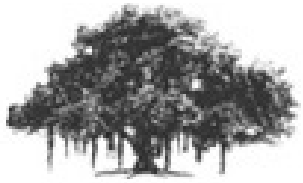


IIIT-H

Web Mining
Lecture 9: Social Network Analysis (Part 1)

Manish Gupta

31st Aug 2013

Slides borrowed (and modified) from
http://www.stanford.edu/class/cs224w/slides/01-intro.pdf
http://temporalweb.net/2011/files/kumar-twaw2011.pdf
http://www.stanford.edu/class/cs224w/slides/02-gnp.pdf
http://www.stanford.edu/class/cs224w/slides/03-smallworld.pdf

Recap of Lecture 8: Social Recommender Systems (Part 2)

- ~~Recommendation for Groups~~
- The Cold Start Problem
- Trust
- ~~Temporal Aspects in Social Recommendation~~
- Evaluation Methods

Announcements

- Midsem Exam: Sep 6, 1:30pm-3pm
- Assignment 2 is up
  - Submission Deadline is Sep 9, 9pm
  - Any form of copying will lead to 0 marks for everyone with the same answer
- Midsems
  - No cheating
  - Allowed 1 A4 size cheat sheet (both sides)
  - Same format as Assignment 2
  - Covers content covered up to Aug 22 (last lecture)
- Doubt clearing session
  - Sep 2, 7:30pm-9pm
  - TAs will conduct this one at 104 Himalaya
- Next class is on Sep 2, 6-7:30pm.

Today's Agenda

- Introduction to Social Network Analysis
- Structure of the Web Graph
- Erdös-Renyi Model
- Small World Model and Kleinberg's Model
- Power Laws

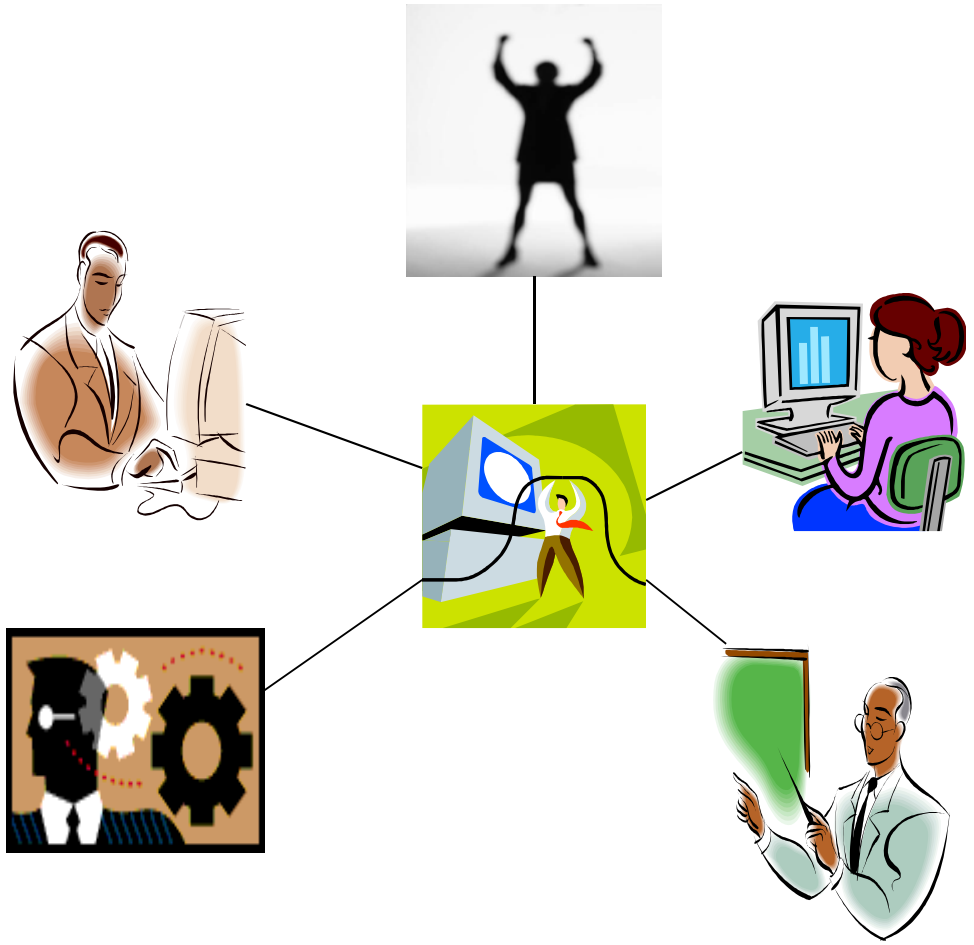
Today's Agenda

- **Introduction to Social Network Analysis**
- Structure of the Web Graph
- Erdös-Renyi Model
- Small World Model and Kleinberg's Model
- Power Laws

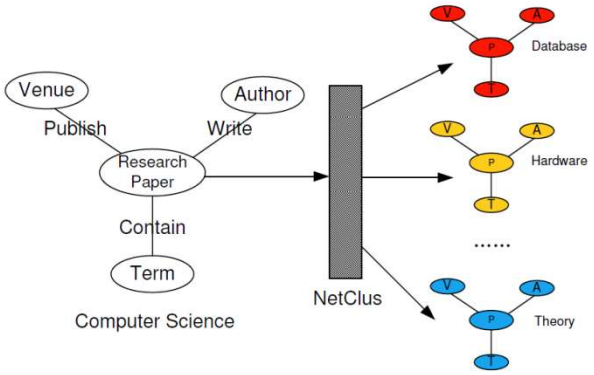
Lots of Social Networks Today

What is a Social Network

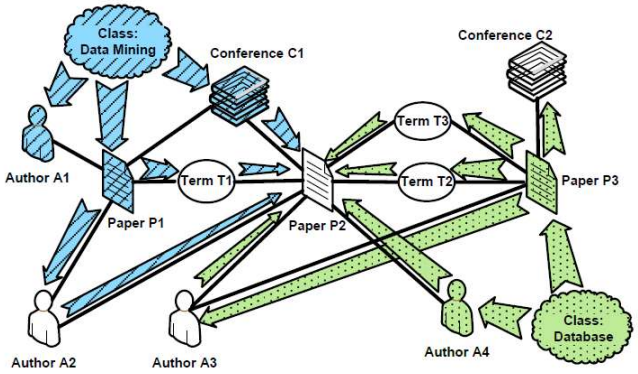
A social network is a description of the social structure between actors, mostly individuals or organizations. It indicates the ways in which they are connected through various social familiarities ranging from casual acquaintance to close familiar bonds.



Network Analysis Tasks

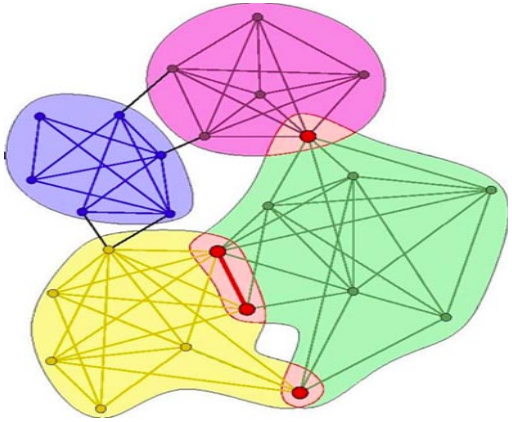


Clustering

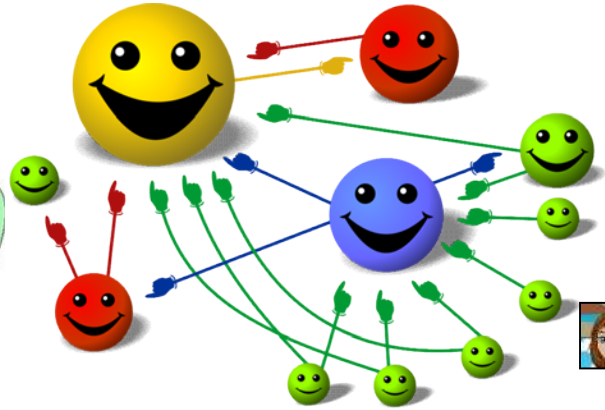


Classification

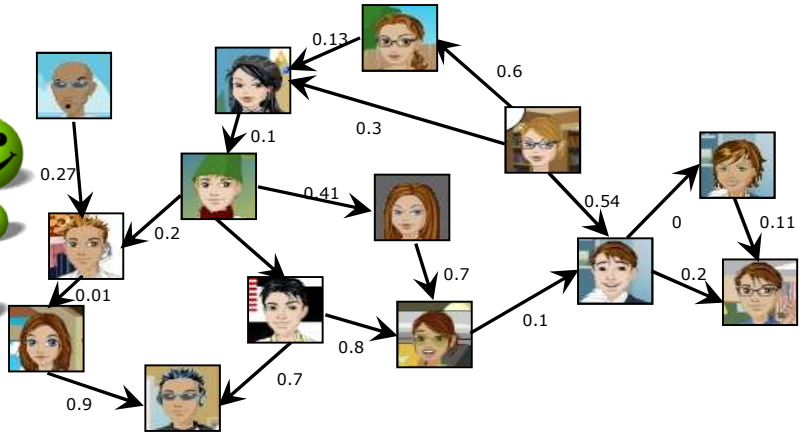
Link Prediction



Community Detection



PageRank



Influence Propagation

Studying Social Networks

- More and more of all interactions are happening online
- The resulting data is a goldmine for studies
  - Massive amounts, even hundreds of millions of nodes
    - Search companies are now working on crawls of 100+ billion pages
    - Facebook has over 600M active users
  - Study phenomena at different scales (eg, interaction of people in focused groups of different sizes, overall structure of the network)
  - Ability to measure, record, and track individual activities at the finest resolution (eg, user befriending another, user buying a dvd, user tagging a photo, user tweeting — when, how, why)
  - Interplay between monadic and dyadic attributes
- A double-edged sword
  - Data is inherently noisy
  - Large scale of data leads to algorithmic challenges
- Graph-theoretic analysis has led to significant impact
  - Link analysis in web search
  - Sophisticated recommendation systems
- Interplay of measurements, modeling, and algorithms

Network Size

- Network data: Orders of magnitude
  - 436-node network of email exchange at a corporate research lab [Adamic-Adar, SocNets '03]
  - 43,553-node network of email exchange at an university [Kossinets-Watts, Science '06]
  - 4.4-million-node network of declared friendships on a blogging community [Liben-Nowell et al., PNAS '05]
  - 240-million-node network of communication on Microsoft Messenger [Leskovec-Horvitz, WWW '08]
  - 800-million-node Facebook network [Backstrom et al. '11]

Complexity of Data in a Social Network

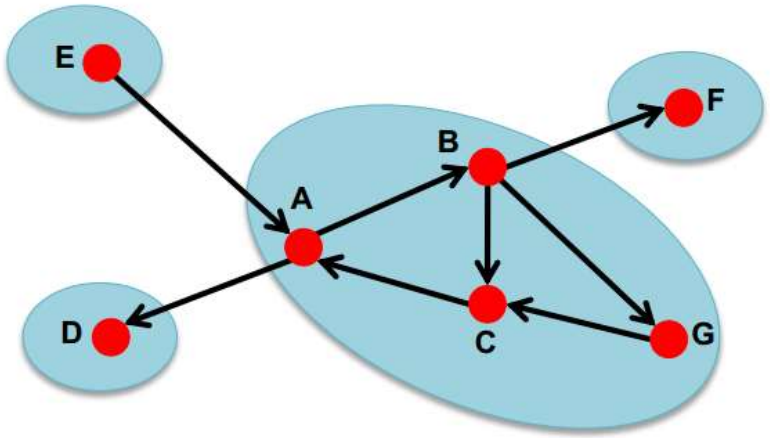
- Nodes
  - Of different types
  - With multi-typed feature values
- Links
  - Of different types
  - With weights
  - With uncertainty
  - Multiple links between nodes
  - Directed/undirected
- Time
  - Temporal graph snapshots
  - Streams of graphs
- Subgraphs and latent communities
  - Of different sizes
  - Latent communities using different features

Today's Agenda

- Introduction to Social Network Analysis
- **Structure of the Web Graph**
- Erdös-Renyi Model
- Small World Model and Kleinberg's Model
- Power Laws

Structure of the Web Graph (1)

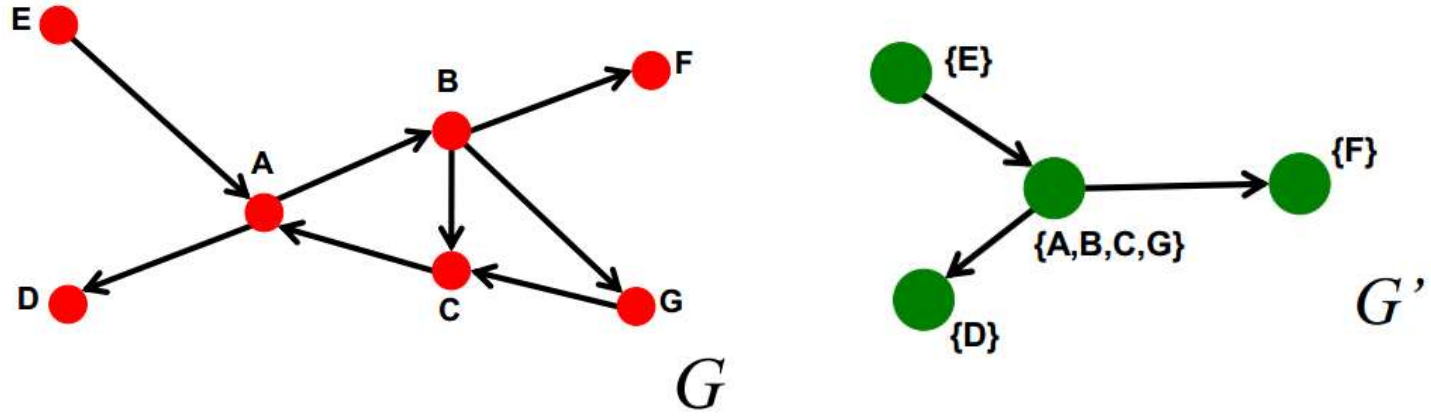
- Strongly connected component (SCC)
  - It is a set of nodes S such that
    - Every pair of nodes in S can reach each other
    - There is no larger set containing S with this property
- Weakly connected component
  - Connected if we disregard the direction



Strongly connected components of the graph: {A,B,C,G}, {D}, {E}, {F}

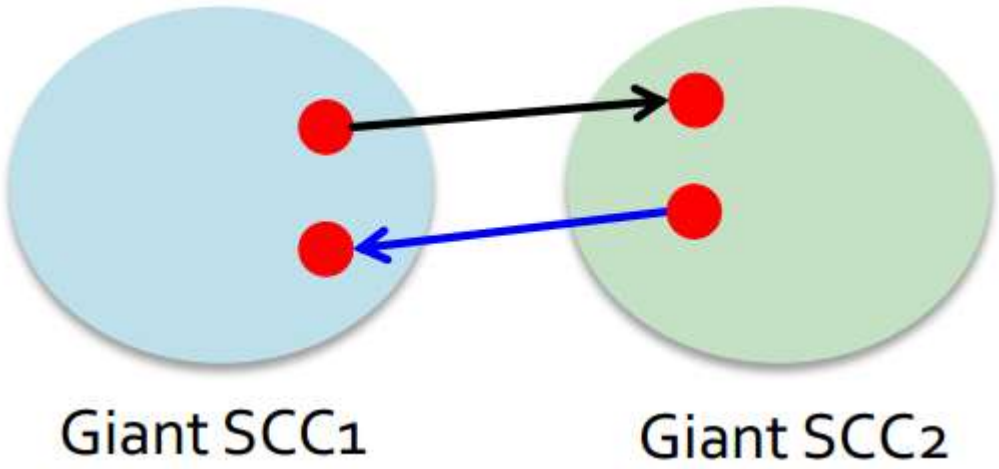
Structure of the Web Graph (2)

- Two types of directed graphs
  - Strongly connected
    - Any node can reach any node via a directed path
  - DAG – Directed Acyclic Graph
    - Has no cycles: if u can reach v, then v can not reach u
- Every directed graph is a DAG on its SCCs
  - SCCs partitions the nodes of G



Structure of the Web Graph (3)

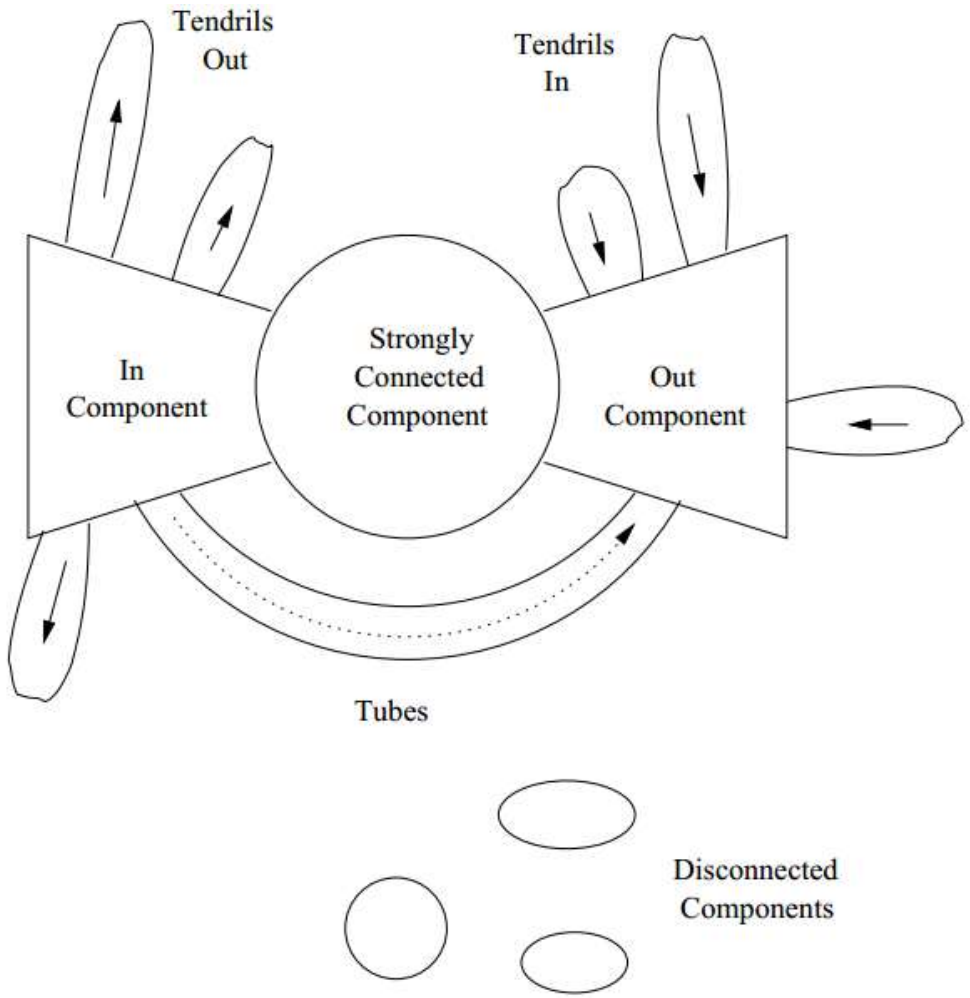
- In the web graph, there is a giant SCC
- There won't be 2 giant SCCs
- Heuristic argument
  - It just takes 1 page from one SCC to link to the other SCC
  - If the 2 SCCs have millions of pages the likelihood of this not happening is very very small



Structure of the Web Graph (4)

- Broder et al., 2000
  - Altavista crawl from October 1999
    - 203 million URLS
    - 1.5 billion links
- Undirected version of the Web graph
  - 91% nodes in the largest weakly connected component
- Directed version of the Web graph
  - Largest SCC: 28% of the nodes (56 million)

Bow-tie Structure of the Web



Today's Agenda

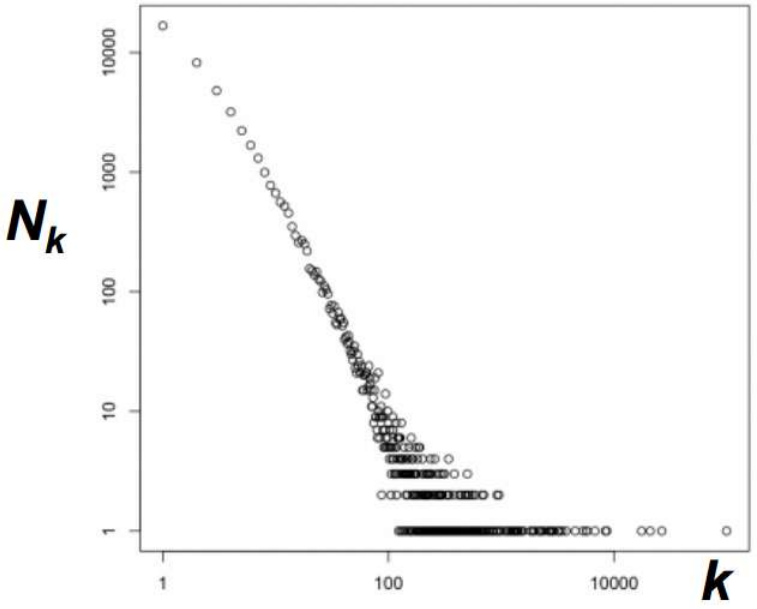
- Introduction to Social Network Analysis
- Structure of the Web Graph
- **Erdös-Renyi Model**
- Small World Model and Kleinberg's Model
- Power Laws

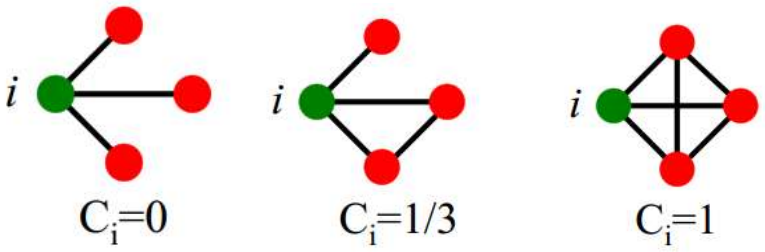
Properties of Graphs

- Degree distributions
  - Heavy tail
- Clustering
  - High clustering coefficient
- Communities and dense subgraphs
  - Abundance; locally dense, globally dense; spectrum
- Connected components
  - Distribution; bow-tie structure
- Connectivity
  - Low diameter; small-world properties
- Compressibility

Properties of Graphs

- Degree distribution P(k): Probability that a randomly chosen node has degree k
  - N_k= # nodes with degree k
- Network Diameter
  - the maximum (shortest path) distance between any pair of nodes in a graph
- Clustering coefficient of a node i
  - What fraction of i's neighbors are connected?
  - Node i with degree k_i
    - $C_i = \frac{2e_i}{k_i(k_i-1)}$
    - $C_i \in [0,1]$
- Average clustering coefficient
  - $C = \frac{1}{N}\sum_{i=1}^{N} C_i$

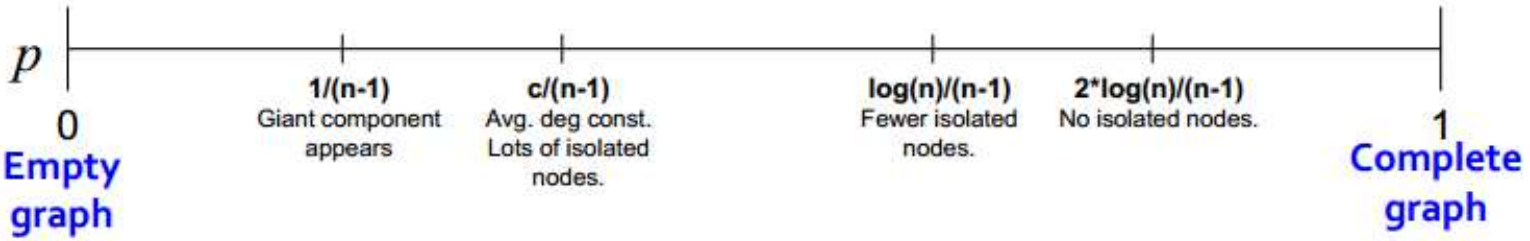




The G_{np} Model

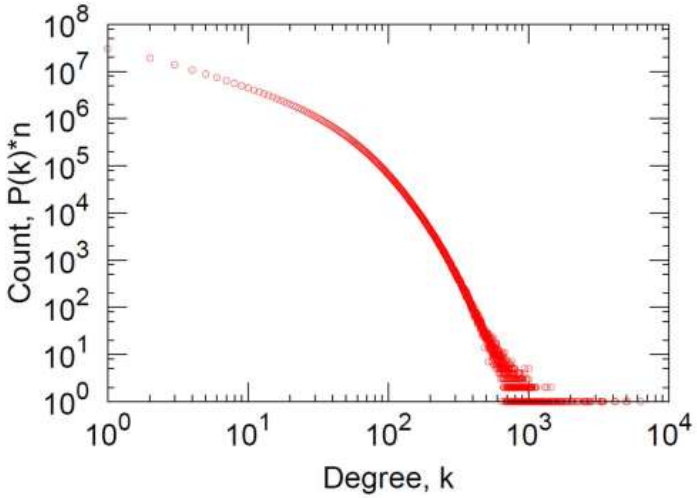
- Erdös-Renyi Random Graphs [Erdös-Renyi, '60]
- Undirected graph on n nodes and each edge (u,v) appears i.i.d. with probability p
- Expected degree of a node?
  - Let random variable X_{uv} denote the presence/absence of edge (u,v)
  - Thus expected degree = $E[X_u] = \sum_{u=1}^{n-1} E[X_{uv}] = (n-1)p$
- Degree distribution of G_{np} is Binomial(n,p)
  - $P(k) = n-1_{C_k} p^k (1-p)^{n-1-k}$
- Clustering coefficient=p
- Path length= O(log n)

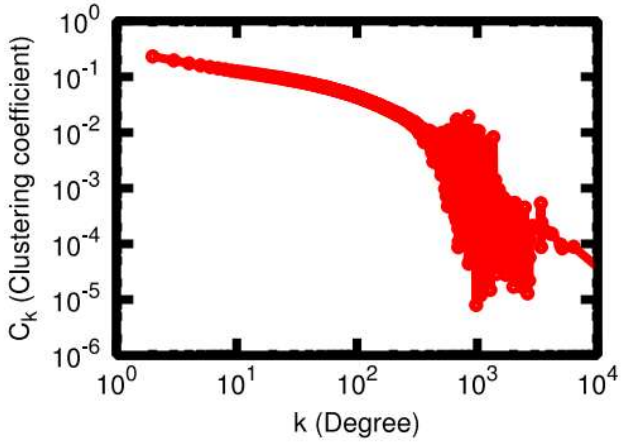
- **Graph structure of G_{np} as *p* changes:**



How well does G_{np} correspond to Real Networks?

- MSN Messenger activity in June 2006:
  - 245 million users logged in
  - 180 million users engaged in conversations
  - More than 30 billion conversations
  - More than 255 billion exchanged messages





Avg. clustering of the MSN:
$C = 0.1140$

Avg. clustering of corresponding G_{np}:
$C = \overline{k}/n \approx 8 \cdot 10^{-8}$

Real Networks vs. G_{np}

- Are real networks like random graphs?
  - Giant connected component, and Average path length match
  - Clustering Coefficient and degree distribution are different
- Problems with the random network model
  - Giant component in most real network does NOT emerge through a phase transition
  - Degree distribution differs from real networks
  - No local structure – clustering coefficient is too low

Today's Agenda

- Introduction to Social Network Analysis
- Structure of the Web Graph
- Erdös-Renyi Model
- **Small World Model and Kleinberg's Model**
- Power Laws

Small World Experiment

- What is the typical shortest path length between any two people?
- Experiment on the global friendship network
- Small-world experiment [Milgram '67]
- Picked 300 people in Omaha,Nebraska and Wichita, Kansas
- Ask them to get a letter to a stock-broker in Boston by passing it through friends
- How many steps did it take?
- 64 chains completed (i.e., 64 letters reached the target)
- It took 6.2 steps on the average,thus "6 degrees of separation"

Criticism of Milgram's Experiment

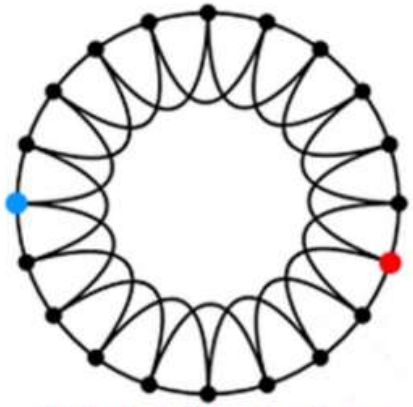
- Funneling
  - 31 of 64 chains passed through 1 of 3 people as their final step. Hence, not all links/nodes are equal
- Starting points and the target were non-random
- People refused to participate (25% for Milgram)
- Some sort of social search: People in the experiment follow some strategy (e.g., geographic routing) instead of forwarding the letter to everyone. They are not finding the shortest path!
- There are not many samples (only 64)
- People might have used extra information resources

Dodds, Muhamad and Watts Experiment

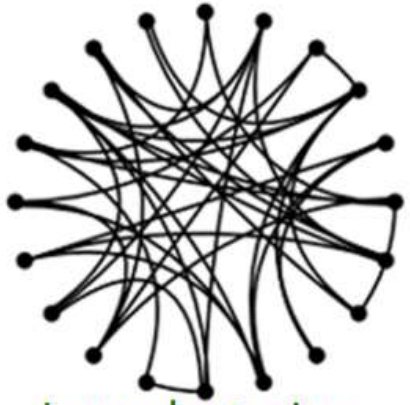
- In 2003 Dodds, Muhamad and Watts performed the experiment using e-mail
- 18 targets of various backgrounds
- 24,000 first steps (~1,500 per target)
- 65% dropout per step
- 384 chains completed (1.5%)
- Avg. chain length= 4.01
- Problem: People stop participating
- After the correction: Typical path length h = 7

High Clustering vs. Small Diameter

- MSN network has 7 orders of magnitude larger clustering than the corresponding G_{np}
- Real networks show clustering behavior
- Clustering implies high diameter (distance between nodes in different clusters could be high)
- Small diameter means there exist shortcuts which bridge clusters
- Could a network with high clustering be at the same time a small world?
- How can we at the same time have high clustering and small diameter?



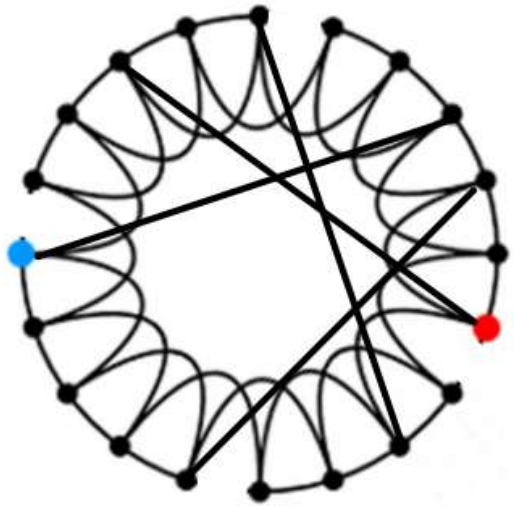
High clustering
High diameter



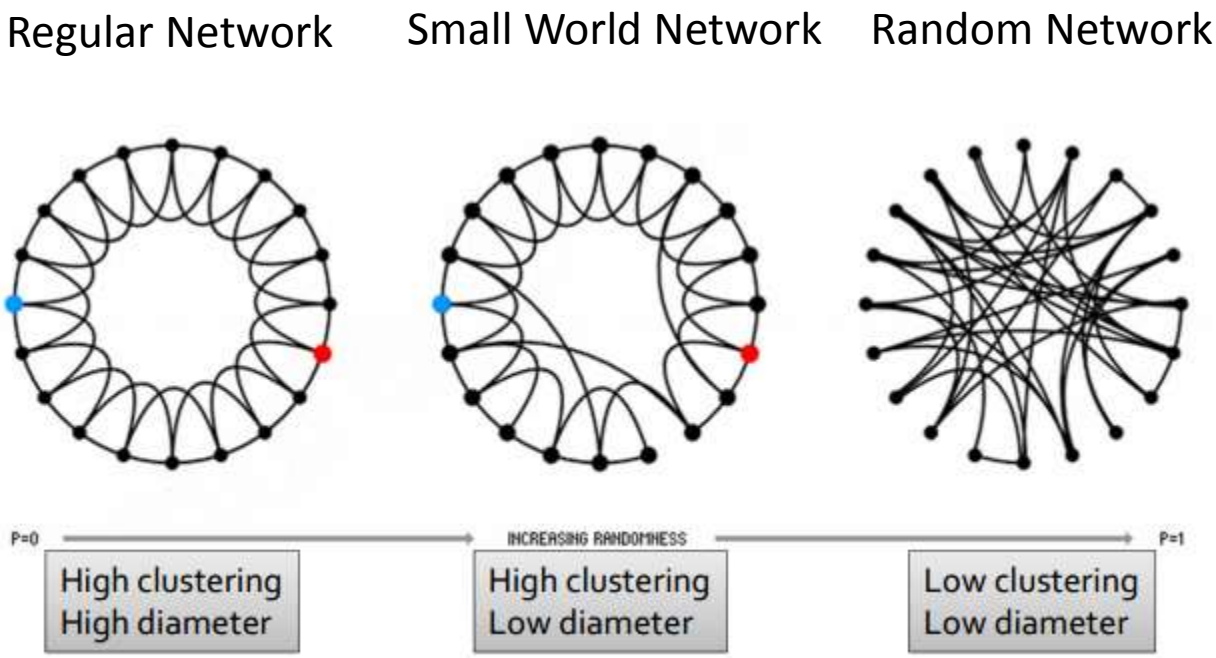
Low clustering
Low diameter

The Small World Model [Watts-Strogatz, '98]

- 2 components to the model
  - Start with a low-dimensional regular lattice
    - Has high clustering coefficient
    - Now introduce randomness ("shortcuts")
  - Rewire
    - Add/remove edges to create shortcuts to join remote parts of the lattice
    - For each edge with prob. p move the other end to a random node



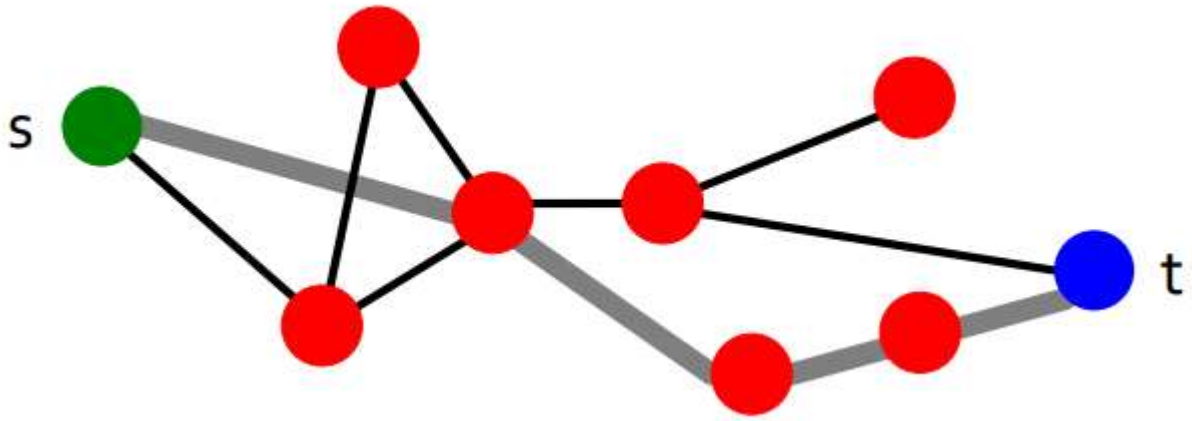
The Small World Model



- Rewiring allows us to “interpolate” between a regular lattice and a random graph
- It takes a lot of randomness to ruin the clustering, but a very small amount to create shortcuts.
- Diameter is log n
- Accounts for the high clustering of real networks
- Does not lead to the correct degree distribution
- Does not enable navigation

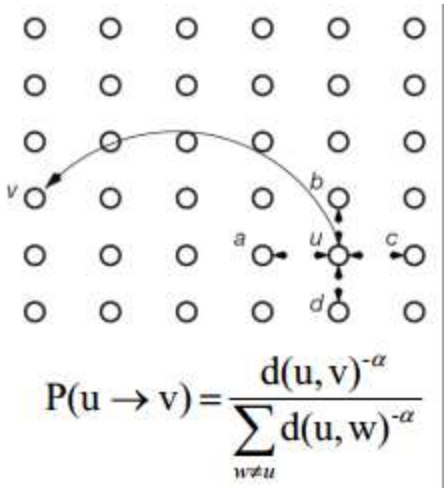
Searchable Graphs

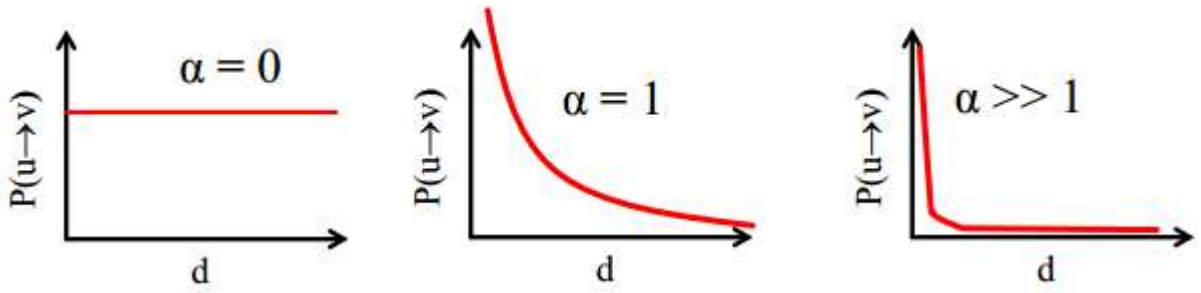
- s only knows locations of its friends and location of the target t
- s does not know links of anyone else but itself
- Geographic Navigation
  - s navigates to the node closest to t
- Search time T: Number of steps to reach t
  - Watts-Strogatz Model: $T = O(n^{\alpha}) \Rightarrow$ Not searchable
  - Erdős–Rényi Model: $T = O(n) \Rightarrow$ Not searchable
  - Kleinberg's model: $T = O((log\ n)^2) \Rightarrow$ Searchable



Kleinberg's Model

- Watts-Strogatz graphs are not searchable
- How do we make a searchable small-world graph?
- Intuition
  - Our long range links are not random
  - They follow geography!
- Model
  - Nodes on a grid
  - Node has one long range link
  - Prob. of long link to node v: $P(u \to v) \sim d(u,v)^{-\alpha}$
    - d(u,v) is the grid distance between u and v





Today's Agenda

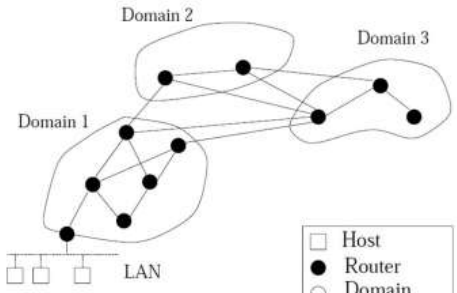
- Introduction to Social Network Analysis
- Structure of the Web Graph
- Erdös-Renyi Model
- Small World Model and Kleinberg's Model
- **Power Laws**

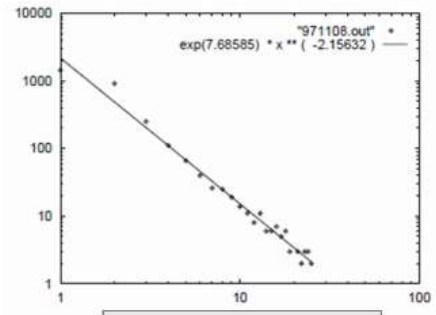
What do we Want?

- Small world
  - Small diameter
  - High Clustering Coefficient
- Searchability
- Power law degree distribution
  - Preferential Attachment
  - $P(k) = k^{-\alpha}$
  - Appears as a straight line on log-log plot

Power Laws Everywhere

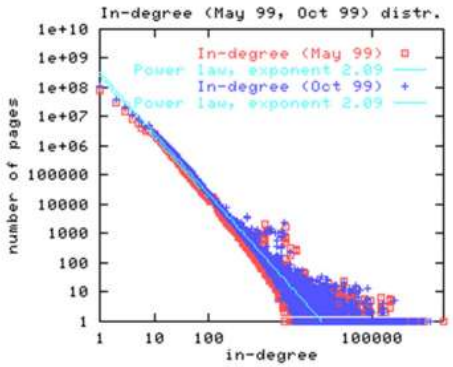
- **Internet Autonomous Systems**
  [Faloutsos, Faloutsos and Faloutsos, 1999]

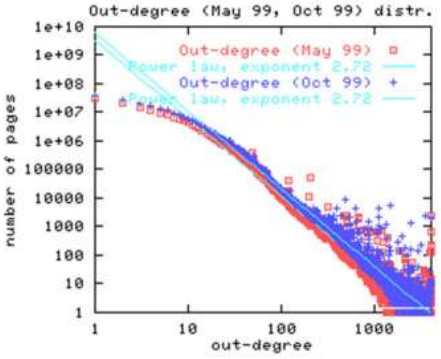




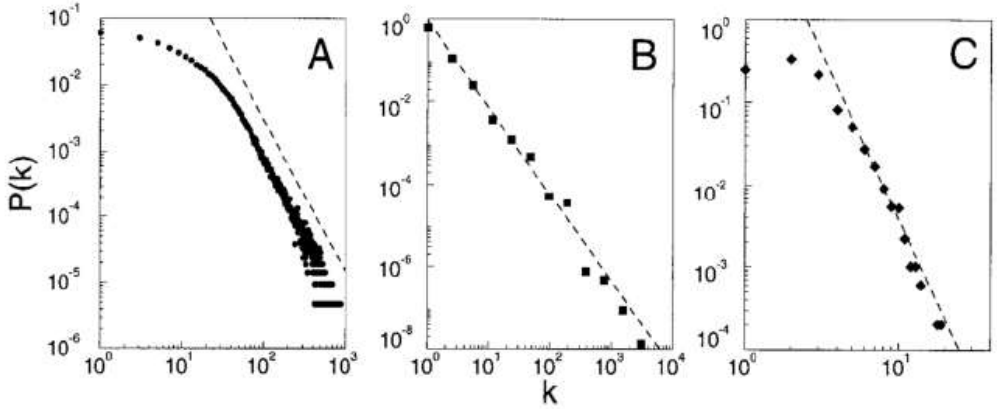
Internet domain topology

- **The World Wide Web** [Broder et al., 2000]

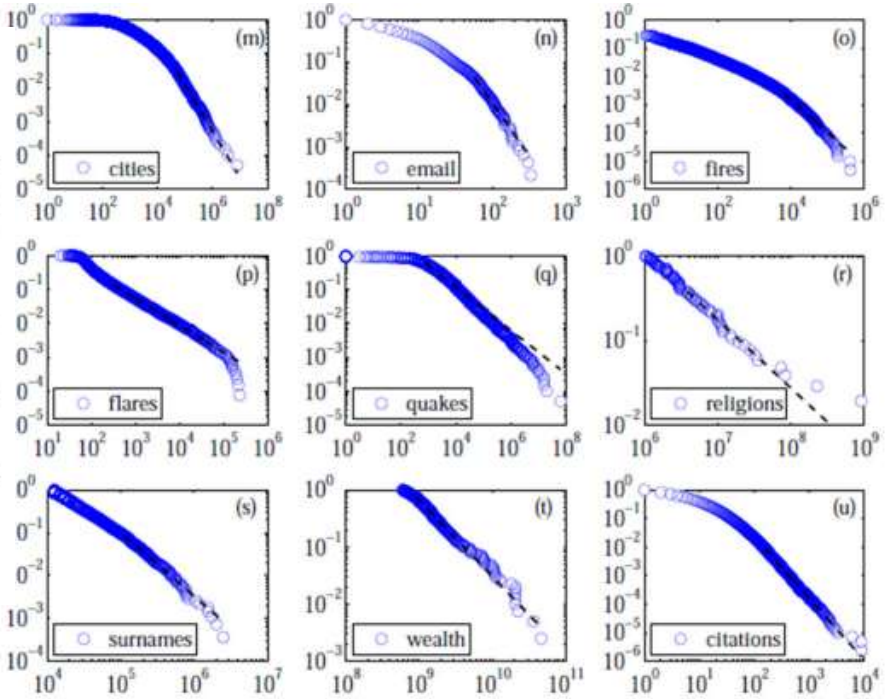




- **Other Networks** [Barabasi-Albert, 1999]



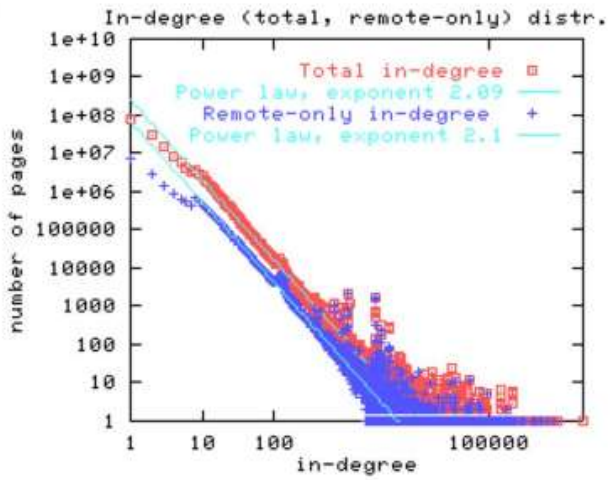
Actor collaborations | Web graph | Power-grid

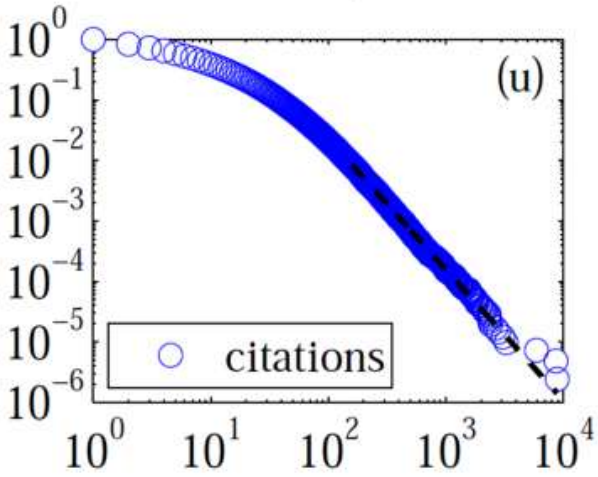


Power Law Degree Exponents

Power-law degree exponent is typically $2 < \alpha < 3$

- Web graph:
  - $\alpha_{in} = 2.1$, $\alpha_{out} = 2.4$ [Broder et al. 00]
- Autonomous systems:
  - $\alpha = 2.4$ [Faloutsos^3, 99]
- Actor-collaborations:
  - $\alpha = 2.3$ [Barabasi-Albert 00]
- Citations to papers:
  - $\alpha \approx 3$ [Redner 98]
- Online social networks:
  - $\alpha \approx 2$ [Leskovec et al. 07]





Scale-free Networks

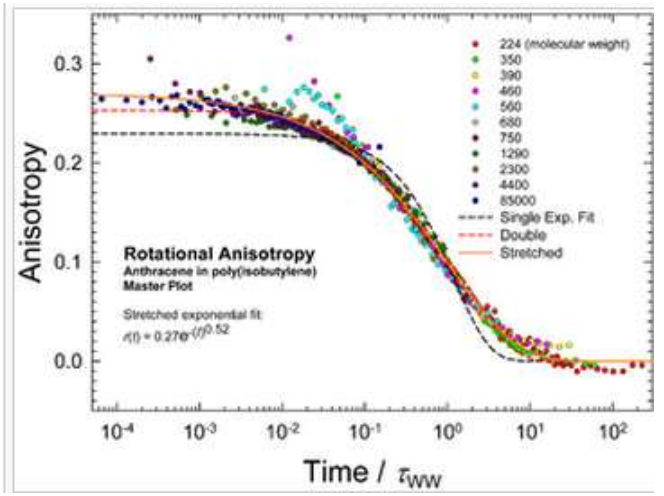
- Networks with a power law tail in their degree distribution are called “scale-free networks”
  - Scale invariance: There is no characteristic scale
  - Scale-free function: $f(ax) = a^{\lambda} f(x)$
    - Power-law function: $f(ax) = a^{\lambda} x^{\lambda} = a^{\lambda} \mathrm{f(x)}$

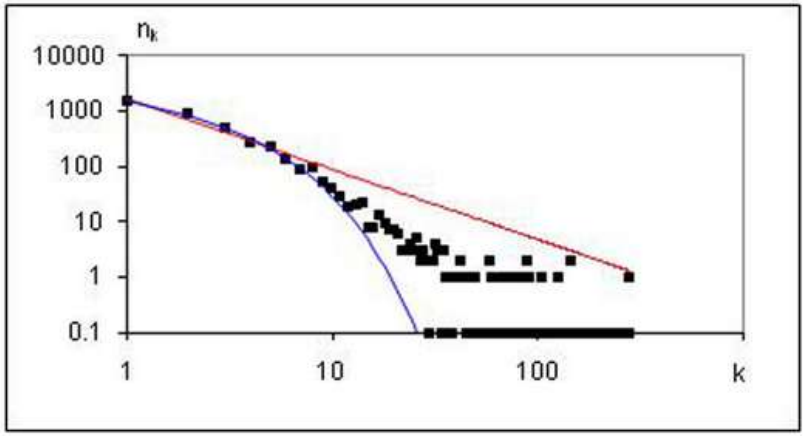
Mathematics of Power Laws (1)

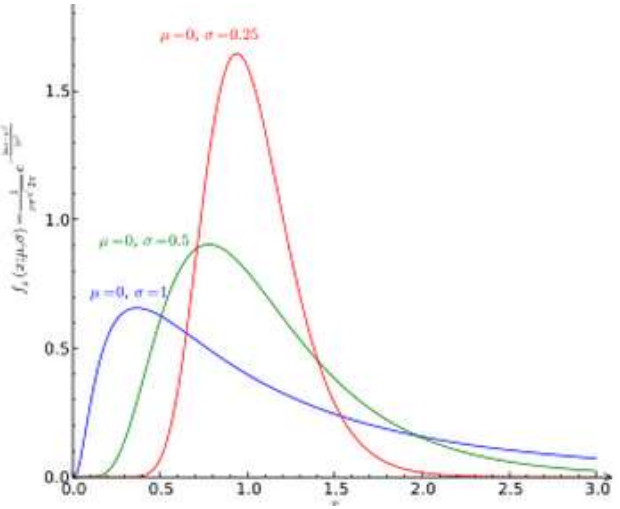
- Heavy Tailed Distributions
  - Degrees are heavily skewed
    - Distribution $P(X > x)$ is heavy tailed if
      - $\lim_{x \to \infty} \frac{P(X>x)}{e^{-\lambda x}} = \infty$
  - These distributions are not heavy tailed
    - Normal PDF: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
    - Exponential PDF: $p(x) = \lambda e^{-\lambda x}$

Mathematics of Power Laws (2)

- Heavy tailed distributions
  - Are also called Long tail, Heavy tail, Zipf's law, Pareto's law
  - The following are heavy tailed distributions
    - Power law: $P(x) \propto x^{-\alpha}$
    - Power law with exponential cutoff: $P(x) \propto x^{-\alpha} e^{-\lambda x}$
    - Stretched exponential: $P(x) \propto x^{\beta-1} e^{-\lambda x^{\beta}}$
    - Log-normal: $P(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$







Mathematics of Power Laws (3)

- What is the normalizing constant?
  - $p(x) = zx^{-\alpha}$ But z=?
  - p(x) is a distribution: $\int p(x)dx = 1$
  - Let x_m be the minimum value of the power kaw distribution. For $x < x_m$, power laws usually diverge.
  - Thus, $z\int_{x_m}^{\infty} x^{-\alpha}dx = 1$
  - Solving, we get, $z = (\alpha - 1)x_m^{\alpha-1}$
  - Thus, $p(x) = \frac{(\alpha-1)}{x_m}\left(\frac{x}{x_m}\right)^{-\alpha}$

Mathematics of Power Laws (4)

- Expectation of a power law random variable
  - $E[x] = \int_{x_m}^{\infty} xp(x)dx = z\int_{x_m}^{\infty} x^{-\alpha+1}\,dx$
  - Thus, $E[x] = \frac{\alpha-1}{\alpha-2}x_m$ if $\alpha > 2$
  - Real graphs have $2 < \alpha < 3$
    - E[x]=constant, Var[x]=∞
  - If $\alpha \leq 2, E[x] = \infty$
  - If $\alpha \leq 3, Var[x] = \infty$
    - Average is meaningless when variance is infinite

Mathematics of Power Laws (5)

- How to generate power-law distributed random numbers?
- We want to generate x: a power-law distributed random number with density $p(x) = \frac{(\alpha-1)}{x_m}\left(\frac{x}{x_m}\right)^{-\alpha}$
- But we have access to r: uniform random number with density $p(r) = 1$ if $0 \leq r \leq 1$ else $p(r) = 0$
- Transforming densities
  - $p(x) = p(r)\frac{dr}{dx} = 1\frac{dr}{dx} = \frac{dr}{dx}$
  - Now $\int_x^\infty p(x')dx' = \left(\frac{x}{x_m}\right)^{-\alpha+1}$ and $\int_r^1 dr' = 1 - r$
  - Solving for x, we get $x = x_m(1-r)^{-\frac{1}{\alpha-1}}$

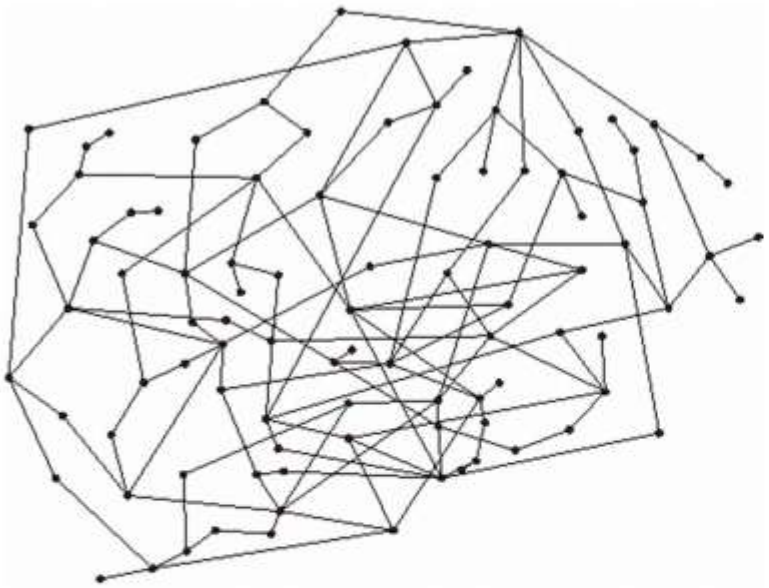
Mathematics of Power Laws (6)

- Estimating the power-law exponent α
  - Fit a line on log-log axis using least squares
    - Solve $argmin_{\alpha}\ [\log y - \alpha \log x]^2$
    - Bad because the estimate is biased
  - Estimating using CDF
    - $P(X \geq x) = \int_x^{\infty} zj^{-\alpha} dj = \frac{z}{\alpha}[j^{1-\alpha}]_x^{\infty} = \frac{z}{\alpha} x^{-(\alpha-1)}$
    - Thus, estimated $\alpha = 1 + \alpha'$ where α' is the slope of P(X>x)
  - Estimating using MLE
    - $L(\alpha) = \ln(\prod_{i=1}^{n} p(d_i)) = \sum_{i=1}^{n} \ln p(d_i) = \sum_{i=1}^{n} \ln(\alpha - 1) - \ln(x_m) - \alpha \ln\left(\frac{d_i}{x_m}\right)$
    - To max $L(\alpha)$, set $\frac{dL(\alpha)}{d\alpha} = 0$
    - This gives $\hat{\alpha} = 1 + n\ \left[\sum_{i=1}^{n} \ln \frac{d_i}{x_m}\right]^{-1}$

Mathematics of Power Laws (7)

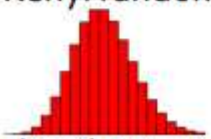
- What is the expected maximum degree K in a scale-free network?
  - The expected number of nodes with degree >K should be less than 1
    - $\int_K^\infty p(x)dx \approx \frac{1}{n}$
    - $\int_K^\infty p(x)dx = Z\int_K^\infty x^{-\alpha}dx = \frac{Z}{1-\alpha}[x^{1-\alpha}]_K^\infty = \frac{(\alpha-1)x_m^{\alpha-1}}{-(\alpha-1)}[0-K^{1-\alpha}] = x_m^{\alpha-1}.K^{-(\alpha-1)}$
    - Thus $K = x_m n^{\frac{1}{\alpha-1}}$

Random vs. Scale-free Networks

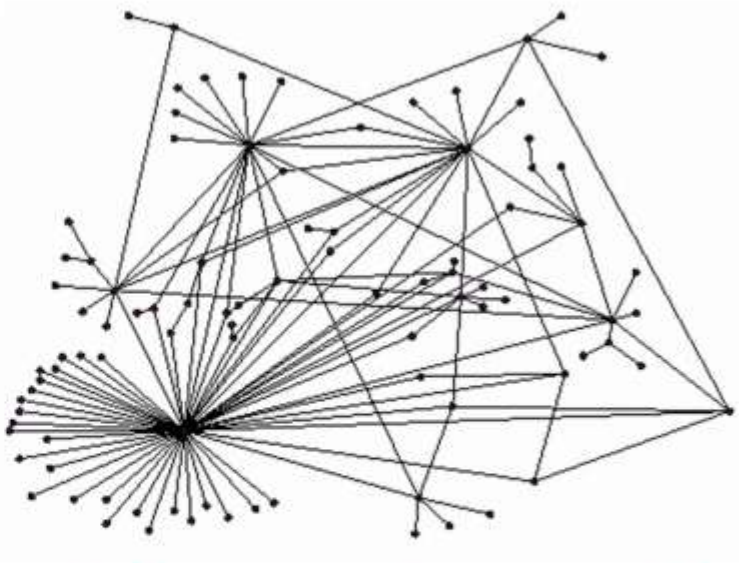


Random network

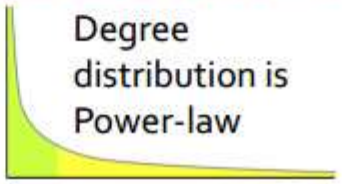
(Erdos-Renyi random graph)



Degree distribution is Binomial



Scale-free (power-law) network



Take-away Messages

- More and more of all interactions are happening online. Hence social network analysis is important.
- Social network data is complex because of its structure, both static and temporal, and scale.
- Web graph looks like a bow-tie
- We recognized a few important properties of graphs
- We studied random graph model, small world model, Kleinberg's model
- We also studied Power laws

Further Reading

- http://www.stanford.edu/class/cs224w/handouts.html
- https://www.coursera.org/course/sna
- Book by David Easley and Jon Kleinberg
  - http://www.cs.cornell.edu/home/kleinber/networks-book/
- Barabási, A.-L.; R. Albert (1999). "Emergence of scaling in random networks". *Science* **286** (5439): 509–512
- R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. 2000. Stochastic models for the Web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science* (FOCS '00). IEEE Computer Society, Washington, DC, USA

Preview of Lecture 10: Social Network Analysis (Part 2)

- Preferential Attachment Model
- Copying Model, Forest Fire Model
- Model with Network Components
- Evolving Network Model
- Compressible Graph Model

Disclaimers

- This course represents opinions of the instructor only. It does not reflect views of Microsoft or any other entity (except of authors from whom the slides have been borrowed).
- Algorithms, techniques, features, etc mentioned here might or might not be in use by Microsoft or any other company.
- Lot of material covered in this course is borrowed from slides across many universities and conference tutorials. These are gratefully acknowledged.

Thanks!