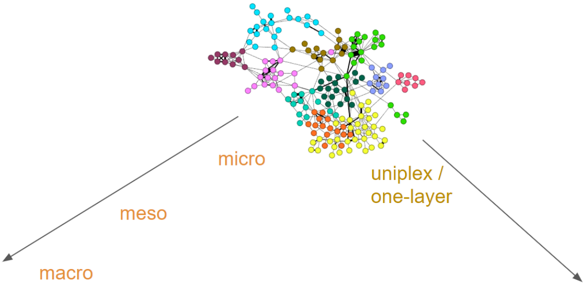


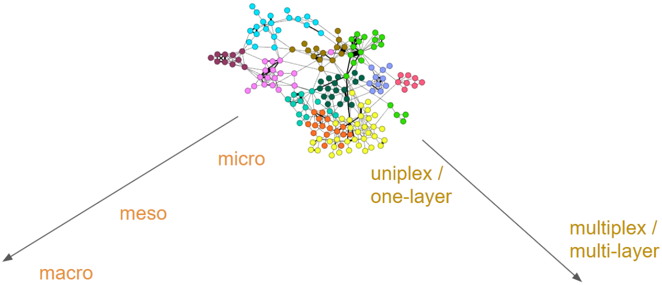


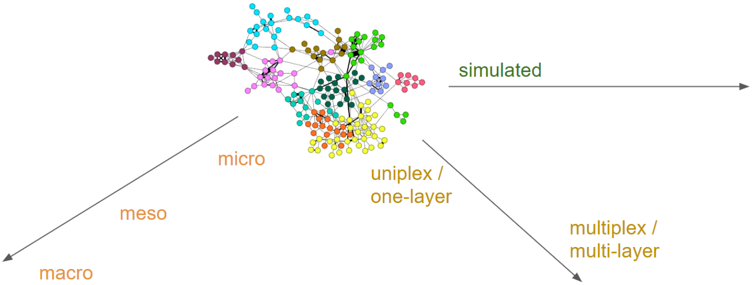

micro

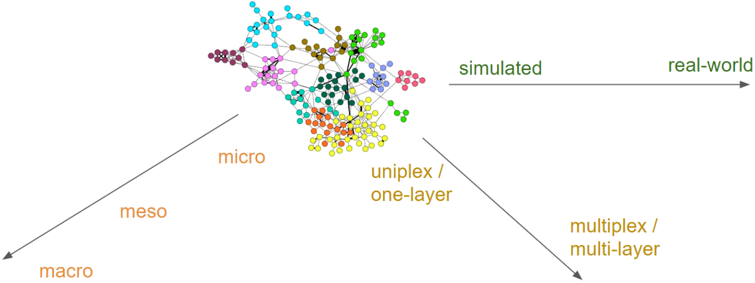

micro
macro

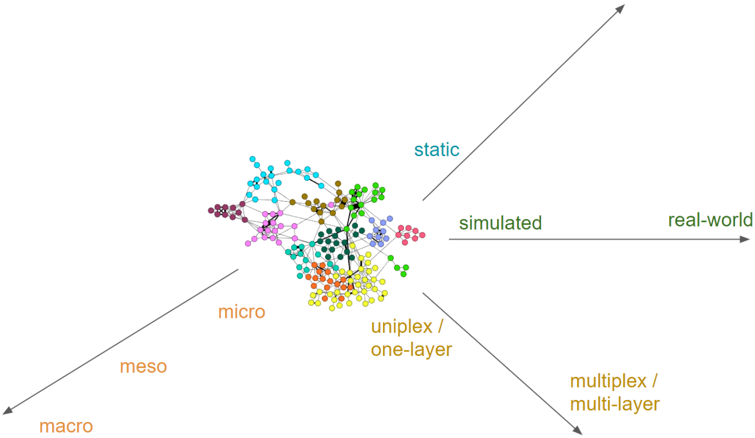

micro
meso
macro

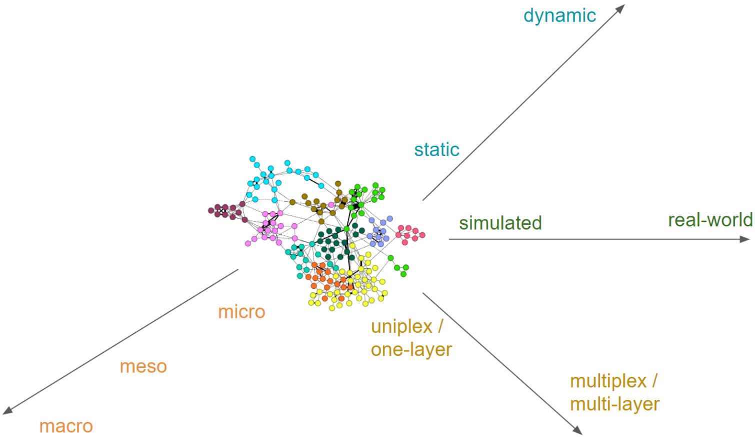

micro
meso
macro
uniplex /
one-layer

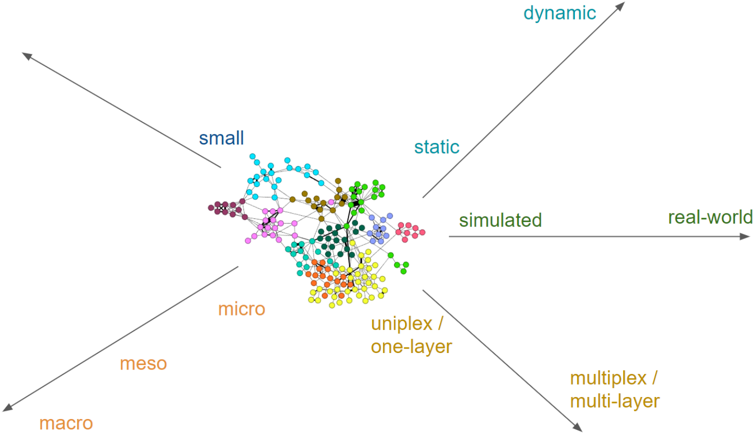

micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer

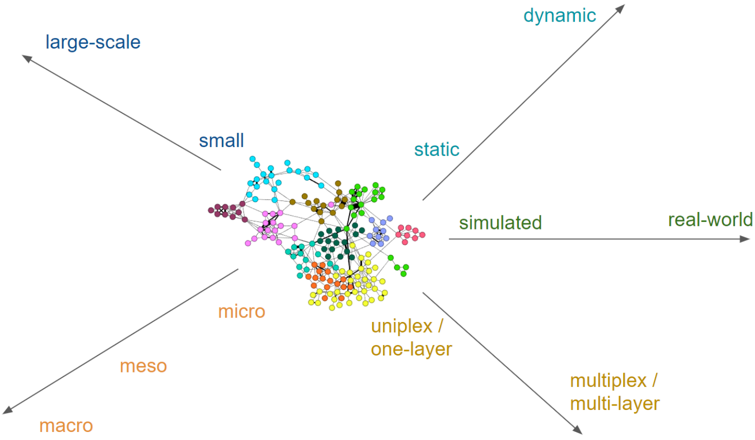

simulated
micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer

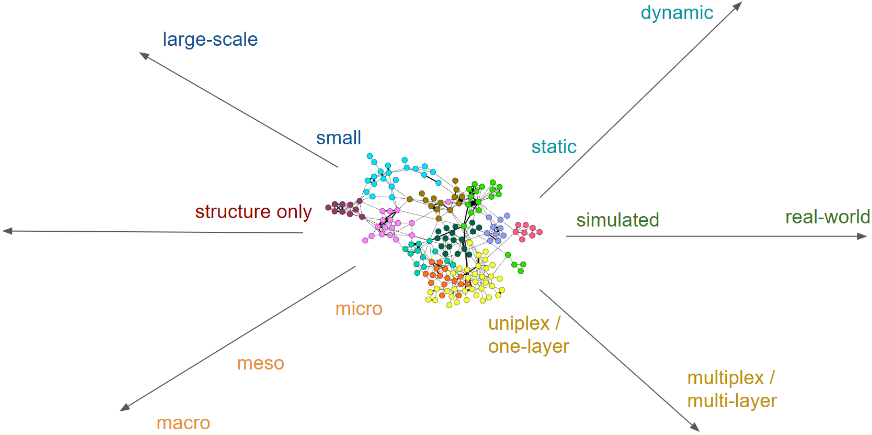

simulated
real-world
micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer

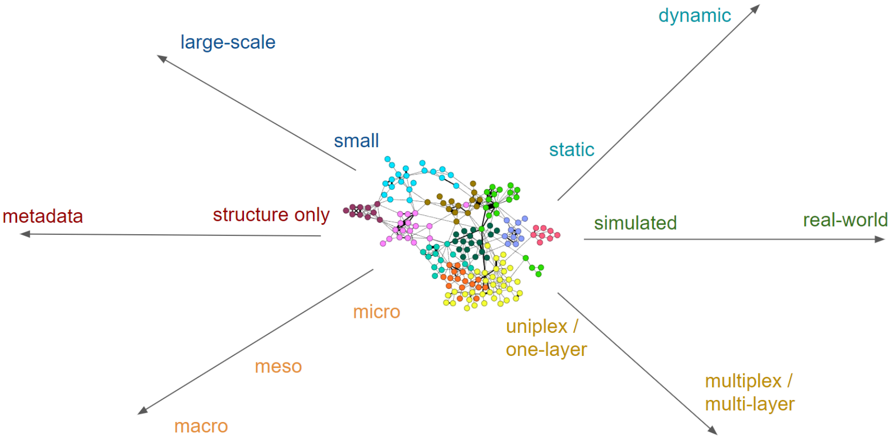

static
simulated
real-world
micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer


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

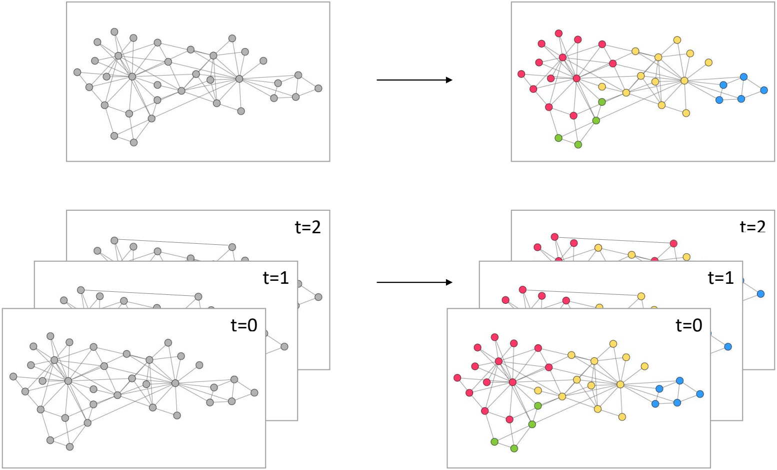

dynamic
small
static
simulated
real-world
micro
meso
macro
uniplex /
one-layer
multiplex /
multi-layer


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

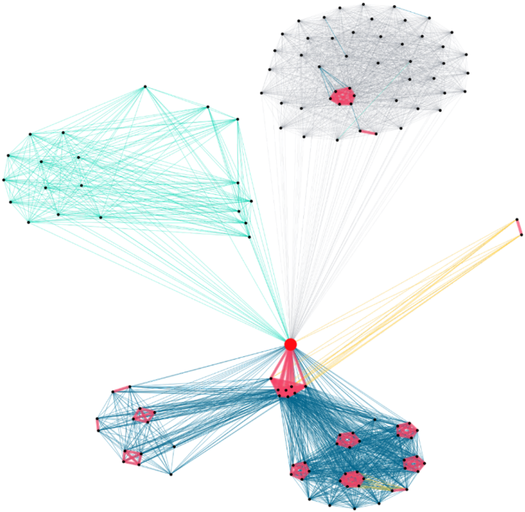

dynamic
large-scale
small
static
structure only
simulated
real-world
micro
uniplex /
one-layer
meso
multiplex /
multi-layer
macro


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

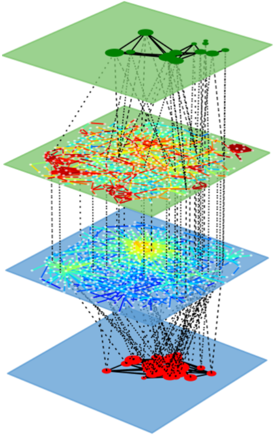
Network (community) dynamics



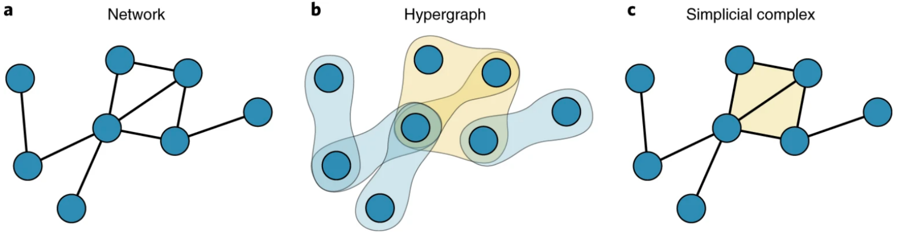
Multilayer networks



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.

Upcoming week

- Next week: last lab session to work on Assignment 2
- Next week: no lecture; from Oct 27 onwards: student presentations
- Be sure you know the following:
  - your track letter (A/B/C/D)
  - with whom (Frank or Hanjo) you are presenting
  - the time of your session: differs per week; 11:00 or 12:10; see website
- Presenting? On the Tuesday before your Friday presentation, drop by Frank (157b; agree on a time the lab session before) or Hanjo (126; Tuesdays between 15-17h).
- From now on, use the time between 9:00 and 11:00 to work on your course project; we are there to help.