Introduction to Social Media Analytics (Lec 10)

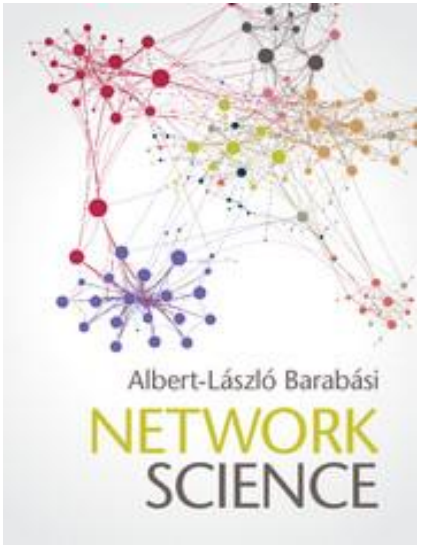
Hao PENG
Department of Data Science
City University of Hong Kong
https://haoopeng.github.io/

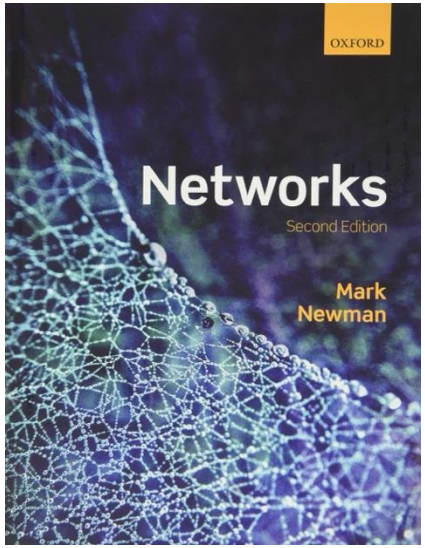
Topics for this week

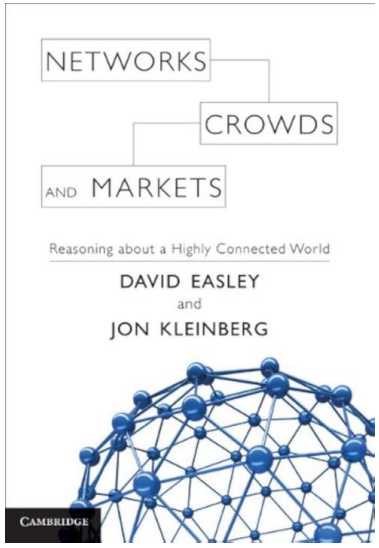
- Network Science
  - Background
- Characteristics of Social Networks
  - Short average distance
  - High clustering coefficient
  - Power-law degree distribution
  - Social influence/hubs
  - Strong and weak ties
  - Community structure
- Applications
  - Community detection
  - Cascade prediction
  - Information diffusion
  - Network visualization

Network Science

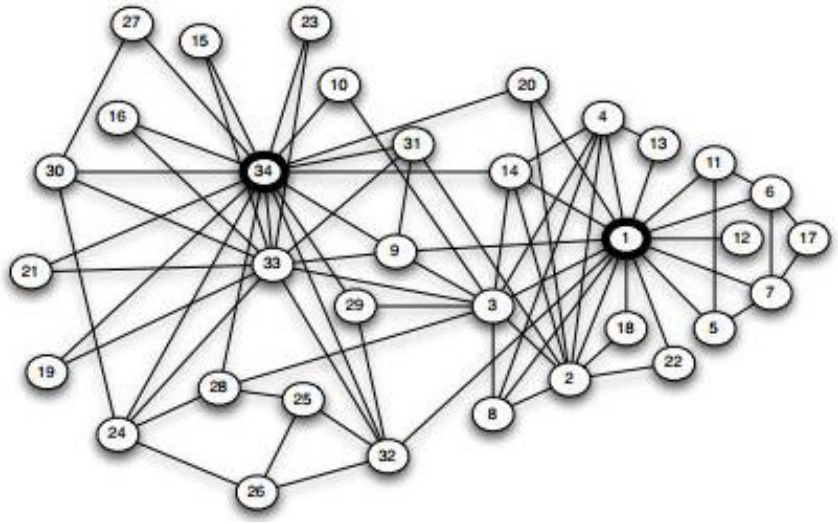
- In the domain of network science, researchers don't study networks in the abstract, but instead, they study numerous real-world complex networks to understand their properties, dynamics, and behaviors.
- Examples include social networks, transportation networks, WWW, gene regulatory networks, citation networks, collaboration networks, etc.
- Key scholars: Mark Newman, Albert-László Barabási, Jon Kleinberg, Duncan Watts...







Networks before ``Big Data”



Friendship network in a karate club (Zachary 1977)

FIGURE 3

QUANTIFIED MATRIX OF RELATIVE STRENGTHS OF THE RELATIONSHIPS IN THE KARATE CLUB: THE MATRIX C

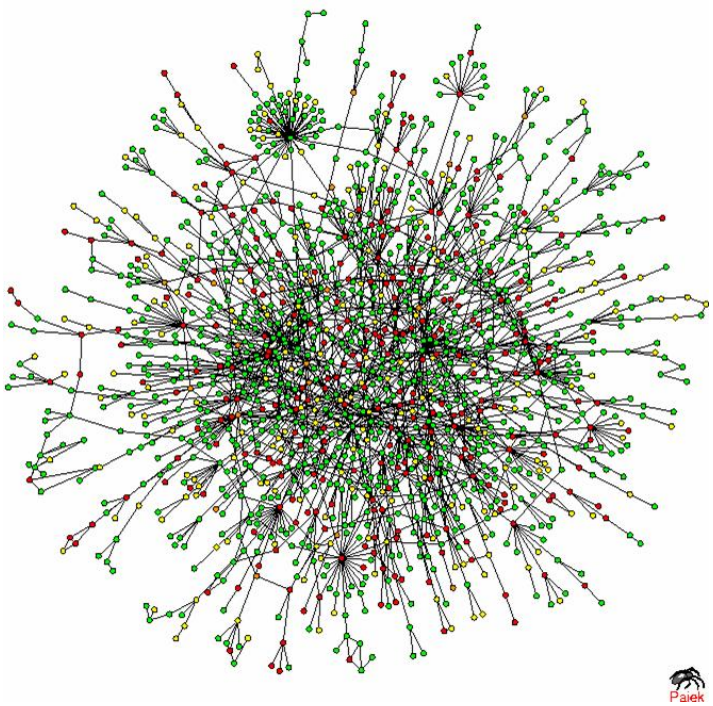
Individual Number

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	0	4	5	3	3	3	3	2	2	0	2	3	2	3	0	0	0	2	0	2	0	2	0	0	0	0	0	0	0	0	2	0	0
2	4	0	6	3	0	0	0	4	0	0	0	0	0	5	0	0	0	1	0	2	0	2	0	0	0	0	0	0	0	0	2	0	0	0
3	5	6	0	3	0	0	0	4	5	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	3	0
4	3	3	3	0	0	0	0	3	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	3	0	0	0	0	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	3	0	0	0	0	0	5	0	0	0	3	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	3	0	0	0	2	5	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	2	4	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	2	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	4	3	
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	
11	2	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	3	5	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2		
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	4		
17	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
18	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2			
20	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1			
22	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0			
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	4	0	2	0	5	4		
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	0	0	2	0	0		
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	2	0	0	0	7	0	0		
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	2		
28	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	0	0	0	0	0	4		
29	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2		
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	4	0	0	0	0	2		
31	0	2	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3		
32	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	7	0	0	2	0	0	4	4	
33	0	0	2	0	0	0	0	0	3	0	0	0	0	0	3	3	0	0	1	0	3	0	2	5	0	0	0	0	0	4	3	4	0	5
34	0	0	0	0	0	0	0	0	4	2	0	0	0	3	2	4	0	0	2	1	1	0	3	4	0	0	2	4	2	2	3	4	5	0

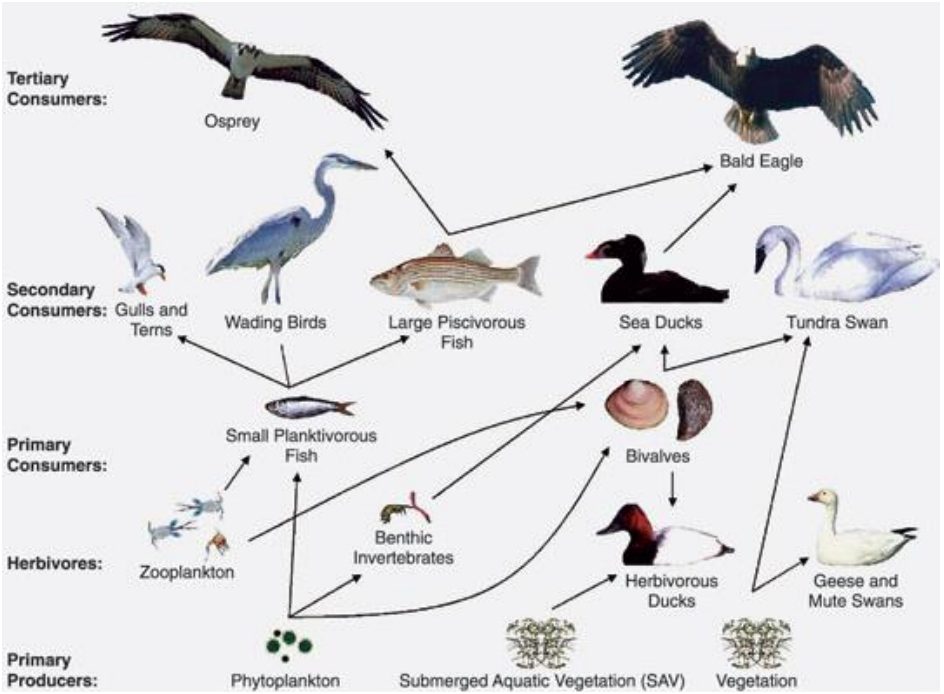
The scale given in the text (p. 461) was used to assign values of relative strength/weakness of the relationships in the club. This matrix gives the values assigned to each edge specified in matrix E (Figure 2). The value represents the number of contexts in which interaction took place between the two individuals involved. The row/column ordering is the same as in Figure 2.

Adjacency Matrix of tie strength in karate club network

Biological Networks

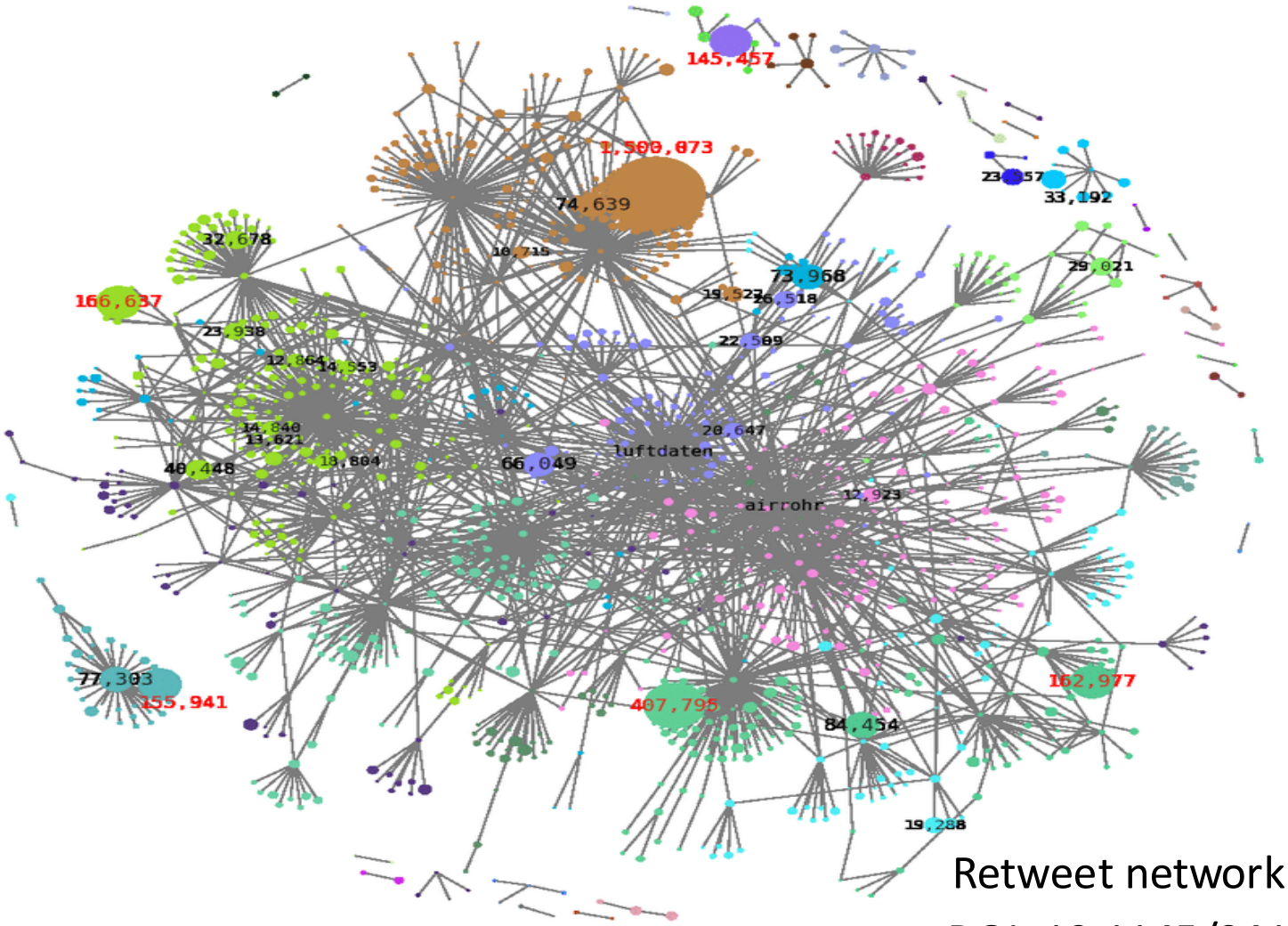


Protein-protein interactions (Jeong)



Chesapeake Bay Waterbird Food Web

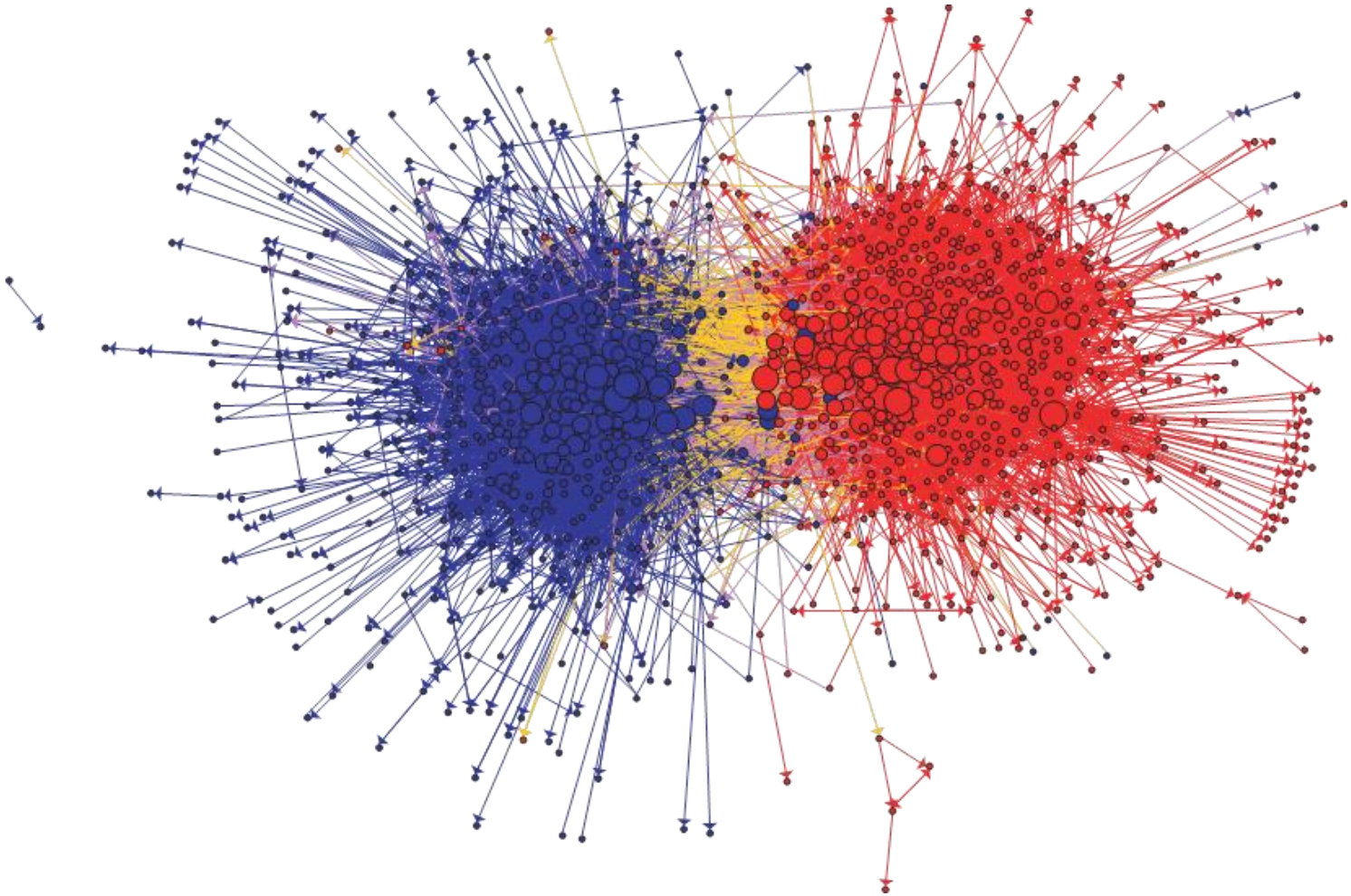
Networks after ``Big Data”



Retweet network (Hamm 2021)

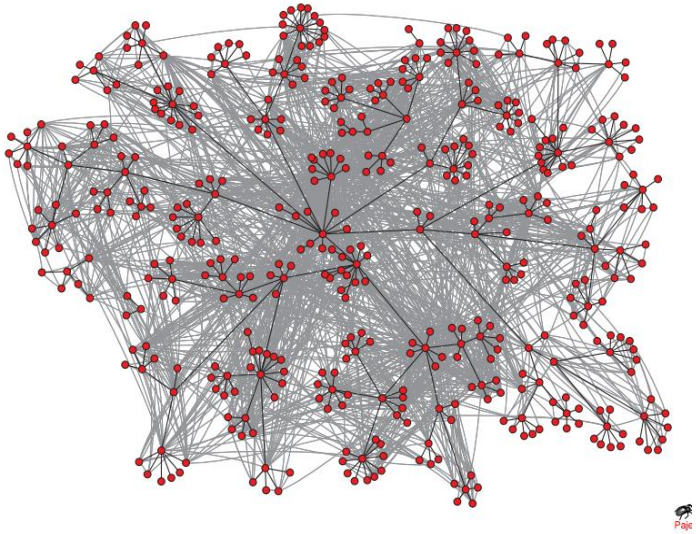
DOI: 10.1145/3411764.3445667

Information Networks

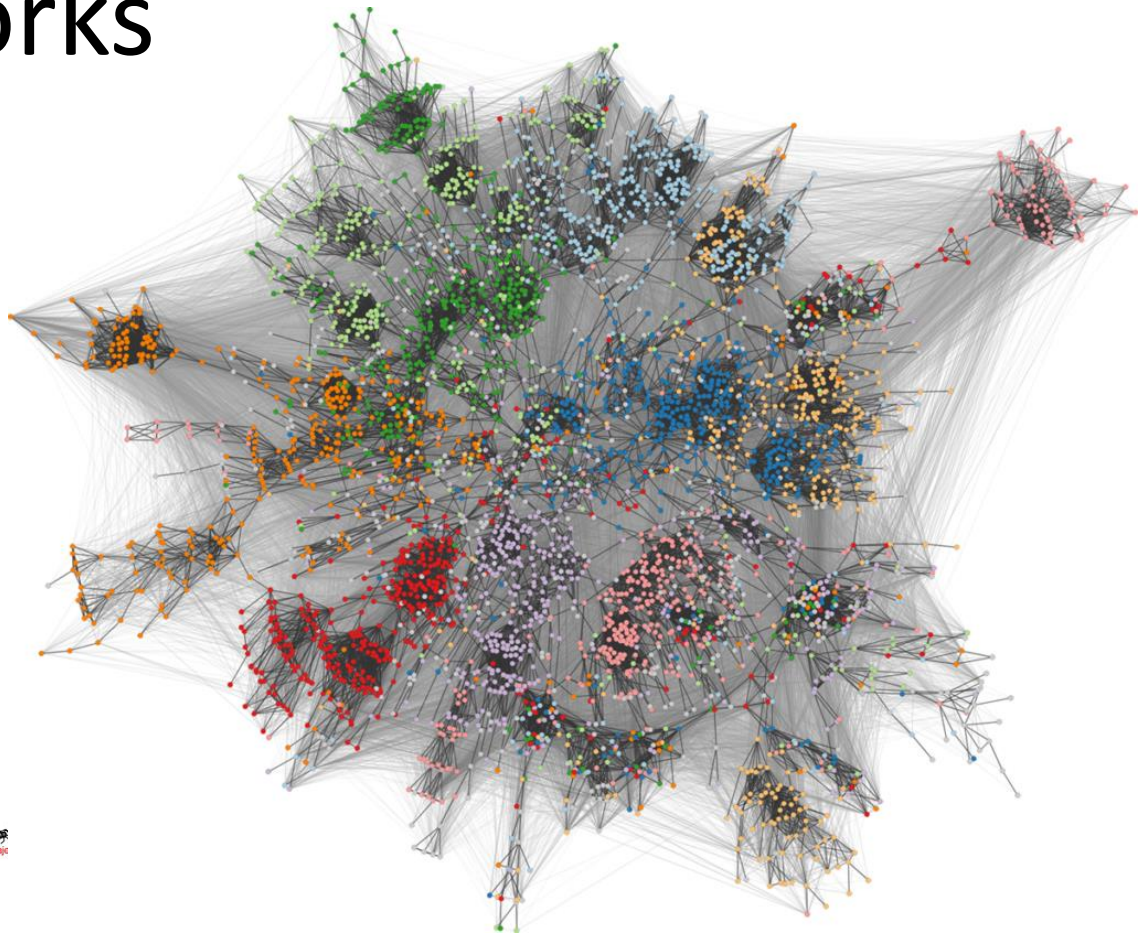


Communication between left-wing and right-wing political blogs (Adamic and Glance)

Social Networks



E-mail communication network among 436 HP employees

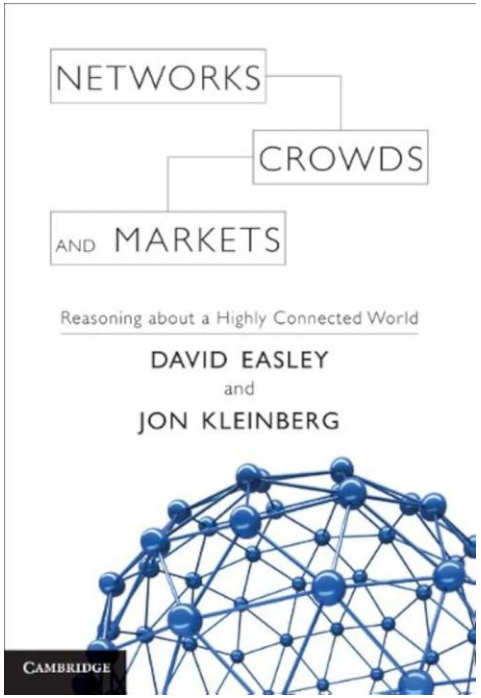


Facebook network of university students (color shows dormitory)

And more…

This is why we need statistical models and algorithms to study networks!

- Trade networks
- Financial networks
- Citation networks
- Collaboration networks
- Mentorship networks
- Epidemic networks
- …

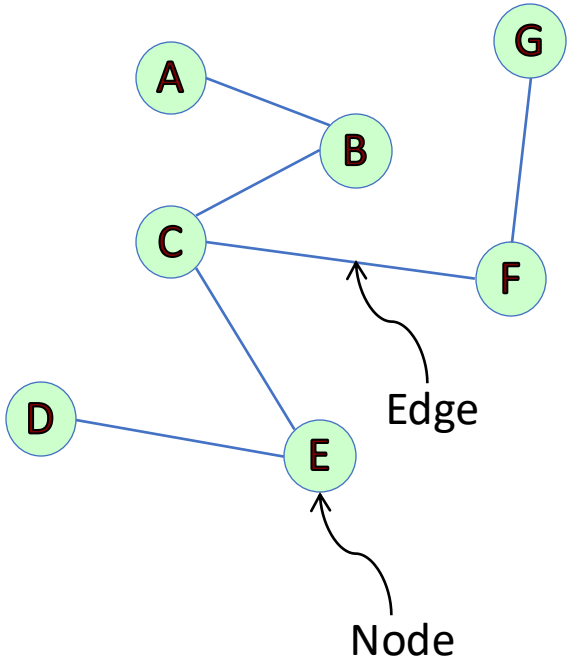


Free online: https://www.cs.cornell.edu/home/kleinber/networks-book/

Network definition & vocabulary

A representation of **connections** among a set of **items**.

- Items are called **nodes** (vertices)
- Connections are called **edges** (or link or ties)

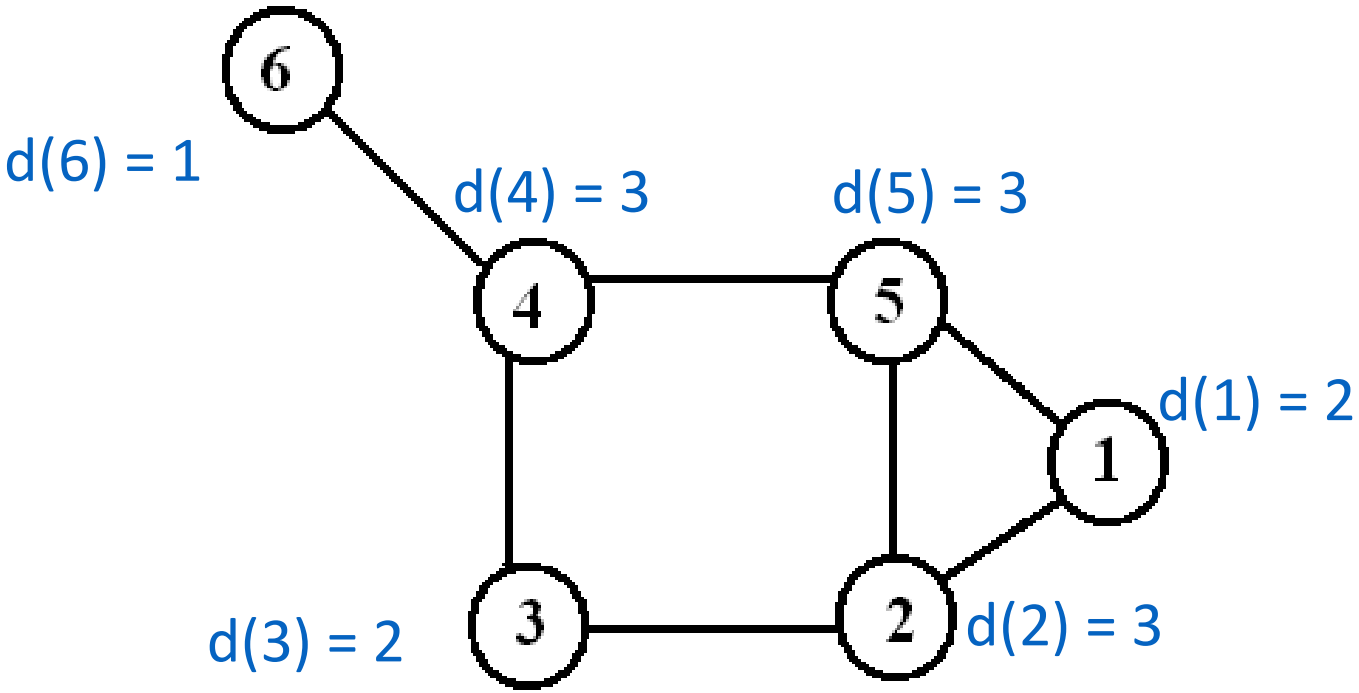


Different types of networks:

- Undirected, directed
- Unweighted, weighted
- Connected, disconnected

Degree of nodes

- Number of edges incident on a node
- The degree of node i: d(i)



Average distance

The distance $d_{i,j}$ between nodes *i* and *j* is the length of the shortest path from *i* to node *j*.

Mean distance: $l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i>j} d_{i,j}$

Note that if there are disconnected nodes $l = \infty$. Hence, disconnected nodes are typically ignored in the mean.

(Somewhat) Surprising finding: many real networks are navigable (connected) & the distance tends to be small (e.g., < 6 in many social networks).

The small-world experiment

On average, it took only six steps to reach the target.

Fig: Baruch Barzel

J. Travers and S. Milgram, *Sociometry* **32**,425–443 (1969)

https://en.wikipedia.org/wiki/Small-world_experiment

Six-degree replicated

Peter Sheridan Dodds, Roby Muhamad, and Duncan Watts replicated Milgram's experiment with email:

- 18 targets in 13 countries
- 60,000 participants from 166 countries

"Targets included a professor at an Ivy League university, an archival inspector in Estonia, a technology consultant in India, a policeman in Australia, and a veterinarian in the Norwegian army."

Average: 7 steps

Even shorter distance on FB

The Anatomy of the Facebook Social Graph

Johan Ugander[1,2*], Brian Karrer[1,3*], Lars Backstrom[1], Cameron Marlow[1†]
1 Facebook, Palo Alto, CA, USA
2 Cornell University, Ithaca, NY, USA
3 University of Michigan, Ann Arbor, MI, USA
∗ These authors contributed equally to this work.
† Corresponding author: cameron@fb.com

Average distance on FB: 4.74 (world-wide) and 3.57 (U.S.)

Clustering (transitivity)

There are many triangles in social networks, why?

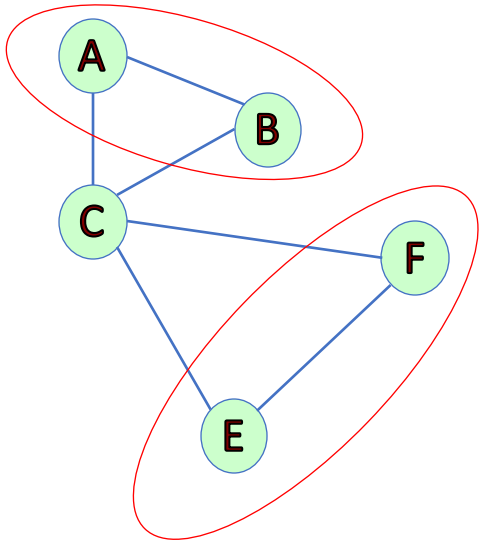
(Local) clustering coefficient C_i of node *i:*

$$C_i = \frac{\#\text{ of pairs of C's neighbors who are connected}}{\#\text{ of pairs of C's neighbors}}$$

Nodes with degree < 2 have zero C_i

Node C has local clustering coefficient 2/6.

Mean clustering coefficient (MCC): $C = \frac{1}{n}\sum_i C_i$



<u>Interpretation of *C*:</u>

The average probability of two nodes being connected, conditioning that they are both connected to a common neighbor.

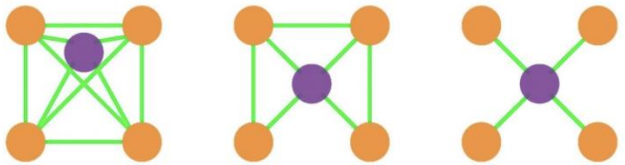
High clustering coefficient

In real-world social networks, the clustering coefficient tends to be very high — my friends tend to be friends with each other.

Example:

- Facebook friendship network
  Clustering coefficient: 0.3-0.5 or even higher

- Film actors from IMDB
  Clustering coefficient: 0.79



Properties of real social networks

- Short average distance
- High clustering coefficient
- Power-law degree distribution
- Homophily, (dis)assortative mixing
- Social influence and hubs
- Strong ties and weak ties
- Community structure
- And many more….

Models of network formation

- Erdős–Rényi model (ER random networks)
- Barabási–Albert model (scale-free networks)
- Watts–Strogatz model (small-world networks)
- Many others…

These network generation models help us understand how the real-world networks come into current shape.

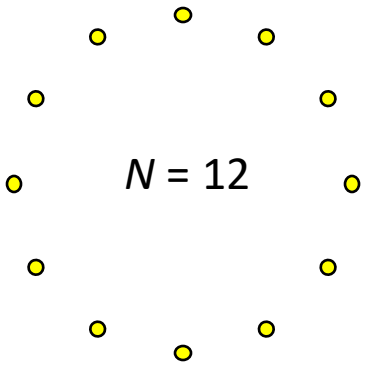
Random graph (ER model)

- *N* nodes
- Each pair of nodes has a probability *p* of being connected
- Average degree, $k = pN(N-1)/N \approx pN$

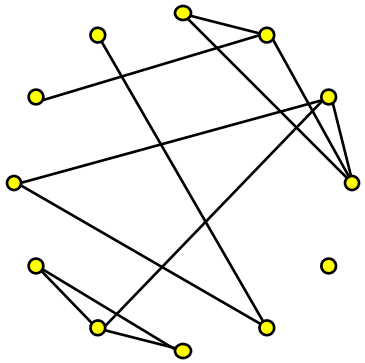
The network undergoes a phase transition in connectivity when the average degree reaches 1.

When k < 1, it is almost disconnected.

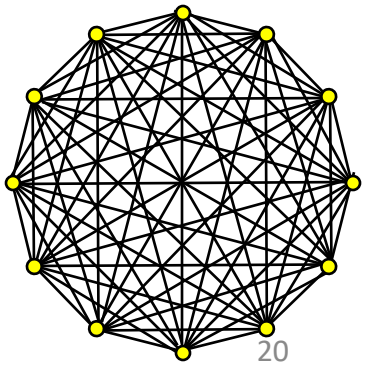
$p = 0.0$; $k = 0$ $N = 12$



$p = 0.09$; $k = 1$



$p = 1.0$; $k = 11$



Erdős and Renyi (1959)

Random graph phase transition

We can prove that the structure of a *random network* changes in "sudden" ways as p changes (phase transition).

- When $k < 1$, there is almost surely no existence of a "giant component" larger than $O(\log(n))$
- When $k > 1$, there is almost surely exist a "giant component" *containing a large fraction of nodes (*proportional to $O(n)$*), and all other components are small.*
- What about its average distance and clustering coefficient?
  (it has short avg distance: log N / log k)
  (it has a low clustering coefficient which equals to *p*, why?)

  Avg. degree: $k = pN = 2M/N \Rightarrow p = 2M/N^2$; *p is close to 0 when N is large*

Properties of the ER model

The Erdős–Rényi (random graph) model is easy to study mathematically. We understand many of its properties.

Yet, random graphs are still very different from real networks:

1. Low clustering coefficient (→ 0 as *n* gets larger)
2. Poisson degree distribution
3. No super-star nodes
4. No community structure
5. No homophily / (dis)assortative in degree
6. Except, it has short average distance

Overall, not very useful to model real social networks.

High clustering in social networks

Example:

- Film actors from IMDB

  Clustering coeff.: 0.79

- Random graph with the same number of N nodes and M links

  Clustering coeff.: 0.00027 (equals to *p* in the ER model)

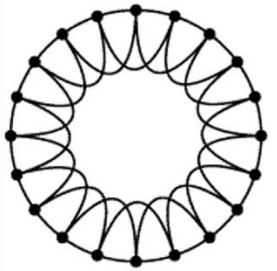
  avg. degree: $k = pN = 2M/N$

  $\Longrightarrow p = 2M/N^2$, *close to 0 when N is large*

Can we design a network generation model (other than the ER graph) to have both (i) short avg distance (ii) high clustering?

Watts–Strogatz model

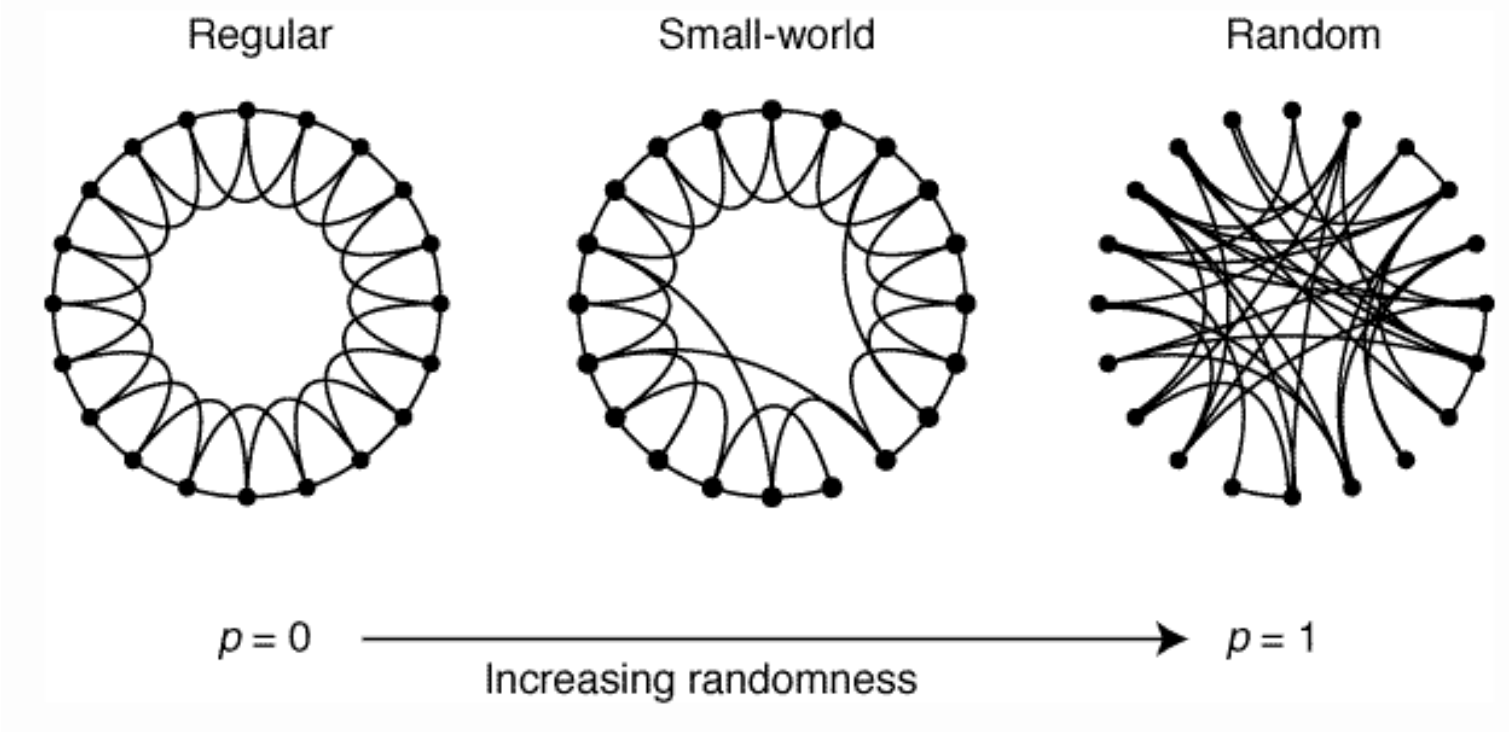
Motivation: Social networks often exhibit both short avg. distance and high clustering coefficient. However, the two properties seem to go against each other (e.g., in the following regular lattice network, what is its clustering coefficient and avg distance?).



WS Network generation process:

- Start with a ring lattice of n nodes; each node is connected to its $2r$ nearest nodes. (think of each node as a tribe in ancient times; each only contacts nearby tribes)
- Fix a parameter $p \in [0,1]$.
- For each node u, consider the edges to its r nearest (clockwise) neighbors. With prob. p, rewire each edge so it connects u to a random node in the network.
- Continue this process until each edge has been considered for rewiring once.

WS model (Small-world networks)

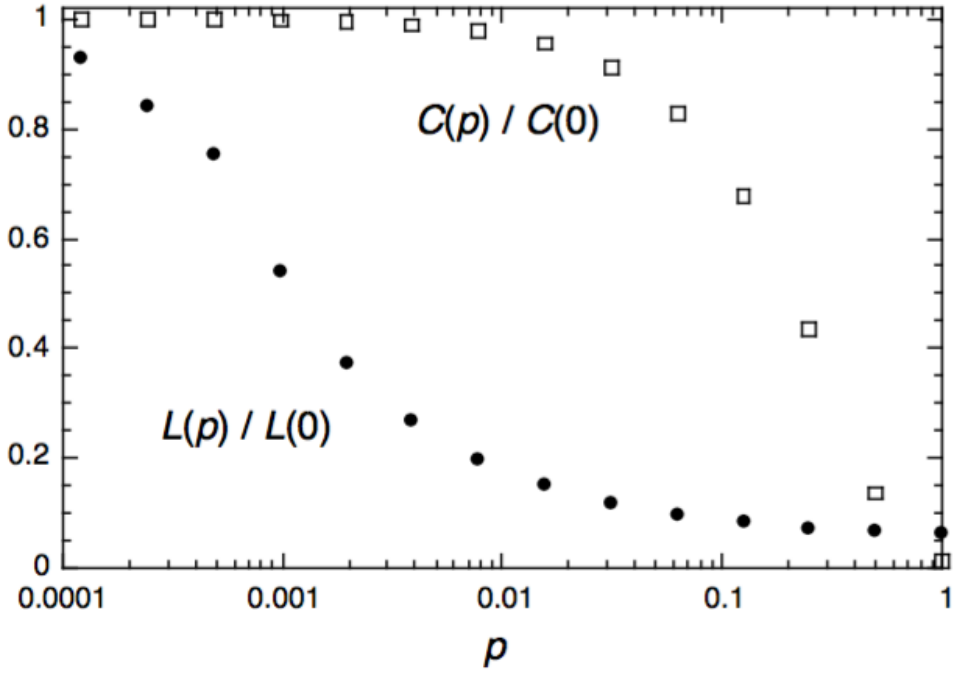


What is the mean distance and clustering coefficient for different p?

Watts and Strogatz. *Nature*, 1998.

Model simulation

Simulation with $k = 10$ and $n = 1000$ averaged over 20 runs.
$C(p)$ = mean clustering coefficient of network with parameter p.
$L(p)$ = mean distance of network with parameter p.
L(0), C(0) are for the regular lattice before rewiring.



L drops very fast as p increase.
C remains high for relatively large p.

This simple model can reproduce certain features of social networks.
But it has fixed degree for all nodes!

What about the degree distribution in ER random networks?

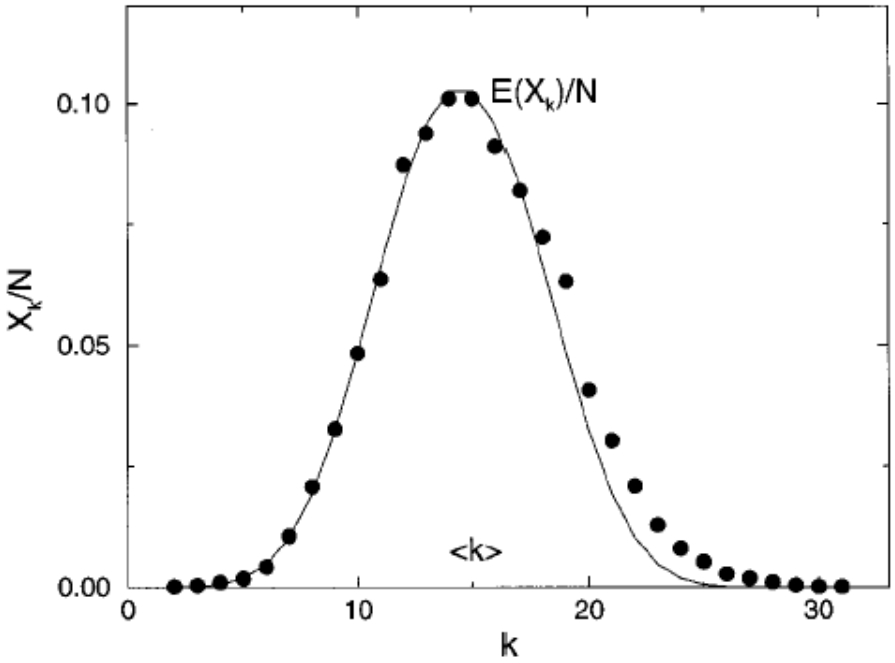
Watts and Strogatz. *Nature*, 1998.

Poisson degree dist. in ER graphs

k_i = degree of node i

p_k = fraction of nodes that have degree k

ER random graph results in a Poisson degree distribution.



Degree distribution of a random graph, with N = 10,000; p = 0.0015; k = 15. (Curve is a Poisson curve.)

How many nodes in this ER network have > 30 connections to others?

What about the degree distribution in real social networks?

Power-law degree distribution

Most social networks have a power-law degree distribution:

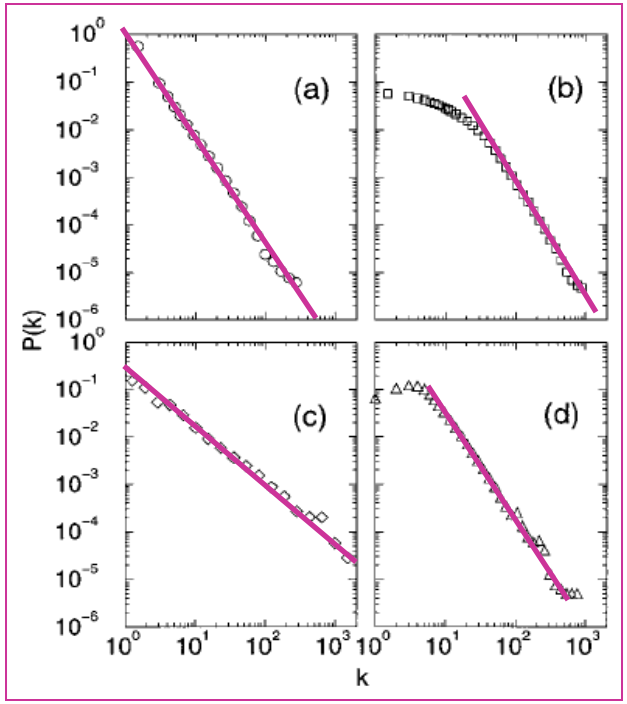
$$p_k \sim k^{-\alpha}, \text{with } \alpha \; in \; [2, 3.5],$$

Power-laws are straight lines in log-log space: $\log(p_k) \sim -\alpha \log(k)$

Power laws in real networks:

(a) WWW hyperlinks
(b) co-starring in movies
(c) co-authorship of physicists
(d) co-authorship of neuroscientists

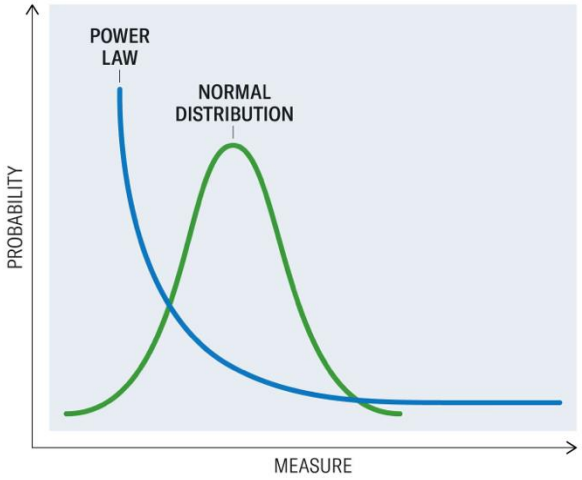
This means that there exists many "super-star" nodes in real networks! (missing in both ER & SW networks)

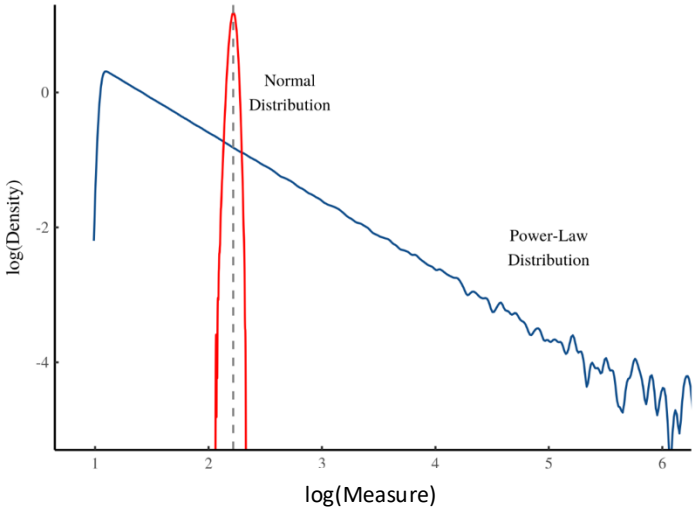


Power-law distribution

A statistical phenomenon where a small num. of items occupy a large fraction of outcomes (wealth, popularity, success, etc.).

It is perhaps the most well-know distribution in social & economic sciences.
Also known as the “80/20 rule”, “long-tail / scale-free distribution”:





How can we design a network generation model to produce the power-law degree distribution?

Barabási–Albert (scale-free) model

A model with preferential attachment:

- Start with a small number m_o of nodes.
- At each time step, add a new node with m edges to existing nodes in the network.
- The probability that the new node connects to node u is $k_u / \sum_j k_j$. (This is the preferential attachment ingredient.)
- After t steps, the network has $m_o + t$ nodes and mt edges.

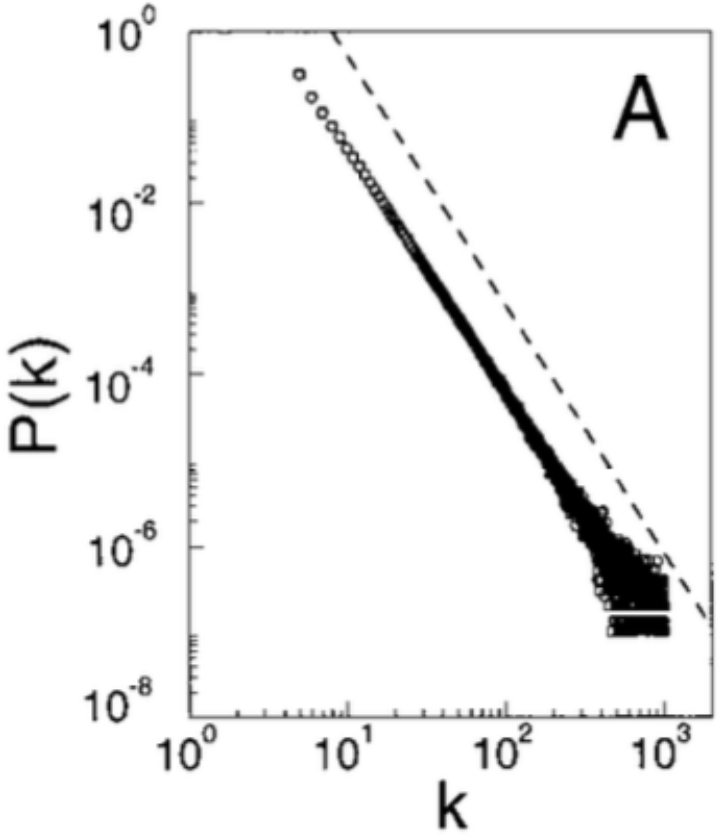
Emergence of Scaling in Random Networks

ALBERT-LÁSZLÓ BARABÁSI AND RÉKA ALBERT Authors Info & Affiliations

SCIENCE • 15 Oct 1999 • Vol 286, Issue 5439 • pp. 509-512 • DOI: 10.1126/science.286.5439.509

Simulations of the BA model

- $m = m_0 = 5$
- The model produces a power-law degree distribution with an exponent -2.9 +/- 0.1.
- t = 150k and t = 200k are shown.
- The degree distribution is independent of t or size (a.k.a., scale-free).

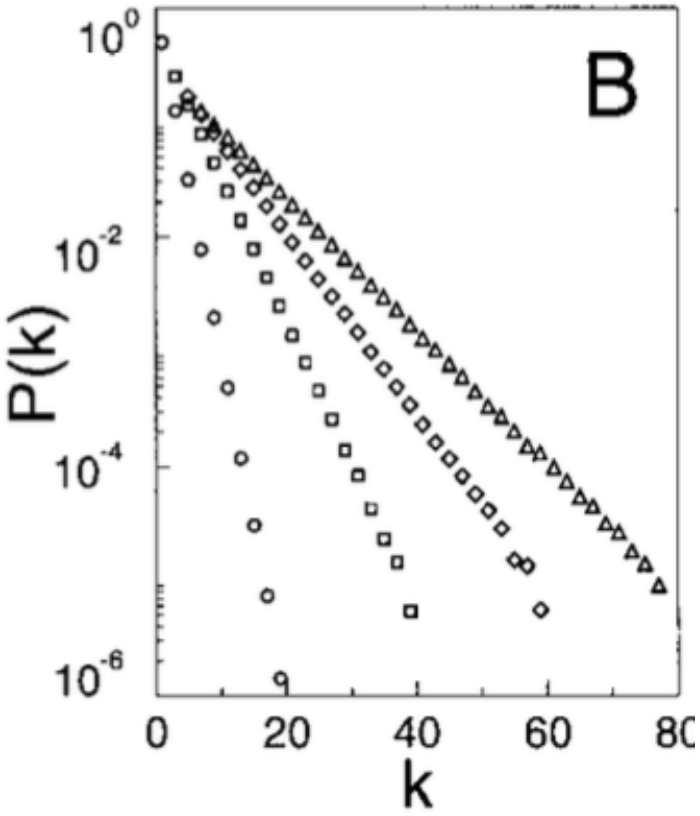


Barabási and Albert. *Science*, 1999.

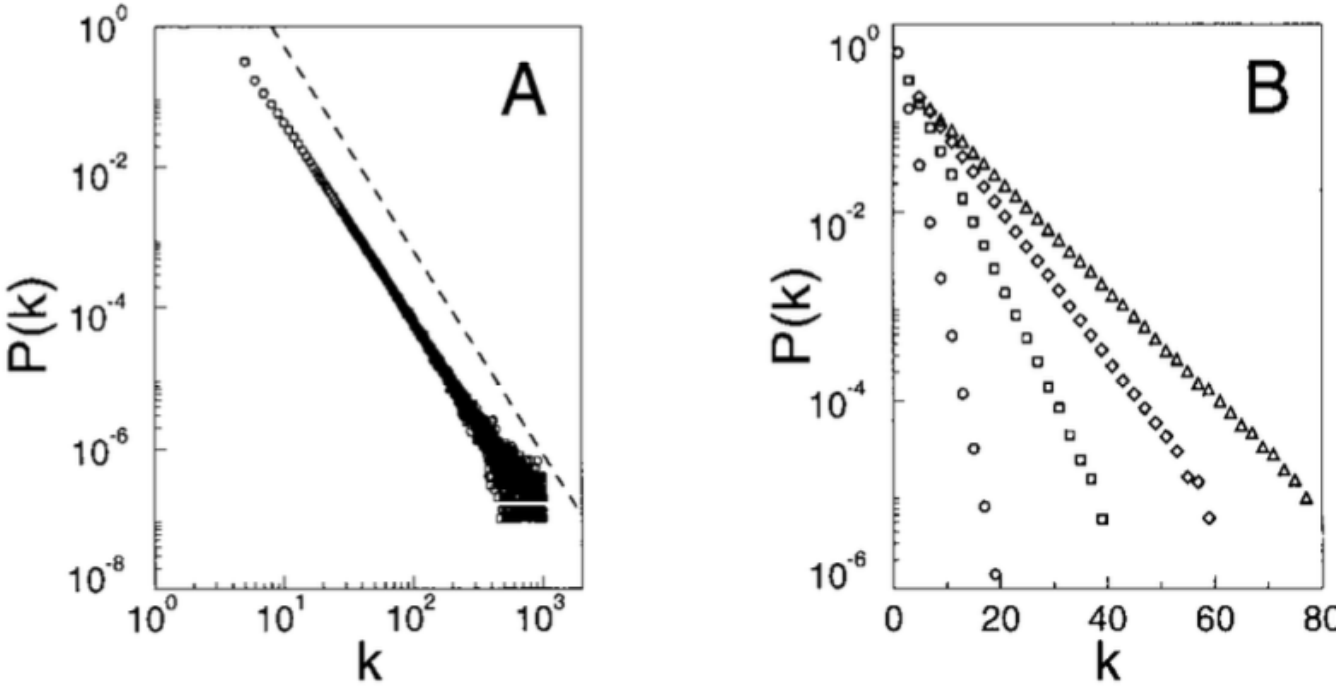
Preferential attachment is the key

One alternative model:

- Nodes still arrive one at a time, but they attach to existing nodes uniformly at random (with no preferential attachment)
- Resulting degree distribution is exponential $p_k \sim e^{-\beta k}$.
- Figure shows degree distribution for $m = m_0 = 1, 3, 5, 7$



More hubs in scale-free networks

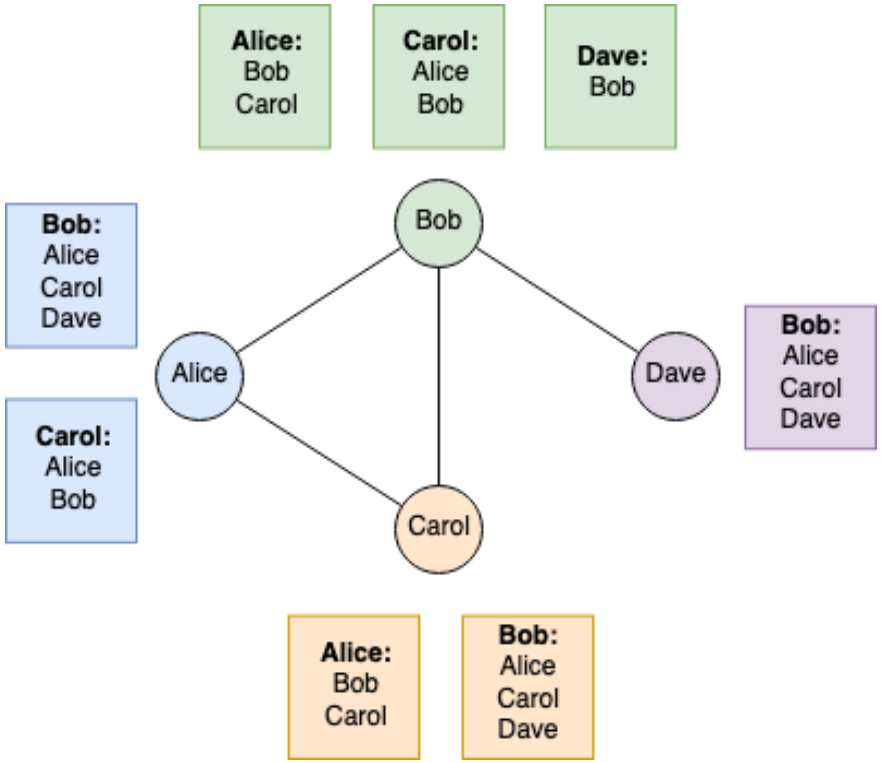


This simple model can reproduce the power-law degree distribution, where a few nodes (hubs) have a disproportionately high number of connections, while most have few.

Limitations of the BA model?

Friendship paradox

The phenomenon that an individual's friends have, on average, more friends than that individual has.



Finding hubs and influentials

Which nodes are the most important?

- Suppose you are going to give away free tickets to your show
  Who should you give them too?
- Suppose you can pay someone to TikTok about your product
  Who should you pay?
- Suppose you want to be elected to a position
  Whose endorsement will help you the most?

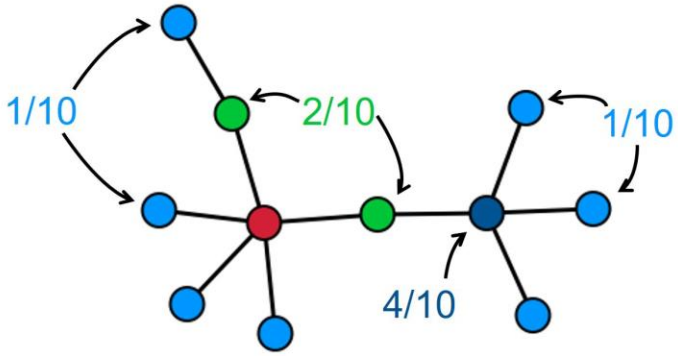
Measures of node centrality

- Degree centrality
- PageRank centrality
- Betweenness centrality
- Closeness centrality
- Eigenvector centrality
- …

Which is best depends on the problem you are studying.

Degree centrality

The degree centrality of a node is simply its degree divided by $n - 1$, where n is the total number of nodes in the network.



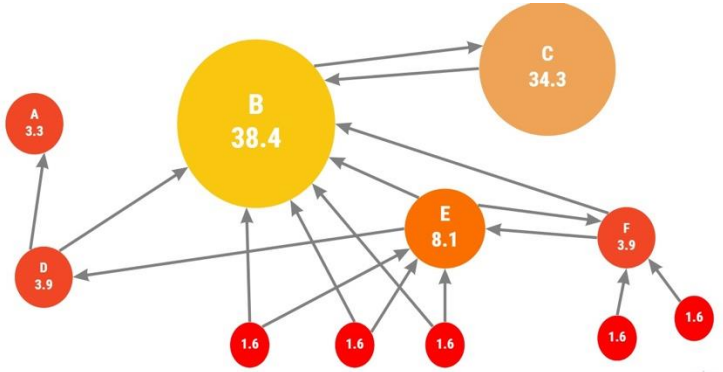
- *When would this be a good measure of influence?*
- *When/why might it not be a good measure of influence?*

PageRank algorithm

- PageRank is the algorithm behind the Google search engine.
- It measures the importance of a webpage by analyzing both the quantity and quality of the links that point to this page.

Basic idea:

- Stage 0: Assign each node an initial PageRank value $1/n$
- Stage k: Each node divides its PageRank equally among all its outgoing links and gives these shares to its neighbors
- Take the limit as $k \to \infty$



PageRank algorithm

https://www.ted.com/talks/cedric_villani_what_s_so_sexy_about_math

Watch video starting at minute 6:50

Betweenness centrality

The betweenness centrality of a node is calculated as the number of shortest paths of all pairs of nodes that pass through that node. It can measure the importance of nodes connecting different communities.

Notes:

- Computation is very costly;
- Nodes with high betweenness centrality scores are often called “broker”;
- Broker plays a crucial role in exchanging information between communities;

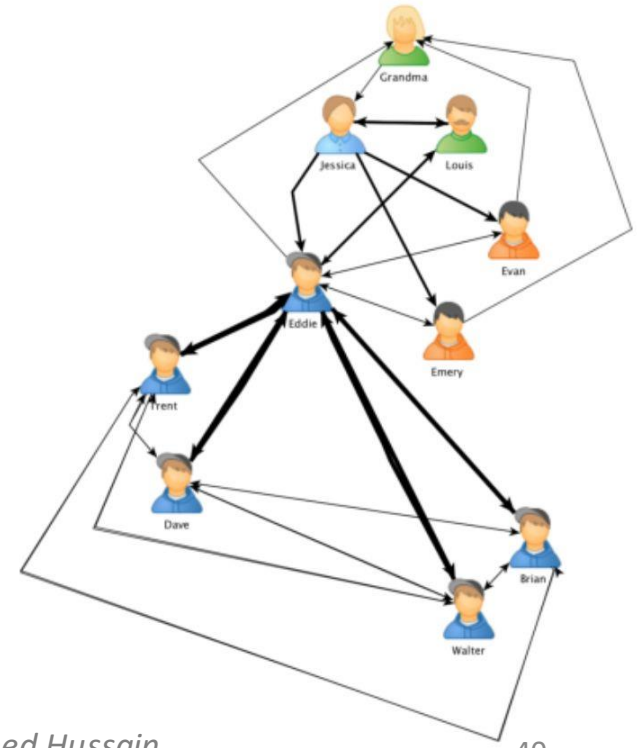
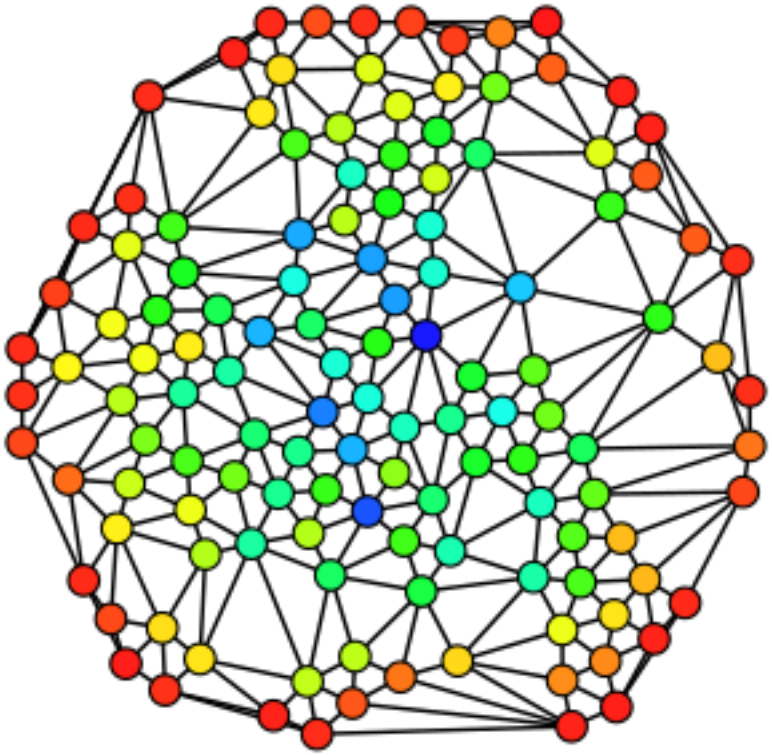


Fig: Zaki Mohamed Hussain

Betweenness centrality



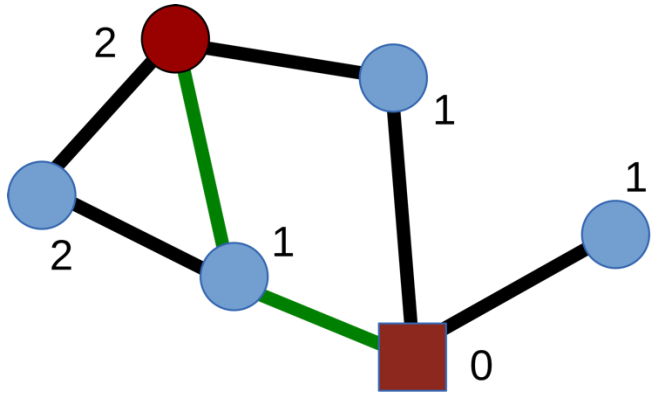
- Sometimes called “structural hole”, “bridge”.
- Can be used to reflect power, social capital.
- Real networks are vulnerable to attacks on brokers and hubs.

Fig: A directed graph colored based on the betweenness centrality of each node from least (red) to greatest (blue).

Closeness centrality

The closeness centrality of a node is calculated as the inverse of its avg shortest distance between the node and all other nodes in the network. The more central a node is, the closer it is to all other nodes.

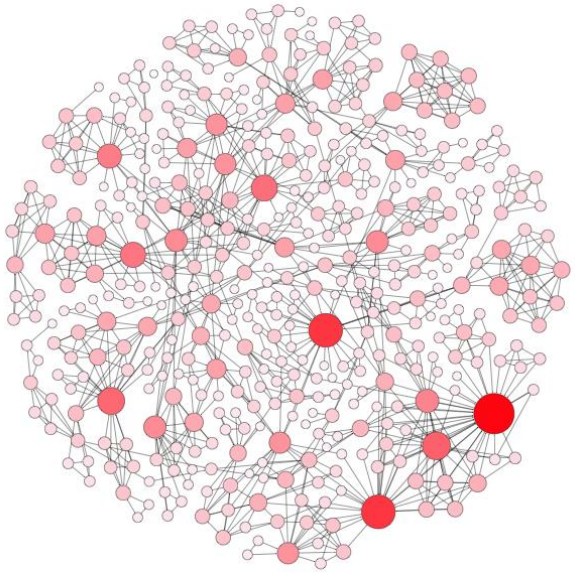
$$C(x) = \frac{N-1}{\sum_y d(y,x)}.$$



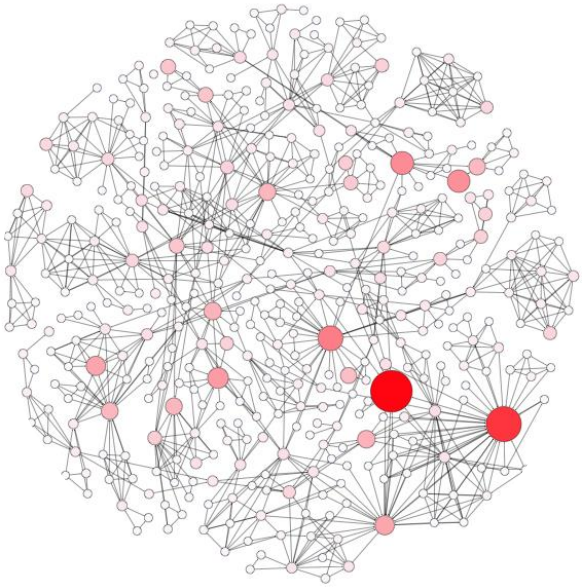
- The green edges illustrate one of the two shortest paths between the red square node and the red circle node.
- Closeness of the square node: 5/(1+1+1+2+2) = 5/7

Comparing different measures

Degree centrality



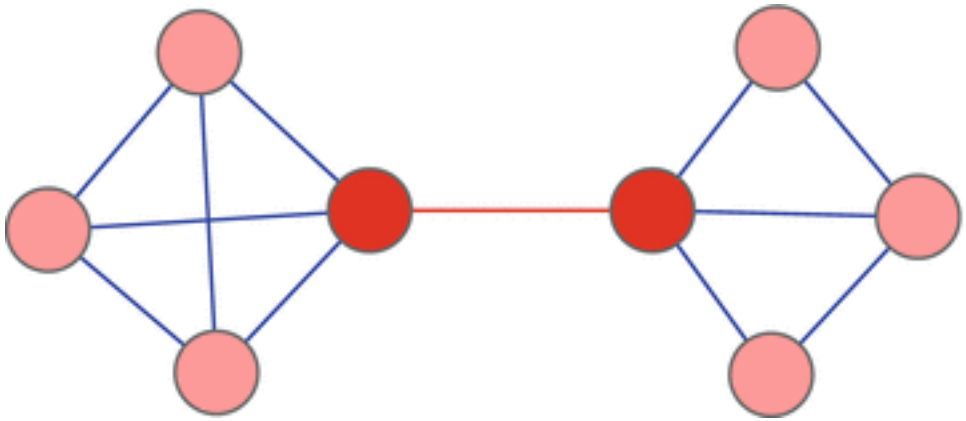
PageRank



- In what social scenarios should we use degree centrality?
- What about the PageRank centrality? Give real examples.

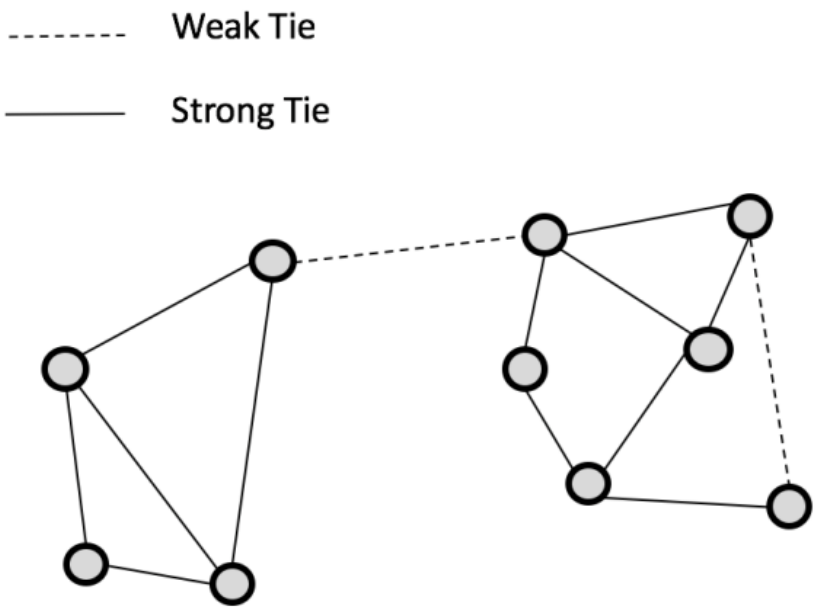
Edge betweenness centrality

Betweenness centrality of an edge is the number of all-pairs shortest paths that pass through the edge.



Strong ties and weak ties

- Measured by edge weights such as the contact frequency, the number of mutual friends, instead of structural positions.
- Weak ties often have high betweenness centrality.
- Strong ties are often within communities; weak ties often connect communities.



Quick survey

- How did you hear about our program?
- Who will you reach out for finding a good job?
- Who provide the opportunities in both scenarios?
  - Family members
  - Close friends
  - Acquaintance
  - Someone you don't really know

Strength of weak tie (classic theory)

- The idea that people often land jobs via weak connections.
- One of the most (missed) influential ideas in social sciences.

The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

https://www.jstor.org/stable/2776392
https://www.youtube.com/watch?v=g3bBajcR5fE

Evidence (observational) on Facebook

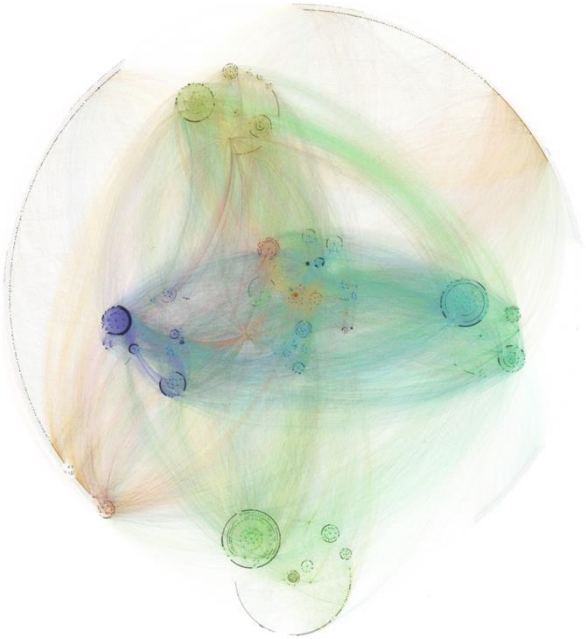
The Role of Social Networks in Information Diffusion

Eytan Bakshy*
Facebook
1601 Willow Rd.
Menlo Park, CA 94025
ebakshy@fb.com

Itamar Rosenn
Facebook
1601 Willow Rd.
Menlo Park, CA 94025
itamar@fb.com

Cameron Marlow
Facebook
1601 Willow Rd.
Menlo Park, CA 94025
cameron@fb.com

Lada Adamic
University of Michigan
105 S. State St.
Ann Arbor, MI 48104
ladamic@umich.edu



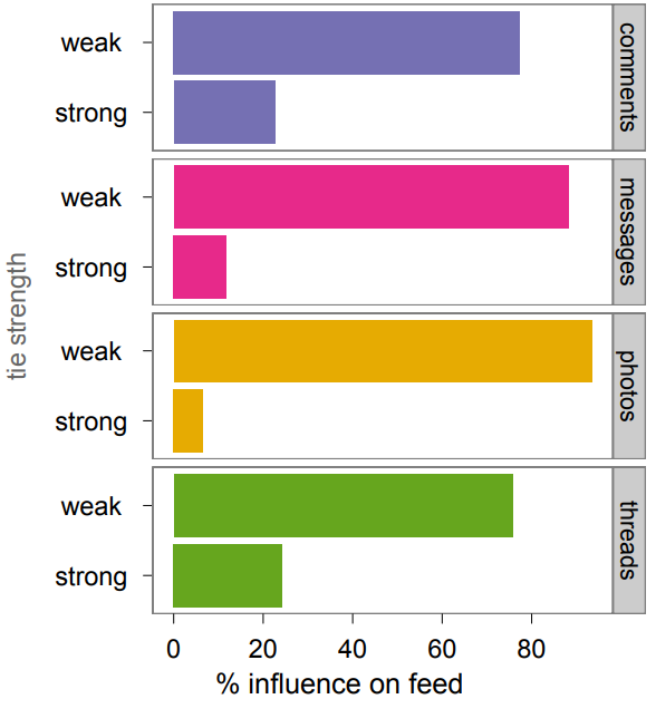


Figure 7: Weak ties are collectively more influential than strong ties. Panels show the percentage of information spread by strong and weak ties for all four measurements of tie strength. Although the probability of influence is significantly higher for those that interact frequently, most contagion occurs along weak ties, which are more abundant.

Replicated by LinkedIn experiment

- A “delayed” causal test of this paradoxical theory;
- Largest randomized experiments conducted on LinkedIn’s “People You May Know” algorithm;
- The weakest ties had the greatest impact on job mobility, whereas the strongest ties had the least.

REPORT | SOCIAL NETWORKS

A causal test of the strength of weak ties

KARTHIK RAJKUMAR, GUILLAUME SAINT-JACQUES, IAVOR BOJINOV, ERIK BRYNJOLFSSON, AND SINAN ARAL

SCIENCE • 15 Sep 2022 • Vol 377, Issue 6612 • pp. 1304-1310 • DOI: 10.1126/science.abl4476

Link prediction

- Given a network snapshot, the task is to predict new links among all nodes that are likely to occur in the future.

- Often use topological features to measure node proximity:
  - Network distance
  - Common neighbors
  - Jaccard's coefficient: score(x, y) = $|\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)|$
  - Adamic and Adar score: $\text{score}(x, y) := \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$
  - And more …

- Can also include node-level characteristics:
  - Gender, age, political orientation
  - Activity level, language use, profile image, …

Nowell and Kleinberg. "The Link-Prediction Problem for Social Networks." *CIKM* 2003.

Course notes

- HW2 due today!
- HW3 released
  - Due on Nov 28.
- Social network v2 next week
- Data visualization next week

LOQ course evaluation survey

- Appreciate your insightful feedback!
- LOQ system: https://onlinesurvey.cityu.edu.hk/
- Also available on Canvas course site.