# Social Network Analysis for Computer Scientists

Frank Takes

LIACS, Leiden University

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

## Lecture 5 — Network evolution and human traversal

# Assignment feedback

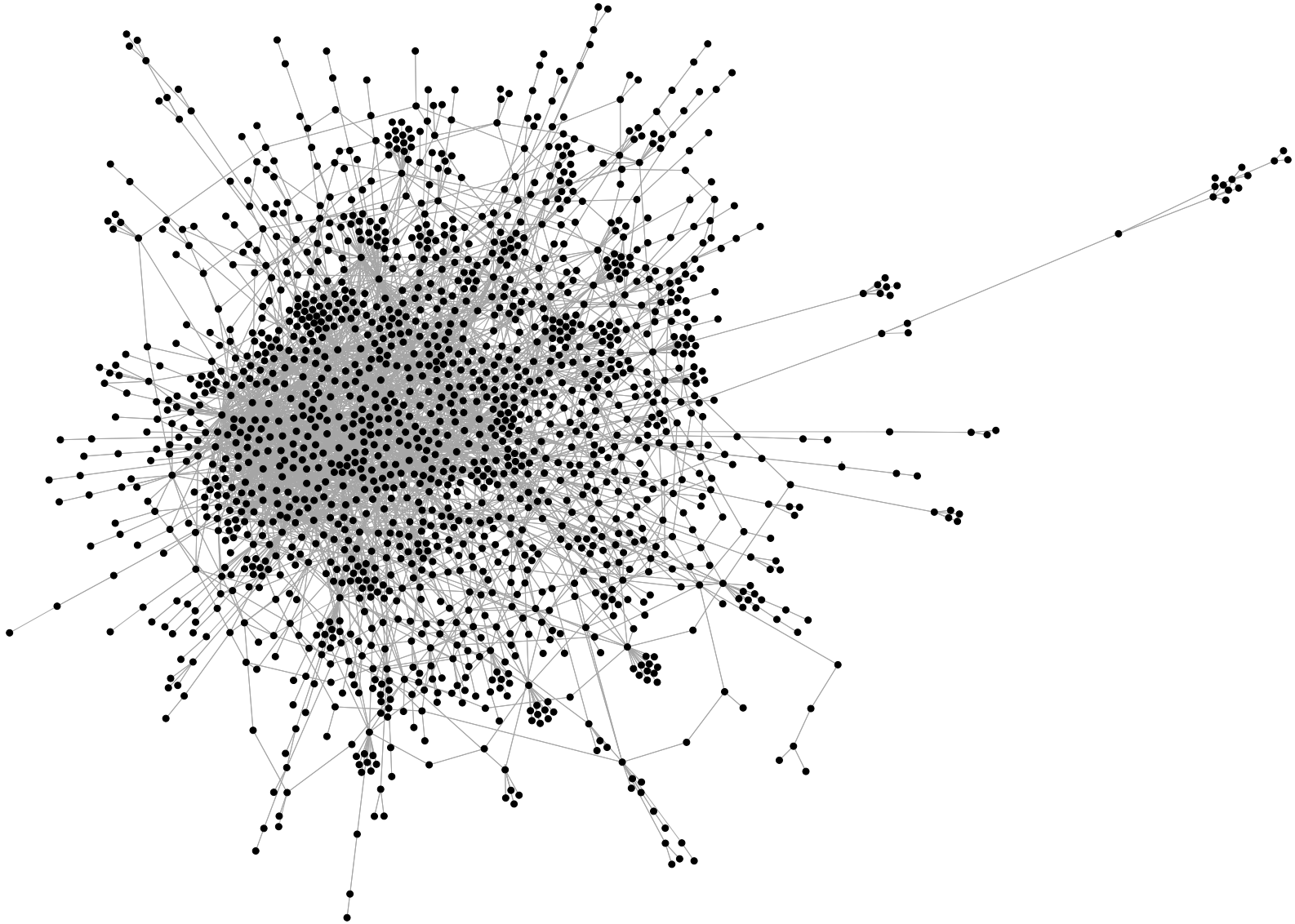
- Please study and utilize detailed assignment feedback
- Grade $< 5.0$: insufficient; compensate with extra assignment
- $5.0 \leq$ Grade $< 5.5$: insufficient, unless compensated with Assignment 2 to average of two assignments $\geq 5.5$
- Grade $\geq 5.5$: sufficient
- Questions? Ask your grader during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

# Today

- Recap
- Temporal networks
- Network models
- Network dynamics and evolution
- Network science challenges
- Break: Presentation schedule

# 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 — $\lvert V \rvert$ | $n$ |
| Number of edges — $\lvert E \rvert$ | $m$ |
| 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

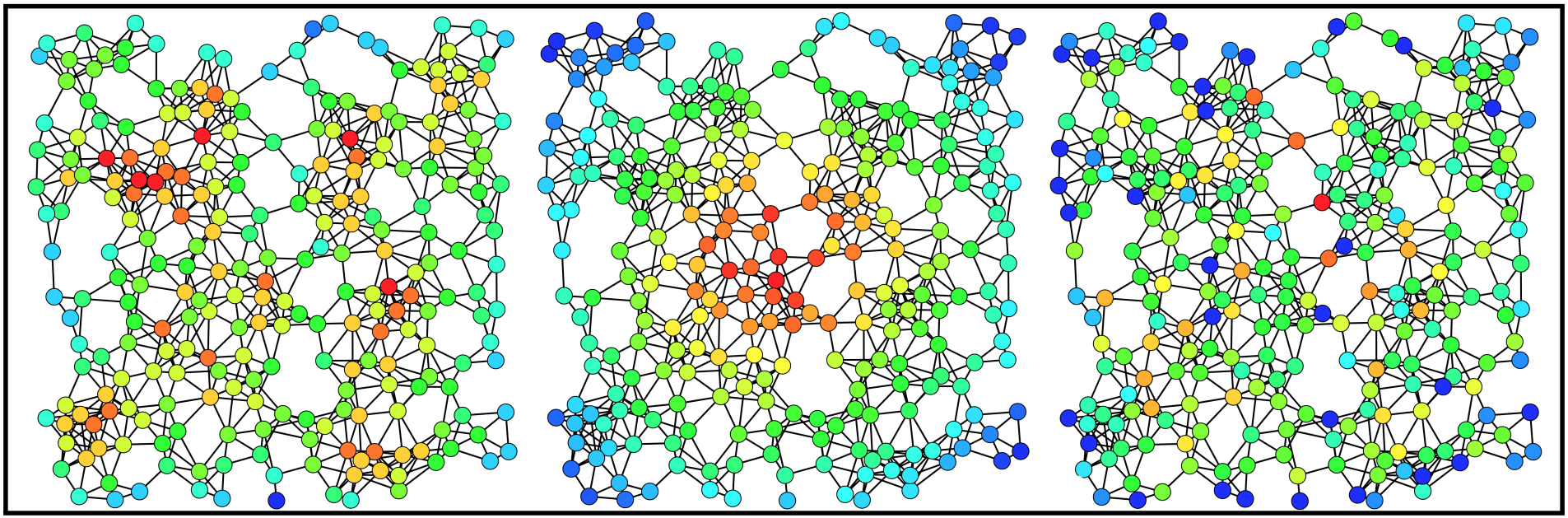
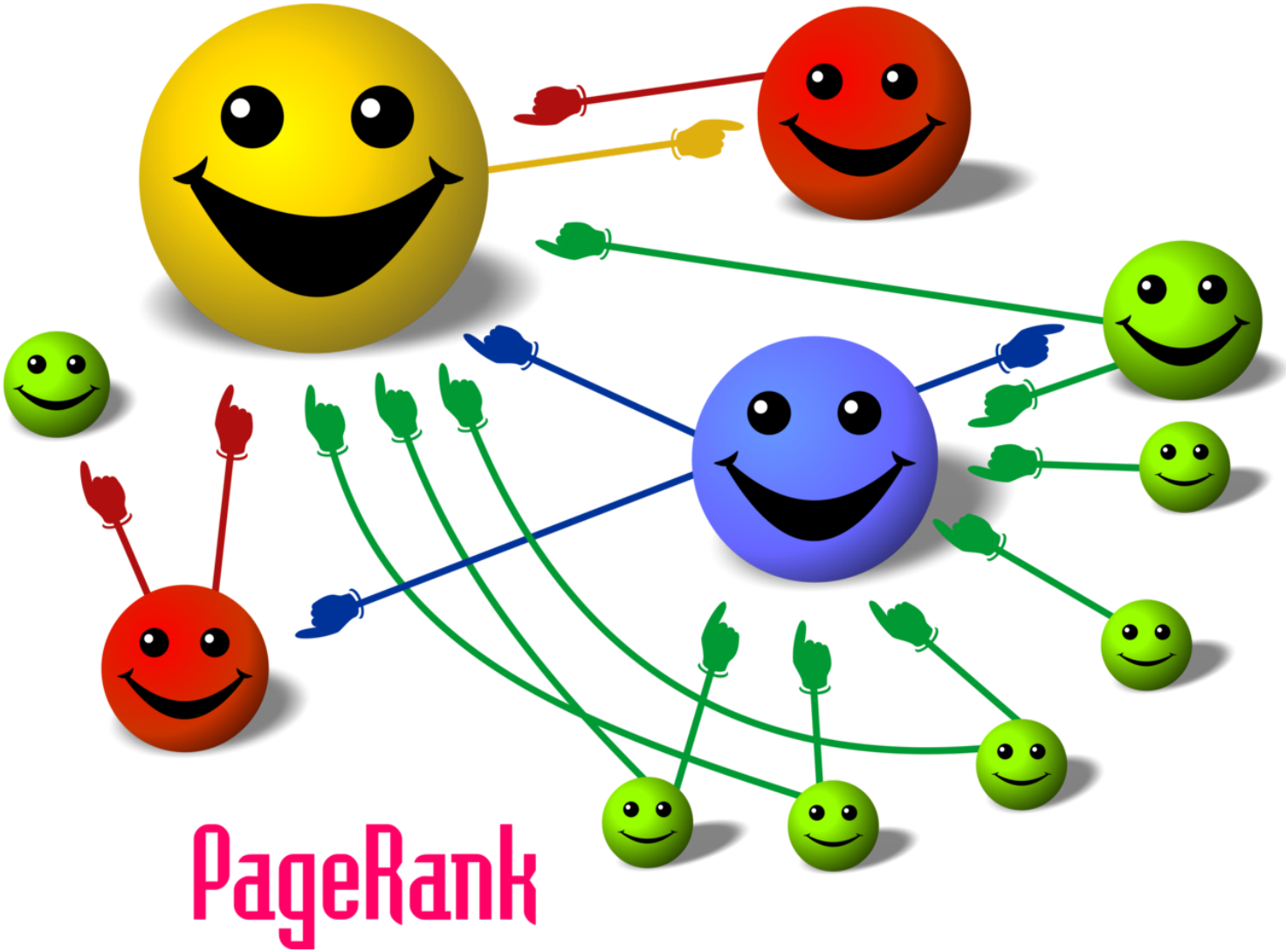


Figure: Degree, closeness and betweenness centrality

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

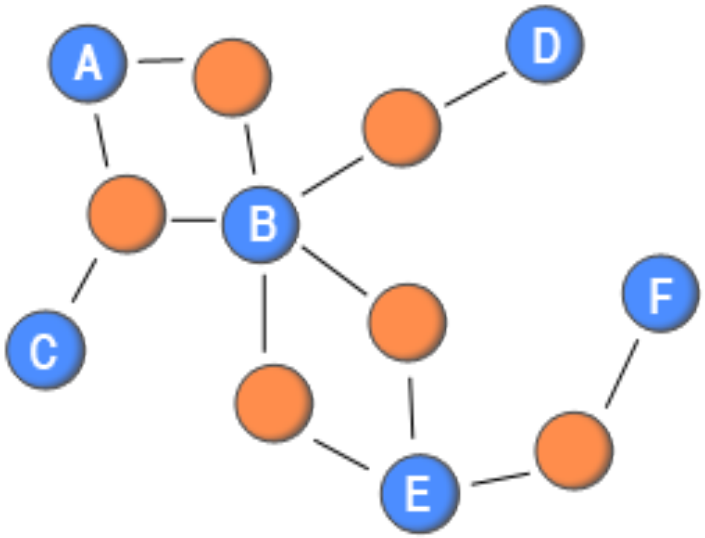
# Centrality measures: PageRank



# 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

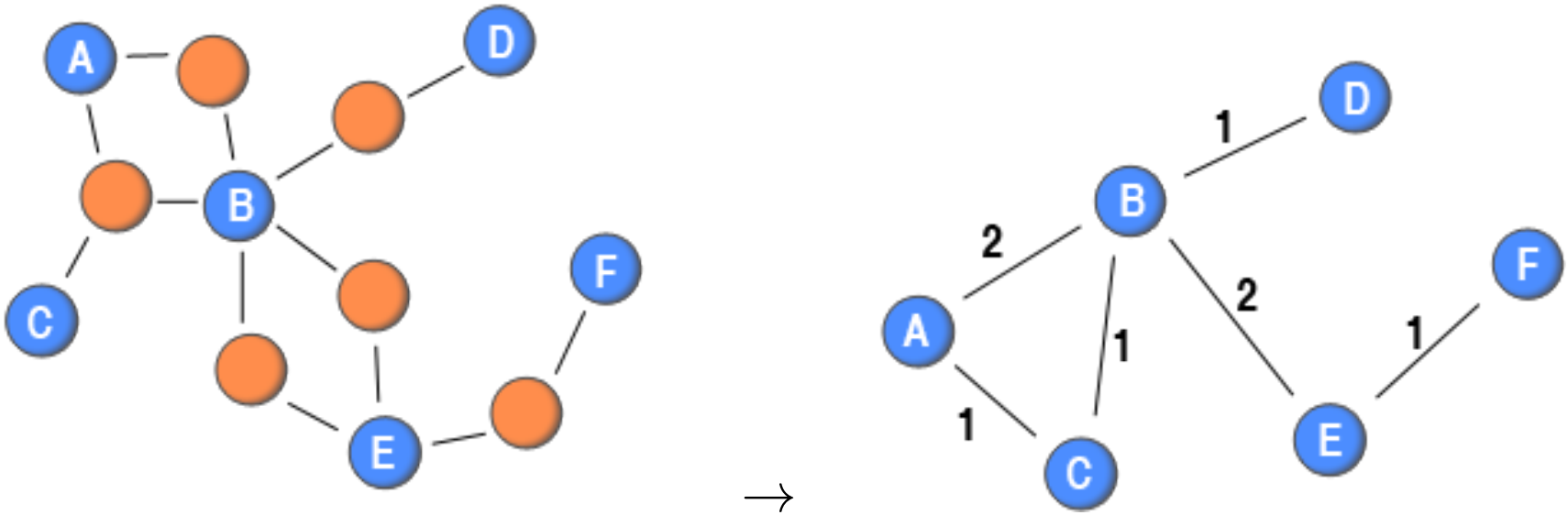



Image: `http://toreopsahl.com`

# Community detection

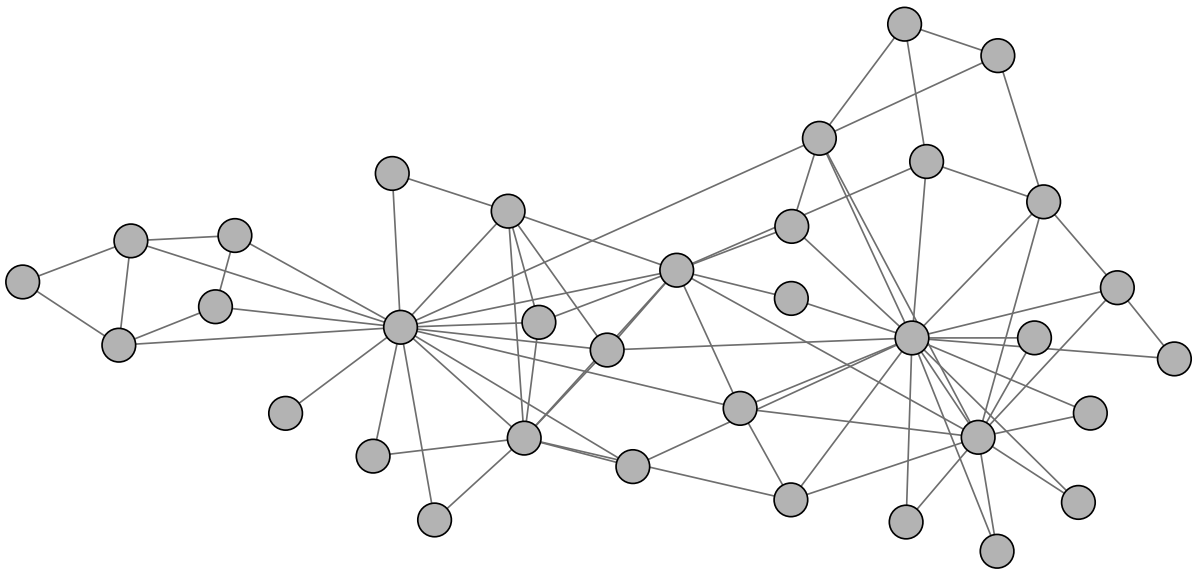


Figure: Communities: node subsets connected more strongly with each other

# Community detection

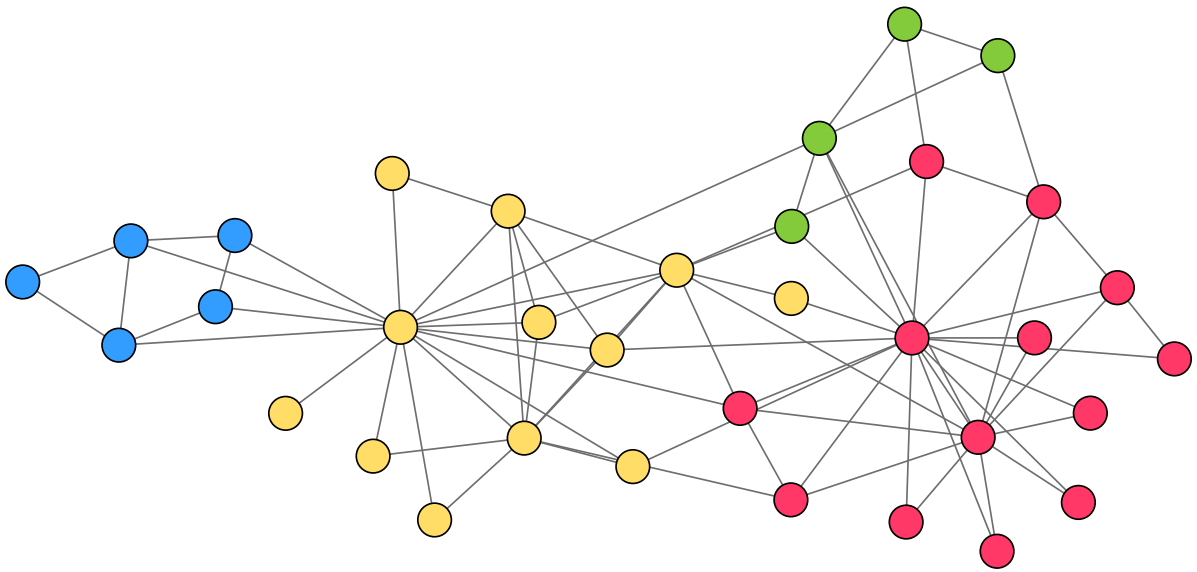
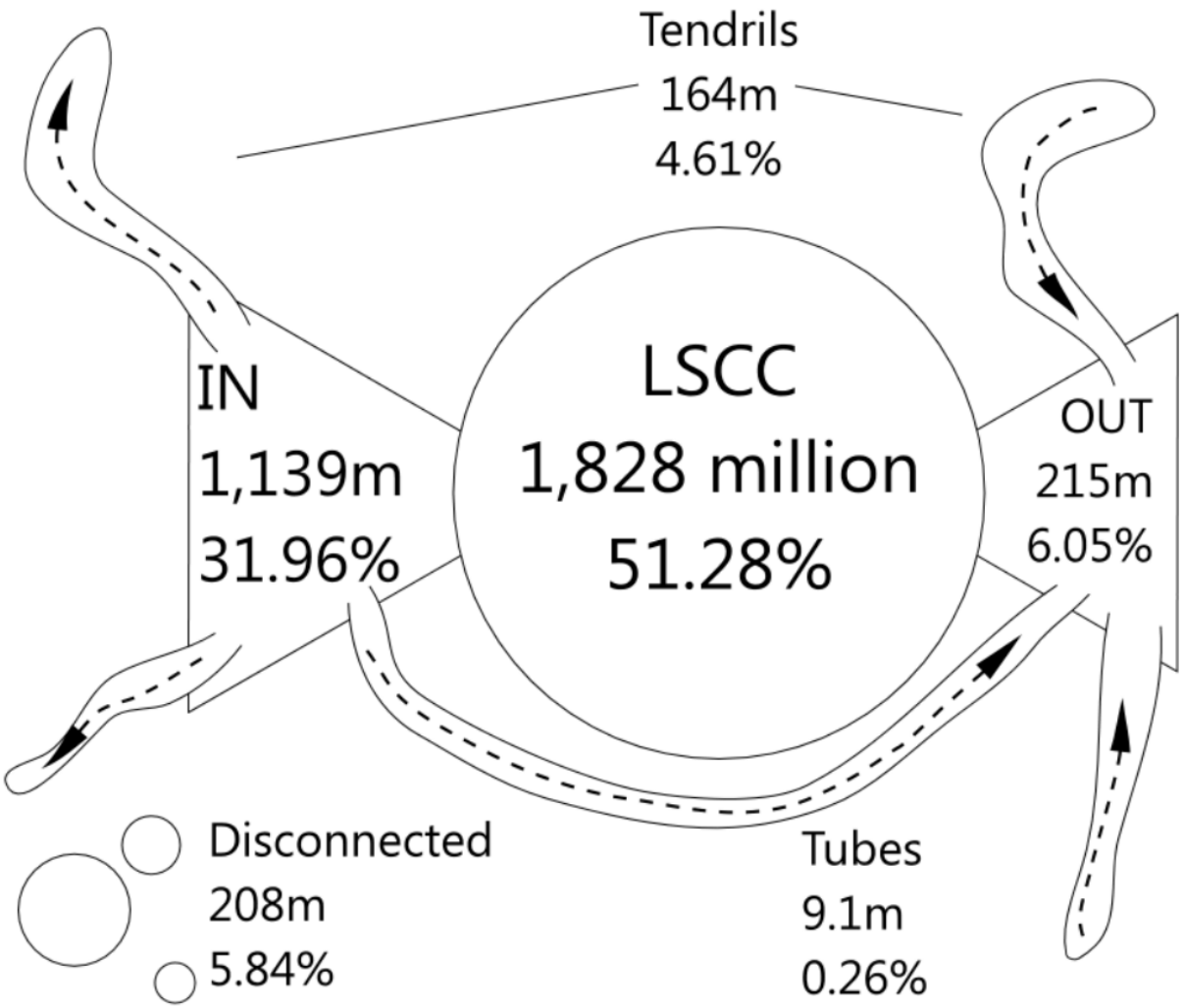


Figure: Communities: node subsets connected more strongly with each other

# Bow-tie structure of the web




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

# News last week

Tech & Innovation • Sep 29, 3:03 PM • Updated Sep 29, 3:03 PM

## Inventor of the World Wide Web warns: 'We are at a new crossroads'

Author : BNR Web Editor

**In 1990, Tim Berners-Lee conceived the idea for the World Wide Web. The goal was to make the World Wide Web freely available to everyone. But now, 35 years later, he sees his invention being used to gather data that leads to "real-life violence, disinformation, seriously harms psychological well-being, and undermines social cohesion."**

# 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
- What happens to their structure over time?

# 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 (**most common**)
- 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 approaches to (temporal) network analysis

- **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
  - Focus on empirical data from an actual network
  - Compute and interpret structural properties (over time)
  - (Compare with null model or similar networks)
  - Interpret computational analyses

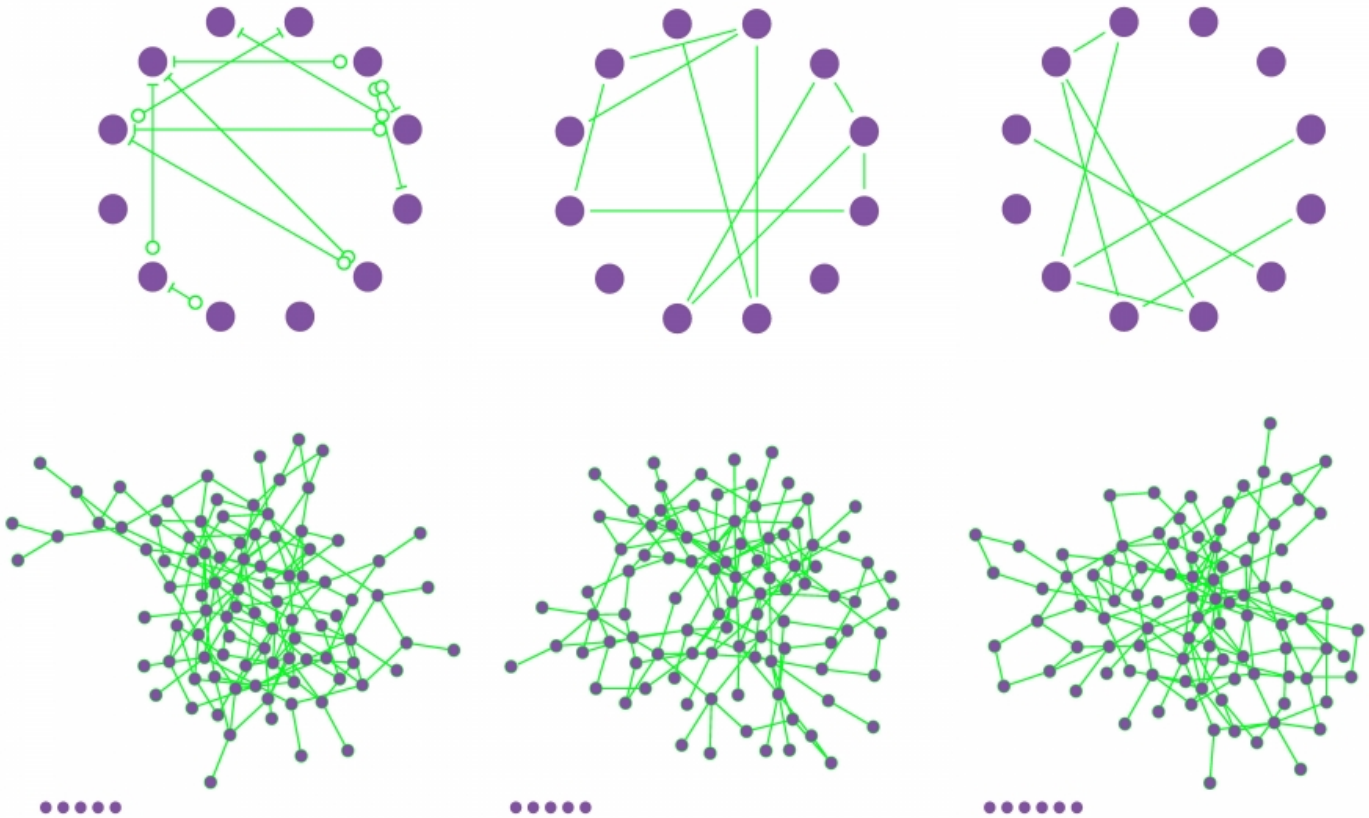
# Three models

- 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

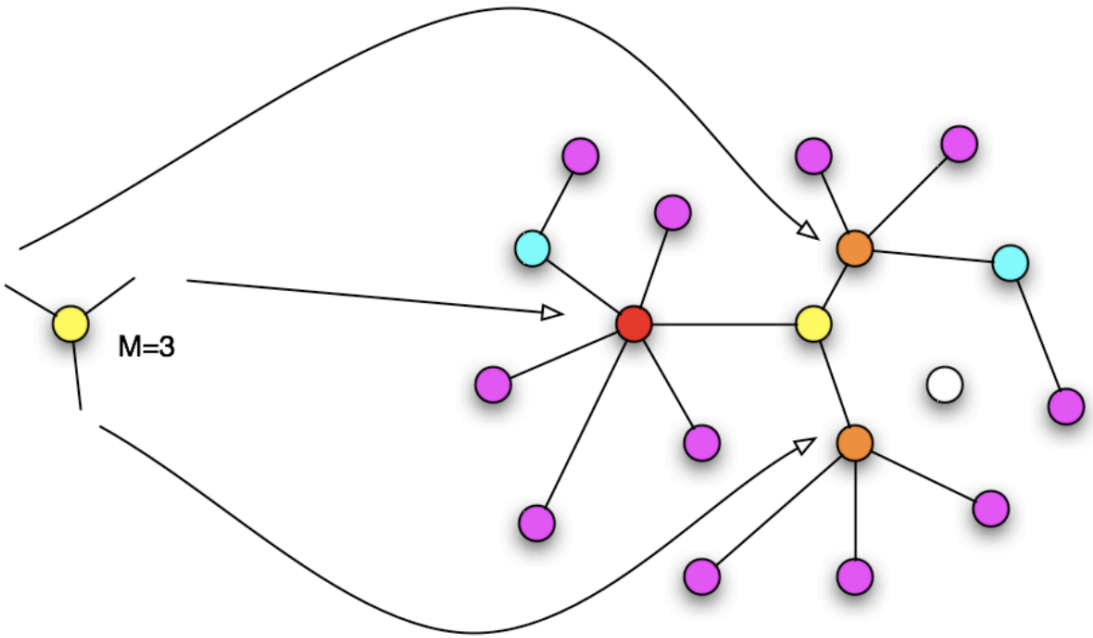
# Barábasi-Albert model (1999)

- "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:

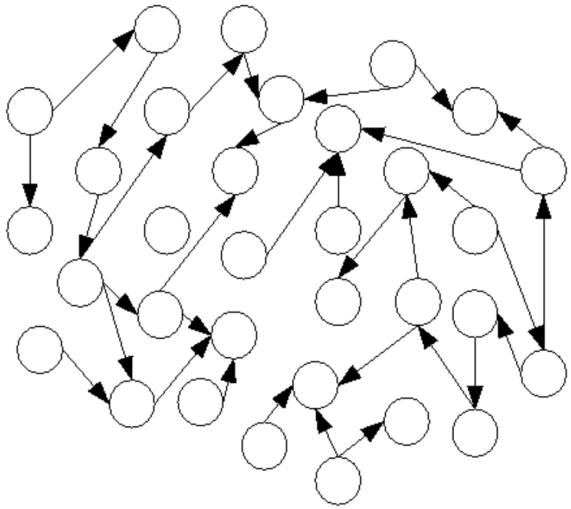
$$p(v) = \frac{deg(v)}{\sum_{w \in V} deg(w)}$$

- (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

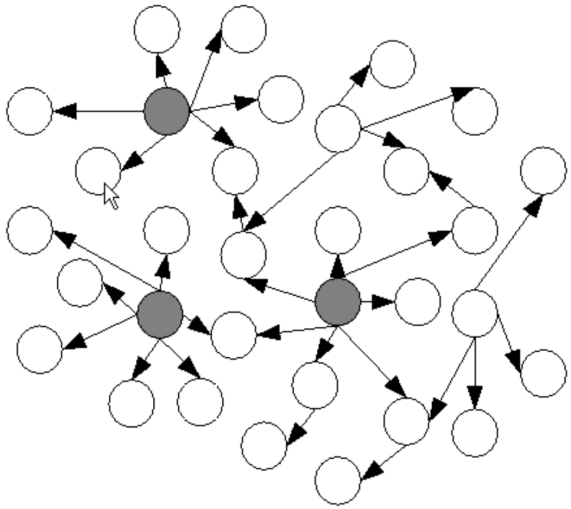
# Barábasi-Albert model (1999)



# Random vs. scale-free



(a) Random network



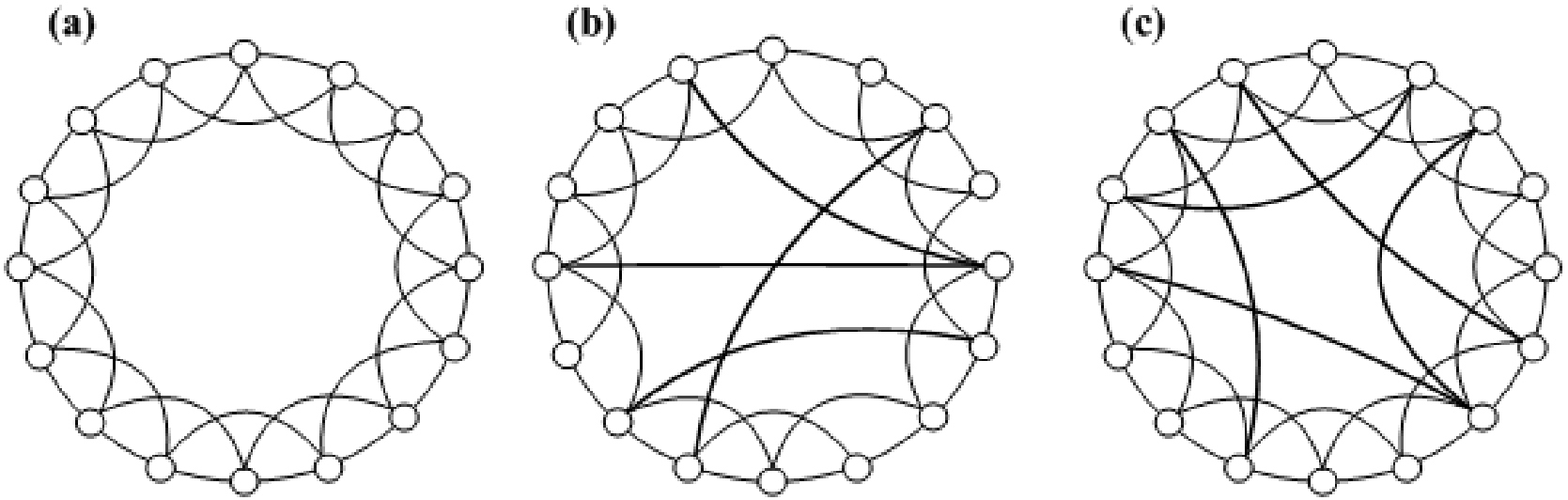
(b) Scale-free network

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

# Watts-Strogatz model (1998)

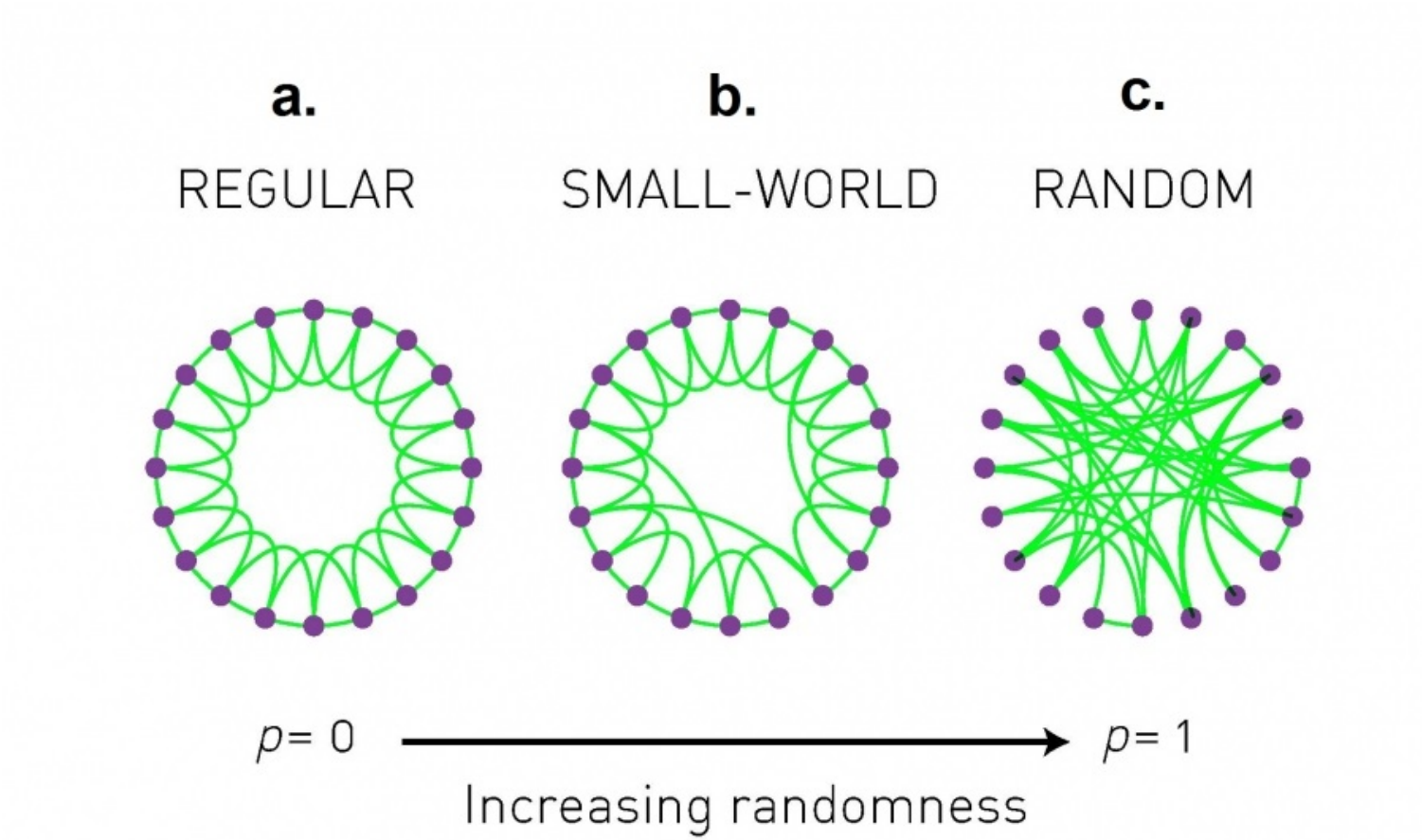
- Input number of nodes $n$, average degree $k$ and parameter $p$
- Constructs undirected graph with $n$ nodes and $\frac{1}{2} \cdot n \cdot k$ edges
- Start with "circle-shaped" graph connecting each node to its $k$ nearest neighbors
- Until each edge has been considered, in clock-wise order, **Rewire** 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

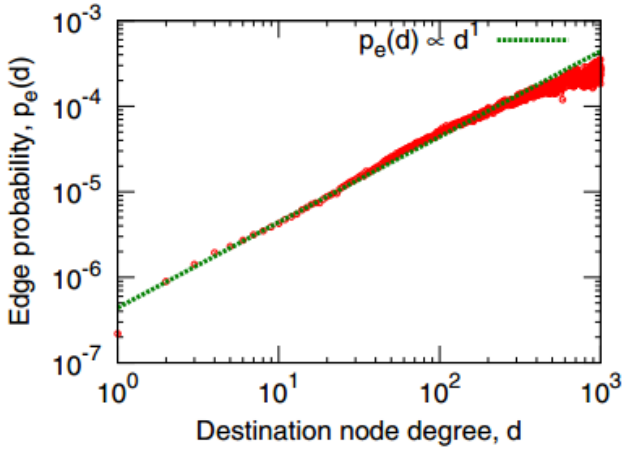
Leskovec et al., Microscopic Evolution of Social Networks, in Proceedings of KDD, pp. 462-470, 2008.

# Datasets

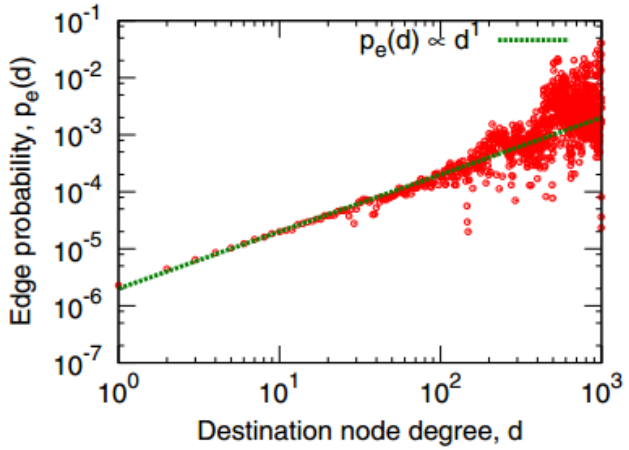
| Network | $T$ | $N$ | $E$ | $E_b$ | $E_u$ | $E_\Delta$ | % | $\rho$ | $\kappa$ |
|---|---|---|---|---|---|---|---|---|---|
| FLICKR (03/2003–09/2005) | 621 | 584,207 | 3,554,130 | 2,594,078 | 2,257,211 | 1,475,345 | 65.63 | 1.32 | 1.44 |
| DELICIOUS (05/2006–02/2007) | 292 | 203,234 | 430,707 | 348,437 | 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, $E_\Delta$ is the number of edges that close triangles, % is the fraction of triangle-closing edges, $\rho$ is the densification exponent ($E(t) \propto N(t)^\rho$), and $\kappa$ is the decay exponent ($E_h \propto \exp(-\kappa h)$) of the number of edges $E_h$ closing $h$ hop paths**

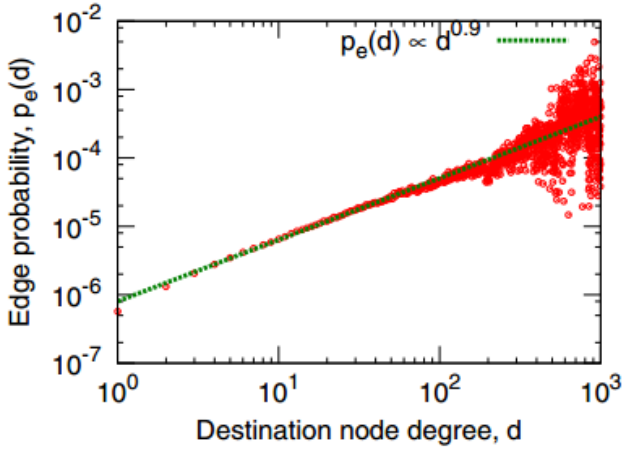
# Preferential attachment: degree



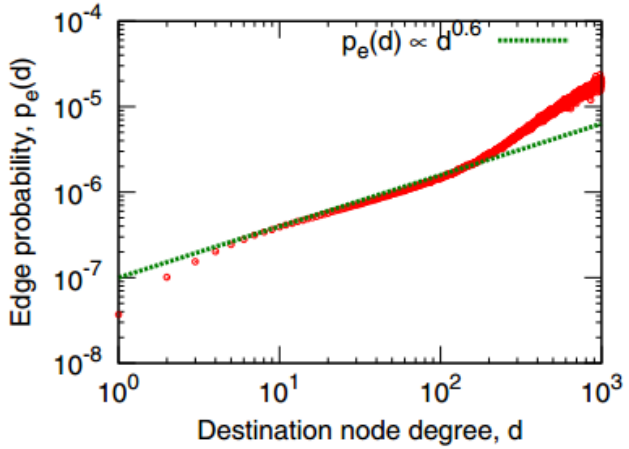

(c) FLICKR




(d) DELICIOUS

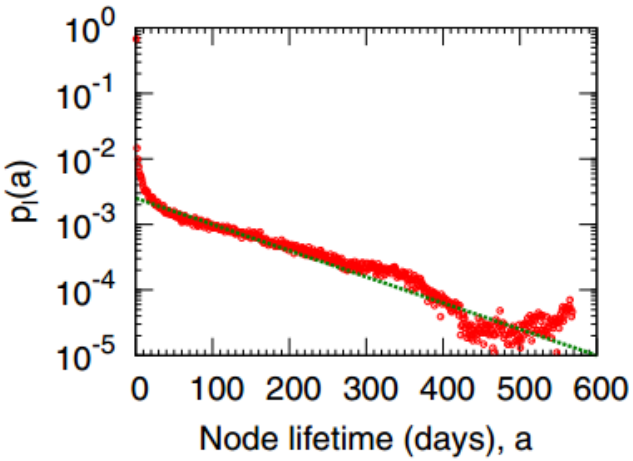



(e) ANSWERS

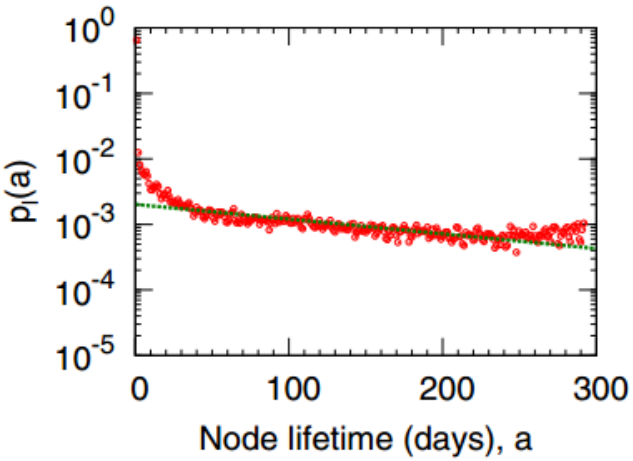



(f) LINKEDIN

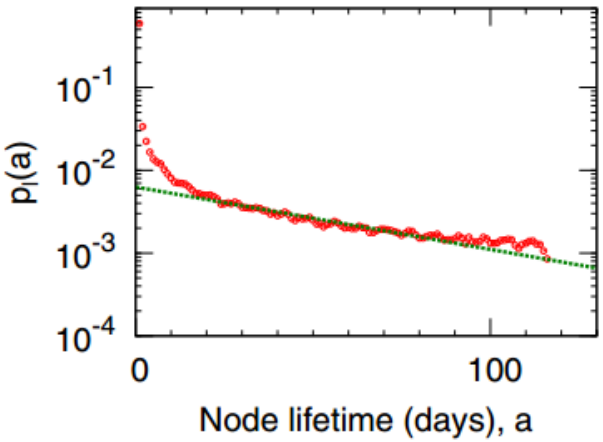
# Preferential attachment: age



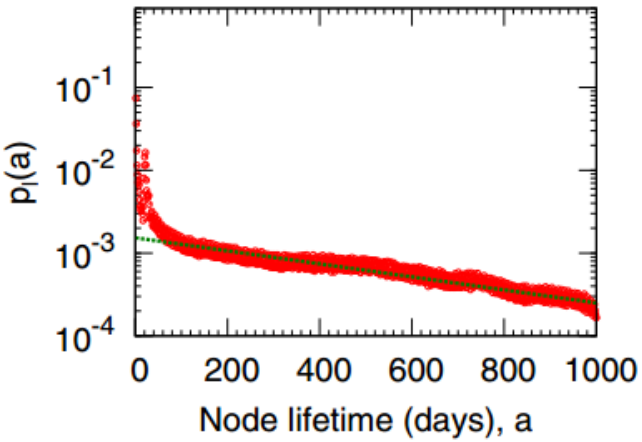

(a) FLICKR




(b) DELICIOUS

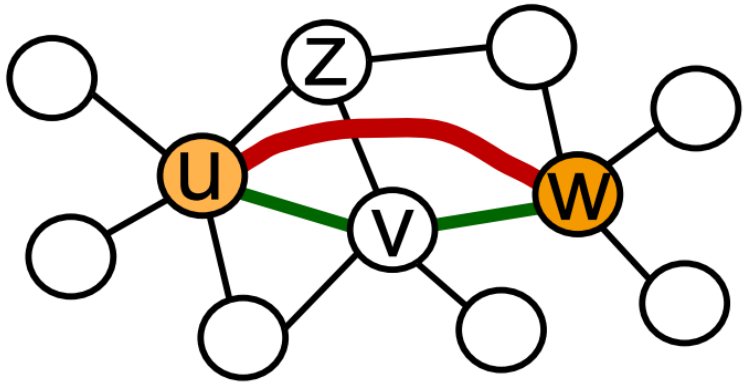



(c) ANSWERS

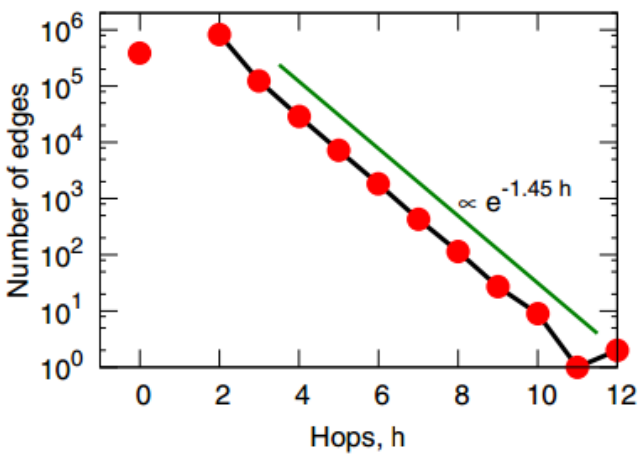



(d) LINKEDIN

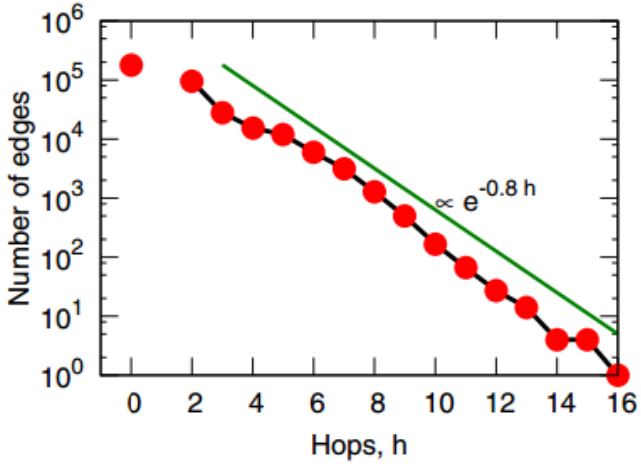
# Triadic closure



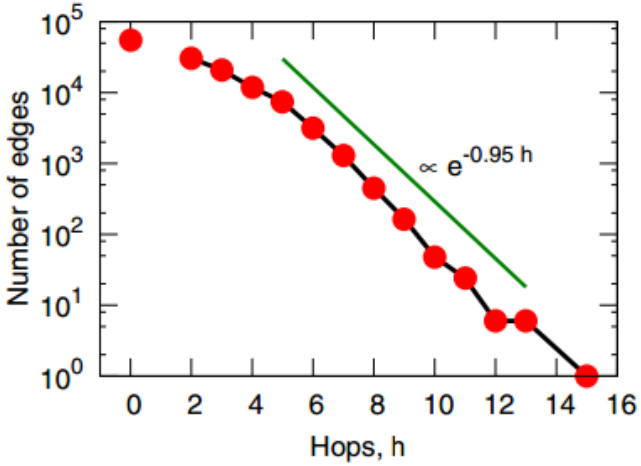
# Preferential attachment: hops



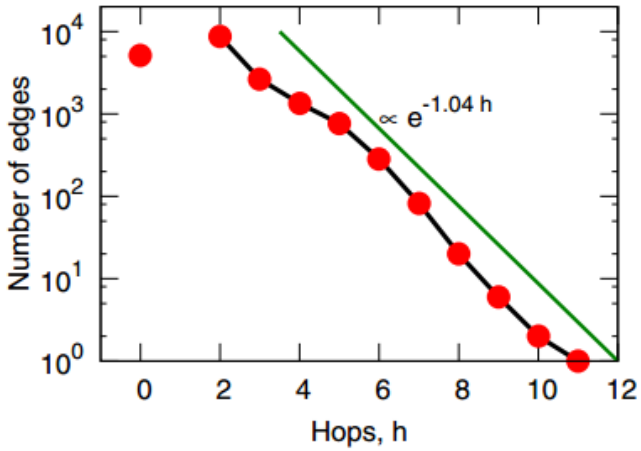

(c) FLICKR




(d) DELICIOUS




(e) ANSWERS




(f) LINKEDIN

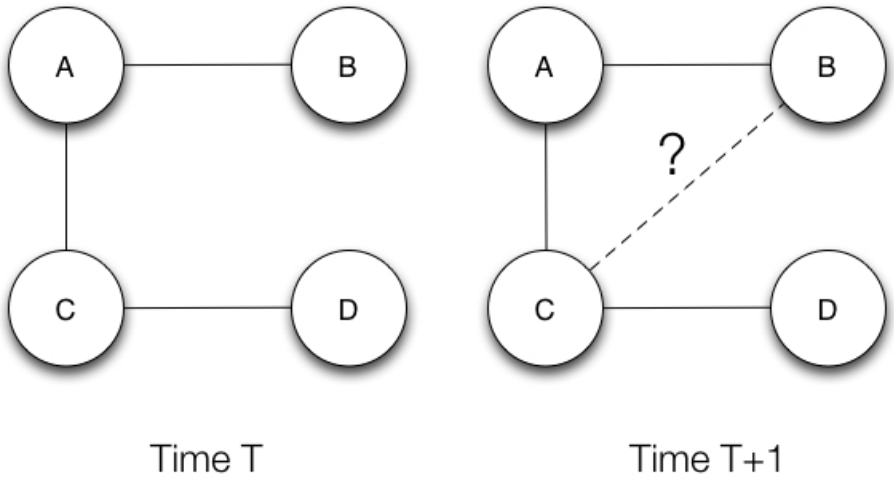
# Microscopic evolution model

- 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

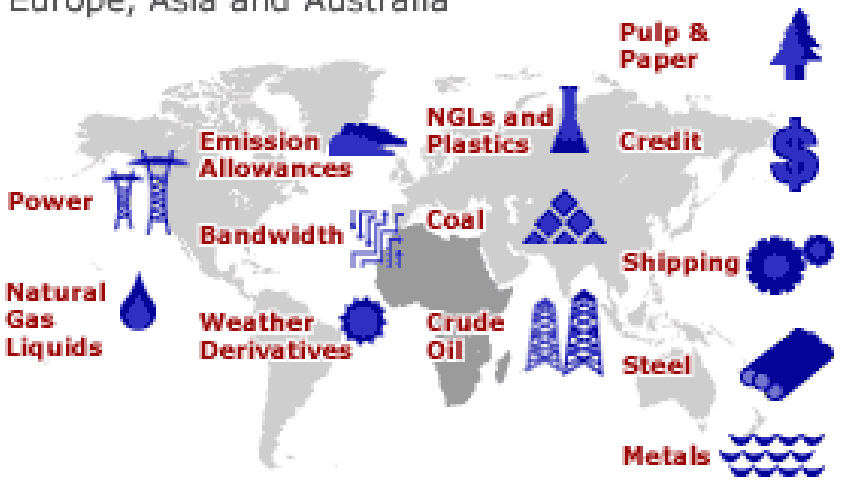
| Dataset | Nodes | Edges | Time | DPL exponent |
|---|---|---|---|---|
| Arxiv HEP–PH | 30,501 | 347,268 | 124 months | 1.56 |
| Arxiv HEP–TH | 29,555 | 352,807 | 124 months | 1.68 |
| Patents | 3,923,922 | 16,522,438 | 37 years | 1.66 |
| AS | 6,474 | 26,467 | 785 days | 1.18 |
| Affiliation ASTRO–PH | 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

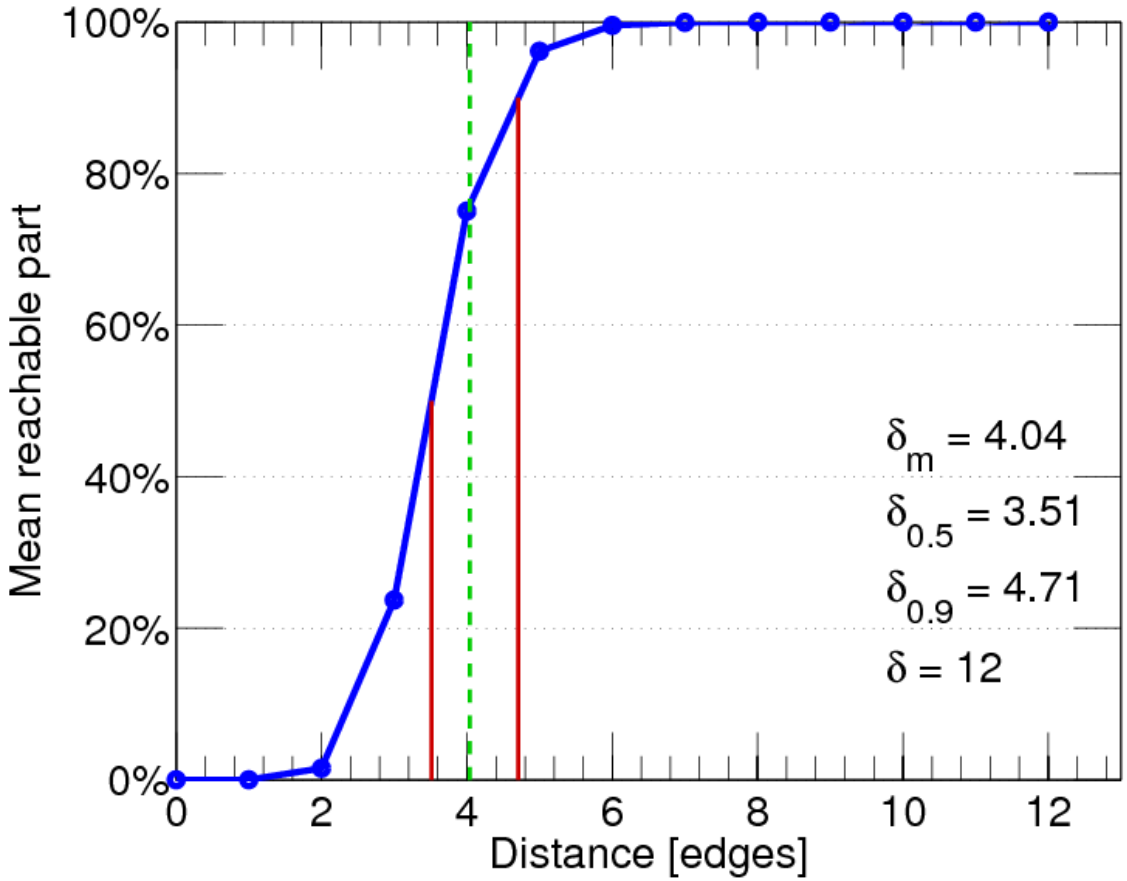

2001: Enron trading in hundreds of commodities
Interests in: USA, South America, Europe, Asia and Australia



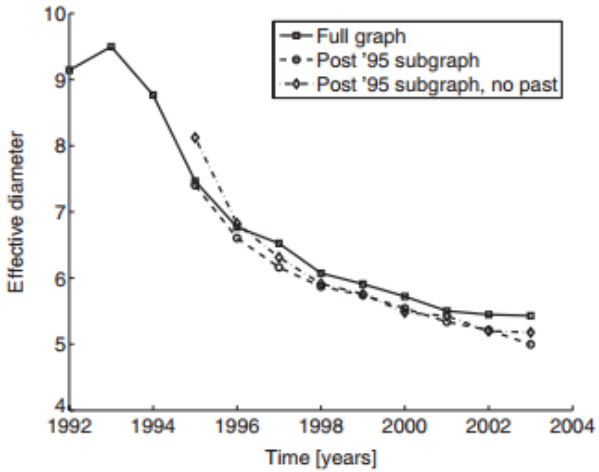
# Macroscopic patterns

- 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

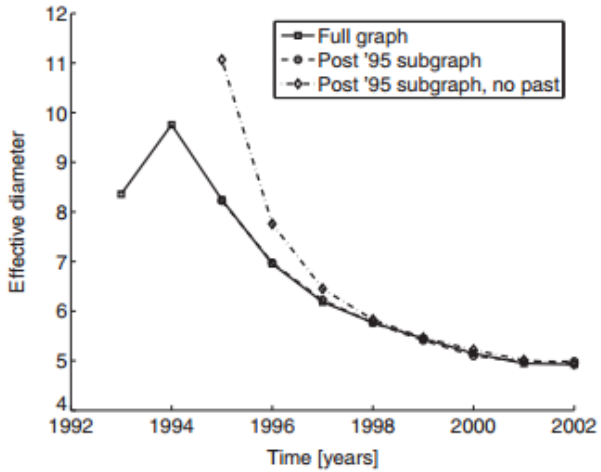
# Effective diameter



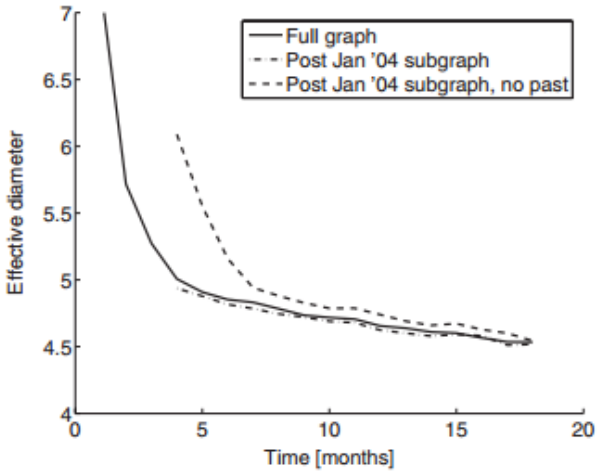
# Effective diameter



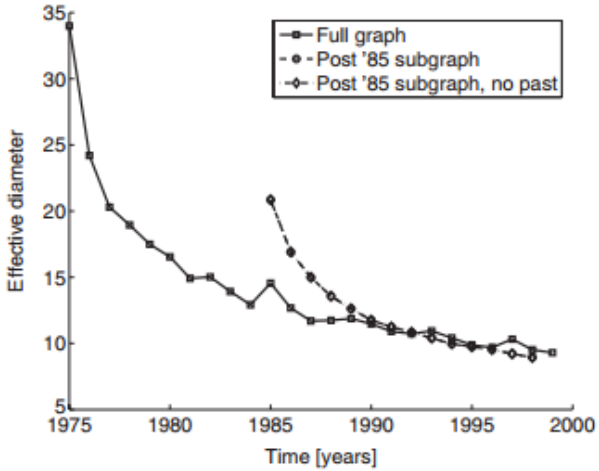

(a) arXiv citation graph



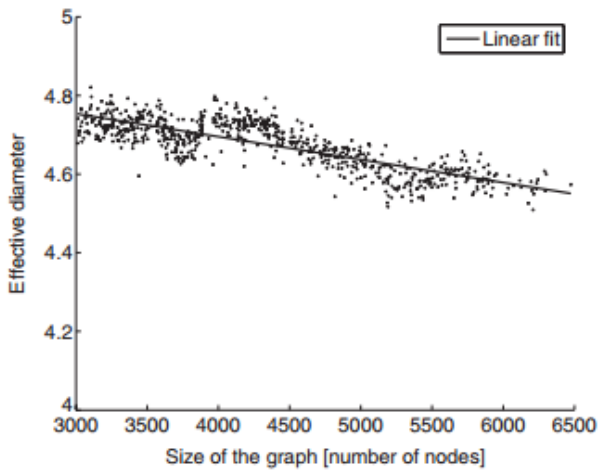

(b) Affiliation network



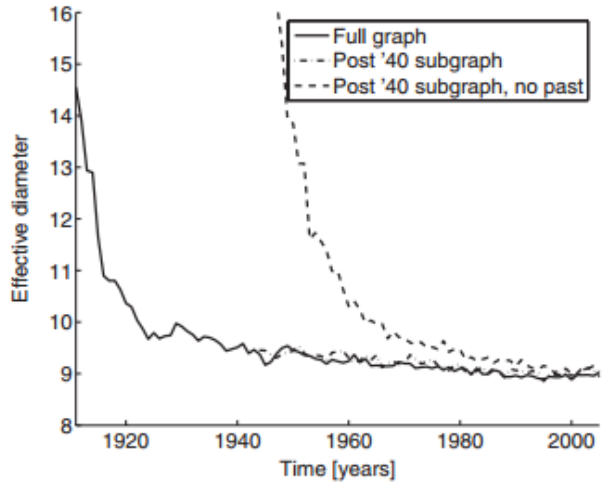

(e) Email network




(c) Patents citation graph

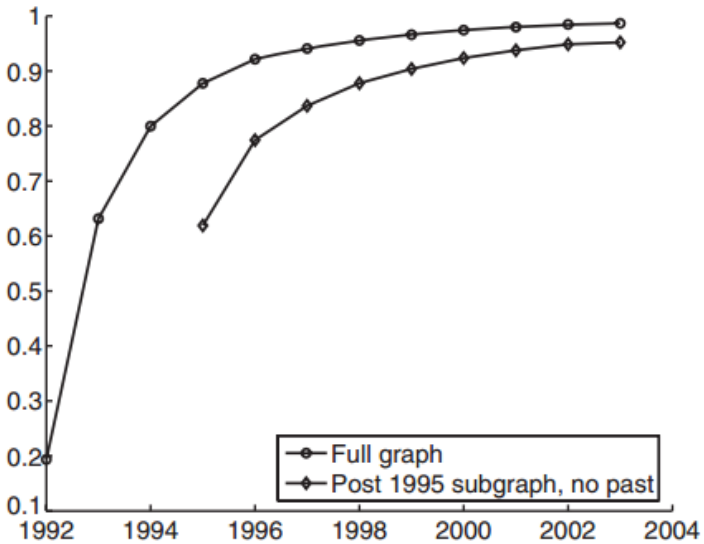



(d) Autonomous Systems

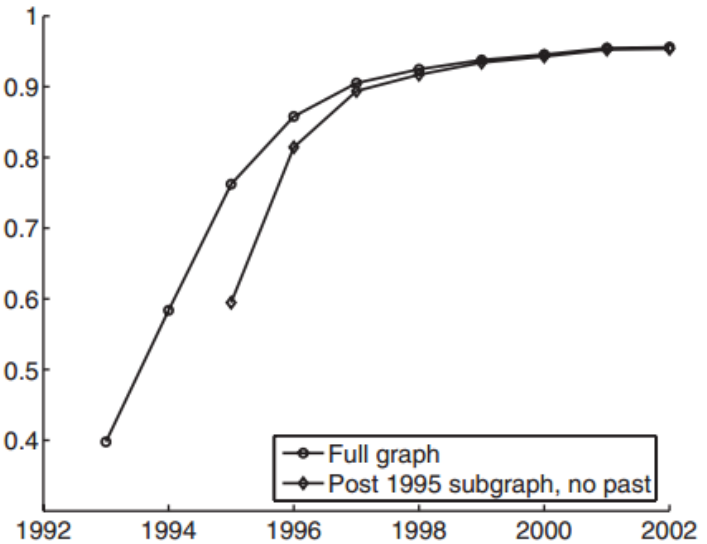



(f) IMDB actors to movies network

# Giant component

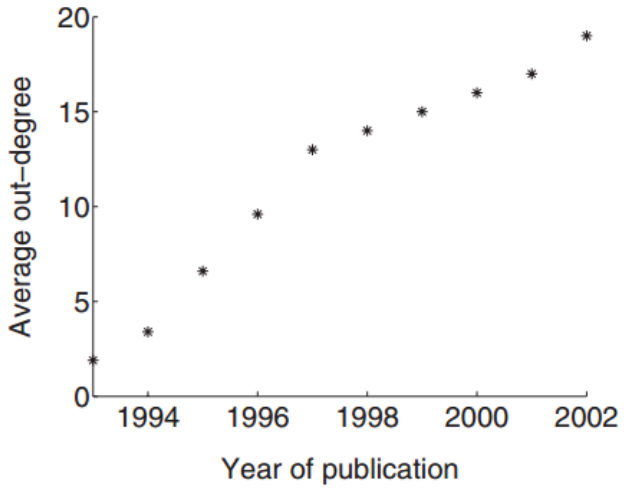



(a) arXiv citation graph

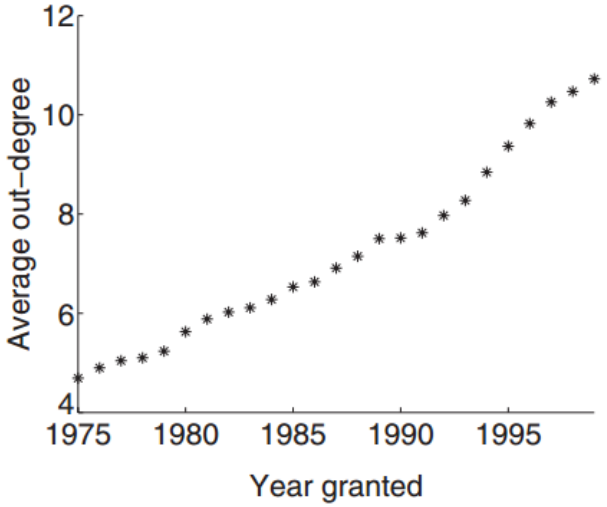



(b) Affiliation network

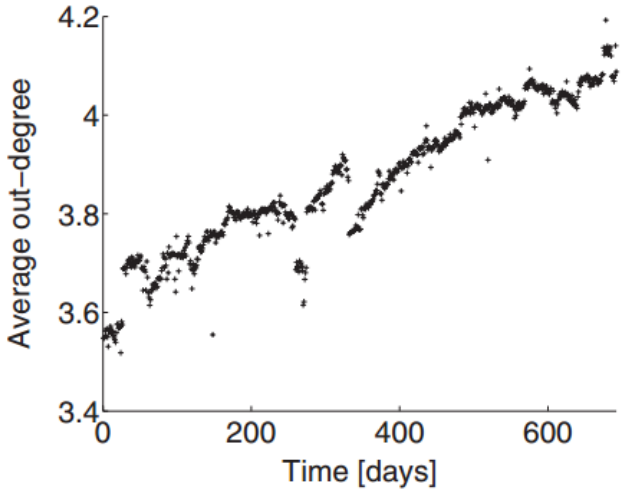
# Densification



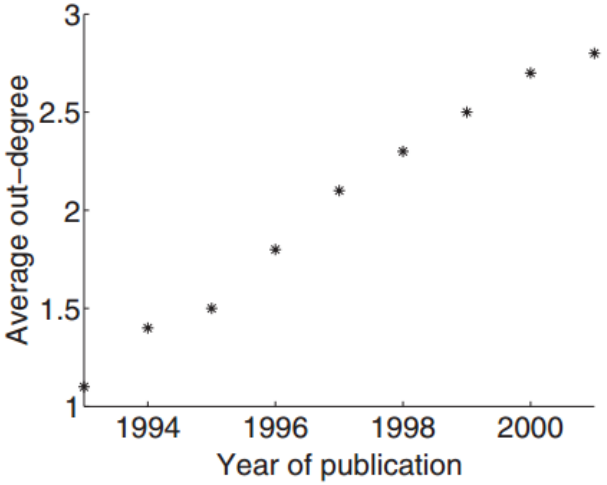

(a) arXiv




(b) Patents




(c) Autonomous Systems



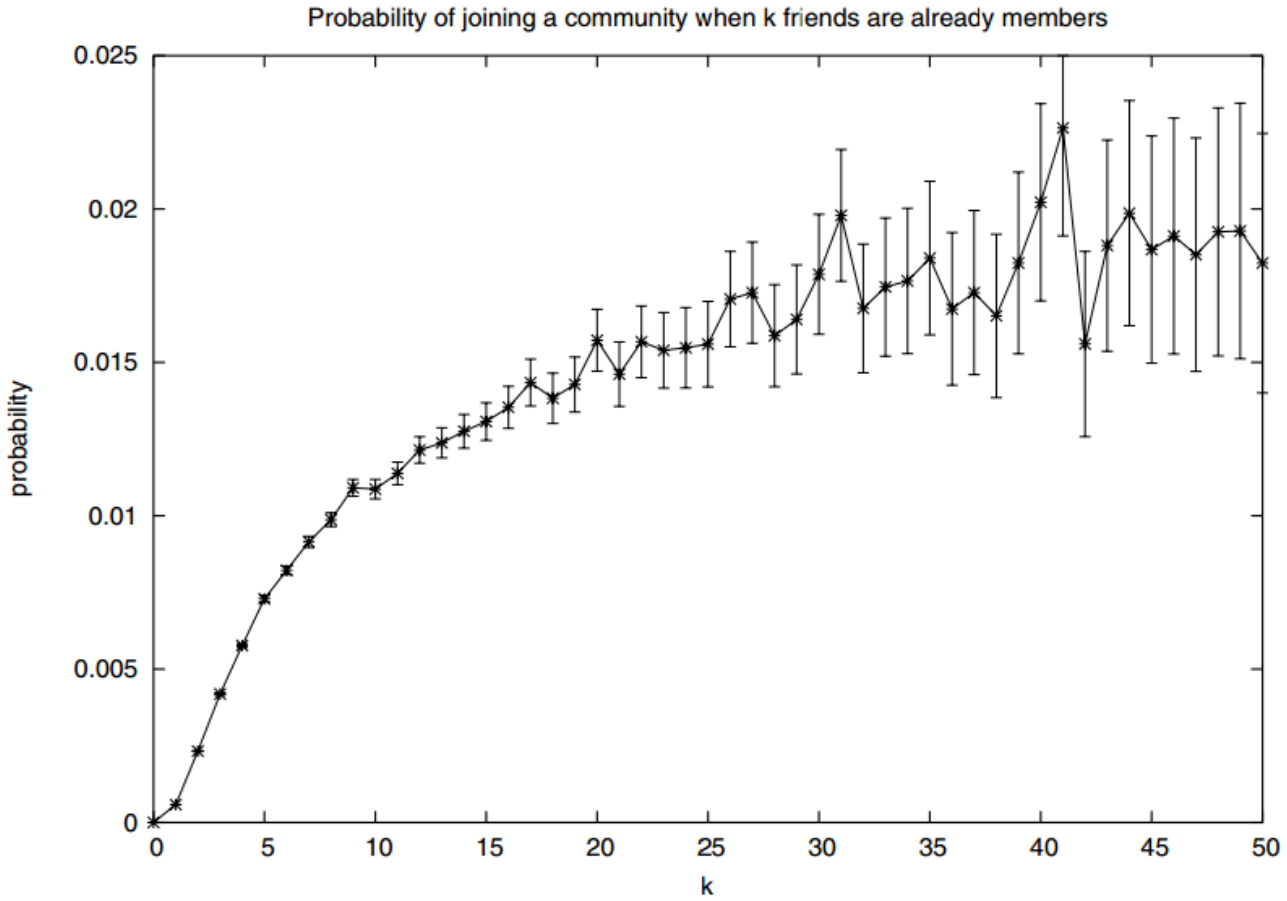

(d) Affiliation network

# Community evolution

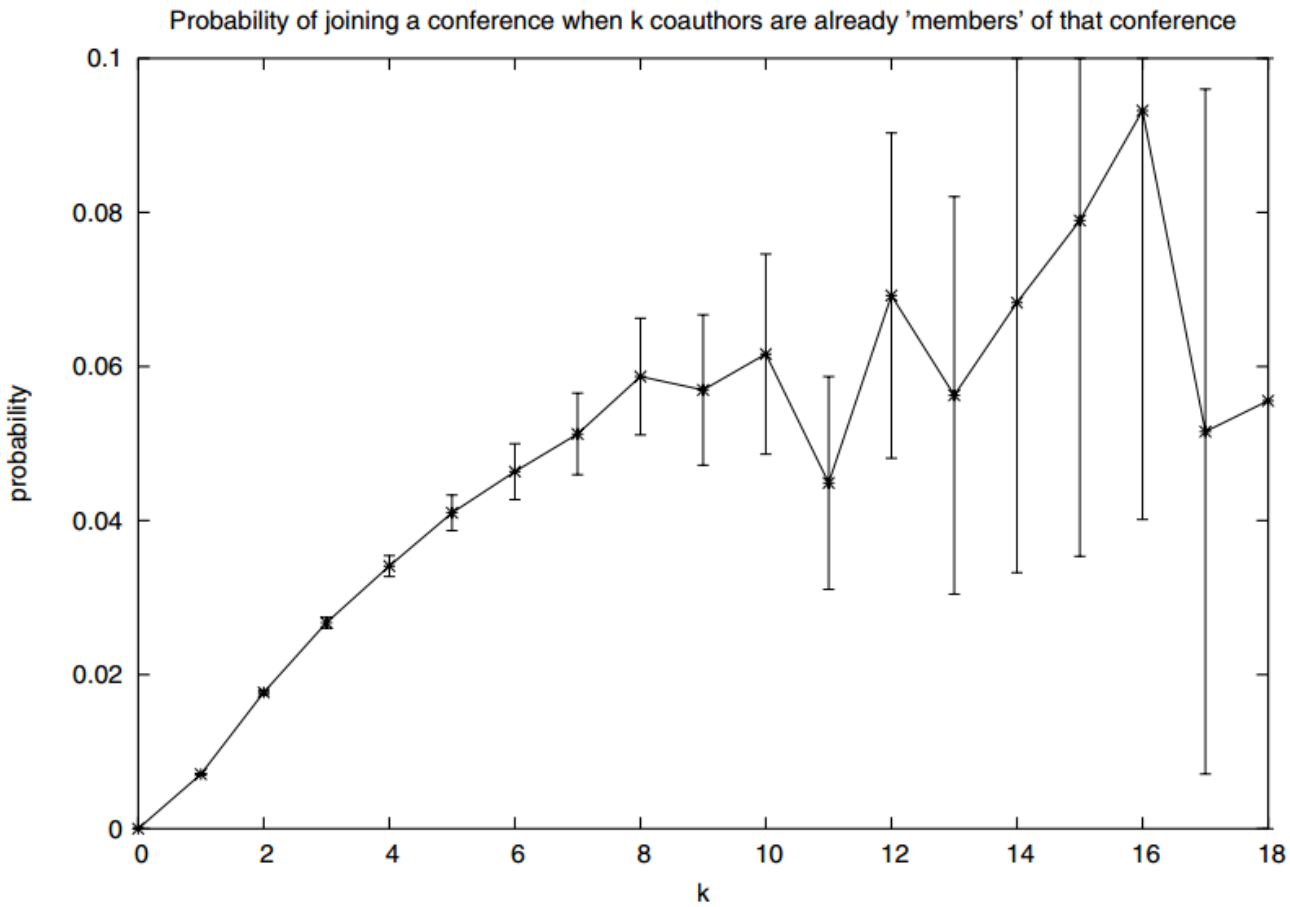
- Slightly different: user-defined communities
- DBLP: scientific collaboration network where communities are conferences that authors visit
- LiveJournal: online social network with explicit groups based on common interest
- What motivates nodes to join a community?
- What causes nodes to switch between 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 ($\|C\|$).<br>Number of individuals with a friend in $C$ (the *fringe* of $C$) .<br>Number of edges with one end in the community and the other in the fringe.<br>Number of edges with both ends in the community, $\|E_C\|$.<br>The number of open triads: $\|\{(u,v,w)\|(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \notin E_C \wedge u \neq w\}\|$.<br>The number of closed triads: $\|\{(u,v,w)\|(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \in E_C\}\|$.<br>The ratio of closed to open triads.<br>The fraction of individuals in the fringe with at least k friends in the community for $2 \leq k \leq 19$.<br>The number of posts and responses made by members of the community.<br>The number of members of the community with at least one post or response.<br>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 ($\|S\|$).<br>Number of adjacent pairs in $S$ ($\|\{(u,v)\|u,v \in S \wedge (u,v) \in E_C\}\|$).<br>Number of pairs in $S$ connected via a path in $E_C$.<br>Average distance between friends connected via a path in $E_C$.<br>Number of community members reachable from $S$ using edges in $E_C$.<br>Average distance from $S$ to reachable community members using edges in $E_C$.<br>The number of posts and response made by individuals in $S$.<br>The number of individuals in $S$ with at least 1 post or response. |

# Decision tree (LiveJournal)

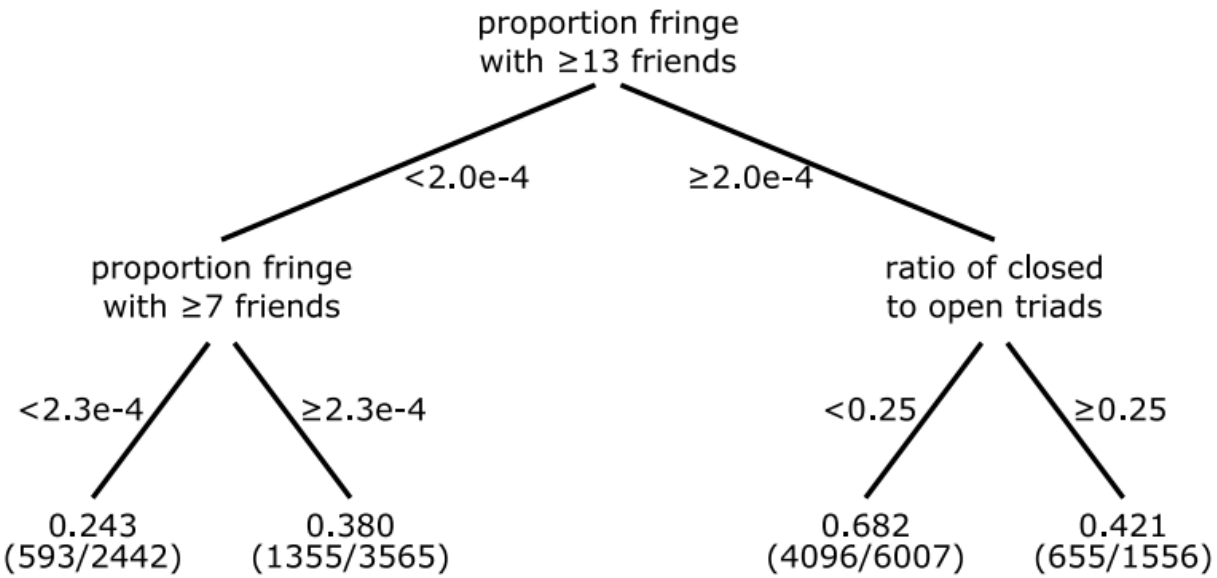



**Figure 5: The top two levels of decision tree splits for predicting community growth in LiveJournal.**

# Decision tree (LiveJournal)

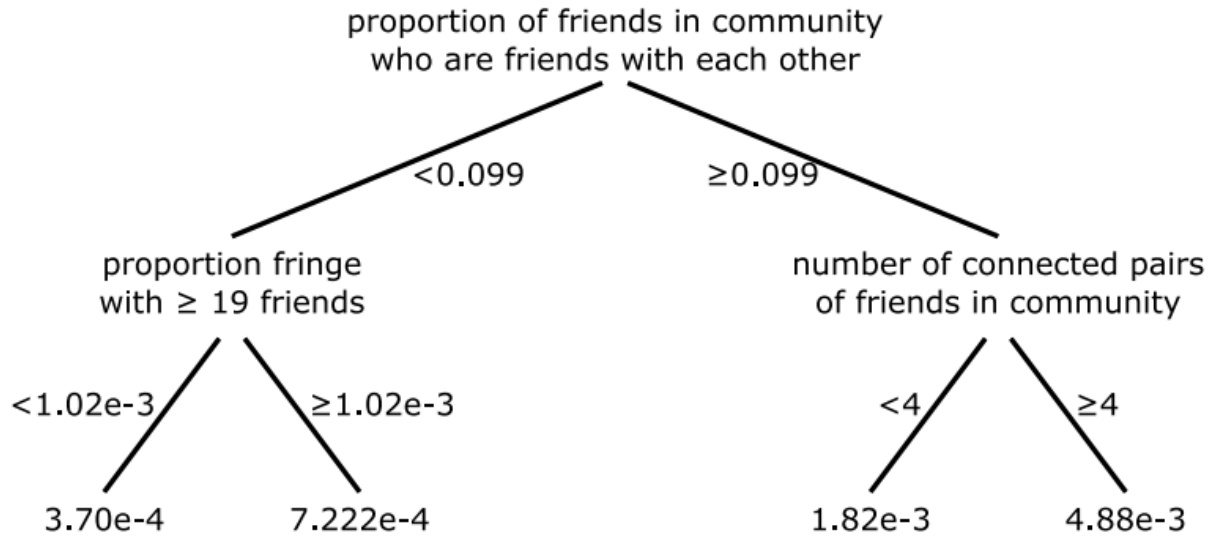
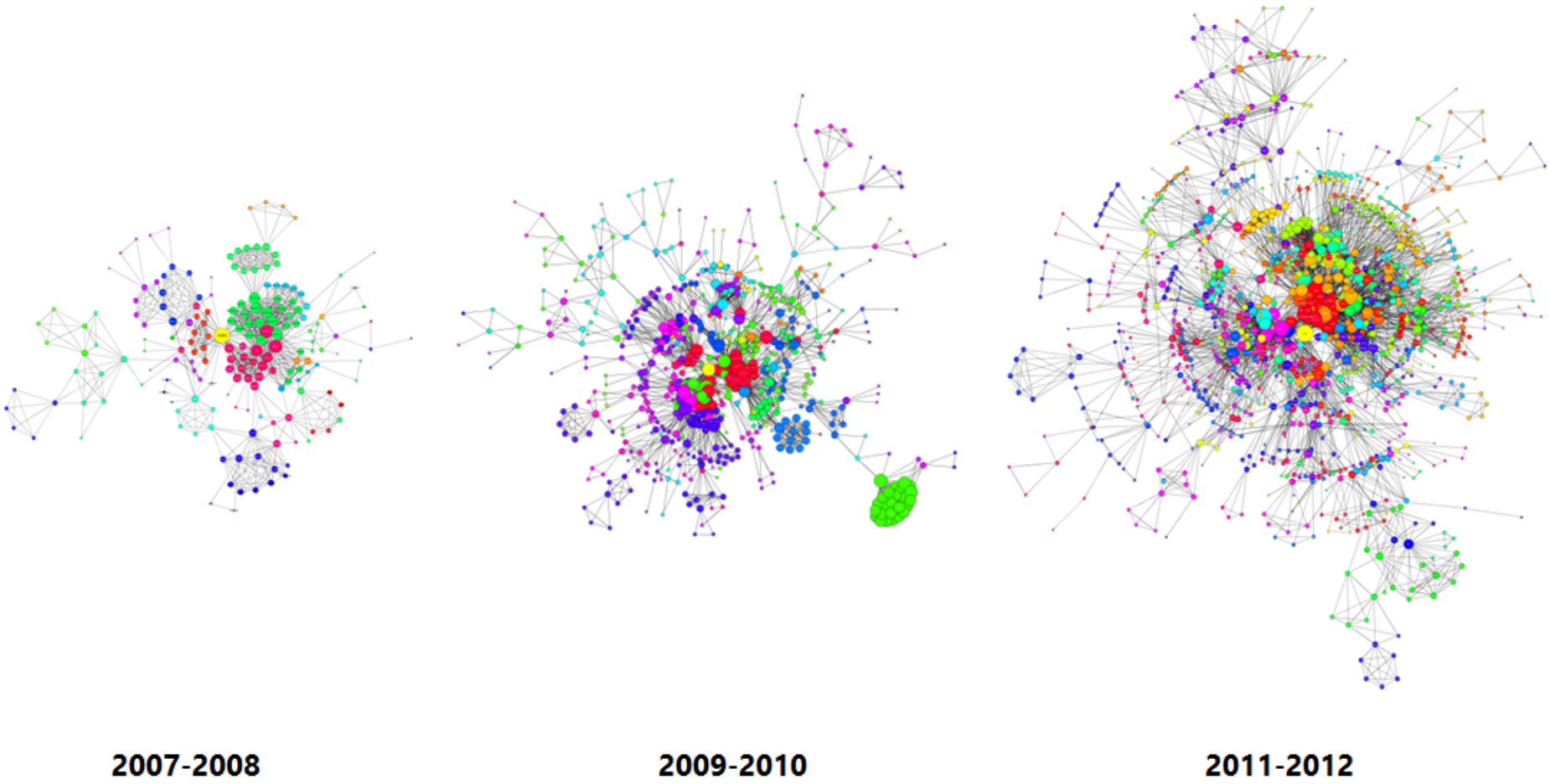



**Figure 3: The top two levels of decision tree splits for predicting single individuals joining communities in LiveJournal. The overall rate of joining is 8.48e-4.**

# 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?

# Walks in information networks

F.W. Takes and W.A. Kosters, Mining User-Generated Path Traversal Patterns in an Information Network, in Proceedings of the 12th IEEE/ACM International Conference on Web Intelligence (WI 2013), pp. 284-289, IEEE, 2013.
F.W. Takes and W.A. Kosters, The Difficulty of Path Traversal in Information Networks, in Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2012), pp. 138-144, 2012.

# Walks

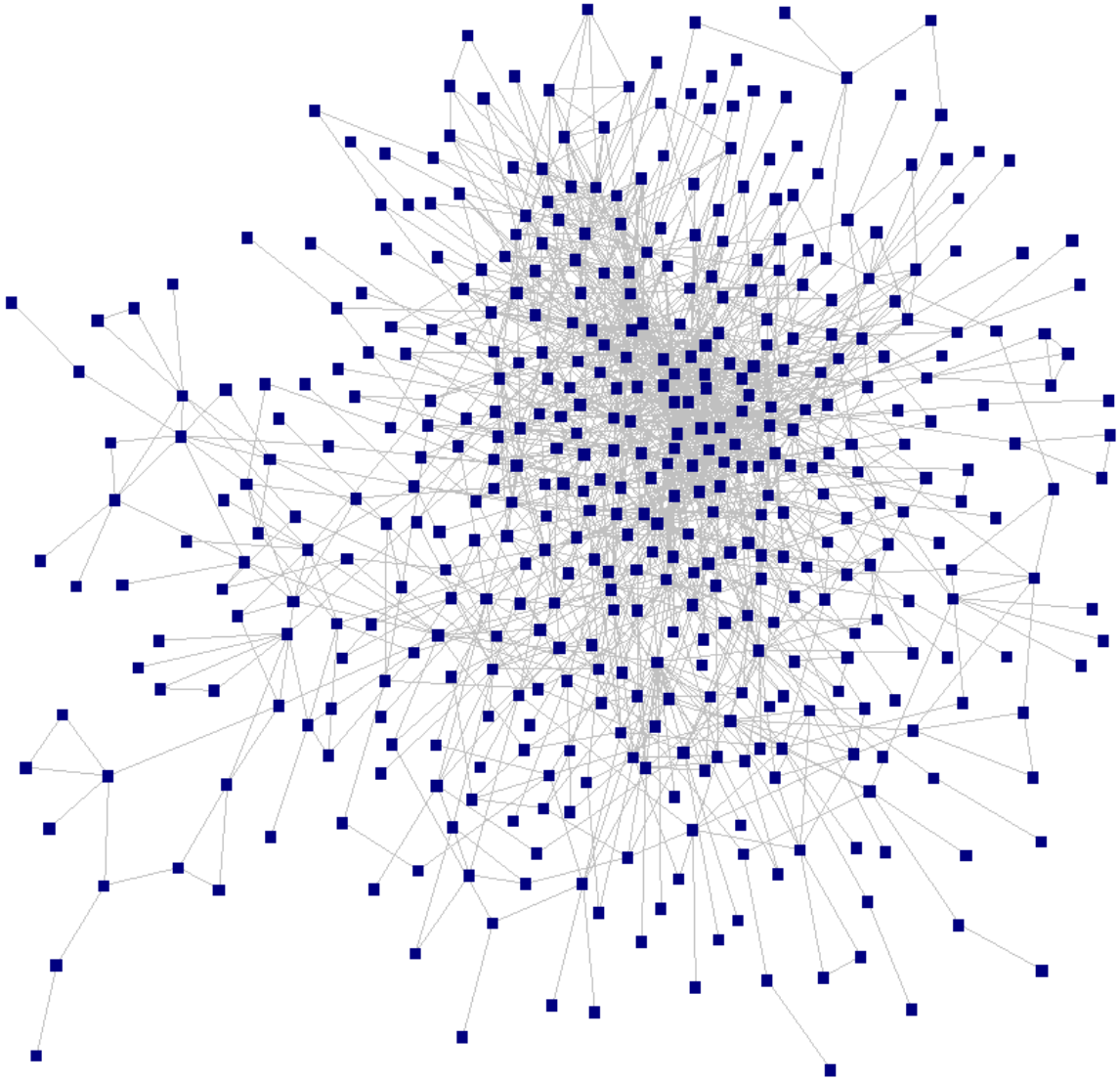
- Random walk
- Biased random walk
- **Targeted walk**

# Walks

- Random walk
- Biased random walk
- **Targeted walk**
  - By humans
  - On Wikipedia

# The Wiki Game

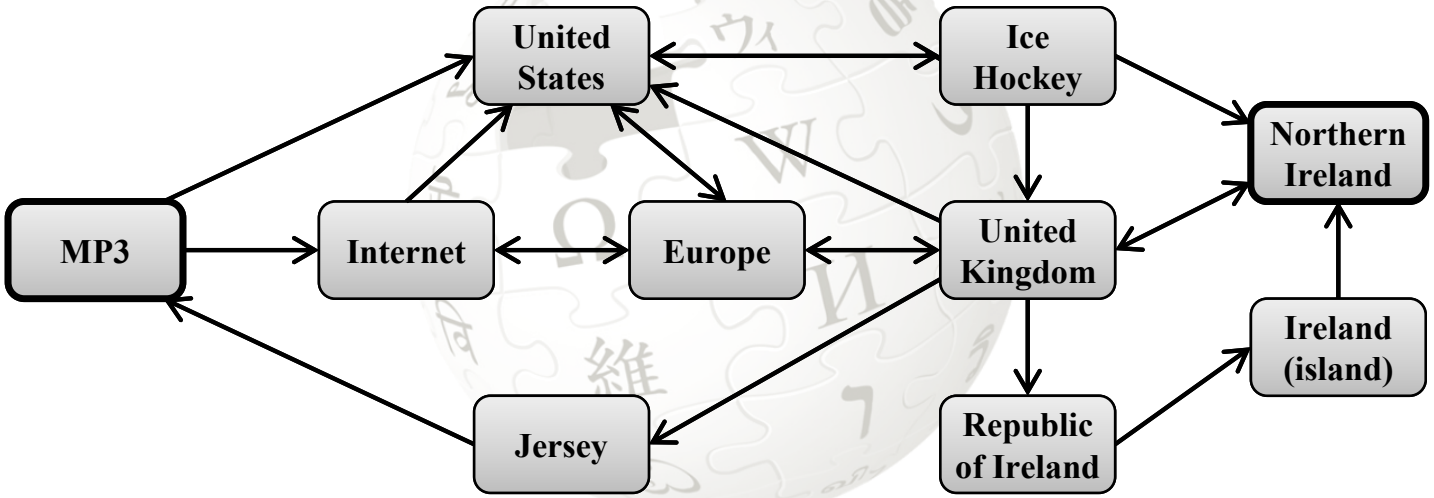
# Information Networks



# The topic explained

**The Difficulty of Path Traversal in Information Networks**

How hard is it for a human to find a path between two Wikipedia pages?



# Motivation & Goals

- Search strategies
- Human search behavior
- Navigating the "Deep Web"
- Path traversal patterns
- Pattern mining

# Concepts

- **Information network**: directed graph $G = (V, E)$ with $n$ nodes and $m$ links (Wikipedia)
- Path traversal task: navigate from $u$ to $v$ (with $u, v \in V$)
- Path: sequence of nodes connected by edges $(u, a, b, c, a, d, v)$
- Path length: number of traversed edges (6)
- Shortest path: minimum number of traversed edges $(u, a, d, v)$
- Distance: shortest path length $d(u, v) = 3$
- Failed path: $(u, a, b, c, u, a, e, a, d)$

# Wikipedia dataset

- DBpedia 3.7 (`http://dbpedia.org`) from August 2011
- Small world network

| | |
|---|---|
| Articles ($n$) | $3,464,902$ |
| Directed links ($m$) | $82,019,786$ |
| Largest WCC | 99.9% |
| Average indegree | 26 |
| Average outdegree | 22 |
| Average distance ($\overline{d}$) | 4.81 |
| Effective diameter | 7 |
| Diameter | 11 |

Table: Wikipedia dataset

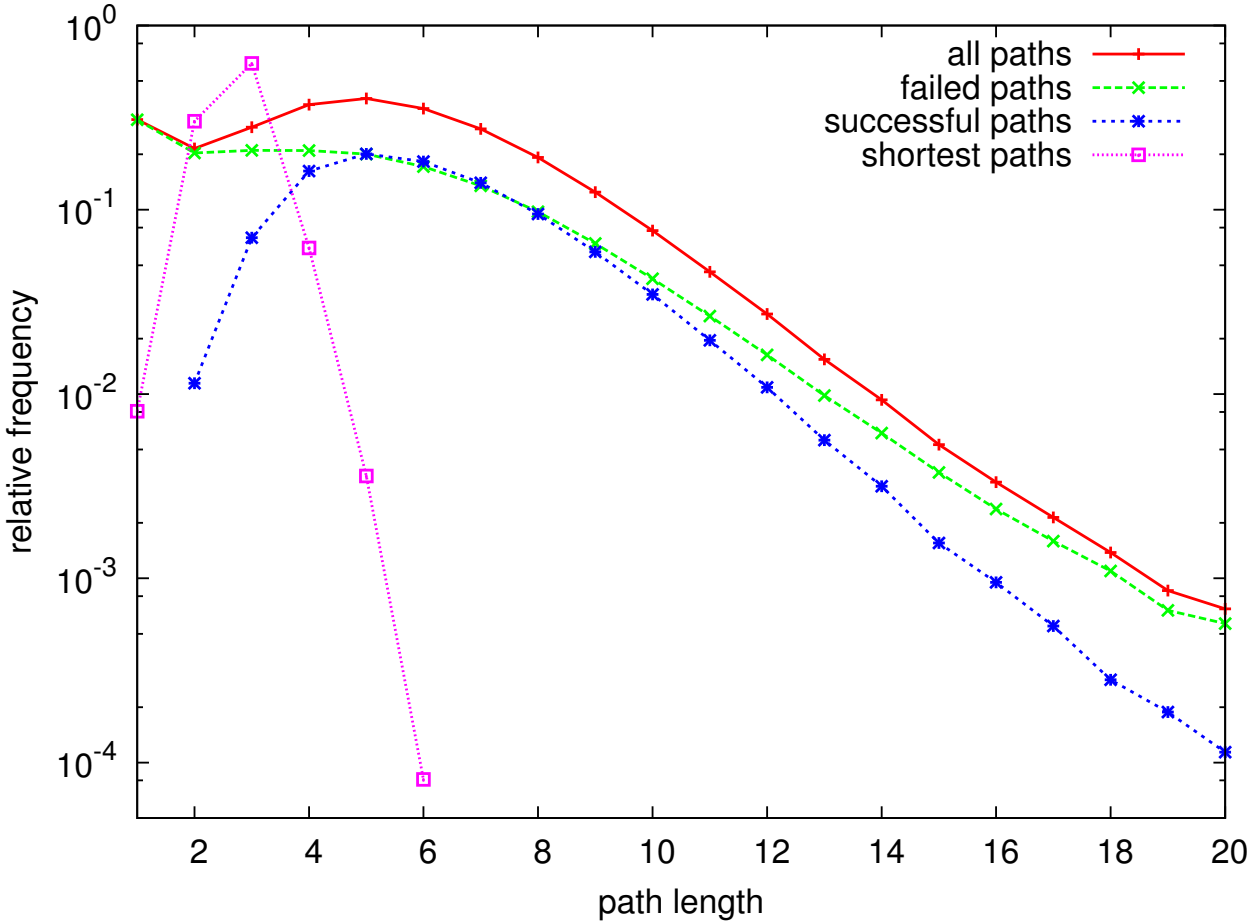
# The Wiki Game dataset

- The Wiki Game (`http://thewikigame.com`)
- User-generated paths from 2009–2012
- Original set: $3,219,641$ paths consisting of $17,151,824$ clicks
- Only consider succesful paths (33.8%) of length 3–20.

| | |
|---|---|
| User-generated paths | $1,137,337$ |
| Clicks | $7,135,060$ |

Table: Wiki Game dataset

# Successful and failed paths



# Difficulty measures

- Start outdegree:

$$f_{outdeg}(u, v) = outdeg(u)$$

- Goal indegree:

$$f_{indeg}(u, v) = indeg(v)$$

- start 2-neighborhood size:

$$f_{|N_2|}(u, v) = |N_2(u)|$$

- goal reversed 2-neighborhood size:

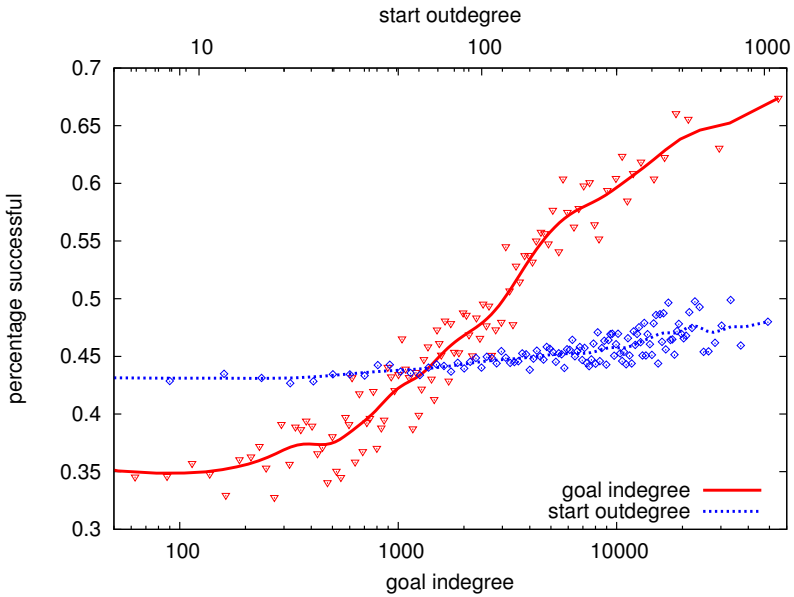
$$f_{|N_2'|}(u, v) = |N_2'(v)|$$

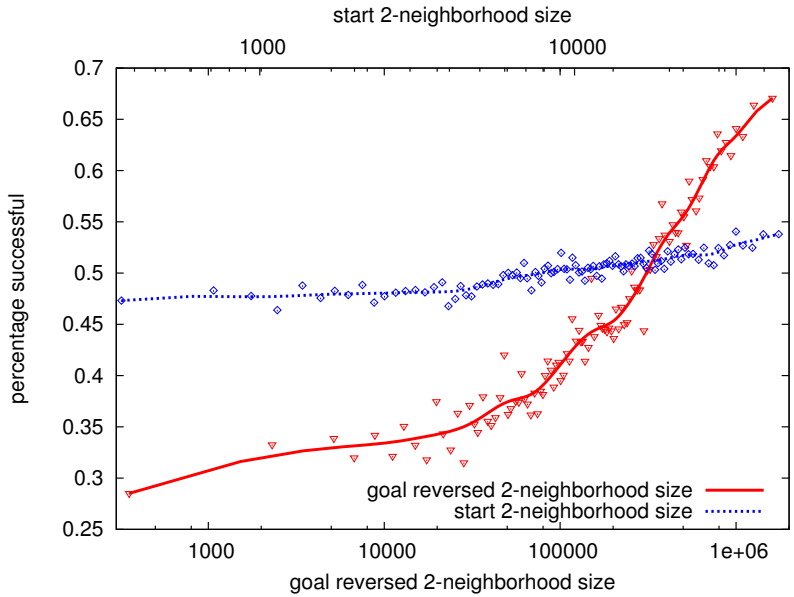
- Path length (distance):

$$f_d(u, v) = d(u, v)$$

# Local difficulty measures

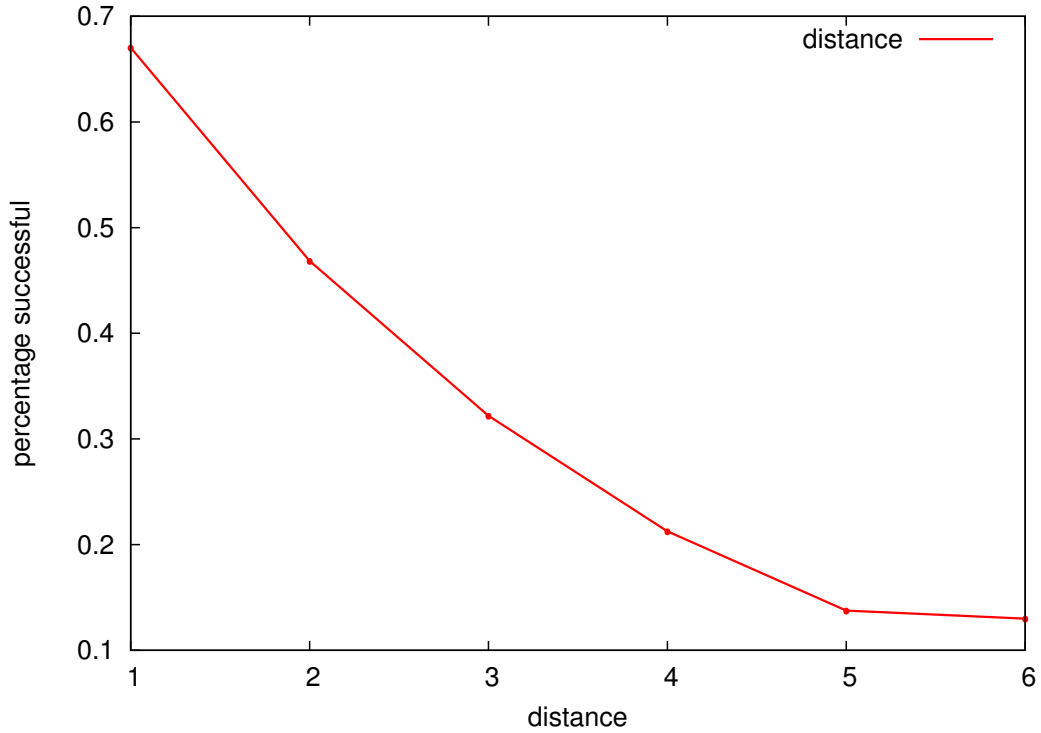
- Start page is of little influence
- Neighborhood of goal page is of high influence



# Global measure: path length

- Path length (distance) is directly related to path formation difficulty



# Subgraph centrality

- Users are able to select a central set of nodes for reaching their navigation goals.
  But is this a good and efficient set?
- Node centrality: the importance of a single node
- **Subgraph centrality**: the importance of a set of nodes
- Subgraph consisting of the top-$k$ user-defined central nodes
- Determine subgraph's closeness centrality, for top-$k$ nodes of different centrality measures

# Results: subgraph centrality

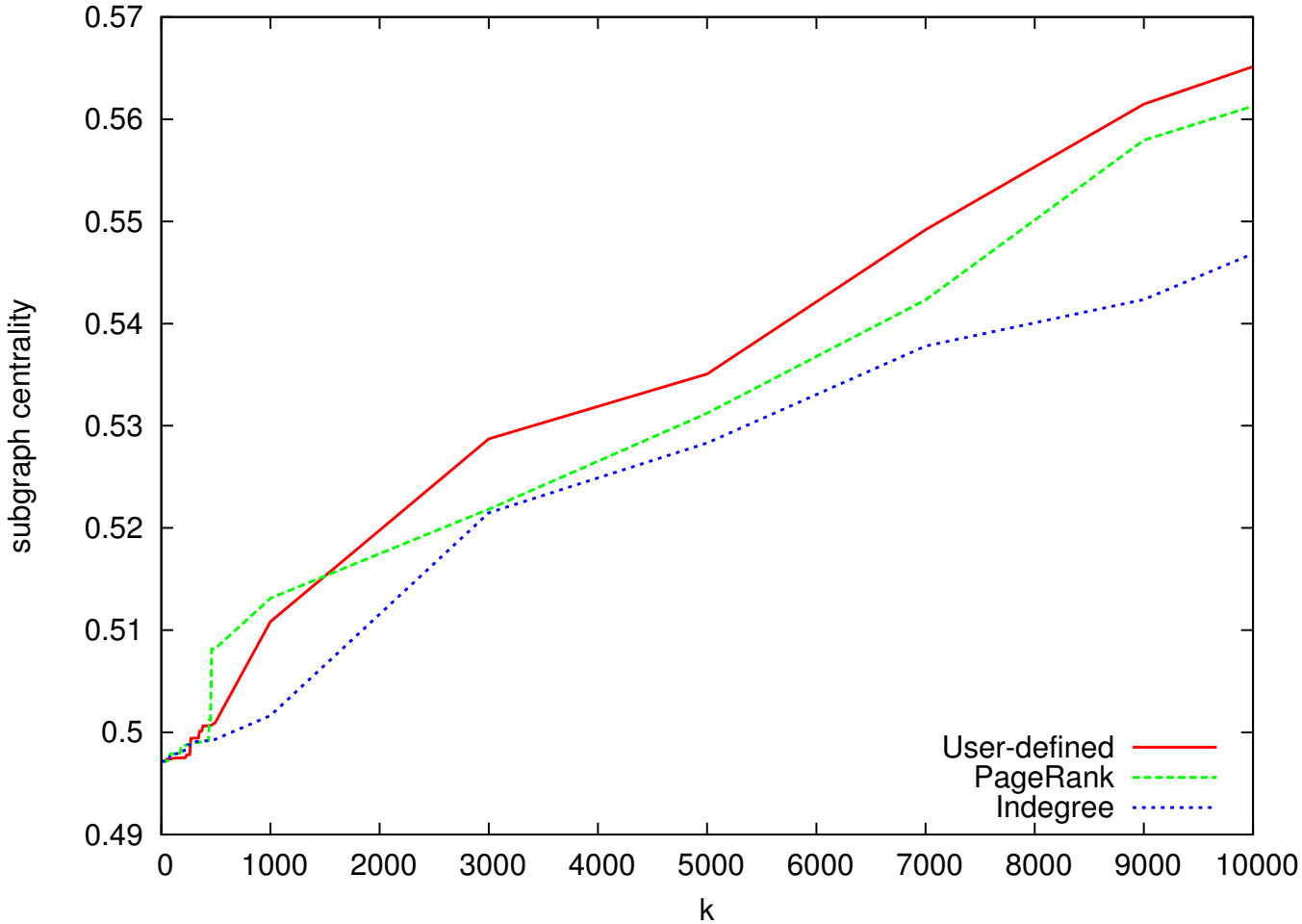



Figure: Comparison of subgraph centrality (vertical axis) of various centrality

# Findings

- Human paths are roughly twice as long as shortest paths
- Progress is easy close to start and goal
- High-level reasoning causes users to miss opportunities
- Hubs are crucial in the beginning
- Local properties of goal page explain difficulty
- Users are apparently able to select an efficient portion of the graph that is useful for traversing it

# Upcoming week

- Learn more about network dynamics (thanks for the tip, Marc!): `https://www.youtube.com/watch?v=CYlon2tvywA`
- Next week: last lecture; then 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
- Make serious progress with Assignment 2
- Note: SysAdmin REL downtime announcement: Oct 20–22