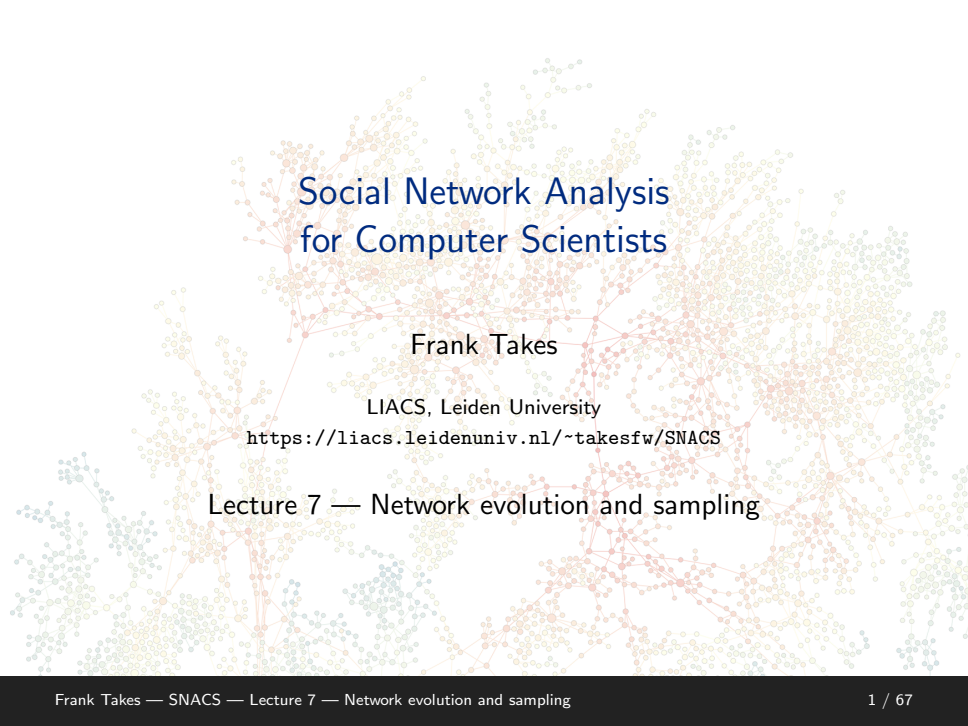




Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

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

Lecture 7 — Network evolution and sampling

Today

- Recap
- Micro, meso and macro scale
- Network centralization
- Network evolution
- Measurement error
- Network sampling algorithms

Recap

Notation

Concept

- Network (graph)
- Nodes (objects, vertices, ...)
- Links (ties, relationships, ...)
 - Directed — $E \subseteq V \times V$ — “links”
 - Undirected — “edges”
- Number of nodes — $|V|$
- Number of edges — $|E|$
- Degree of node u
- Distance from node u to v

Symbol

$G = (V, E)$

V

E

n

m

$\deg(u)$

$d(u, v)$

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
- Many examples: social networks, communication networks, citation networks, collaboration networks, protein interaction networks, information networks, webgraphs, financial networks, ...

Centrality measures

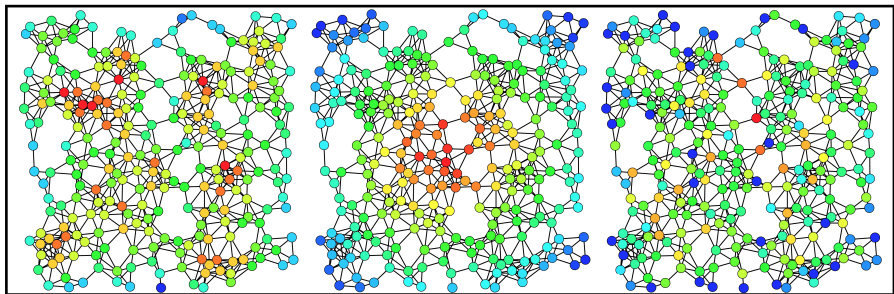


Figure: Degree, closeness and betweenness centrality.

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

Attribute assortativity

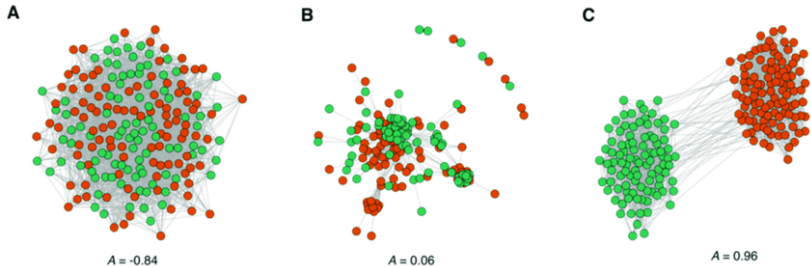


Figure: Attribute assortativity.

Image: Moya-García, A. et al. Identification of New Toxicity Mechanisms ... *Genes*, 13(7), 1292, 2022.

Projection

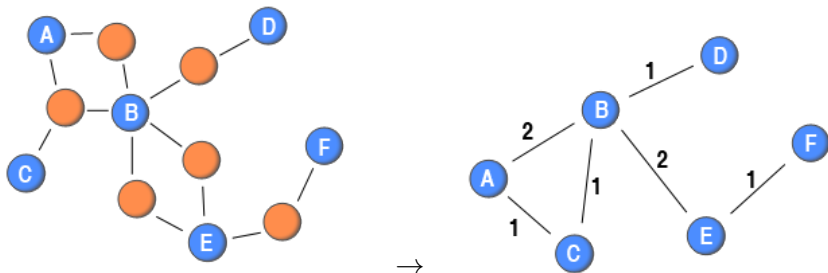


Image: <http://toreopsahl.com>

Community detection

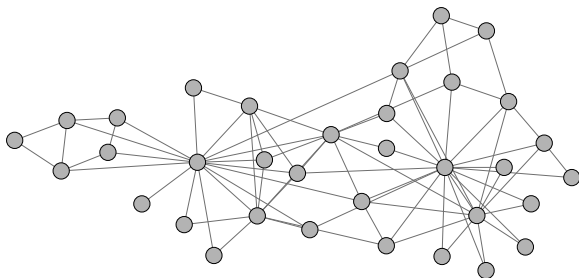


Figure: Communities: node subsets connected more strongly with each other.

- Partitioning a given graph into densely connected groups of nodes

Community detection

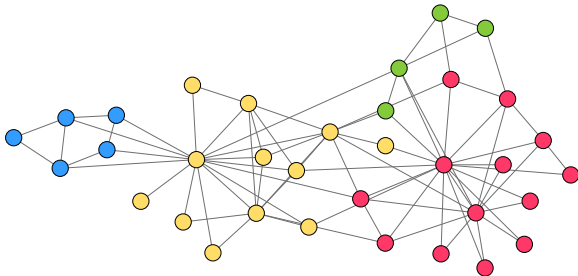


Figure: Communities: node subsets connected more strongly with each other.

- Divisive algorithms (Girvan–Newman), local search algorithms (Louvain) optimizing a quality function (e.g., modularity), label propagation and random walk based approaches.

Core/periphery structure

- Many large networks have a densely connected **core** and a **periphery** of nodes attaching to only a few neighbors
- **Core/periphery structure**: there is **one** dense area
- **Community structure**: there are **several**, with little periphery

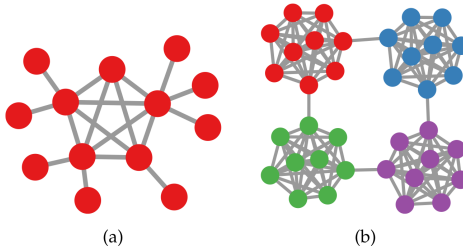
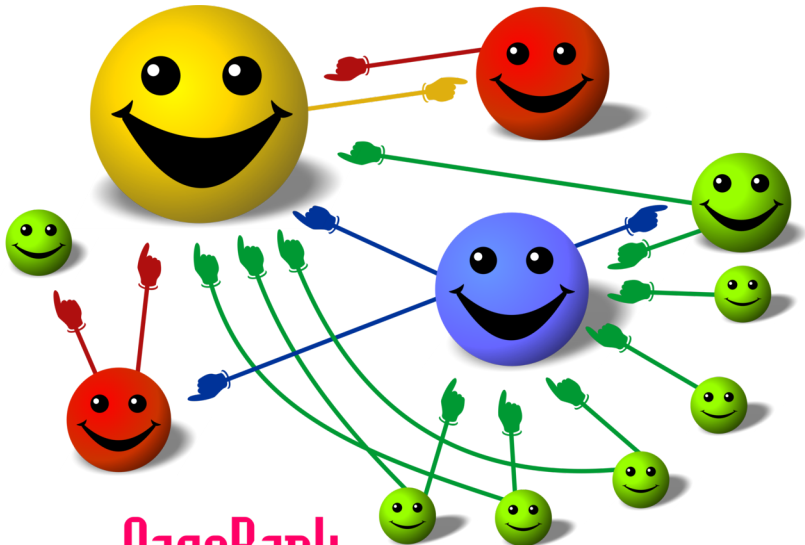
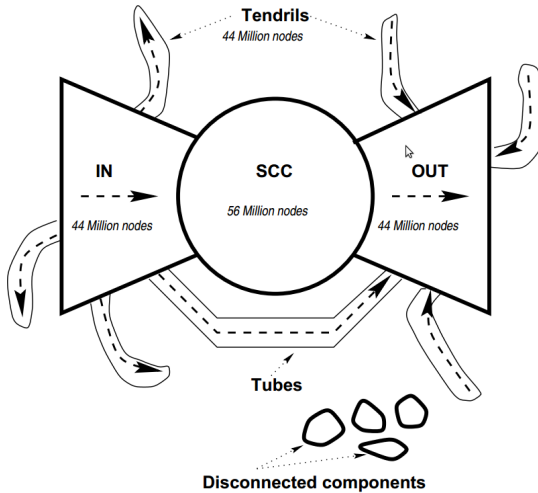


Figure: (a) Core/periphery structure and (b) Community structure.



PageRank

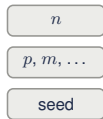
Webgraph



Broder et al., Graph structure in the web, Computer Networks 33(1): 309–320, 2000.

Network models

Input parameters



Graph model

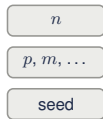
generation mechanism
ER, WS, BA, CM, ...

Synthetic network



Network models

Input parameters



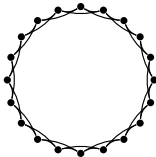
Graph model

generation mechanism
ER, WS, BA, CM, ...

Synthetic network

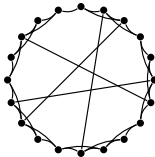


Regular

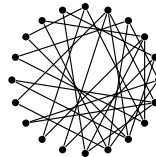


$p_r = 0$

Small-world



Random



$p_r = 1$

Increasing randomness

Based on: Watts, D., Strogatz, S. Collective dynamics of 'small-world' networks. Nature 393, 440–442 (1998).

Lancichinetti–Fortunato–Radicchi (LFR) graphs

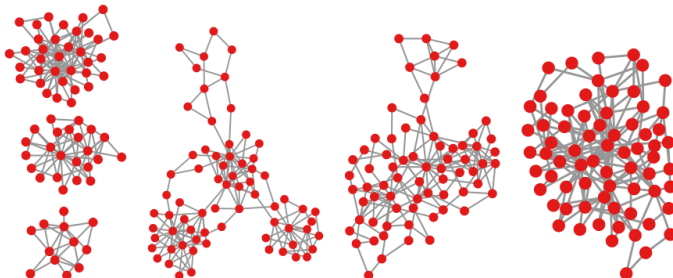
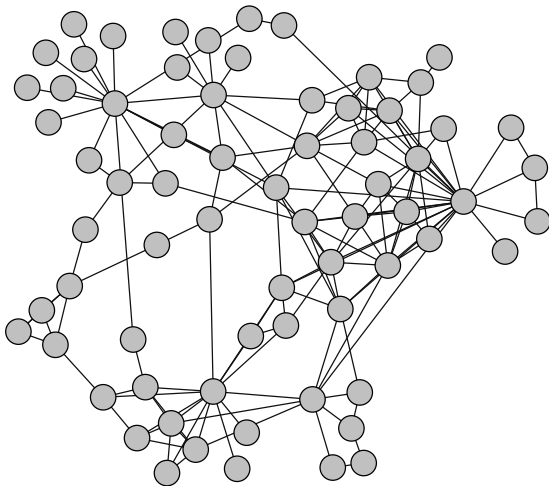


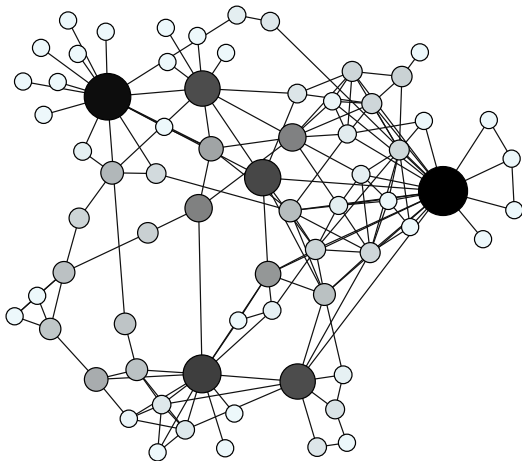
Figure: LFR models with community mixing parameter $\mu = 0, 0.05, 0.2, 1$.

Micro, meso and macro scale

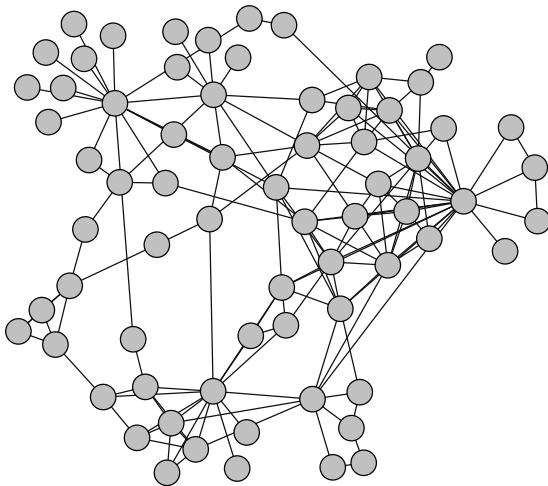
Levels of analysis in network science



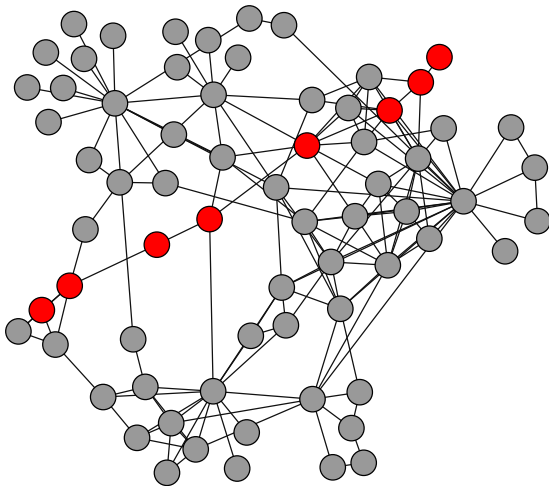
Micro scale



Macro scale



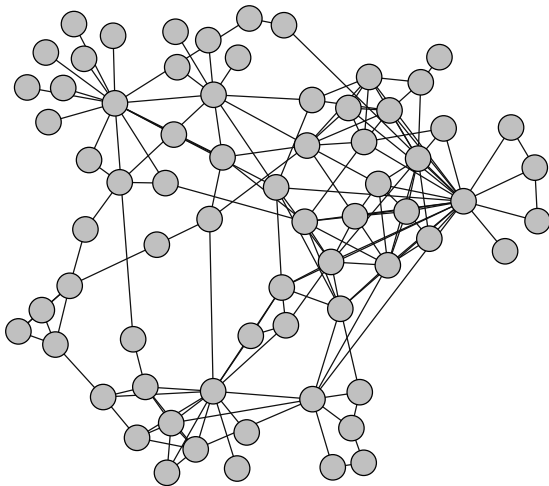
Macro scale



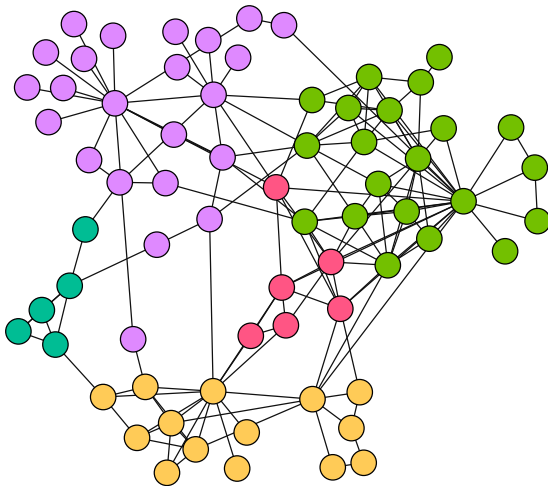
Levels of analysis in network science

- Level $\approx$ Scale
- **Micro scale:** individual nodes local
Degree, clustering, node centrality, ...
- **Meso scale:** network substructures
Community structure, core/periphery, motifs, ...
- **Macro scale:** network as a whole global
Density, reciprocity, average distance, diameter, ...

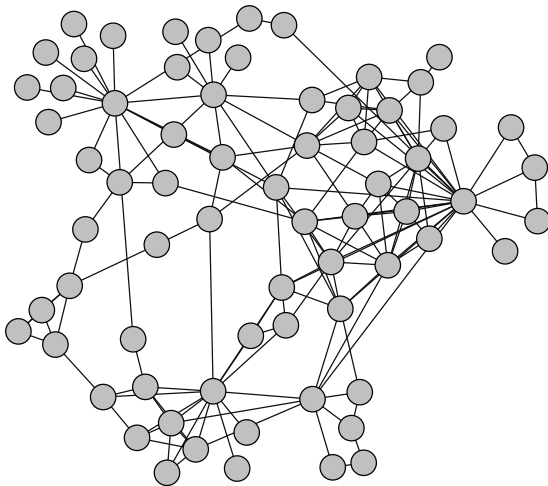
Meso scale: communities



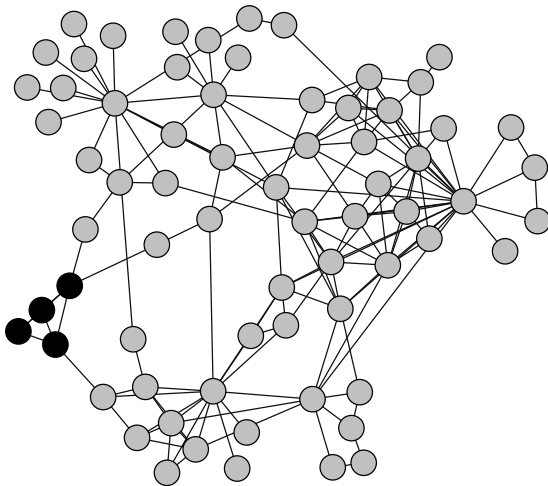
Meso scale: communities



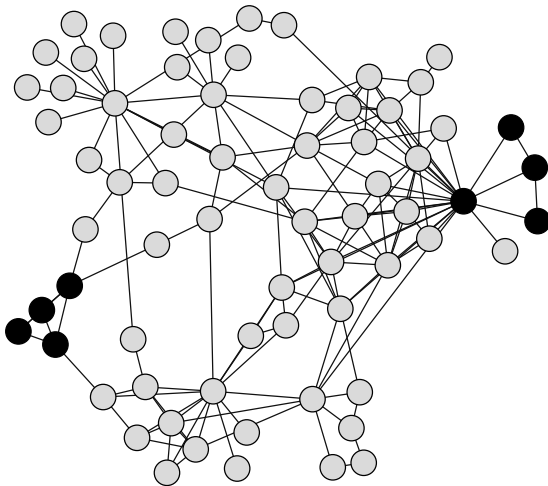
Meso scale: motifs



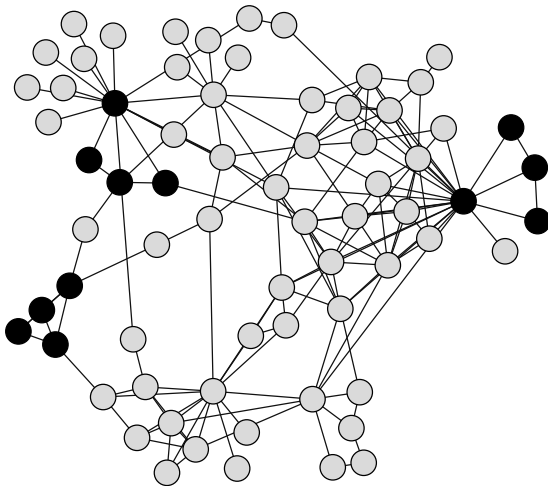
Meso scale: motifs



Meso scale: motifs



Meso scale: motifs



Network centralization

Centrality: micro scale

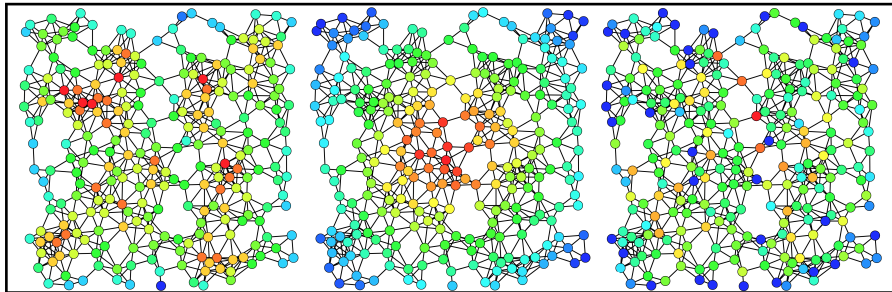


Figure: Degree, closeness and betweenness centrality.

- Centrality says something about **individual nodes**
- Can we say something about the **network as a whole**?

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

Centralization: macro scale

- **Centralization**: extent to which one node is much more central than all other nodes

Centralization: macro scale

- **Centralization**: extent to which one node is much more central than all other nodes
- Centralization is macro-scale counterpart of micro-scale centrality

Centralization: macro scale

- **Centralization**: extent to which one node is much more central than all other nodes
- Centralization is macro-scale counterpart of micro-scale centrality
- Most centralized network is **star** network: one node connected to all other nodes and no other edges
- Two approaches:

Centralization: macro scale

- **Centralization**: extent to which one node is much more central than all other nodes
- Centralization is macro-scale counterpart of micro-scale centrality
- Most centralized network is **star** network: one node connected to all other nodes and no other edges
- Two approaches:
 - 1 Information-theoretic: entropy of the degree distribution
 - 2 Centrality-based: compare the most central node with all other nodes

Freeman, Centrality in social networks: conceptual clarification, Social Networks 1(3): 215–239, 1978.

Information-theoretic centralization

- Degree sequence as probability distribution: $p_v = \deg(v)/2m$
- Example: degrees $[4, 2, 2, 1, 1] \rightarrow [0.4, 0.2, 0.2, 0.1, 0.1]$
- Shannon entropy $H = -\sum_{v \in V} p_v \log p_v$
- **Lower entropy** $\rightarrow$ fewer bits needed to describe the distribution, i.e., it compresses well $\rightarrow$ a few nodes attract most edges $\rightarrow$ **more centralized**
- Problem: network size-dependent and no meaning to value itself

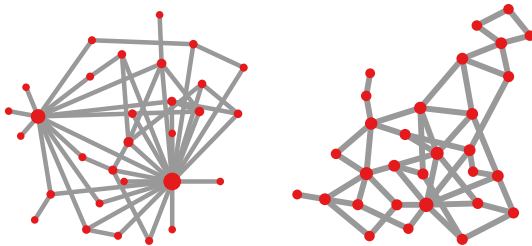


Figure: Size proportional to degree. Entropy: (left) 2.98, (right) 3.30.

Centrality-based centralization

- Given centrality measure C and most central node v^* in graph $G = (V, E)$
- Sum the differences between v^* and all other nodes, and divide by the largest possible sum over star graph $G' = (V', E')$ with the same number of nodes:

$$Cent_C(G) = \frac{\sum_{v \in V} (C(v^*) - C(v))}{\sum_{v' \in V'} (C'(v'^*) - C'(v'))}$$

- (Assuming the star graph has the highest of such differences)
- Value between 0 (all nodes equally central) and 1 (star)
- Any centrality measure can be used: degree, closeness, betweenness, ...

Centrality-based centralization

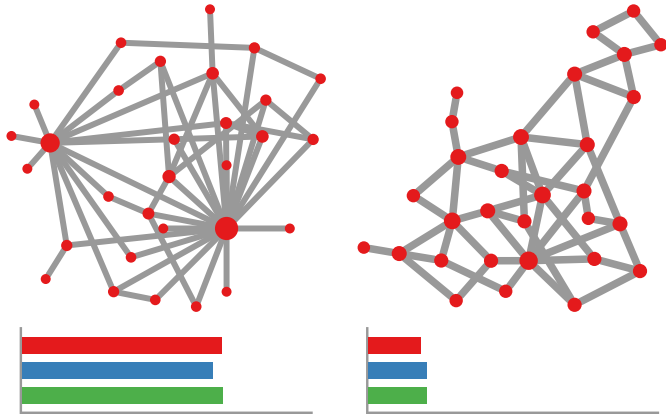


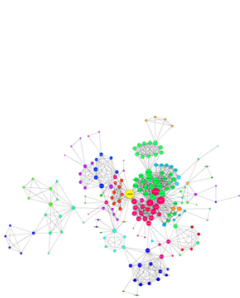
Figure: Centralization of two networks according to degree (red), betweenness (blue) and closeness (green) centrality; bar length is the centralization value.

Network evolution

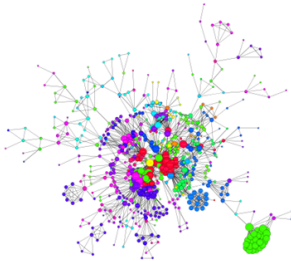
Network evolution

- Examples of **network evolution**
 - Online 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, linking the authors
 - Human communication networks: people send messages to new contacts
- Also called **temporal network analysis**: how does the network structure change over time?

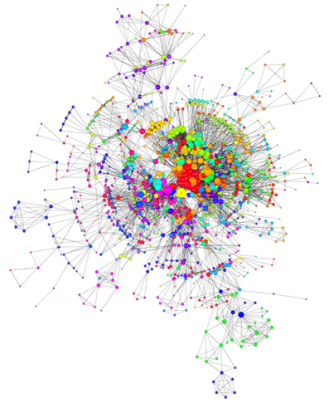
Apple collaboration network



2007-2008



2009-2010



2011-2012

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

Temporal networks

- **Temporal network** $G^t = (V^t, E^t)$ with 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 over time **most common**
- Timestamp on node $v \in V$:
 $\tau(v) \in [0, T]$
- Timestamp on edge $e \in E$: **most common**
 $\tau(e) \in [0, T]$, 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

Temporal/dynamic networks and snapshots

- **Temporal network**: network in which each edge e has a timestamp $\tau(e)$, the moment it appears; once added, edges remain
- **Dynamic network**: network in which edges (and nodes) can appear and also disappear again; each edge has one or more intervals during which it is active, usually stored as edge attributes
- **Snapshot**: the network as it is at time t
 - Cumulative snapshot $G_{\leq t}$: edges with $\tau(e) \leq t$ **most common**
 - Sliding window snapshot $G_{(t-\Delta, t]}$ of width Δ : only edges with $t - \Delta < \tau(e) \leq t$

Levels of network 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 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 (recap)

- Preferential attachment (PA)
- Cumulative advantage
- Rich get richer

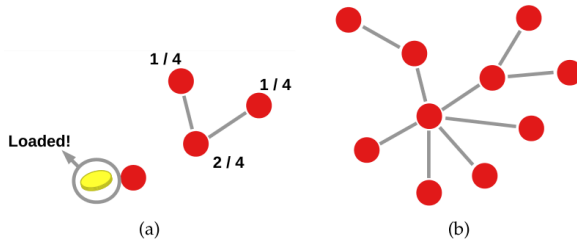
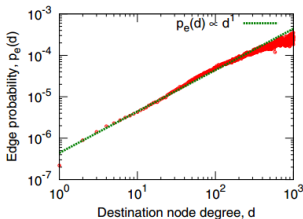
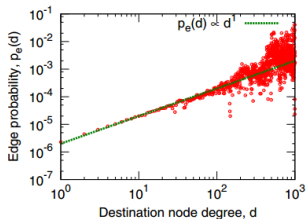


Figure: PA: target node based on degree (a), graph formed through PA (b).

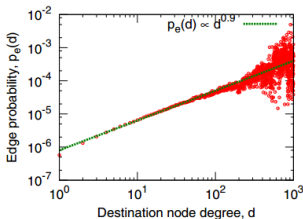
Preferential attachment: degree



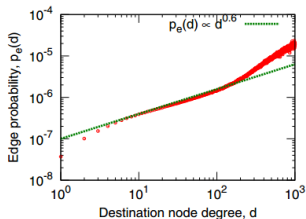
(c) FLICKR



(d) DELICIOUS

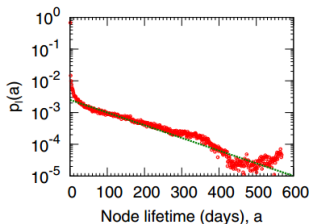


(e) ANSWERS

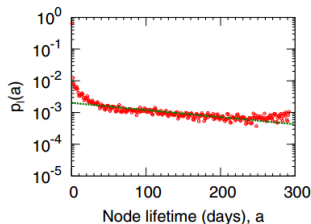


(f) LINKEDIN

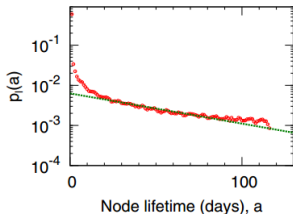
Preferential attachment: age



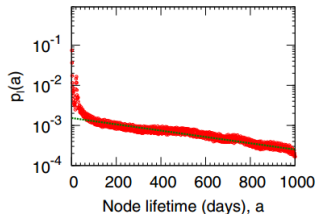
(a) FLICKR



(b) DELICIOUS

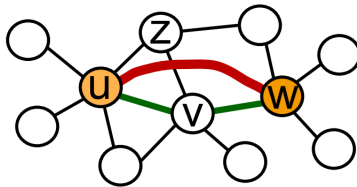


(c) ANSWERS

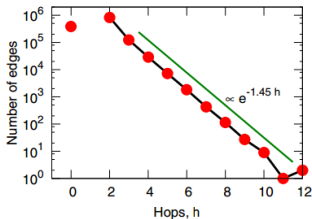


(d) LINKEDIN

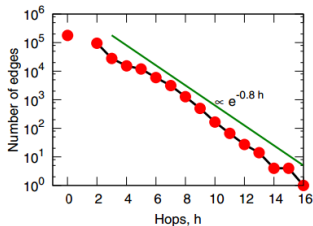
Triadic closure



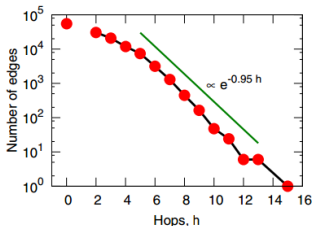
Preferential attachment: hops



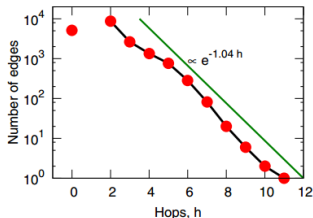
(c) FLICKR



(d) DELICIOUS



(e) ANSWERS



(f) LINKEDIN

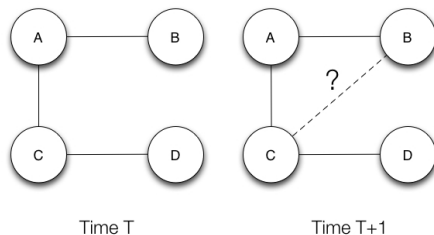
Microscopic evolution model

- 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 (topic will be covered later)



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

Levels of network evolution

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

Macroscopic evolution

- Look at evolution of network as a whole
- Observe different characteristic graph 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



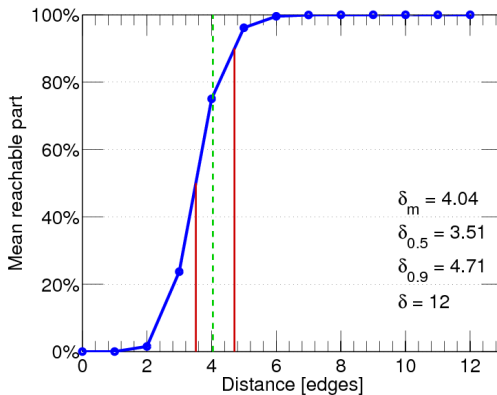
2001: Enron trading in hundreds of commodities
Interests in: USA, South America, Europe, Asia and Australia



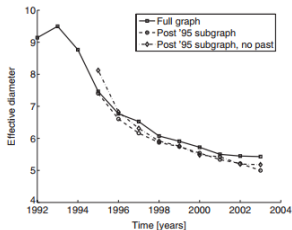
Macroscopic patterns

- Densification: density increases over time
- The fraction of nodes in the giant component grows over time
- 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 connected node pairs are at distance $\delta_{0.9}$ or less
 - Diameter: longest shortest path length

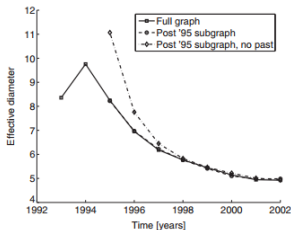
Effective diameter



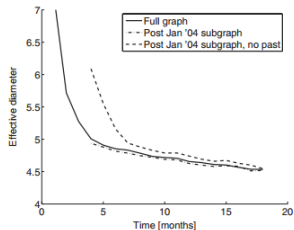
Effective diameter



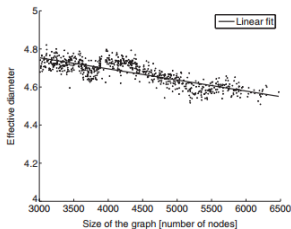
(a) arXiv citation graph



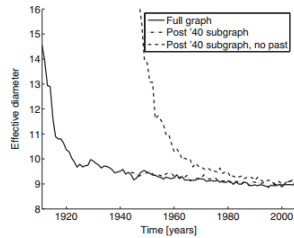
(b) Affiliation network



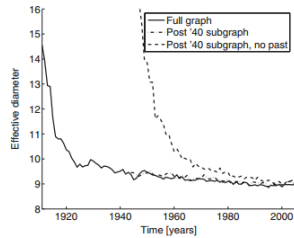
(c) Patents citation graph



(d) Autonomous Systems

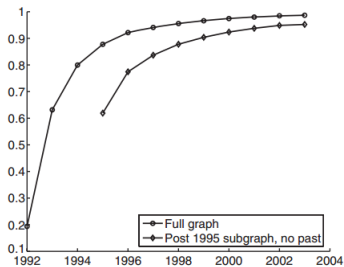


(e) Email network

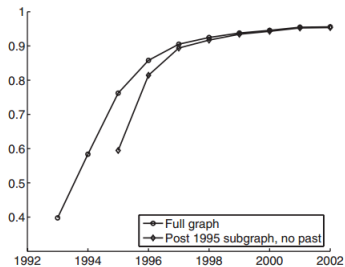


(f) IMDB actors to movies network

Giant component

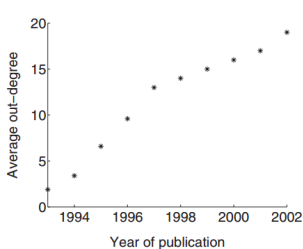


(a) arXiv citation graph

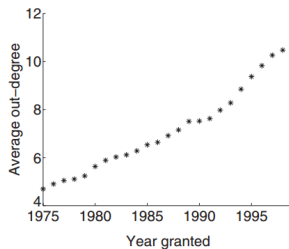


(b) Affiliation network

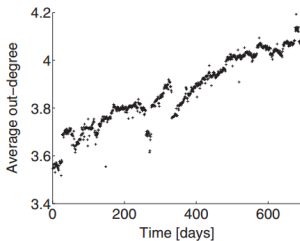
Densification



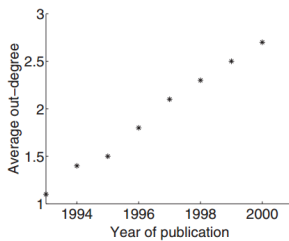
(a) arXiv



(b) Patents



(c) Autonomous Systems



(d) Affiliation network

Measurement error

Using real data to construct networks

- Real-world data is messy
- **Zachary's karate club**
 - Nodes: 34 members of a karate club
 - Edges: members who interacted outside the club
 - Club split after conflict between coach and president

Using real data to construct networks

- Real-world data is messy
- **Zachary's karate club**
 - Nodes: 34 members of a karate club
 - Edges: members who interacted outside the club
 - Club split after conflict between coach and president
- 77 or 78 edges? The adjacency matrix in the original paper is not symmetric

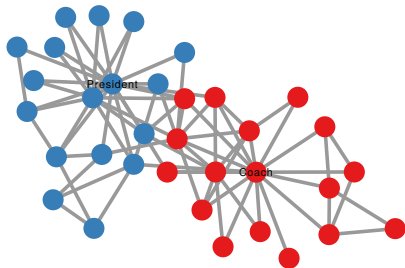


Figure: Zachary's karate club; node color shows which side each member joined after the split.

Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* 33(4): 452–473, 1977.

Measurement errors in networks

Measurement errors in networks

- 1 **Missing nodes/edges:** entities or relations that exist, but were not recorded (false negatives)

Measurement errors in networks

- 1 **Missing nodes/edges:** entities or relations that exist, but were not recorded (false negatives)
- 2 **Additional nodes/edges:** recorded entities or relations that do not exist (false positives)

Measurement errors in networks

- 1 **Missing nodes/edges:** entities or relations that exist, but were not recorded (false negatives)
- 2 **Additional nodes/edges:** recorded entities or relations that do not exist (false positives)
- 3 **Duplicated nodes:** one entity split into multiple nodes, e.g., an actor who also uses a pseudonym

Measurement errors in networks

- 1 **Missing nodes/edges:** entities or relations that exist, but were not recorded (false negatives)
- 2 **Additional nodes/edges:** recorded entities or relations that do not exist (false positives)
- 3 **Duplicated nodes:** one entity split into multiple nodes, e.g., an actor who also uses a pseudonym
- 4 **Aggregated nodes:** multiple entities merged into one node, e.g., a band instead of its musicians

Measurement errors in networks

- 1 **Missing nodes/edges:** entities or relations that exist, but were not recorded (false negatives)
- 2 **Additional nodes/edges:** recorded entities or relations that do not exist (false positives)
- 3 **Duplicated nodes:** one entity split into multiple nodes, e.g., an actor who also uses a pseudonym
- 4 **Aggregated nodes:** multiple entities merged into one node, e.g., a band instead of its musicians
- 5 **Misestimated features:** wrong edge weight or direction, or wrong node or edge attributes

Measurement errors in networks

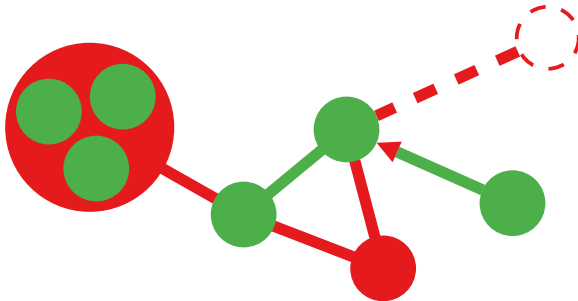


Figure: A network with all types of measurement error. Correct nodes and edges are **green**, errors are **red**: a missing node and edge (dashed), an additional node and edges (solid **red**), an aggregated node hiding three nodes (large), and an edge with the wrong direction (**red** arrowhead).

Dealing with measurement error

- At least two approaches:
 - 1 **Probabilistic networks**: model the uncertainty in the network itself
 - 2 **Assessing robustness empirically**: measure how much errors affect the outcome

Probabilistic networks

- **Probabilistic network** $G = (V, E, \Pi)$:
each edge $(u, v) \in E$ exists with probability $p_{u,v}$
- Each edge exists or not: $2^{|E|}$ **possible worlds**, each with its own probability
- Compute a property in each world and weigh it by the world's probability: a **distribution** instead of a single number

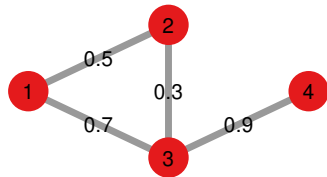


Figure: Edge labels give the probability that the edge exists.

Assessing robustness empirically

- Input: a network and a network analysis task, e.g., ranking nodes by centrality or detecting communities
- Perturb the network by introducing one of the five types of measurement error, e.g., remove 10% of the edges at random
- Perform the task on both the original and the perturbed network
- Check whether the results remain the same, e.g., the overlap of the top- k nodes or the similarity of the partitions
- Repeat for different error types and amounts of error: how much error can the result tolerate before metric values change too much?

Network sampling

The sampling conundrum

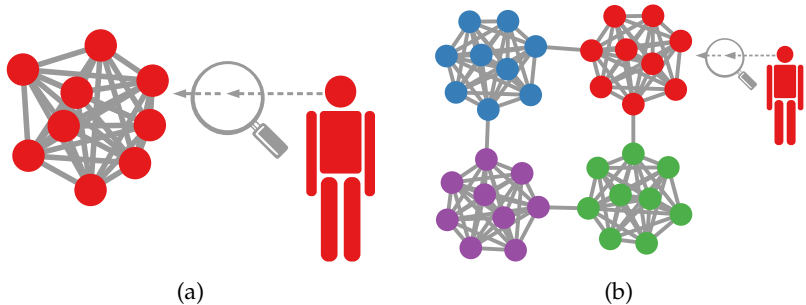


Figure: (a) The sample we are able to observe looks like a clique-ish hairball. (b) The full network has a clear community structure that the sample hides.

Why sample?

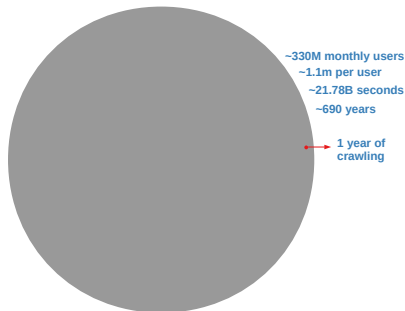


Figure: All Twitter users (gray circle) versus what a non-stop one-year crawl under the API rate limits would collect (tiny red circle): crawling everything would take about 690 years.

Network sampling

- **Network sampling:** to extract from your network of n nodes a smaller version with n_s nodes that is a representative subset of the data
- Relevant structural network properties should be preserved: the property you study should be roughly the same in the sample as in the full network
- Sample size $n_s \ll n$

Network sampling

- **Network sampling:** to extract from your network of n nodes a smaller version with n_s nodes that is a representative subset of the data
- Relevant structural network properties should be preserved: the property you study should be roughly the same in the sample as in the full network
- Sample size $n_s \ll n$
- Two approaches:
 - 1 **Induced sampling techniques:** specify the elements (nodes or edges) that must be in the sample, then collect everything directly connected to them
 - 2 **Topological sampling techniques:** start from a seed node and explore the network along its edges, possibly far away from the seed

Induced sampling

- **Induced subgraph:** a subset of the nodes plus all edges between them (remember?)

Induced sampling

- **Induced subgraph**: a subset of the nodes plus all edges between them (remember?)
- 1 **Node-induced sampling**: choose a set of nodes, e.g., uniformly at random, and collect all their neighbors

Induced sampling

- **Induced subgraph**: a subset of the nodes plus all edges between them (remember?)
- 1 **Node-induced sampling**: choose a set of nodes, e.g., uniformly at random, and collect all their neighbors
 - Problems: multiple components and most nodes have a low degree, so hubs are underrepresented
- 2 **Edge-induced sampling**: choose a set of edges, e.g., uniformly at random, and collect their nodes and neighbors
 - Counteracts the degree bias of node-induced sampling: a random edge is likely attached to a hub
 - Impractical: APIs rarely let you start from a random edge

Induced sampling

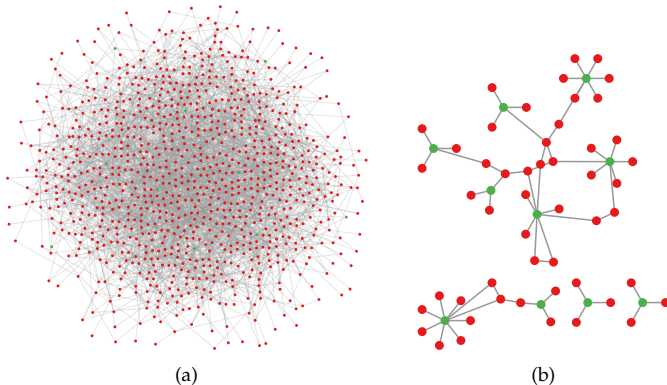


Figure: Node-induced sampling. (a) A network with 1,000 nodes, of which 1% is selected uniformly at random (green). (b) The induced graph of the selected nodes, their neighbors and all edges between them.

Topological sampling

- Topological sampling starts from a single seed node and explores the graph along its edges, possibly far from the seed
- Several approaches:
 - 1 Snowball sampling BFS variant
 - 2 Forest fire sampling BFS variant
 - 3 Random walk sampling random walk
 - 4 Metropolis–Hastings random walk random walk variant

Snowball sampling

- **Snowball sampling:** BFS with a cap of k neighbors per node until sample size n_s is reached
- Start from a seed node and ask it to name k of its neighbors (not all of them). Ask each newly named node the same question, in BFS order
- Reaches hubs, but each node reports at most k neighbors: the sampled degree distribution gets an unrealistic sharp cutoff
- One query per sampled node, returning at most k edges: $O(k \cdot n_s)$ for a sample of n_s nodes

Snowball sampling

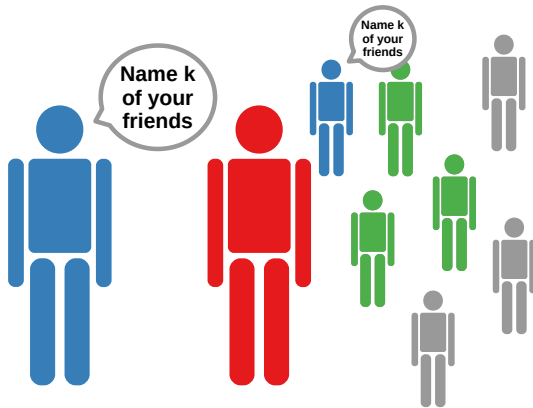


Figure: Snowball sampling. The sampler (blue) asks the seed (red) for $k = 3$ of its neighbors: these (green) are explored next in BFS order, the others (gray) are not.

Forest fire sampling

- **Forest fire sampling:** BFS in which each neighbor is explored only with **burning probability** p until sample size n_s is reached
- Start from a seed node and ask it for all its neighbors. Each neighbor **“burns”** (is explored next, in BFS order) with probability p ; otherwise it is skipped and ignored
- Problem: with a low p the fire can die out before the budget is used up
- Advantage: clustering is estimated well

Leskovec and Faloutsos, Sampling from large graphs, KDD, 631–636, 2006.

Forest fire sampling

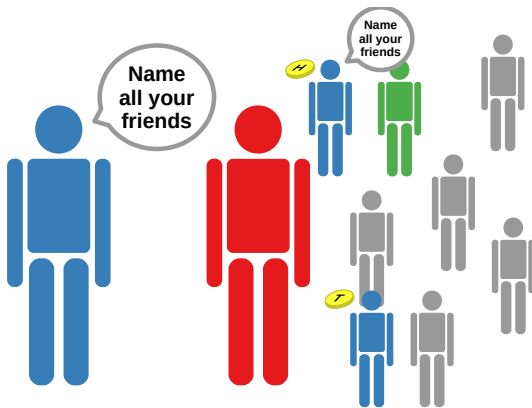


Figure: Forest fire sampling. The seed (red) names all its neighbors; each neighbor is explored only if a coin flip with the burn probability succeeds (green).

Random walk sampling

- **Random walk sampling:** sample the nodes visited by a random walk, until a total budget (sample size n_s) is used up
- Start from a seed node (red) and repeatedly move to one of its neighbors, chosen uniformly at random (green).
- To avoid getting “stuck”, incorporate random jumps
- Degree bias: higher degree nodes visited more often

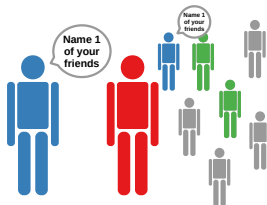


Figure: Random walk sampling.

Metropolis–Hastings random walk

- **Metropolis–Hastings random walk** (MHRW): random walk sampling in which a move to a higher-degree neighbor may be rejected, until sample size n_s is reached
- At current node v , pick a neighbor u uniformly at random and
 - Accept the move with probability $\min(1, \deg(v)/\deg(u))$; moves to lower-degree nodes are always accepted!
 - If u is rejected, try another neighbor.
- Example: a move from degree 3 to degree 100 is accepted with probability 3%
- No degree bias
- Rejected proposals also use budget: more queries than a plain random walk for the same n_s

Metropolis–Hastings random walk

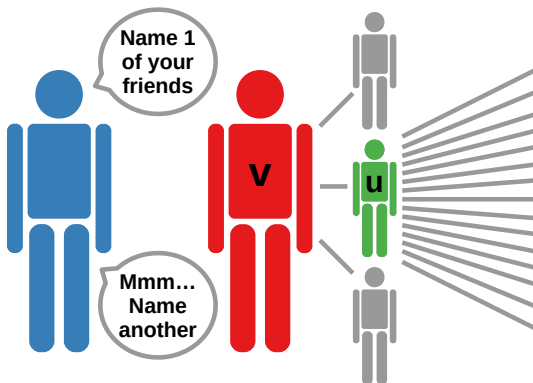


Figure: Metropolis–Hastings random walk (MHRW). A move from v to a high-degree neighbor u is accepted only with probability $\deg(v)/\deg(u)$, which removes the walk's bias towards hubs.

Comparison of algorithms

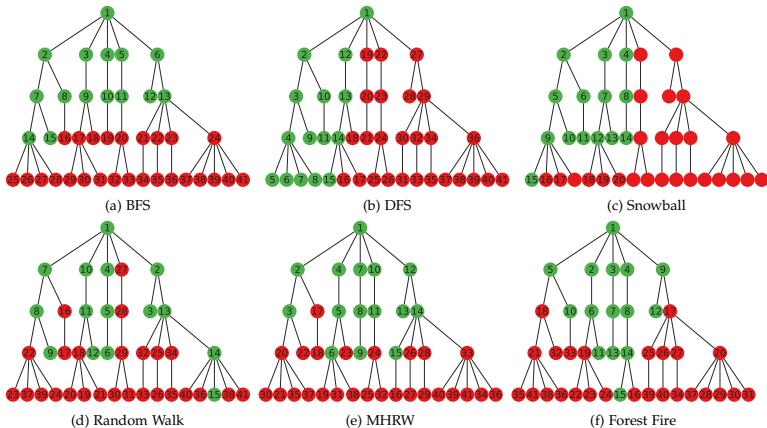


Figure: Six sampling strategies on the same network with a budget of 15 nodes. Labels give the exploration order; nodes are sampled (green) or not (red).

Summary of sampling algorithms

	Giant component	Low-degree nodes	High-degree nodes	Degree distribution	Clustering
<i>Induced</i>					
Node-induced		✓			
Edge-induced			✓		
<i>Topological</i>					
Snowball	✓		✓		
Forest fire	✓				✓
Random walk	✓		✓		*
Metropolis–Hastings random walk	✓	✓	✓	✓	?

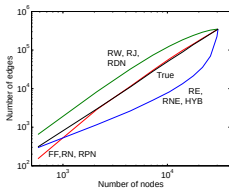
* = with low teleportation

Sampling from large graphs

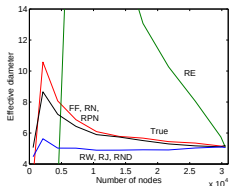
- **Scale-down goal:** sample has similar properties to original graph (what we did so far)
- **Back-in-time goal:** sample should look like the graph did in the past when it had n_s nodes
- Remember: densification, shrinking diameter, growing giant component
- Agreement between a property in sample and original graph:
Kolmogorov–Smirnov D-statistic $D = \max_x |F'(x) - F(x)|$, for cumulative distributions F and F' ; lower is better

Leskovec and Faloutsos, Sampling from large graphs, KDD, 631–636, 2006.

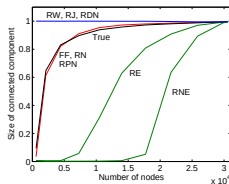
Back-in-time sampling



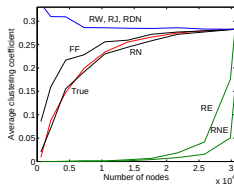
(a) T1: DPL



(b) T2: Diameter vs. time



(c) T3: Size of CC vs. time



(d) T5: Clustering vs. time

Figure: Samples of a citation network compared to the true network as it grew over time ("True"). Forest fire (FF) performs well for (a) densification, (b) shrinking diameter, (c) growing giant component and (d) clustering.

Leskovec and Faloutsos, Sampling from large graphs, KDD, 631–636, 2006.

Sampling from large graphs: empirical findings

- **Scale-down goal:** random walk based methods work best: biased towards high-degree nodes, and the sample is connected
- **Back-in-time goal:** forest fire works best (average $D = 0.13$)
- The best burning probability for forest fire depends on the goal:
 - $p \approx 0.7$ for scale-down: large fires that explore much of a node's vicinity
 - $p \approx 0.2$ for back-in-time: small fires, closer to random node sampling
- A sample of about 15% of the nodes is usually enough
- No single best method: the choice depends on the goal and the properties of interest

Leskovec and Faloutsos, Sampling from large graphs, KDD, 631–636, 2006.

Assignment feedback

- Please study and utilize detailed assignment feedback
- Grade < 5.0 : insufficient; compensate with extra assignment
- Grade ≥ 5.5 : sufficient

Assignment feedback

- Please study and utilize detailed assignment feedback
- Grade < 5.0 : insufficient; compensate with extra assignment
- Grade ≥ 5.5 : sufficient
- $5.0 \leq \text{Grade} < 5.5$: insufficient, unless compensated with Assignment 2 to average of two assignments ≥ 5.5

Assignment feedback

- Please study and utilize detailed assignment feedback
- Grade < 5.0 : insufficient; compensate with extra assignment
- Grade ≥ 5.5 : sufficient
- $5.0 \leq \text{Grade} < 5.5$: insufficient, unless compensated with Assignment 2 to average of two assignments ≥ 5.5
- Questions? Ask your grader during the lab session (name present on your graded work)
Use the feedback to your advantage in Assignment 2!

Upcoming lab session and “Homework”

- Practice session 2 in ANS
- Get started with Assignment 2; remember you are allowed to work in teams of exactly two students

Credits

- Several images are gratefully reused from Michele Coscia, The Atlas for the Aspiring Network Scientist, v3 of <https://arxiv.org/abs/2101.00863>