NETWORK SCIENCE OF ONLINE INTERACTIONS

Chapter 3: Hubs

Joao Neto

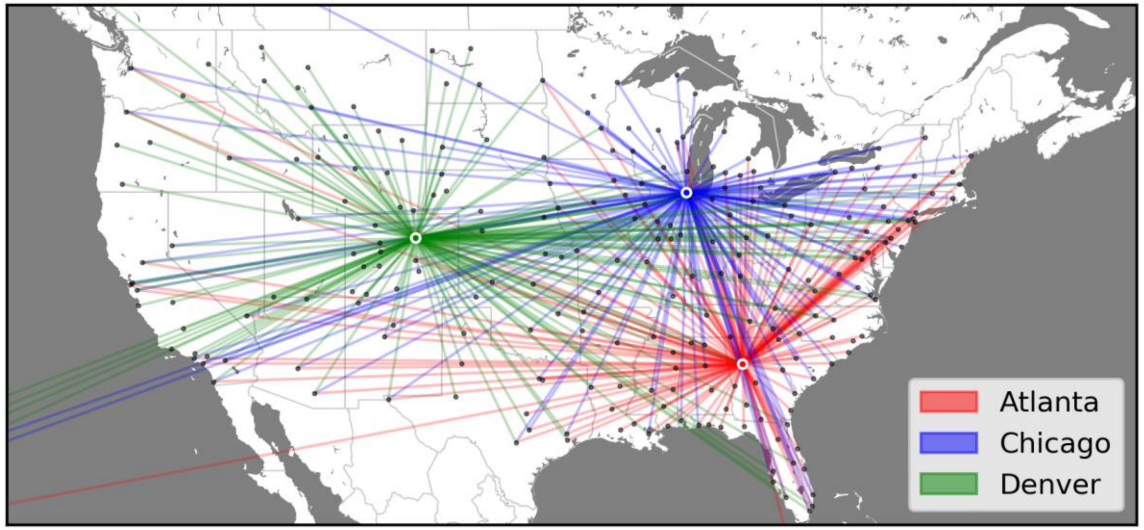
05/May/2023

SUMMARY

- Networks can be **assortative**
- **Path length** is an important characteristic of a network
- **Connected component** and **giant component** are two others
- Most networks have low **average path length** and are **small-world**

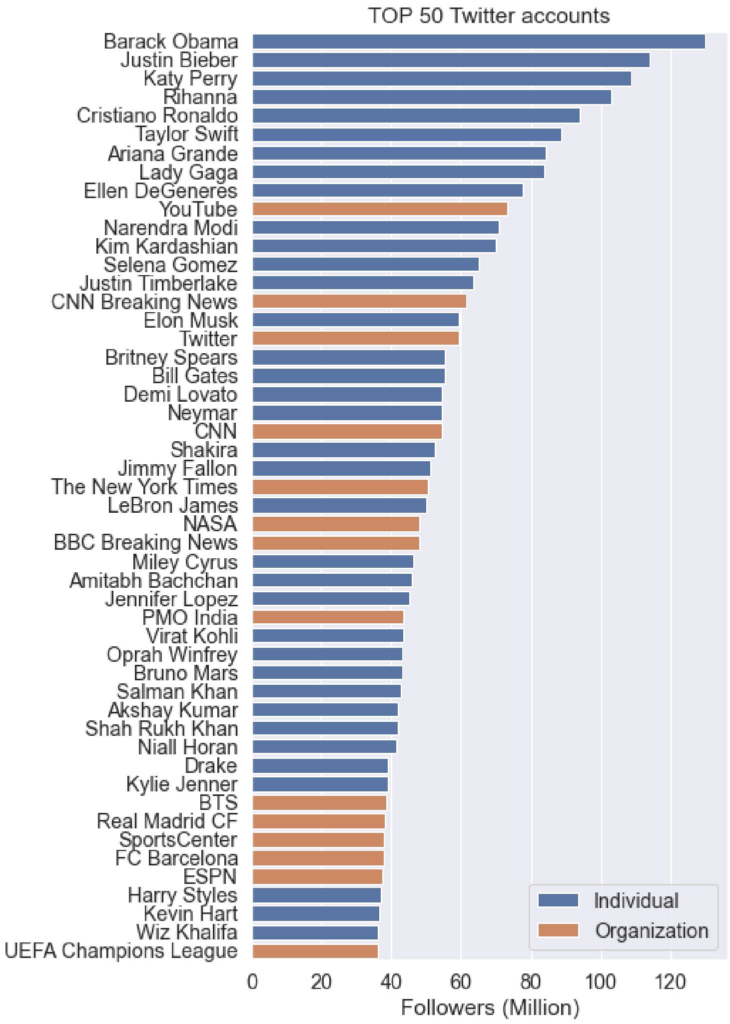
CHAPTER 3 - HUBS

- Networks are heterogeneous
- Some nodes are much important than others
- Node-centric analysis
  - Node **centrality**
  - The friendship paradox
- Network-centric analysis
  - Ultra-small worlds
  - Robutness
  - Core decomposition



3.1 CENTRALITY MEASURES

- Centrality: measure of node importance
  - Degree centrality
  - Closeness centrality
  - Betweness centrality
- Degree centrality
  - Ranks nodes by their degree
  - High-degree nodes are **hubs**
  - Straightforward, but not comparable between networks
  - What about $k_i/\langle k\rangle$?
    - Bad because $\langle k\rangle = \sum_i k_i/N$ is usually not informative



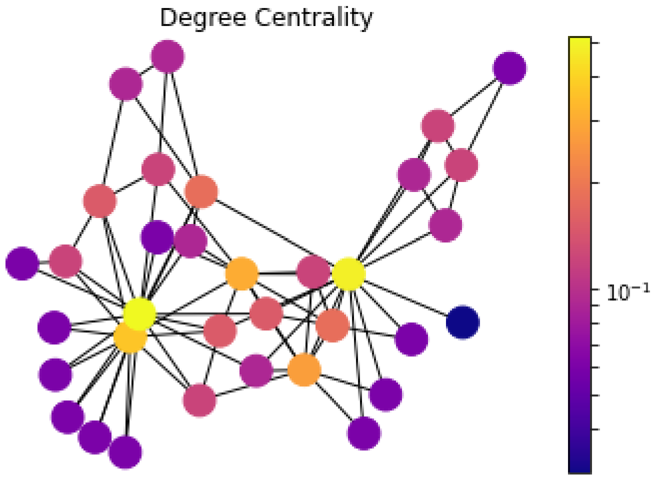
Twitter followers as of 2021

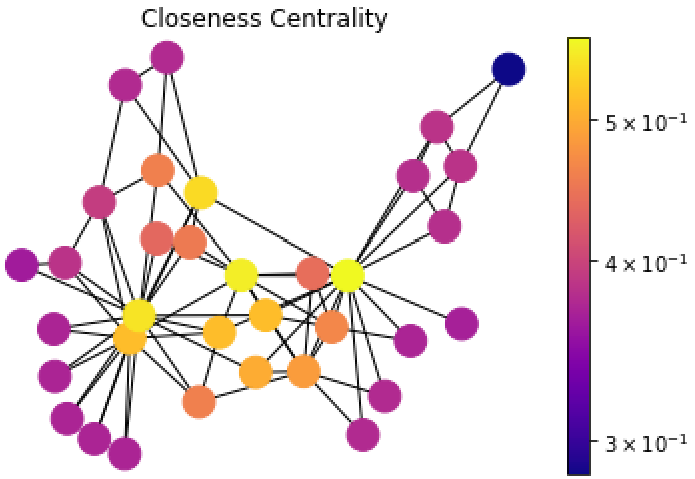
3.1 CENTRALITY MEASURES

- Closeness centrality
  - A central node has short paths to other nodes
  - Definition: $g_i = 1/\sum_j \ell_{ij}$

```
nx.closeness_centrality(G, node)
nx.closeness_centrality(G)
```

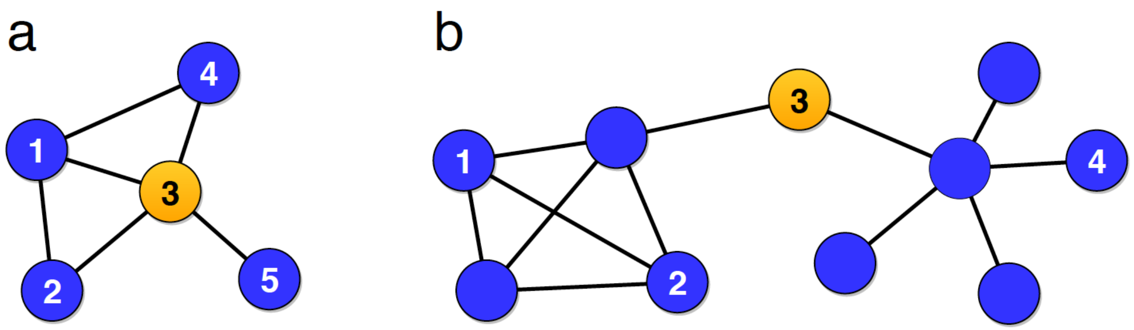
- Comparison
  - Closeness given higher rank to low-degree nodes that are friends of hubs





3.1 CENTRALITY MEASURES

- Betweeness centrality
  - A central node is part of *many paths*
  - Useful if measuring some diffusion process
    - Implementation depends on the process
  - Simplest: processes follow the shortest path
  - Definition:
    - sum of fraction of shortest paths between all pairs of nodes that pass through i
    - $b_i = \sum_{h,j} \frac{\sigma_{hj}(i)}{\sigma_{hj}}$
  - More about position in the network than degree



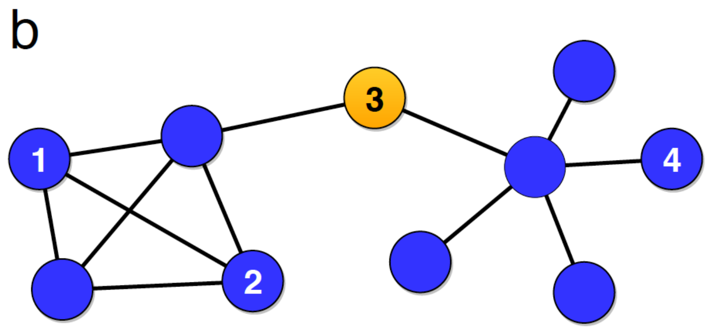
$k_3 = 4$
$b_3 = 3.5$

$k_3 = 1$
$b_3 = 20$

3.1 CENTRALITY MEASURES

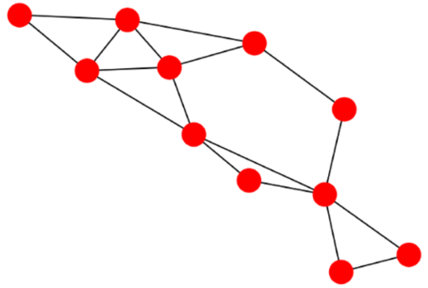
- Link betweenness
  - Focuses on links instead of nodes
  - Fraction of shortest paths passing through the link
  - Central links usually connect communities
    - Cutting central links: community detection
- Max betweenness: $\binom{N-1}{2}$
  - Can normalize b_i

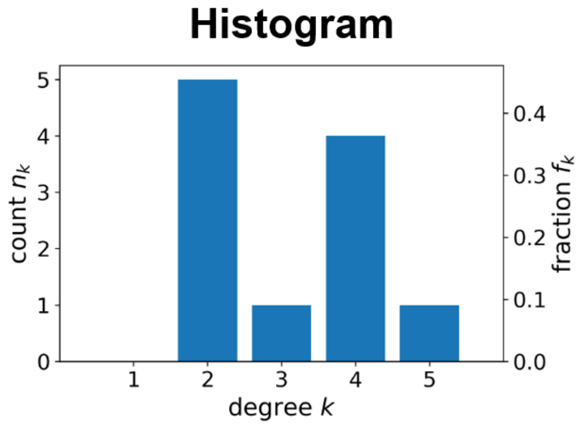
```
nx.betweenness_centrality(G)
nx.edge_betweenness_centrality(G)
```

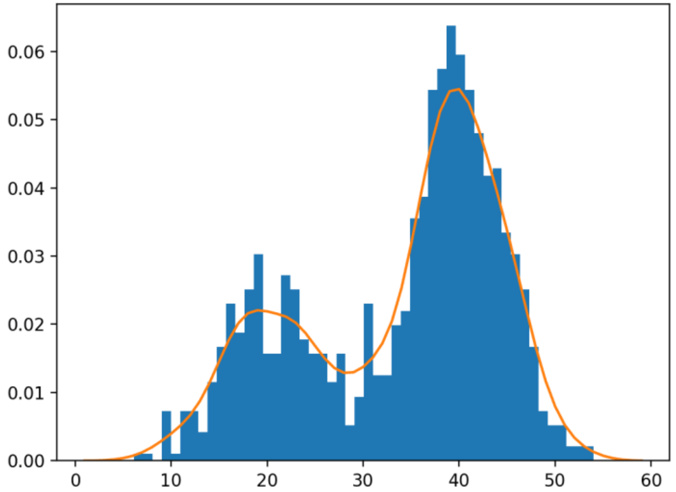


3.2 CENTRALITY DISTRIBUTIONS

- Large networks
  - Can't look into individual nodes/links
  - Solution: **statistical distributions**
- Discrete distributions
  - n_k: number of nodes with degree k
  - $f_k = n_k/N$: frequency of k
  - As $N \to \infty$, f_k converges to a **probability distribution function (PDF)** p_k
    - Check box 3.1
- Continuous distributions
  - More complicated, requires estimation
  - Data binning (linear, logbinning)
  - Kernel density estimators



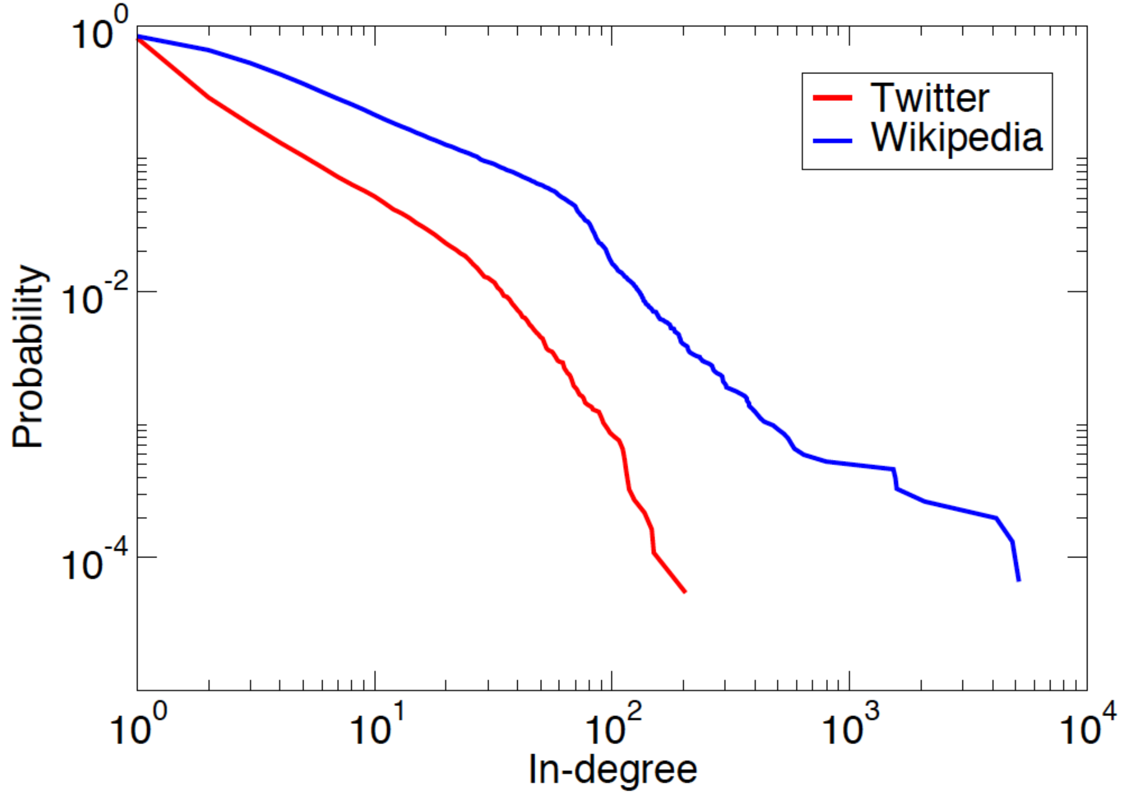




3.2 CENTRALITY DISTRIBUTIONS

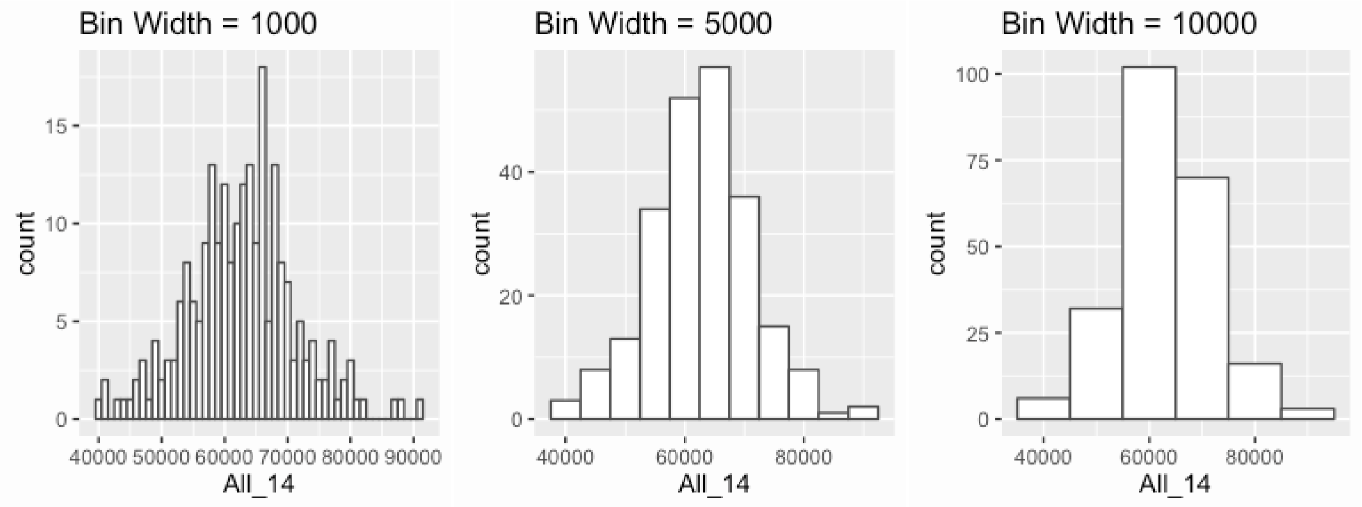
- Social network distributions are usually **heavy-tailed**
  - Few extremal nodes with very large values
  - Typically plotted log-log
  - Large statistical weight in the tail

- Measures of dispersion
  - Heterogeneity parameter κ
  - Definition: $\kappa = \frac{1}{N}\sum_i k_i^2 \times \frac{1}{\langle k \rangle^2}$
  - Low dispersion: $\kappa \approx 1$
  - High dispersion: $\kappa \gg 1$

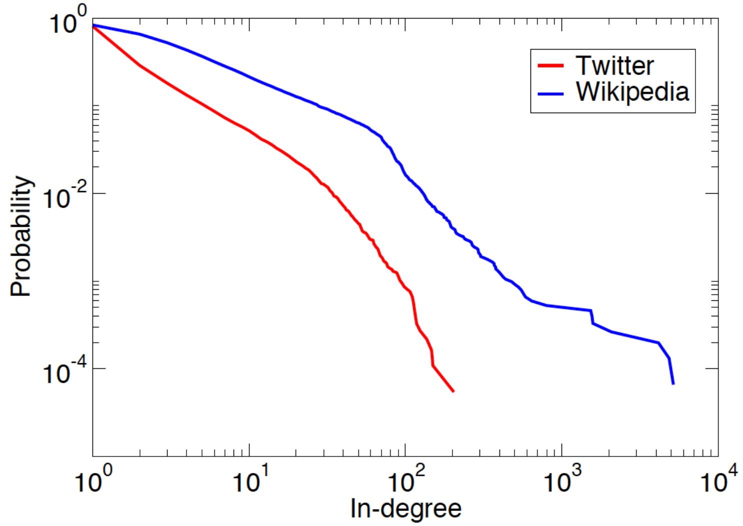


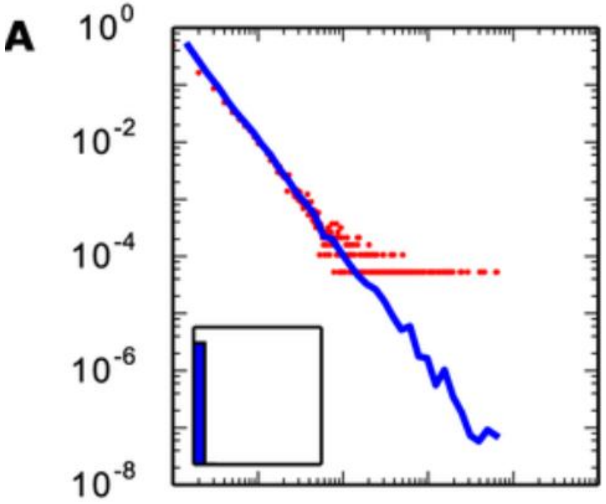
3.2 CENTRALITY DISTRIBUTIONS

- Data binning
  - Discrete with not enough data, or continuous
  - Free parameter: **bin width**
  - Plenty of ways to choose, good ballpark to start is around 20 bins



- The issue here:
  - Binning is bad with heavy tails
  - If linear bin sizes: one giant bin
  - Solution: **log-binning**
  - **The remaining issue: oversmooths data**
    - Makes things look like straight lines
    - Hides peaks





3.2 CENTRALITY DISTRIBUTIONS

- The actual solution: check with **cumulative distributions (CDF)**
- Definition: $P(x) = \sum_{i \geq x} f_i$
- Looks like a sigmoidal
- Sudden jumps represent spikes
- Works with continuous and discrete data
- Harder to interpret, weaker to noise

Pushshift Telegram dataset

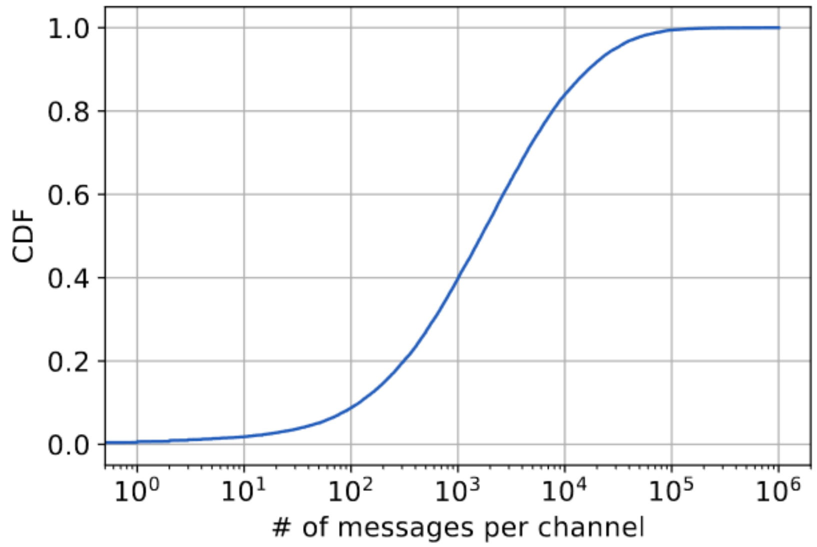
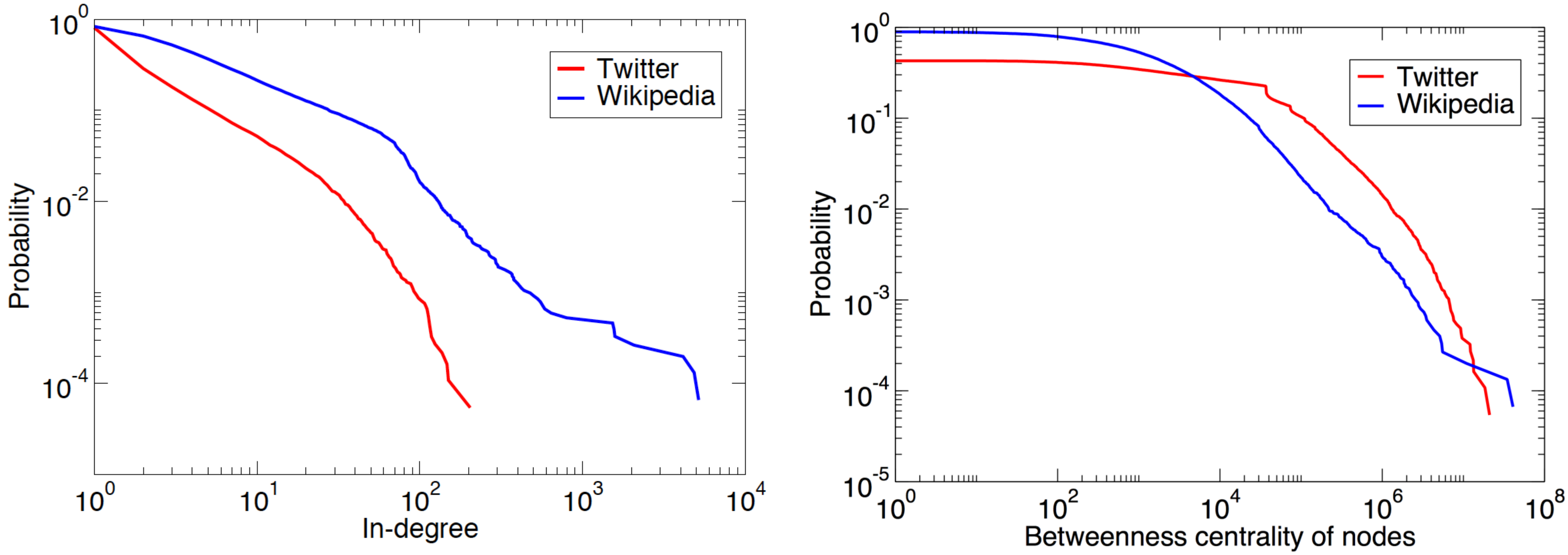


Figure 3: CDF of the number of messages per channel.

3.2 CENTRALITY DISTRIBUTIONS

- Different non-normalized metrics tend to be heavy-tailed
  - Both degree and betweenness of social networks



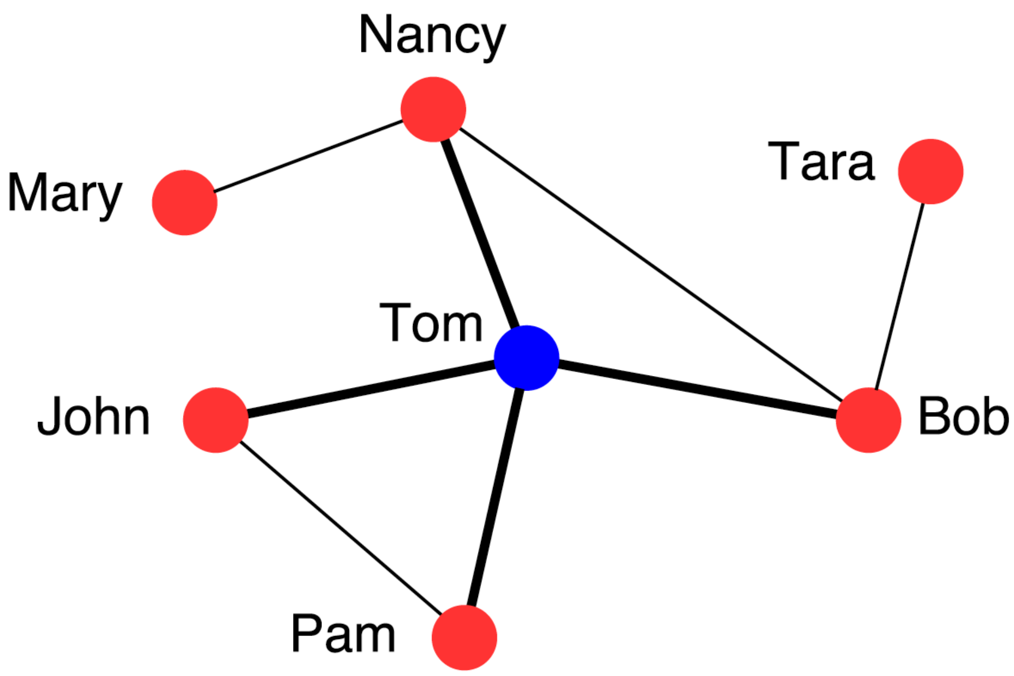
3.2 CENTRALITY DISTRIBUTIONS

- Many networks are heterogeneous/heavy-tailed

Network	Nodes (N)	Links (L)	Average degree ($\langle k \rangle$)	Maximum degree (k_{max})	Heterogeneity parameter (κ)
Facebook Northwestern Univ.	10,567	488,337	92.4	2,105	1.8
IMDB movies and stars	563,443	921,160	3.3	800	5.4
IMDB co-stars	252,999	1,015,187	8.0	456	4.6
Twitter US politics	18,470	48,365	2.6	204	8.3
Enron Email	36,692	367,662	10.0	1,383	14.0
Wikipedia math	15,220	194,103	12.8	5,171	38.2
Internet routers	190,914	607,610	6.4	1,071	6.0
US air transportation	546	2,781	10.2	153	5.3
World air transportation	3,179	18,617	11.7	246	5.5
Yeast protein interactions	1,870	2,277	2.4	56	2.7
C. elegans brain	297	2,345	7.9	134	2.7
Everglades ecological food web	69	916	13.3	63	2.2

3.3 THE FRIENDSHIP PARADOX

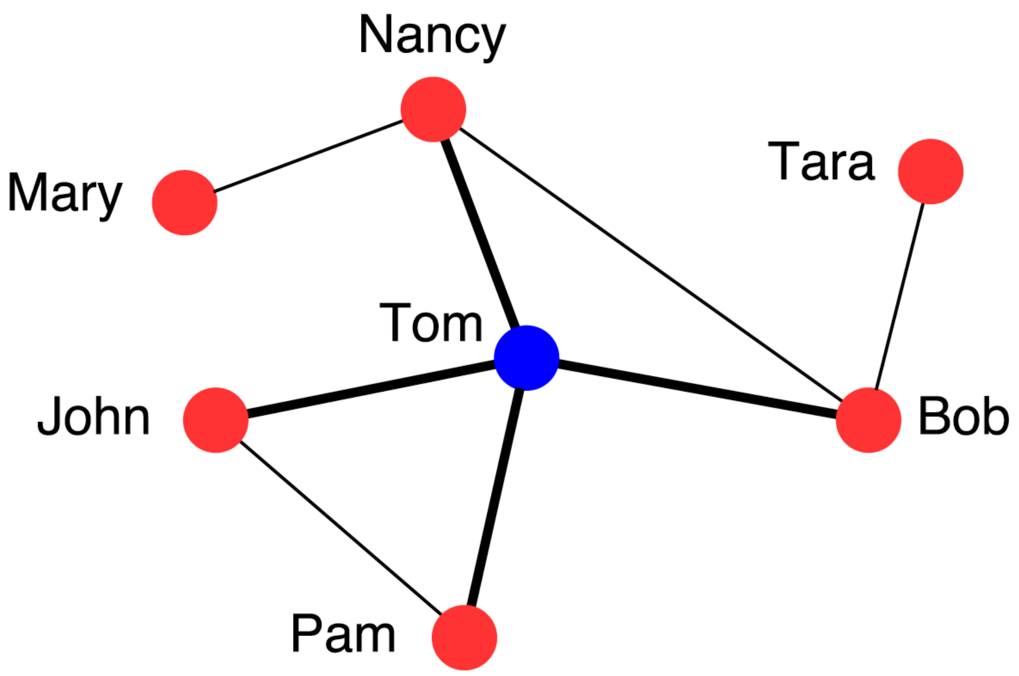
- The paradox: on average, your friends have more friends than you
- Why?
  - Random choice of *nodes*: all equal
  - Random choice of *links*: Tom wins
  - When counting friends, hubs appear disproportionally more
  - This biases the comparison



- Average degree of a node = **2.29**
- Average degree of the neighbors of a node = **2.83 > 2.29**

3.3 THE FRIENDSHIP PARADOX

- More explicitly:
  - Averaging degree of nodes: pick randomly
  - Averaging degree of neighbours: follow links, so nodes with degree k will be counted k times
  - The more hubs, the stronger the effect
- Sampling issue: comparing things calculated differently
- Social sampling issue: people tend to believe the opposite

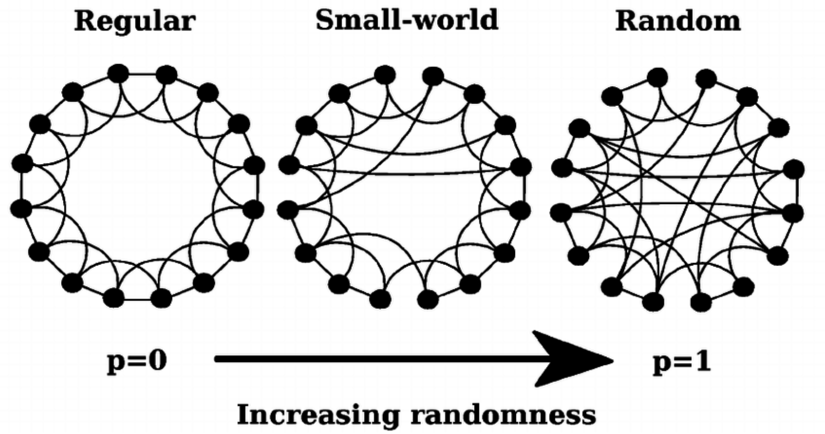


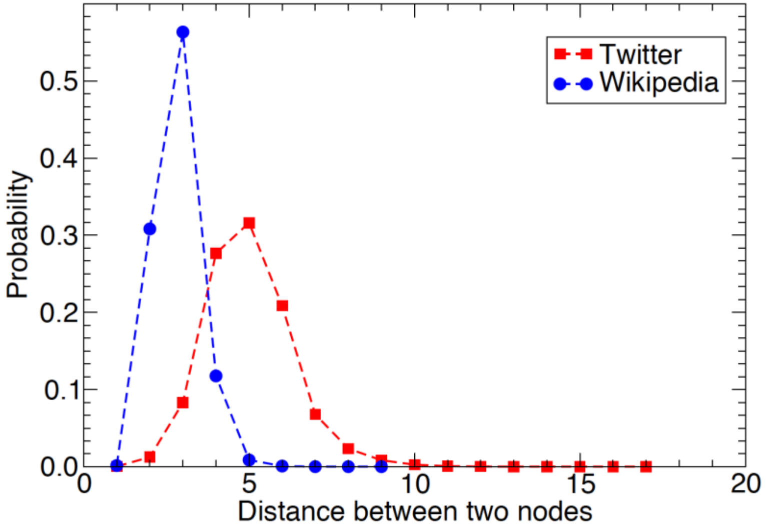
- Average degree of a node = **2.29**
- Average degree of the neighbors of a node = **2.83 > 2.29**

3.3 ULTRA-SMALL WORLDS

- Small-world: many networks have short average paths
- **Ultra-small world**: hubs can create *very* short paths
- Examples:
  - Air transportation networks (hub airports)
  - Social networks
  - Communication networks

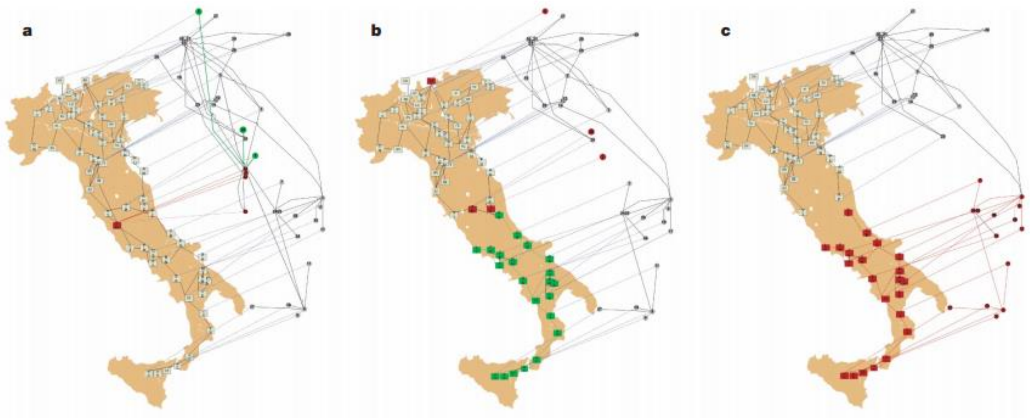
Watts-Strogatz model (Chapter 5)



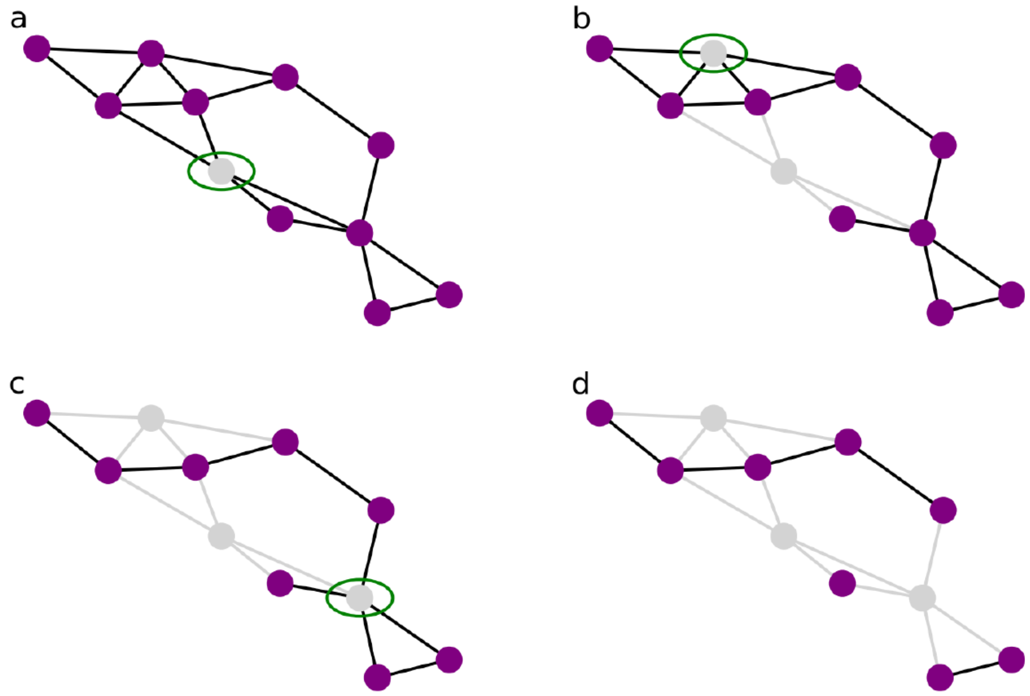


3.5 ROBUSTNESS

- A system is **robust** if the failure of some of its components does not affect its function
- On a network:
  - What happens to its *connectivity* as we remove nodes/links?
  - Components get disconnected
  - On multilayer networks: cascade failure

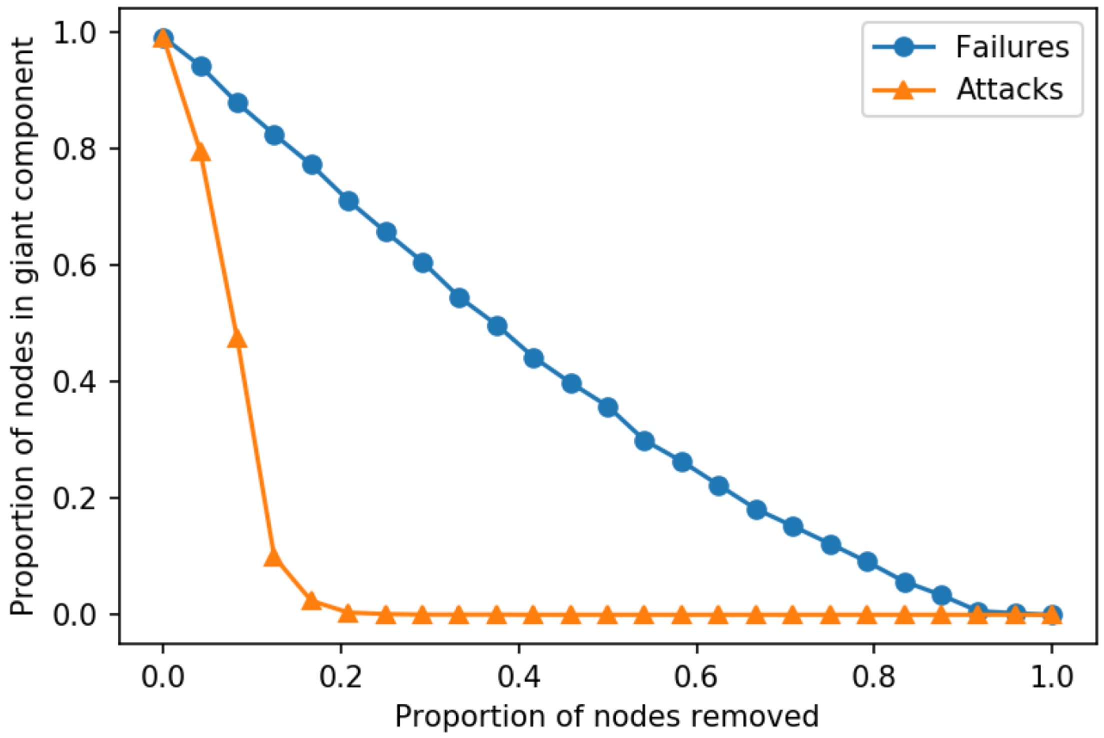


(2003 Italy blackout)



3.5 ROBUSTNESS

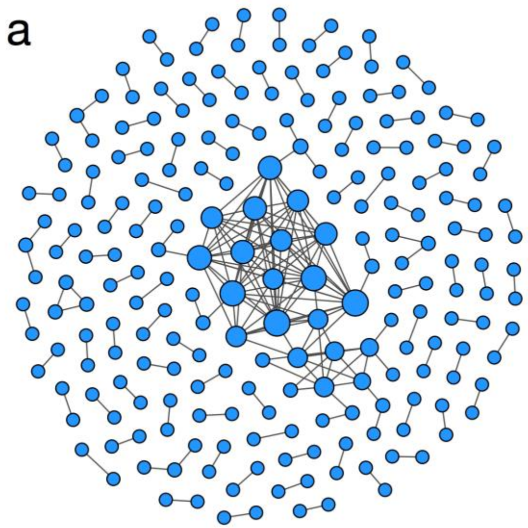
- How do we test robustness?
  - **Random failures**: choose nodes/links randomly
  - **Attacks**: target nodes/links by importance
    - E.g. remove nodes based on degree
- Measure robustness in terms of size of the giant component S vs fraction of removed nodes
- Most networks are **robust against random failure** but **fragile against targeted attacks**
  - Reason: hubs
- Designing networks: trade-off between efficiency and robustness to attacks

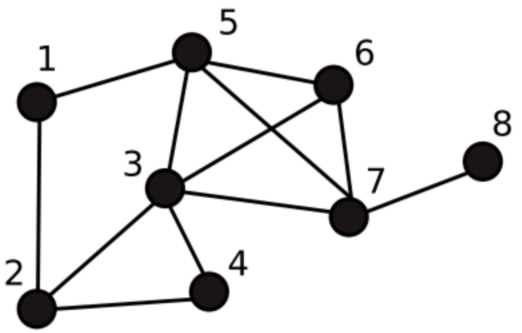


Flight network

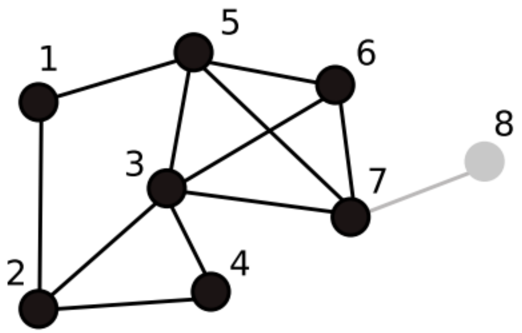
3.6 CORE DECOMPOSITION

- Chapter 2: Core-periphery structure
- **K-core decomposition**: remove low-degree nodes recursively
  1. Start with $k = 0$
  2. Remove all nodes with degree k. Removed nodes are in the **k-shell**, remaining nodes are in the **k-core**
  3. Recompute degrees
  4. Set $k \to k + 1$ and iterate from 2 until there are no more nodes in the k-core.

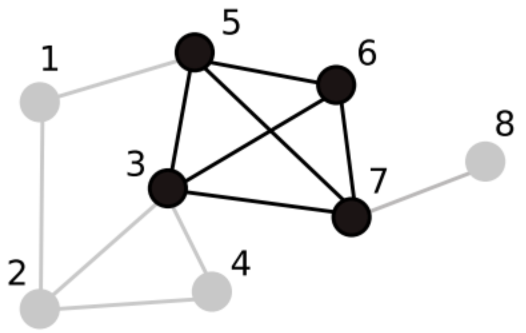




(a) 1-core



(b) 2-core



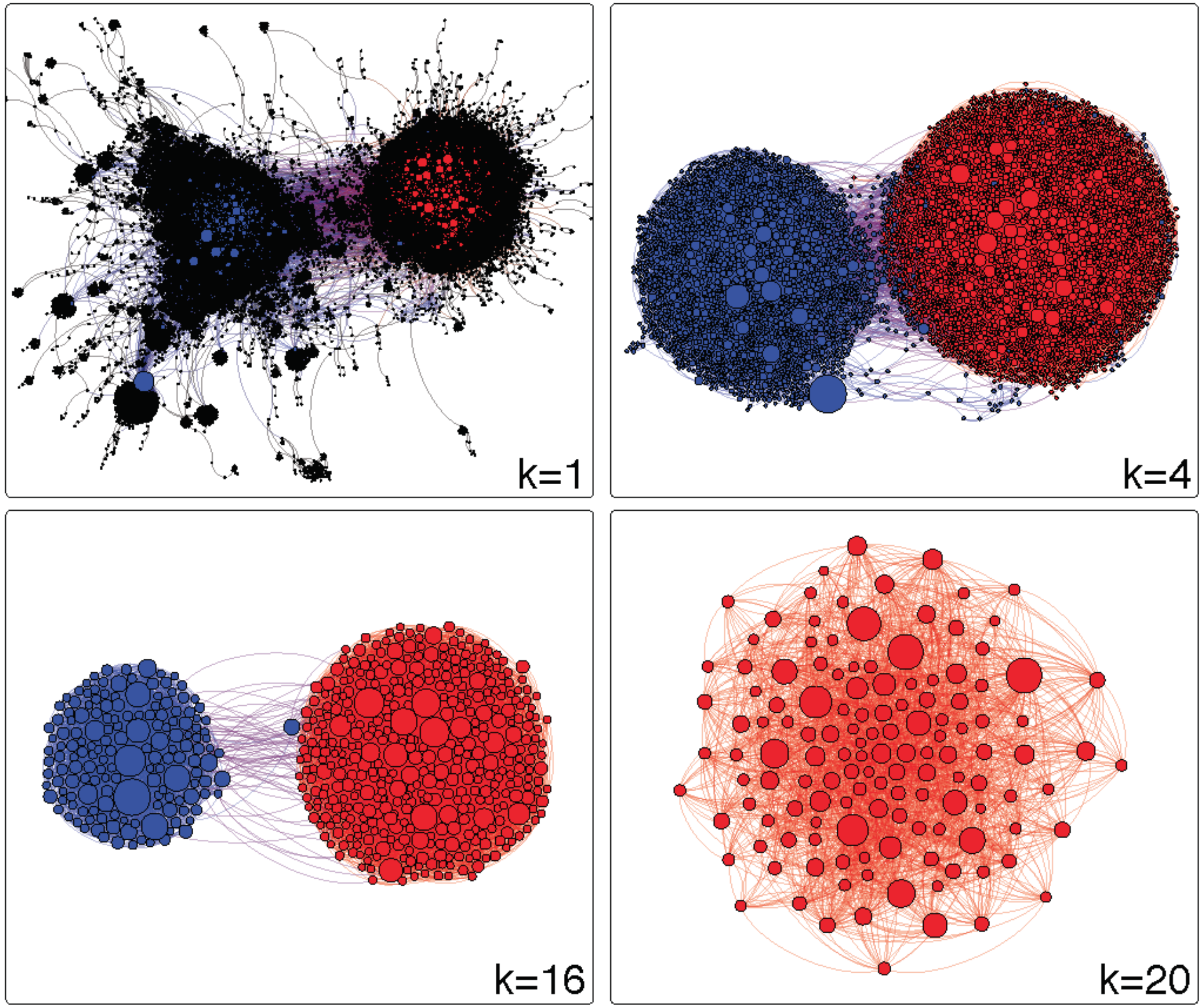
(c) 3-core

3.6 CORE DECOMPOSITION

- Why?
  - Filter out peripherical nodes
  - Useful for visualization
  - Useful for in-depth analysis of subnetworks

- On NetworkX:

```
nx.core_number(G)
list(nx.k_shell(G,k))
list(nx.k_core(G,k))
# innermost (max-degree) core subnetwork
list(nx.k_core(G))
```



SUMMARY

- Different node/link centrality measures with different goals
  - Degree, closeness and betweenness centrality
- Statistical distributions of social networks are usually heavy-tailed
  - Careful when analysing/plotting it
- Networks can have non-intuitive properties (friendship paradox)
- Networks are robust against random failure, but weak against target attack