Social Media Mining

Network Models

Dear instructors/users of these slides:

Please feel free to include these slides in your own material, or modify them as you see fit. If you decide to incorporate these slides into your presentations, please include the following note:

R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*, Cambridge University Press, 2014.
Free book and slides at **http://socialmediamining.info/**

or include a link to the website:
http://socialmediamining.info/

Why should I use network models?

Facebook

May 2011:

- **721 millions** users.
- Average number of friends: **190**
- A total of **68.5 billion** friendships

September 2015:

- **1.35 Billion** users

1. What are the principal underlying processes that help initiate these friendships?

2. How can these seemingly independent friendships form this complex friendship network?

3. In social media there are many networks with millions of nodes and billions of edges.
   - **They are complex and it is difficult to analyze them**

So, what do we do?

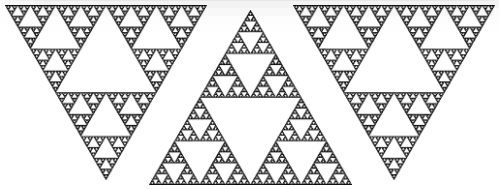
Design models that generate graphs

- The generated graphs should be similar to real-world networks.

If we can guarantee that generated graphs are similar to real-world networks:

1. We can analyze simulated graphs instead of real-networks (**cost-efficient**)
2. We can better understand real-world networks by providing concrete mathematical explanations; and
3. We can perform controlled experiments on synthetic networks when real-world networks are unavailable.

What are properties of real-world networks that should be accurately modeled?



Basic Intuition:

Hopefully! Our complex output [social network] is generated by a simple process

Properties of Real-World Networks

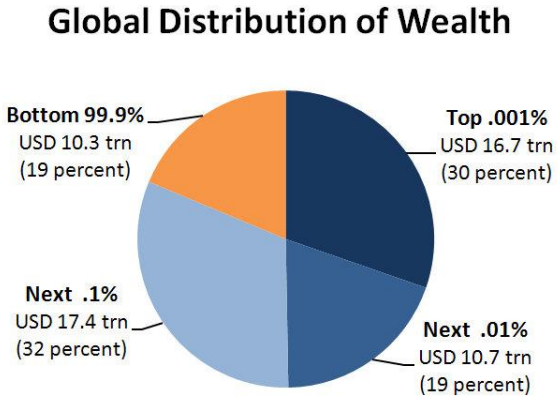
Power-law Distribution
High Clustering Coefficient
Small Average Path Length

Degree Distribution

Distributions

Wealth Distribution:

- Most individuals have average capitals,
- Few are considered wealthy.
- Exponentially more individuals with average capital than the wealthier ones.



James S. Henry, 2012

City Population:

- A few metropolitan areas are densely populated
- Most cities have an average population size.

Herbert A Simon,
On a Class of Skew Distribution Functions, 1955

Social Media:

- We observe the same phenomenon regularly when measuring popularity or interestingness for entities.

The **Pareto principle** (80–20 rule): 80% of the effects come from 20% of the causes

Distributions

Site Popularity:

- Many sites are visited less than a 1,000 times a month
- A few are visited more than a million times daily

User Activity:

- Social media users are often active on a few sites
- Some individuals are active on hundreds of sites

Product Price:

- There are exponentially more modestly priced products for sale compared to expensive ones.

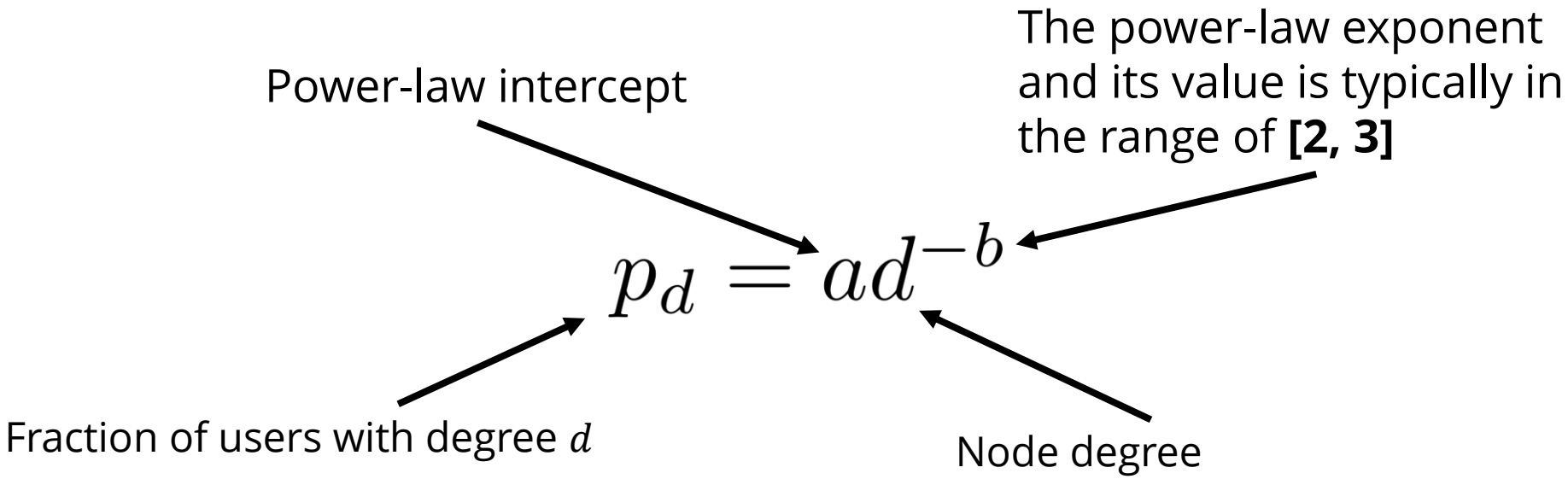
Friendships:

- Many individuals with a few friends and a handful of users with thousands of friends

(**Degree Distribution**)

Power-Law Degree Distribution

- When the frequency of an event changes as a power of an attribute
  - The frequency follows a **power-law**



$$\ln p_d = -b \ln d + \ln a$$

Some Properties of Power-Laws

$$p_d = ad^{-b}$$

We have

$$\sum_{d=0}^{\infty} p_d = 1$$

p_0 has to be zero!

So,

$$a = \frac{1}{\sum_{d=1}^{\infty} d^{-b}} = \frac{1}{\zeta(b)}$$

Riemann Zeta Function

Can be numerically approximated

Usually the very first d values do not exhibit power-law distribution

$$a = \frac{1}{\sum_{d=d_{\min}}^{\infty} d^{-b}} = \frac{1}{\zeta(b, d_{\min})}$$

Generalized (incomplete) Zeta Function

Approximating a

$$p_d = ad^{-b}$$

We can approximate by integrating

$$a = \frac{1}{\sum_{d=d_{\min}}^{\infty} d^{-b}} \approx \frac{1}{\int_{d_{\min}}^{\infty} d^{-b}\mathbf{d}d} = (b-1)d_{\min}^{b-1}$$

Then,

$$p_d \approx \frac{b-1}{d_{\min}}\left(\frac{d}{d_{\min}}\right)^{-b}$$

Moments of Power-Law Degree Distribution

The ith moment of the distribution is
- The first moment is: the mean

$$\sum_{d=0}^{\infty} d^i p_d$$

We can approximate the ith moment by

$$\sum_{d=0}^{\infty} d^i p_d = \sum_{d=0}^{d_{\min}-1} d^i p_d + a \sum_{d_{\min}}^{\infty} d^{i-b}$$

$$\approx \sum_{d=0}^{d_{\min}-1} d^i p_d + a \int_{d_{\min}}^{\infty} d^{i-b} \mathbf{d}d$$

$$= \sum_{d=0}^{d_{\min}-1} d^i p_d + \frac{a}{i-b+1} [d^{i-b+1}]_{d_{\min}}^{\infty}$$

Finite iff. $b > i+1$

We have b in range [2,3],
- First moment is **finite**
- Second is **infinite**

More on Moments

In reality second (or ith, $i > 2$) moment is finite

- You can just compute it from the graph

$$\frac{1}{n}\sum_{j=1}^{n} {d_j}^i$$

- Also the max degree in the graph is $n - 1$

$$\sum_{d=0}^{d_{\min}-1} d^i p_d + \frac{a}{i-b+1}\left[d^{i-b+1}\right]_{d_{\min}}^{\infty}$$

$$\left[d^{i-b+1}\right]_{d_{\min}}^{n-1} \approx (n-1)^{i-b+1} \approx n^{i-b+1}$$

- Finite in finite networks
- 2nd moment is almost n^{3-b}

Power-Law Distribution: Examples

- **Call networks:**
  - The fraction of telephone numbers that receive k calls per day is roughly proportional to $\mathbf{1/k^2}$
- **Book Purchasing:**
  - The fraction of books that are bought by k people is roughly proportional to $\mathbf{1/k^3}$
- **Scientific Papers:**
  - The fraction of scientific papers that receive k citations in total is roughly proportional to $\mathbf{1/k^3}$
- **Social Networks:**
  - The fraction of users that have in-degrees of k is roughly proportional to $\mathbf{1/k^2}$

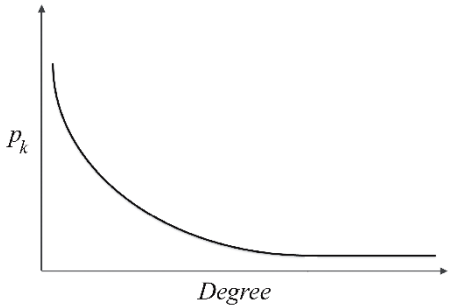
Power-Law Distribution

- Many real-world networks exhibit a *power-law* distribution.

- Power-laws seem to dominate
  - When the quantity being measured can be viewed as a type of popularity.

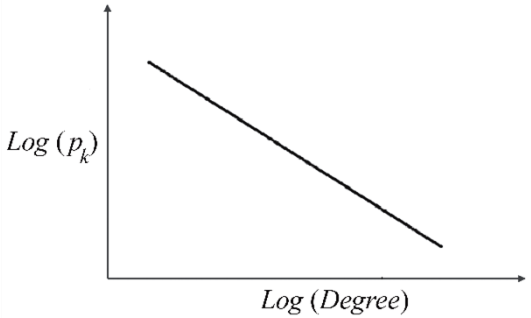
- A power-law distribution
  - **Small occurrences**: common
  - **Large instances**: extremely rare

A typical shape of a power-law distribution



(a) Power-Law Degree Distribution

Log-Log plot



(b) Log-Log Plot of Power-Law Degree Distribution

Power-law Distribution: An Elementary Test

To test whether a network exhibits a power-law distribution

1. Pick a popularity measure and compute it for the whole network
   - Example: number of friends for all nodes
2. Compute p_k, the fraction of individuals having popularity k.
3. Plot a log-log graph, where the x-axis represents $\ln k$ and the y-axis represents $\ln p_k$.
4. If a power-law distribution exists, we should observe a straight line

This is not a systematic approach!

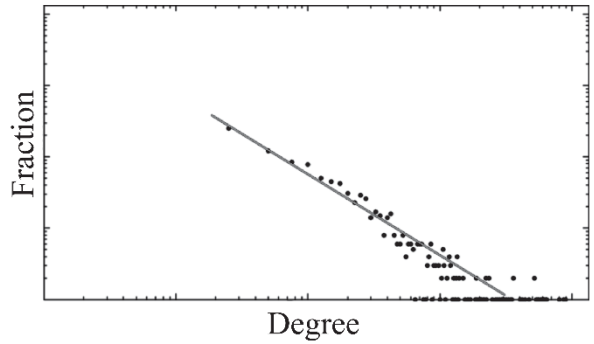
1. Other distributions could also exhibit this pattern
2. The results [estimations for parameters] can be biased and incorrect

For a systematic approach see:

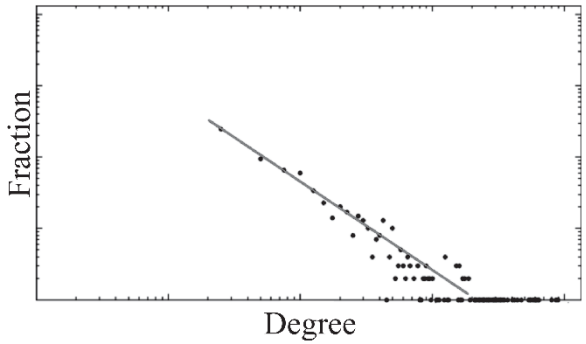
Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." *SIAM review 51(4)* (2009): 661-703.

Power-Law Distribution: Real-World Networks

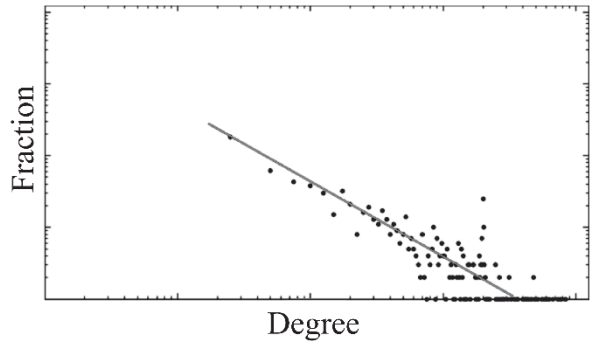
Networks with a power-law degree distribution are called **Scale-Free** networks



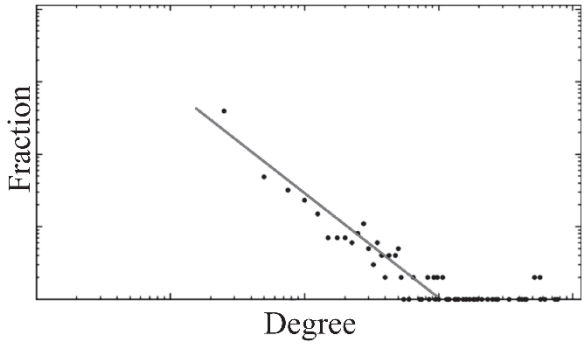
(a) Blog Catalog



(b) My Blog Log



(c) Twitter



(d) My Space

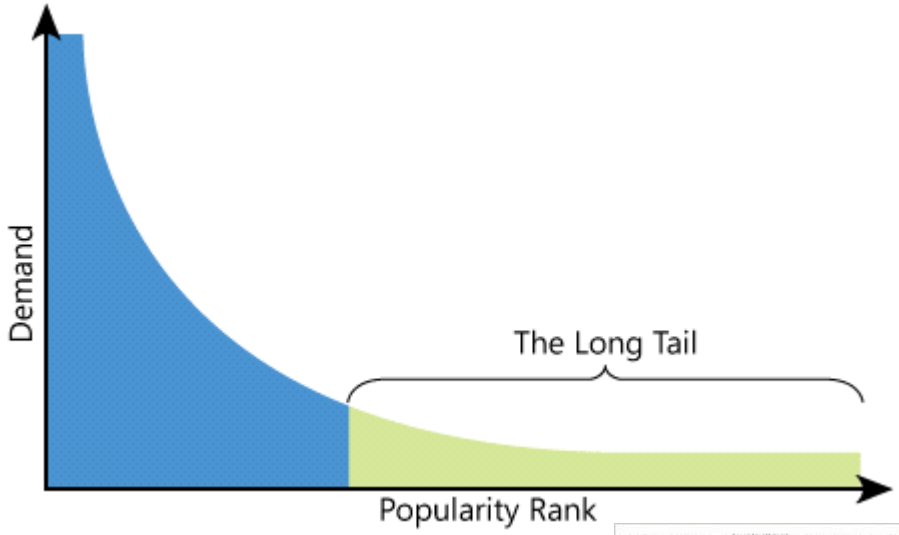
The tail of the power-law distribution is long!

The Looooooong Tail

Are most sales being generated by a small set of items that are enormously popular?

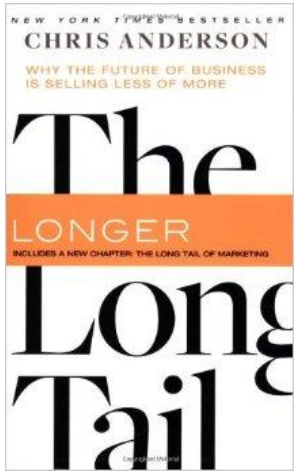
OR

By a much larger population of items that are each individually less popular?



The total sales volume of unpopular items, taken together, is very significant.

- 57% of Amazon's sales is from the long tail



Clustering Coefficient

Clustering Coefficient

- In real-world networks, friendships are highly transitive

Facebook

May 2011:

- Average clustering coefficient of **0.5** for users with **two** friends

  - Friends of a user are often friends with one another
  - These friendships form triads
  - High average [local] clustering coefficient

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
0.081	0.14 (with 100 friends)	0.31	0.33	0.17	0.13

Clustering Coefficient for Real-World Networks

	Network	Type	n	m	C
Social	Film actors	Undirected	449 913	25 516 482	0.20
	Company directors	Undirected	7 673	55 392	0.59
	Math coauthorship	Undirected	253 339	496 489	0.15
	Physics coauthorship	Undirected	52 909	245 300	0.45
	Biology coauthorship	Undirected	1 520 251	11 803 064	0.088
	Telephone call graph	Undirected	47 000 000	80 000 000	
	Email messages	Directed	59 812	86 300	
	Email address books	Directed	16 881	57 029	0.17
	Student dating	Undirected	573	477	0.005
	Sexual contacts	Undirected	2 810		
Information	WWW nd.edu	Directed	269 504	1 497 135	0.11
	WWW AltaVista	Directed	203 549 046	1 466 000 000	
	Citation network	Directed	783 339	6 716 198	
	Roget's Thesaurus	Directed	1 022	5 103	0.13
	Word co-occurrence	Undirected	460 902	16 100 000	
Technological	Internet	Undirected	10 697	31 992	0.035
	Power grid	Undirected	4 941	6 594	0.10
	Train routes	Undirected	587	19 603	
	Software packages	Directed	1 439	1 723	0.070
	Software classes	Directed	1 376	2 213	0.033
	Electronic circuits	Undirected	24 097	53 248	0.010
	Peer-to-peer network	Undirected	880	1 296	0.012
Biological	Metabolic network	Undirected	765	3 686	0.090
	Protein interactions	Undirected	2 115	2 240	0.072
	Marine food web	Directed	134	598	0.16
	Freshwater food web	Directed	92	997	0.20
	Neural network	Directed	307	2 359	0.18

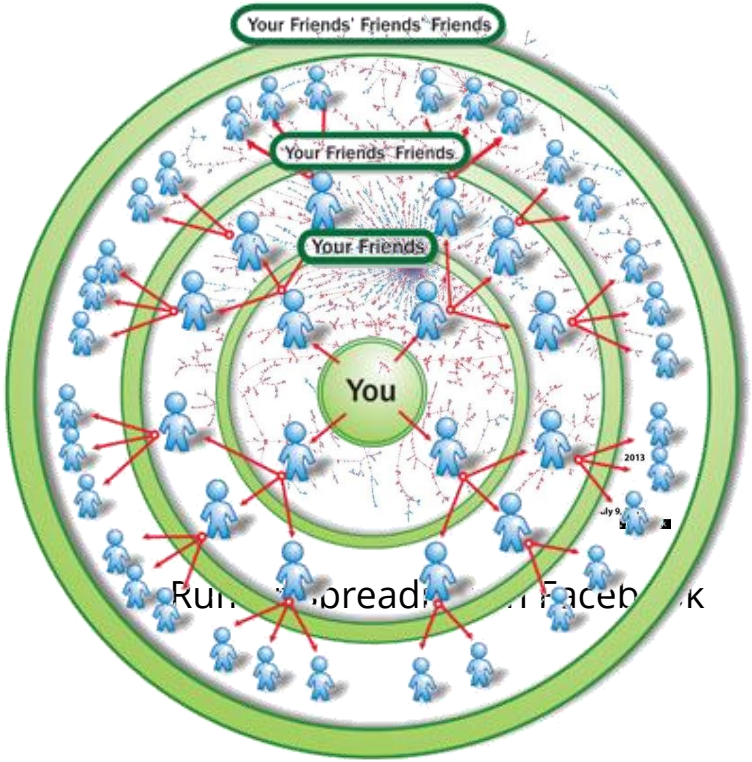
Source: M. E. J Newman

Average Path Length

How Small is the World?

A rumor is spreading over a social network.

- Assume all users pass it immediately to all of their friends



1. How long does it take to reach almost all of the nodes in the network?
2. What is the maximum time?
3. What is the average time?

Milgram's Experiment

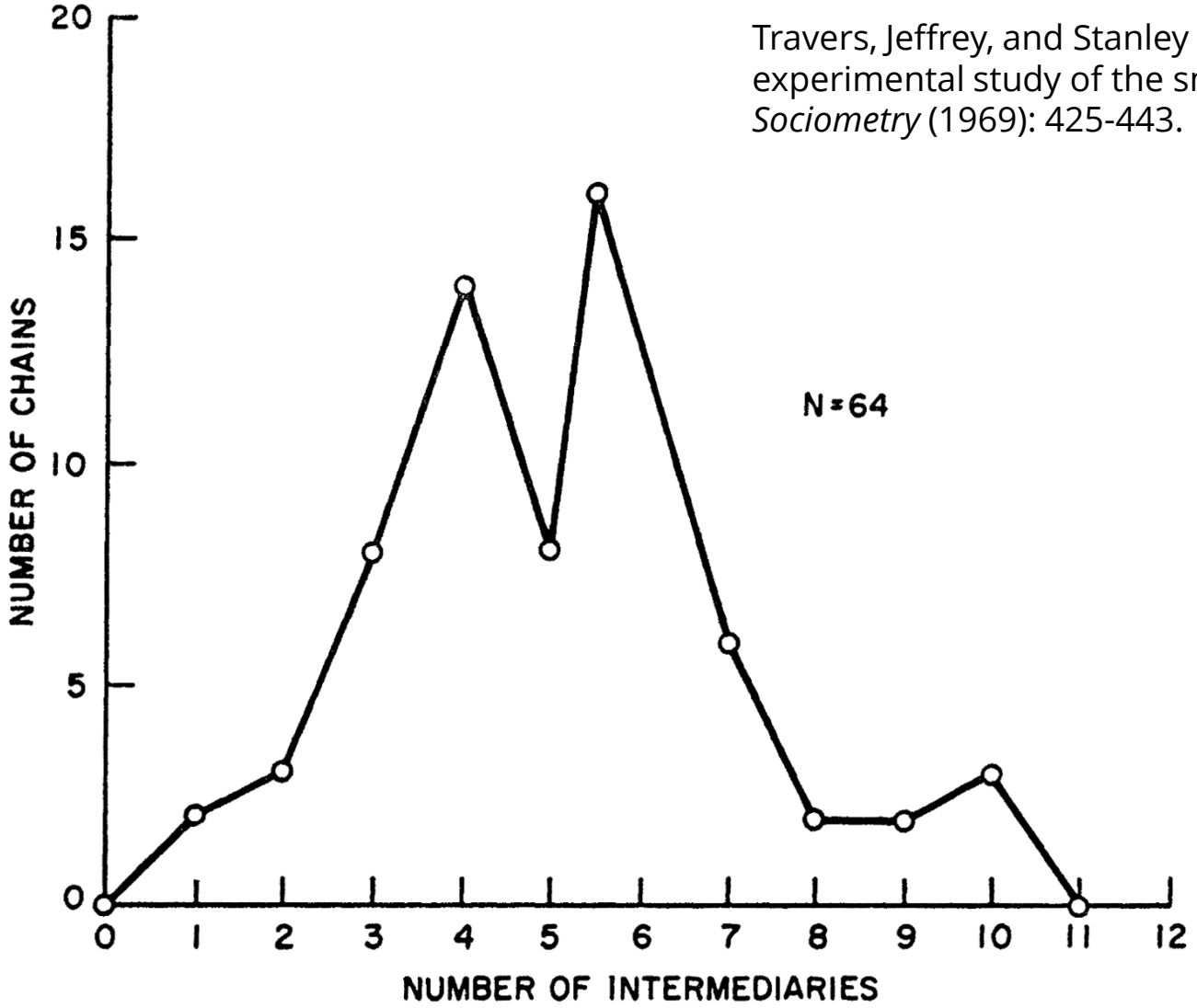
- 296 random people from Nebraska (196 people) and Boston (100 people) were asked to send a letter (via intermediaries) to a stock broker in Boston
- S/he could only send to people they personally knew, i.e., were on a first-name basis

Stanley Milgram (1933-1984)

Among the letters that found the target (64), the average number of links was around **six**.

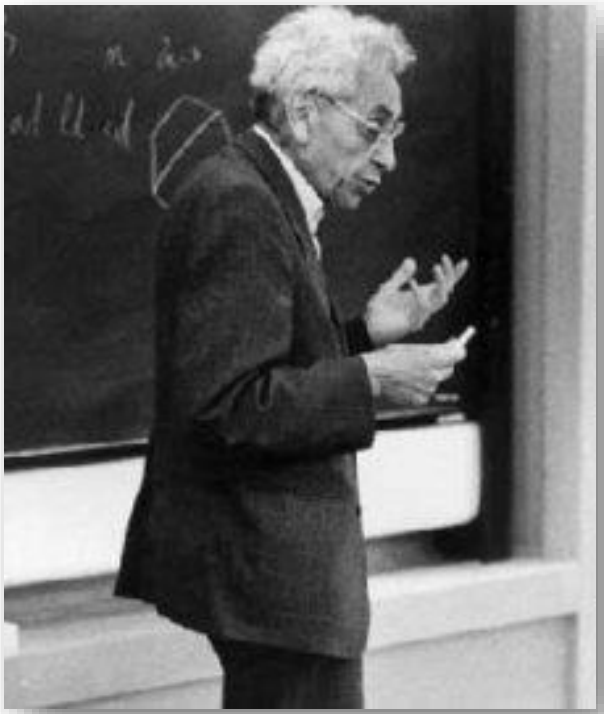
Milgram's Experiment

Travers, Jeffrey, and Stanley Milgram. "An experimental study of the small world problem." *Sociometry* (1969): 425-443.



Average Number of Intermediate people is 5.2

Erdös Number

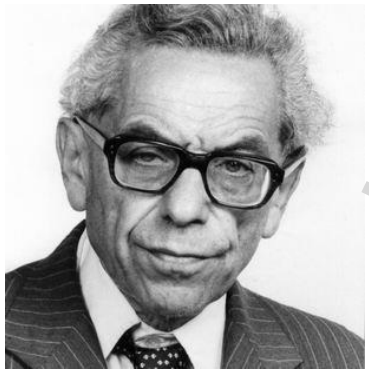


Paul Erdös (1913-1996)

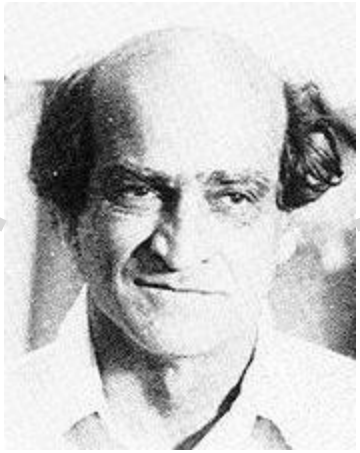
- **Erdös Number:** Number of links required to connect scholars to Erdös, via co-authorship papers
- Erdös wrote 1500+ papers with 507 co-authors.
- The Erdös Number Project allows you to compute your Erdös number:
  - http://www.oakland.edu/enp/
- Connecting path lengths, among mathematicians only:
  - Avg. is **4.65** and Maximum is **13**

Watch Erdös's documentary "*N is a number*" on YouTube

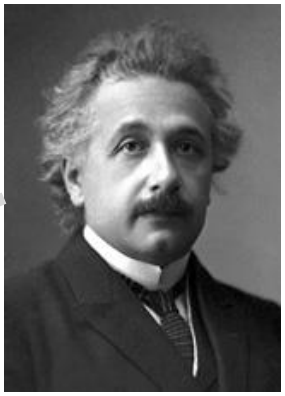
An Example of Erdös number 2 [Einstein]



Paul Erdös (1913-1996)



Ernst Gabor Straus (1922-1983)

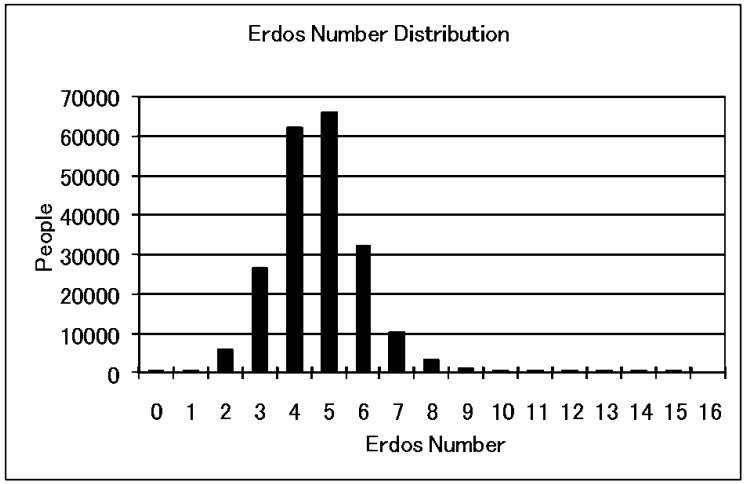


Albert Einstein (1879-1955)

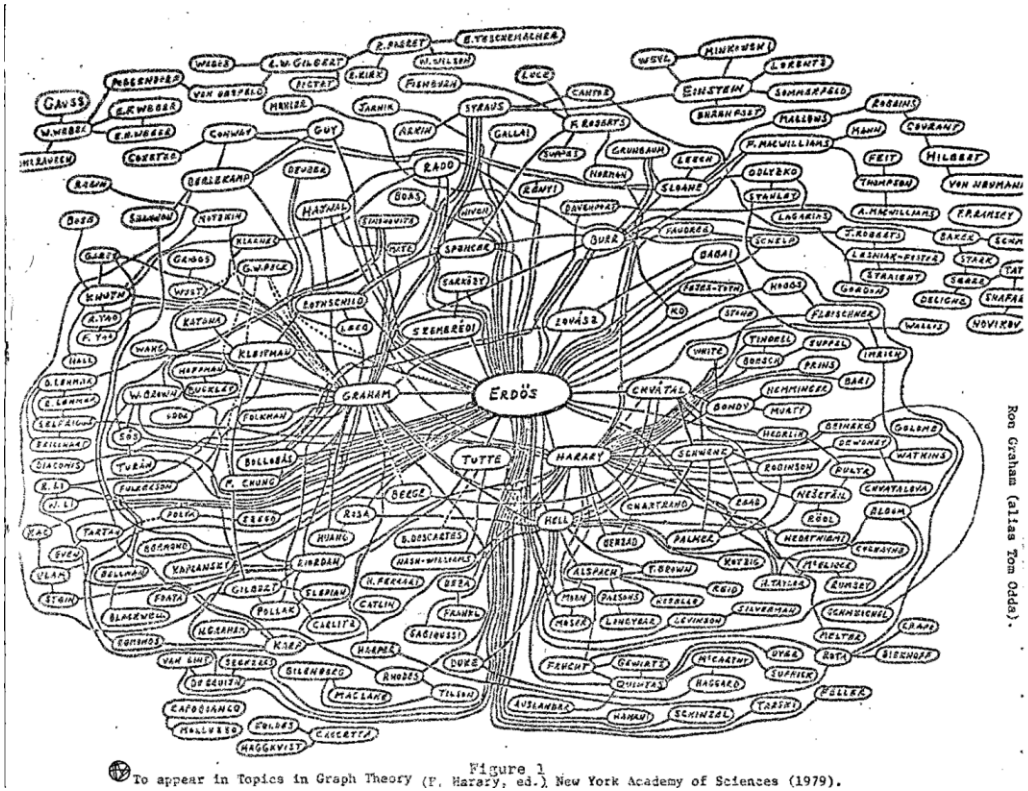
Erdös, Paul, B. Rothschild, and E. G. Straus. "Polychromatic Euclidean Ramsey theorems." *Journal of Geometry* 20.1 (1983): 28-35.

Einstein, Albert, and Ernst Gabor Straus. "A generalization of the relativistic theory of gravitation, II." *Annals of Mathematics* (1946): 731-741.

Erdös number Distribution



- The median Erdös number is **5**
- The mean is **4.65**
- The standard deviation is **1.27**



Erdös Number Project:

http://www.oakland.edu/enp/index.html

The Average Shortest Path

In real-world networks, any two members of the network are usually connected via a short paths.

Facebook

May 2011:

- Average path length was **4.7**
- **4.3** for US users

[Four degrees of separation]

The average path length is small

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
16.12	4.7	5.67	5.88	4.25	5.10

The Average Shortest Path in Sample Networks

	Network	Type	n	m	ℓ
Social	Film actors	Undirected	449 913	25 516 482	3.48
	Company directors	Undirected	7 673	55 392	4.60
	Math coauthorship	Undirected	253 339	496 489	7.57
	Physics coauthorship	Undirected	52 909	245 300	6.19
	Biology coauthorship	Undirected	1 520 251	11 803 064	4.92
	Telephone call graph	Undirected	47 000 000	80 000 000	
	Email messages	Directed	59 812	86 300	4.95
	Email address books	Directed	16 881	57 029	5.22
	Student dating	Undirected	573	477	16.01
	Sexual contacts	Undirected	2 810		
Information	WWW nd.edu	Directed	269 504	1 497 135	11.27
	WWW AltaVista	Directed	203 549 046	1 466 000 000	16.18
	Citation network	Directed	783 339	6 716 198	
	Roget's Thesaurus	Directed	1 022	5 103	4.87
	Word co-occurrence	Undirected	460 902	16 100 000	
Technological	Internet	Undirected	10 697	31 992	3.31
	Power grid	Undirected	4 941	6 594	18.99
	Train routes	Undirected	587	19 603	2.16
	Software packages	Directed	1 439	1 723	2.42
	Software classes	Directed	1 376	2 213	5.40
	Electronic circuits	Undirected	24 097	53 248	11.05
	Peer-to-peer network	Undirected	880	1 296	4.28
Biological	Metabolic network	Undirected	765	3 686	2.56
	Protein interactions	Undirected	2 115	2 240	6.80
	Marine food web	Directed	134	598	2.05
	Freshwater food web	Directed	92	997	1.90
	Neural network	Directed	307	2 359	3.97

l: average path length

Source: M. E. J Newman

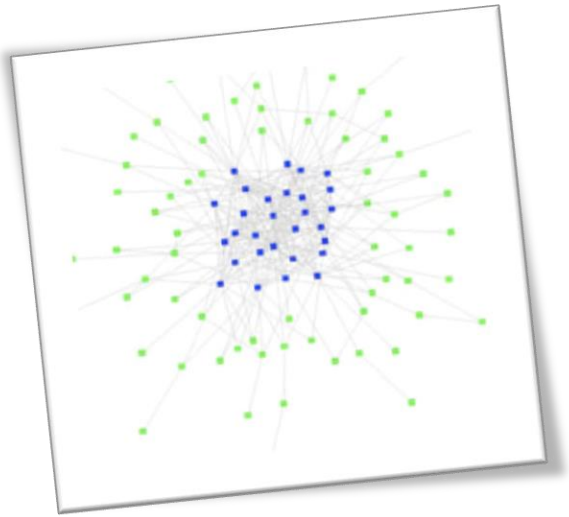
More Properties of Real-World Networks

Friendship Paradox [Feld 1991]

- i.e., you friends, on average, have more friends than you
- **Why?**
  - High degree nodes appear in many averages when averaging over friends
- It holds for 98% of Twitter Users [Hodas et al. 2013]

Core-Periphery Structure

- Dense Core
- Periphery nodes that connects to the core, but not connected among themselves
- Also known as
  - **Jellyfish** or **Octopus** structures



Network Models

- **Model-Driven Models!**

Random graphs

Small-World Model

Preferential Attachment

Random Graphs

Random Graphs

- We have to assume how friendships are formed
  - The most basic form:

Random Graph assumption:
Edges (i.e., friendships) between nodes (i.e., individuals) are formed randomly.

We discuss two random graph models $G(n,p)$ and $G(n,m)$

Random Graph Model - $G(n,p)$

- Consider a graph with a fixed number of nodes n
- Any of the $\binom{n}{2}$ edges can be formed independently, with probability p
- The graph is called a $G(n,p)$ *random graph*

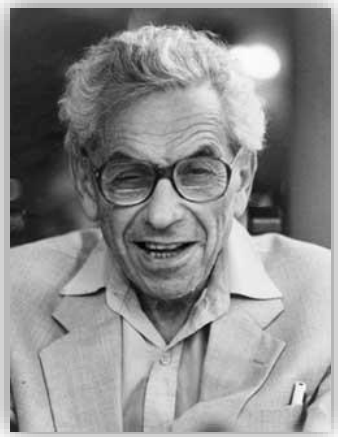
Proposed independently by Edgar Gilbert and by Solomonoff and Rapoport.

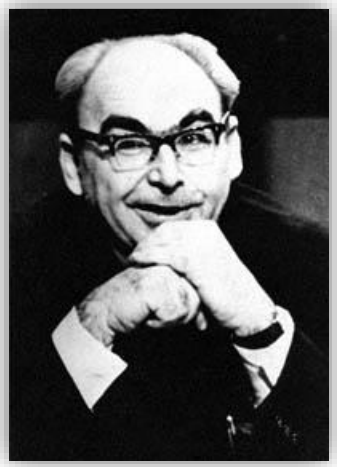
Random Graph Model - $G(n, m)$

- Assume both number of nodes n and number of edges m are fixed.
- Determine which m edges are selected from the set of possible edges
- Let Ω denote the set of graphs with n nodes and m edges
  - There are $|\Omega|$ different graphs with n nodes and m edges

$$|\Omega| = \binom{\binom{n}{2}}{m}$$

- To generate a random graph, we uniformly select one of the $|\Omega|$ graphs (the selection probability is $1/|\Omega|$)





This model was first proposed by *Paul Erdös* and *Alfred Rényi*

Modeling Random Graphs, Cont.

Similarities:

- In the limit (when n is large), both $G(n,p)$ and $G(n,m)$ models act similarly
  - The expected number of edges in $G(n,p)$ is $\binom{n}{2}p$
  - We can set $\binom{n}{2}p = m$ and in the limit, we should get similar results

Differences:

- The $G(n,m)$ model contains a fixed number of edges
- The $G(n,p)$ model is likely to contain none or all possible edges

Expected Degree

The expected number of edges connected to a node (expected degree) in $G(n,p)$ is $c = (n-1)p$

Proof.

- A node can be connected to at most $n-1$ nodes
  - or $n-1$ edges
- All edges are selected independently with probability p
- Therefore, on average, $(n-1)p$ edges are selected

- $c = (n-1)p$ or equivalently,

$$p = \frac{c}{n-1}$$

Expected Number of Edges

The expected number of edges in $G(n,p)$ is $\binom{n}{2}p$

Proof.

– Since edges are selected independently, and we have a maximum $\binom{n}{2}$ edges, the expected number of edges is $\binom{n}{2}p$

The probability of observing m edges

Given the $G(n,p)$ model, the probability of observing m edges is the binomial distribution

$$P(|E| = m) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m}$$

Proof.

- m edges are selected from the $\binom{n}{2}$ possible edges.
- These m edges are formed with probability p^m and other edges are not formed (to guarantee the existence of only m edges) with probability

$$(1-p)^{\binom{n}{2}-m}$$

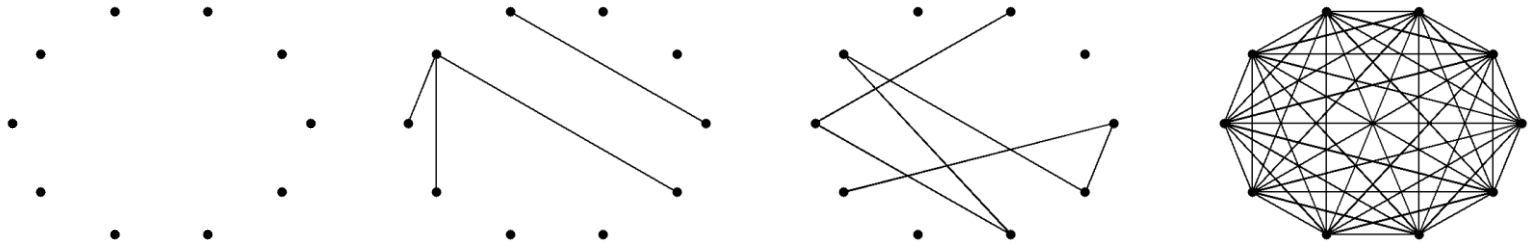
Evolution of Random Graphs

- **Create your own Random Graph Evolution demo: https://github.com/dgleich/erdosrenyi-demo**

The Giant Component

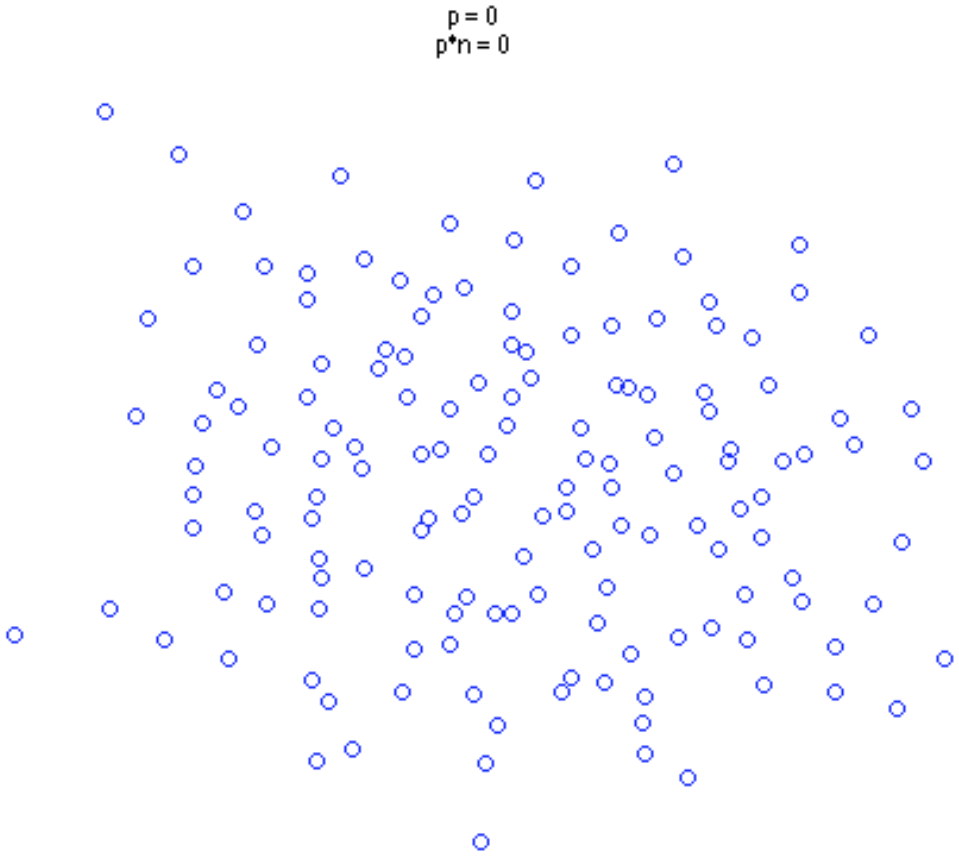
- In random graphs, as we increase p, a large fraction of nodes start getting connected
  - i.e., we have a path between any pair
- This large fraction forms a connected component:
  - **Largest connected component**, also known as the **Giant component**
- In random graphs:
  - $p = 0$
    - the size of the giant component is 0
  - $p = 1$
    - the size of the giant component is n

The Giant Component



Probability (p)	0.0	0.055	0.11	1.0
Average Node Degree (c)	0.0	0.8	≈1	n-1=9
Diameter	0	2	6	1
Giant Component Size	0	4	7	10
Average Path Length	0.0	1.5	2.66	1.0

Demo ($n = 150$)



From *David Gleich*

1st Phase Transition (Rise of the Giant Component)

- **Phase Transition**: the point where diameter value starts to shrink in a random graph
  - We have other phase transitions in random graphs
    - E.g., when the graph becomes connected

- The phase transition we focus on happens when
  - average node degree $c = 1$ (or when $p = 1/(n-1)$)

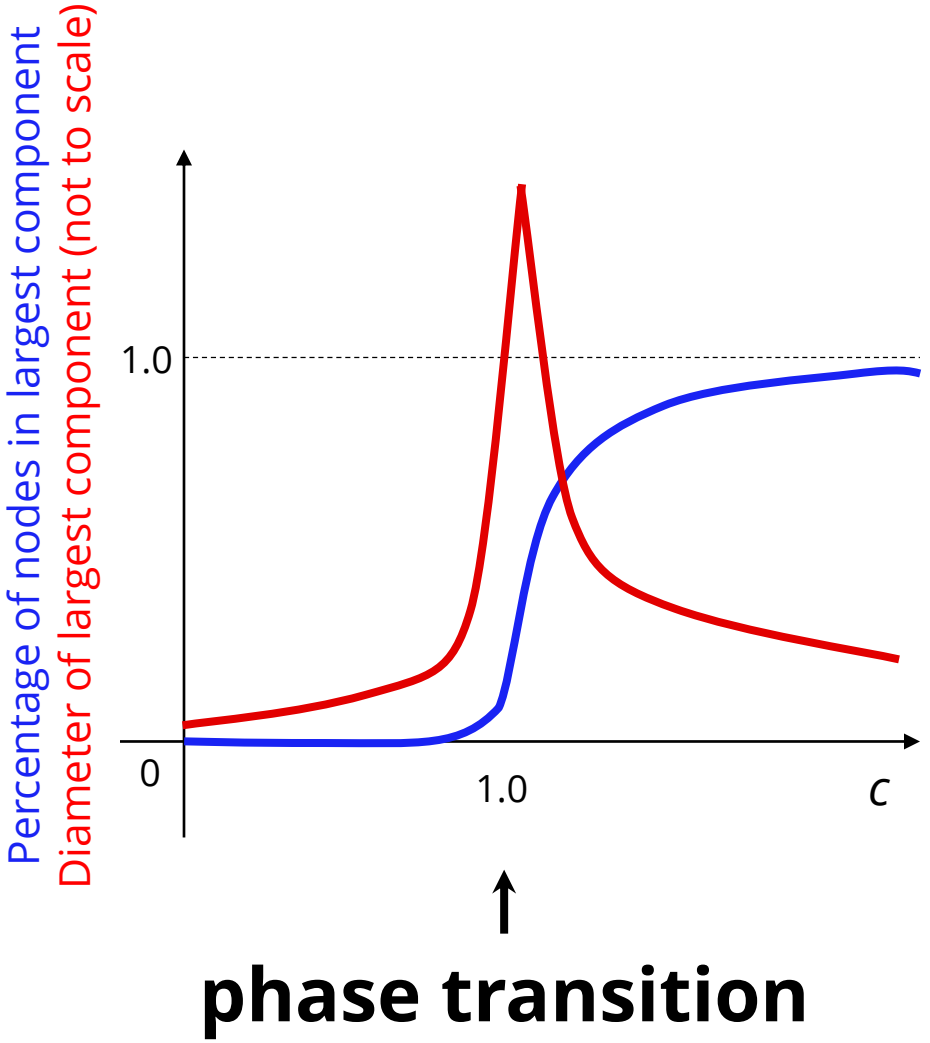
- At this Phase Transition:
  1. The giant component, which just started to appear, starts to grow, and
  2. The diameter, which *just* reached its maximum value, starts decreasing.

Random Graphs

If $c < 1$:
- **small**, isolated clusters
- **small** diameters
- **short** path lengths

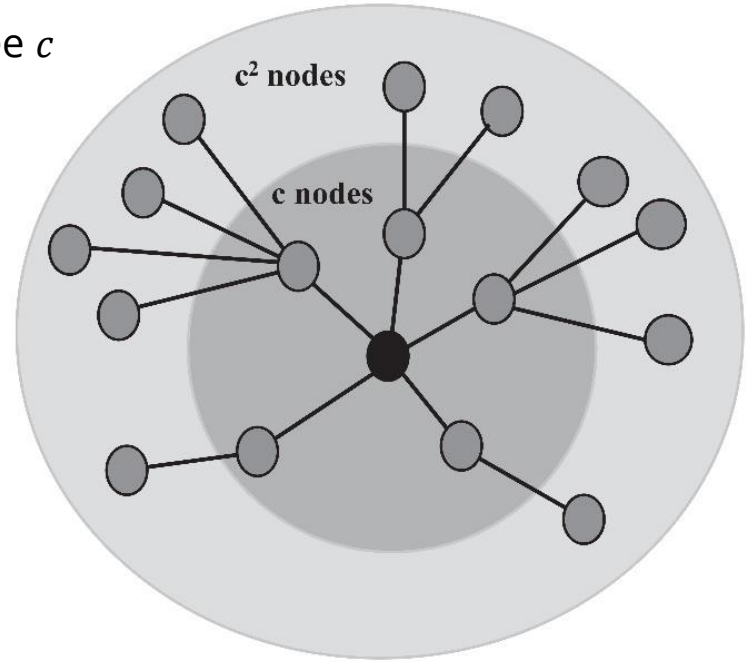
At $c = 1$:
- a ***giant component*** appears
- diameter **peaks**
- path lengths are **long**

For $c > 1$:
- almost **all** nodes **connected**
- diameter **shrinks**
- path lengths **shorten**



Why $c = 1$? [Rough Idea]

Consider a random graph with expected node degree c



- In this graph,
  - Consider any **connected** set of nodes S;
  - Let $S' = V - S$ denote the complement set; and
  - Assume $|S| << |S'|$.
- For any node v in S,
  - If we move one hop away from v, we visit approximately c nodes.
- If we move one hop away from nodes in S,
  - we visit approximately $|S|c$ nodes.

- If S is small, the nodes in S only visit nodes in S' and when moving one hop away from S, the set of nodes *guaranteed to be connected* gets larger by a factor c.
- In the limit, if we want this connected component to become the largest component, then after traveling n hops, its size must grow and we must have

$$c^n \geq 1 \text{ or equivalently } c \geq 1$$

Properties of Random Graphs

Degree Distribution

- When computing degree distribution, we estimate the probability of observing $P(d_v = d)$ for node v

- For a random graph generated by $G(n, p)$, this probability is

$$P(d_v = d) = \binom{n-1}{d} p^d (1-p)^{n-1-d}$$

- This is a binomial degree distribution. In the limit this will become the Poisson degree distribution

Binomial Distribution in the Limit

$$P(d_v = d) = \binom{n-1}{d} p^d (1-p)^{n-1-d}$$

$$\begin{aligned}
\lim_{n\to\infty} \binom{n-1}{d} &= \lim_{n\to\infty} \frac{(n-1)!}{(n-1-d)!\ d!} \\
&= \lim_{n\to\infty} \frac{(\ (n-1)\times(n-2)\times\ldots(n-d)\ )(n-1-d)!}{(n-1-d)!\ d!} \\
&= \lim_{n\to\infty} \frac{(\ (n-1)\times(n-2)\times\ldots(n-d)\ )}{d!} \\
&\approx \frac{(n-1)^d}{d!}.
\end{aligned}$$

$$\begin{aligned}
\lim_{n\to\infty} (1-p)^{n-1-d} &= \lim_{n\to\infty} e^{\ln(1-p)^{n-1-d}} = \lim_{n\to\infty} e^{(n-1-d)\ln(1-p)} \\
&= \lim_{n\to\infty} e^{(n-1-d)\ln(1-\frac{c}{n-1})} = \lim_{n\to\infty} e^{-(n-1-d)\frac{c}{n-1}} = e^{-c}.
\end{aligned}$$

$$\lim_{n\to\infty} \binom{n-1}{d} p^d (1-p)^{n-1-d} = \frac{(n-1)^d}{d!}\left(\frac{c}{n-1}\right)^d e^{-c} = e^{-c}\frac{c^d}{d!}$$

Poison Distribution

2nd Phase Transition (Connectivity)

$$\lim_{n\to\infty} P(d_v = d) = \lim_{n\to\infty} \binom{n-1}{d}\, p^d (1-p)^{n-1-d}$$

$$\lim_{n\to\infty} \binom{n-1}{d}\, p^d (1-p)^{n-1-d} = e^{-c}\frac{c^d}{d!}$$

- When the graph is connected there are no nodes with degree 0

- So, $P(d_v = 0) = e^{-c}$ should be less than $1/n$

$$e^{-c} = \frac{1}{n} \longrightarrow c = \ln n \longrightarrow \boxed{p = \frac{1}{n-1}\ln n}$$

$$p = \frac{c}{n-1}$$

Connectivity Threshold

Expected Local Clustering Coefficient

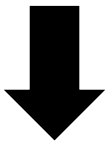
The expected local clustering coefficient for node v of a random graph generated by $G(n,p)$ is p

Proof.

$$C(v) = \frac{\text{number of connected pairs of } v\text{'s neighbors}}{\text{number of pairs of } v\text{'s neighbors}}$$

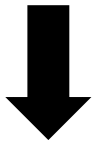
– v can have different degrees depending on the random procedure so the expected value is

$$\mathbf{E}(C(v)) = \sum_{d=0}^{n-1} \mathbf{E}(C(v)|d_v = d)\ P(d_v = d)$$



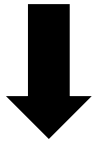
Expected Local Clustering Coefficient, Cont.

$$\mathbf{E}(C(v)) = \sum_{d=0}^{n-1} \mathbf{E}(C(v)|d_v = d)\, P(d_v = d)$$



$$\mathbf{E}(C(v)|d_v = d) = \frac{\text{number of connected pairs of } v\text{'s } d \text{ neighbors}}{\text{number of pairs of } v\text{'s neighbors}}$$

$$= \frac{p\binom{d}{2}}{\binom{d}{2}} = p$$



Sums up to 1

$$\mathbf{E}(C(v)) = p \sum_{d=0}^{d=n-1} P(d_v = d) = p$$

Global Clustering Coefficient

The global clustering coefficient of a random graph generated by $G(n, p)$ is p

Proof.

- The global clustering coefficient defines the probability of two neighbors of the same node being connected.
- In a random graph, for any two nodes, this probability is the same
  - Equal to the generation probability p that determines the probability of two nodes getting connected

The Average Path Length [Rough Idea]

The average path length in a random graph is

$$l \approx \frac{\ln |V|}{\ln c}$$

Proof.

- Assume D is the expected diameter of the graph
- Starting with any node and the expected degree $\boldsymbol{c}$,
  - one can visit approximately $\boldsymbol{c}$ nodes by traveling one edge
  - $\boldsymbol{c^2}$ nodes by traveling 2 edges, and
  - $\boldsymbol{c^D}$ nodes by traveling diameter number of edges
- We should have visited all nodes $c^{\mathcal{D}} \approx |V|$
- The expected diameter size tends to the average path length l in the limit

$$c^{\mathcal{D}} \approx c^{l} \approx |V| \quad \Rightarrow \quad l \approx \frac{\ln |V|}{\ln c}$$

Modeling with Random Graphs

- Compute the average degree c in the real-world graph
- Compute p using $c/(n-1) = p$
- Generate the random graph using p

- How representative is the generated graph?
  - **[Degree Distribution]** Random graphs do not have a power-law degree distribution
  - **[Average Path Length]** Random graphs perform well in modeling the average path lengths
  - **[Clustering Coefficient]** Random graphs drastically underestimate the clustering coefficient

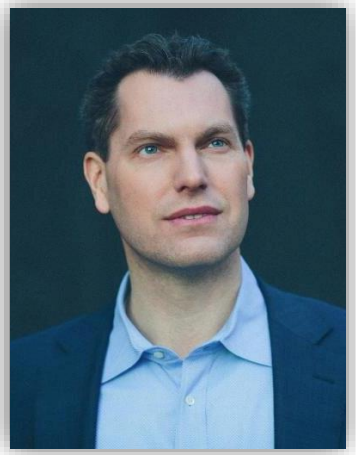
Real-World Networks / Simulated Random Graphs

	Original Network				Simulated Random Graph	
Network	*Size*	*Average Degree*	*Average Path Length*	*C*	*Average Path Length*	*C*
Film Actors	225,226	61	3.65	0.79	2.99	0.00027
Medline Coauthorship	1,520,251	18.1	4.6	0.56	4.91	1.8×10^{-4}
E.Coli	282	7.35	2.9	0.32	3.04	0.026
C.Elegans	282	14	2.65	0.28	2.25	0.05

Small-World Model

Small-world Model

- Small-world model
  - or the **Watts-Strogatz (WS)** model
  - A special type of random graph
  - Exhibits small-world properties:
    - Short average path length
    - High clustering coefficient

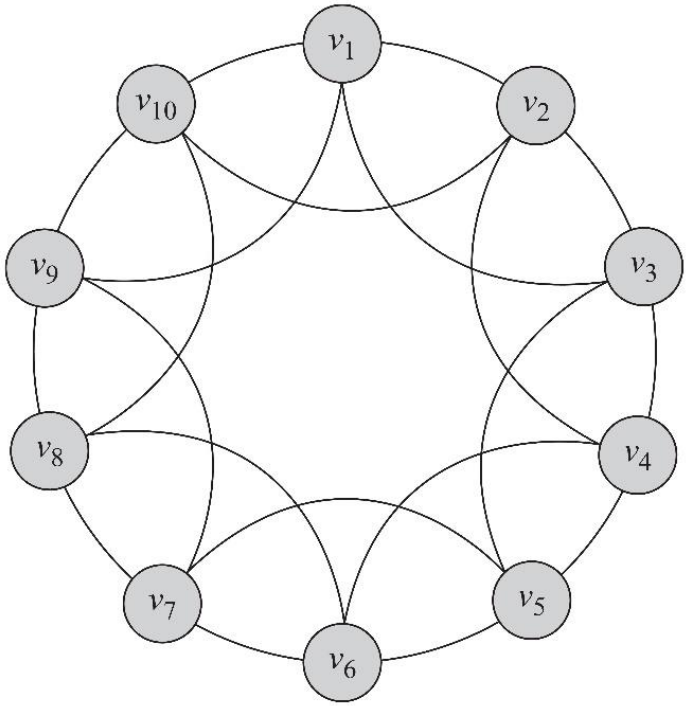


- It was proposed by Duncan J. Watts and Steven Strogatz in their joint 1998 Nature paper

Watts, Duncan J., and Steven H. Strogatz.
"Collective dynamics of 'small-world'networks."
nature 393.6684 (1998): 440-442.

Small-world Model

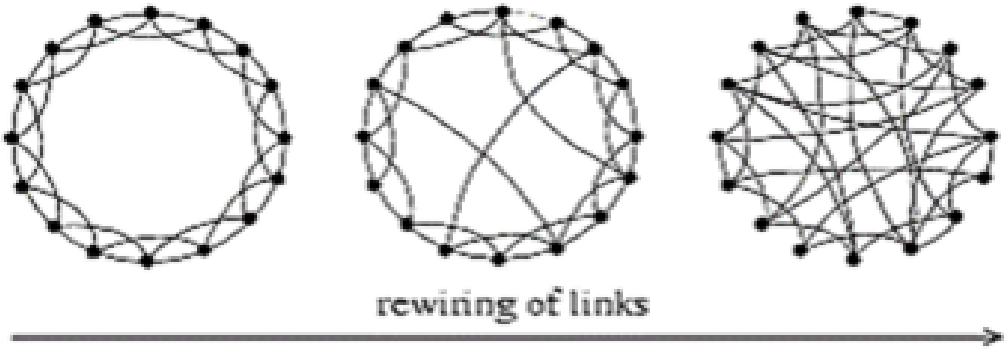
- In real-world interactions, many individuals have a limited and often at least, a fixed number of connections
- In graph theory terms, this assumption is equivalent to embedding users in a regular network
- A regular (ring) lattice is a special case of regular networks where there exists a certain pattern on how ordered nodes are connected to one another
- In a regular lattice of degree c, nodes are connected to their previous $c/2$ and following $c/2$ neighbors
- Formally, for node set V={v_1, v_2, v_3,..., v_n}, an edge exists between node i and j if and only if



$$0 \leq \min(n - |i - j|, |i - j|) \leq c/2$$

Generating a Small-World Graph

- The lattice has a **high**, but **fixed**, clustering coefficient
- The lattice has a **high** average path length



- In the small-world model, a parameter $0 \leq \beta \leq 1$ controls randomness in the model
  - When β is 0, the model is basically a regular lattice
  - When $\beta = 1$, the model becomes a random graph
- The model starts with a regular lattice and starts adding random edges [through **rewiring**]
  - **Rewiring:** take an edge, change one of its end-points randomly

Constructing Small World Networks

Algorithm 4.1 Small-World Generation Algorithm

Require: Number of nodes $|V|$, mean degree c, parameter β

1: **return** A small-world graph $G(V, E)$
2: G = A regular ring lattice with $|V|$ nodes and degree c
3: **for** node v_i (starting from v_1), and all edges $e(v_i, v_j)$, $i < j$ **do**
4: v_k = Select a node from V uniformly at random.
5: **if** rewiring $e(v_i, v_j)$ to $e(v_i, v_k)$ does not create loops in the graph or multiple edges between v_i and v_k **then**
6: rewire $e(v_i, v_j)$ with probability β: $E = E - \{e(v_i, v_j)\}$, $E = E \cup \{e(v_i, v_k)\}$;
7: **end if**
8: **end for**
9: Return $G(V, E)$

As in many network generating algorithms

- Disallow self-edges
- Disallow multiple edges

Small-World Model Properties

Degree Distribution

- The degree distribution for the small-world model is

$$P(d_v = d) = \sum_{n=0}^{\min(d-c/2, c/2)} \binom{c/2}{n} (1-\beta)^n \beta^{c/2-n} \frac{(\beta c/2)^{d-c/2-n}}{(d-c/2-n)} e^{-\beta c/2}$$

- In practice, in the graph generated by the small world model, most nodes have similar degrees due to the underlying lattice.

Regular Lattice vs. Random Graph

- Regular Lattice:
  - Clustering Coefficient (**high**):

$$\frac{3(c-2)}{4(c-1)} \approx \frac{3}{4}$$

  - Average Path Length (**high**): $n/2c$

- Random Graph:
  - Clustering Coefficient (**low**): p
  - Average Path Length (**ok!**) : $\ln |V| / \ln c$

What happens in Between?

- Does smaller average path length mean smaller clustering coefficient?

- Does larger average path length mean larger clustering coefficient?

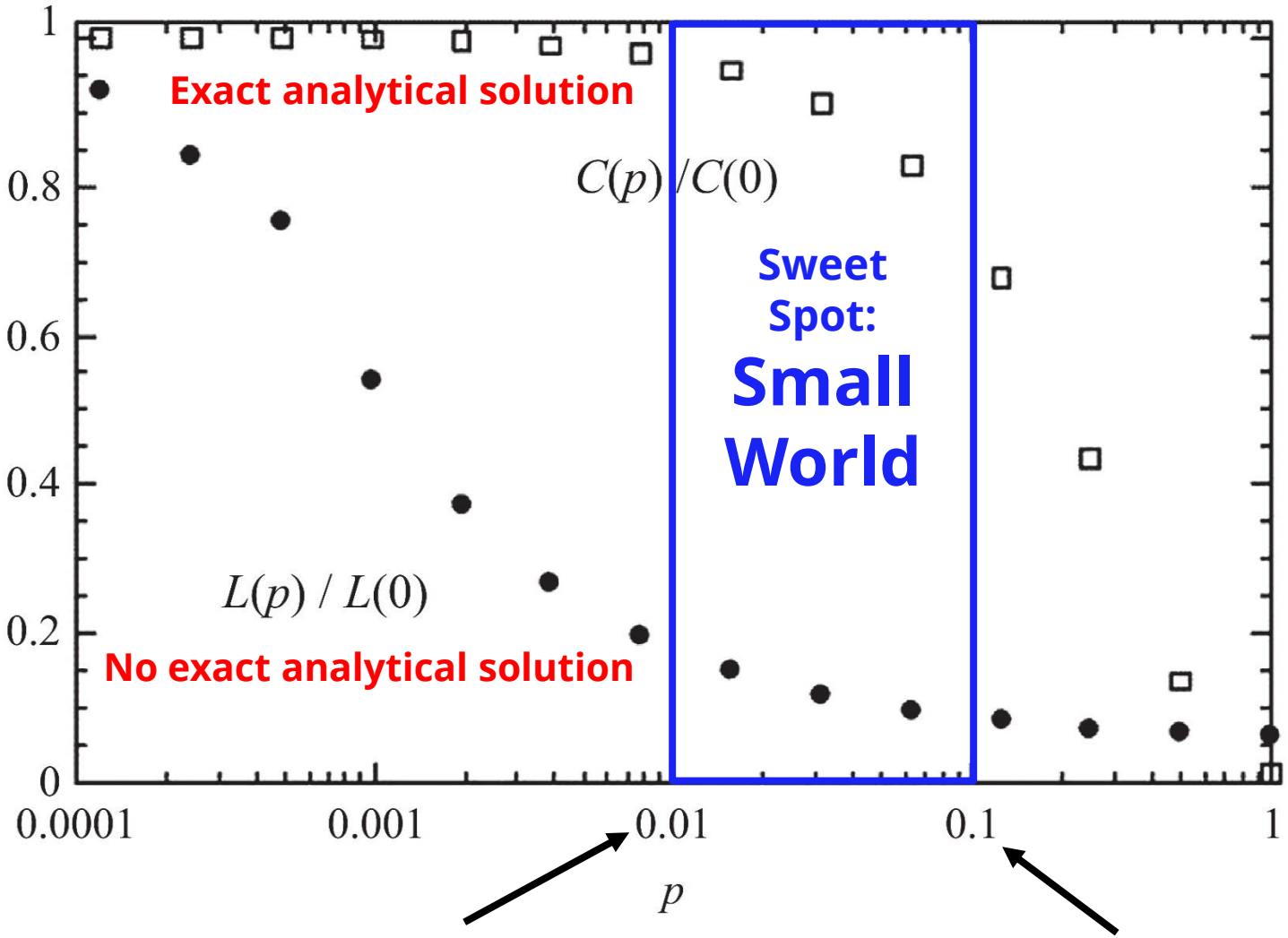
Numerical simulation:

- We increase p (i.e., β) from 0 to 1
- Assume
  - $L(0)$ is the average path length of the regular lattice
  - $C(0)$ is the clustering coefficient of the regular lattice
  - For any p, $L(p)$ denotes the average path length of the small-world graph and $\mathrm{C}(p)$ denotes its clustering coefficient

Observations:

- Fast decrease of average distance $L(p)$
- Slow decrease in clustering coefficient $C(p)$

Change in Clustering Coefficient /Avg. Path Length



Clustering Coefficient for Small-world model

- The probability that a connected triple stays connected after rewiring consists of

  1. The probability that none of the 3 edges were rewired is $(1-p)^3$
  2. The probability that other edges were rewired back to form a connected triple
     - Very small and can be ignored

- Clustering coefficient

$$C(p) \approx (1-p)^3 C(0)$$

Modeling with the Small-World Model

- Given a real-world network in which average degree is c and clustering coefficient C is given,
  - we set $C(p) = C$ and determine $\beta\ (= p)$ using equation

$$C(p) \approx (1-p)^3 C(0)$$

- Given β, c, and n (size of the real-world network), we can simulate the small-world model

Real-World Network and Simulated Graphs

	Original Network				Simulated Graph	
Network	*Size*	*Average Degree*	*Average Path Length*	*C*	*Average Path Length*	*C*
Film Actors	225,226	61	3.65	0.79	4.2	0.73
Medline Coauthorship	1,520,251	18.1	4.6	0.56	5.1	0.52
E.Coli	282	7.35	2.9	0.32	4.46	0.31
C.Elegans	282	14	2.65	0.28	3.49	0.37

Preferential Attachment Model

Preferential Attachment Model

- **Main assumption:**
  - When a new user joins the network, the probability of connecting to existing nodes is proportional to existing nodes' degrees
  - For the new node v
    - Connect v to a random node v_i with probability

$$P(v_i) = \frac{d_i}{\sum_j d_j}$$

  - Proposed by Albert-László Barabási and Réka Albert
    - A special case of the Yule process

Distribution of wealth in the society:

The rich get richer

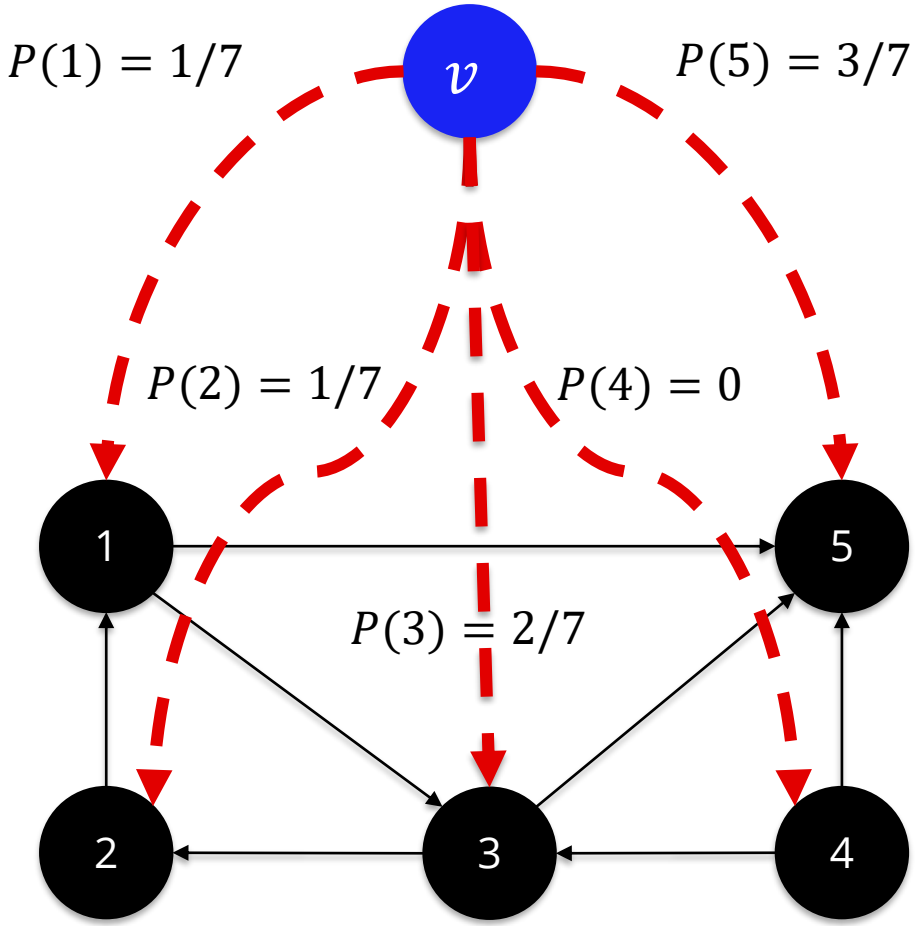
Barabási, Albert-László, and Réka Albert. "Emergence of scaling in random networks." *science* 286.5439 (1999): 509-512.

Preferential Attachment: Example

- Node v arrives

$$P(v_i) = \frac{d_i}{\sum_j d_j}$$

- $P(1) = 1/7$
- $P(2) = 1/7$
- $P(3) = 2/7$
- $P(4) = 0$
- $P(5) = 3/7$



Constructing Scale-free Networks

Algorithm 4.2 Preferential Attachment

Require: Graph $G(V_0, E_0)$, where $|V_0| = m_0$ and $d_v \geq 1 \; \forall \; v \in V_0$, number of expected connections $m \leq m_0$, time to run the algorithm t

1: **return** A scale-free network
2: `//Initial graph with` m_0 `nodes with degrees at least 1`
3: $G(V, E) = G(V_0, E_0)$;
4: **for** 1 to t **do**
5: $V = V \cup \{v_i\}$; `//` `add new node` v_i
6: **while** $d_i \neq m$ **do**
7: Connect v_i to a random node $v_j \in V, i \neq j$ (i.e., $E = E \cup \{e(v_i, v_j)\}$) with probability $P(v_j) = \frac{d_j}{\sum_k d_k}$.
8: **end while**
9: **end for**
10: Return $G(V, E)$

Properties of the Preferential Attachment Model

Properties

- **Degree Distribution:** $P(d) = \frac{2m^2}{d^3}$
- **Clustering Coefficient:** $C = \frac{m_0 - 1}{8} \frac{(\ln t)^2}{t}$
- **Average Path Length:** $l \sim \frac{\ln |V|}{\ln(\ln |V|)}$

Modeling with the Preferential Attachment Model

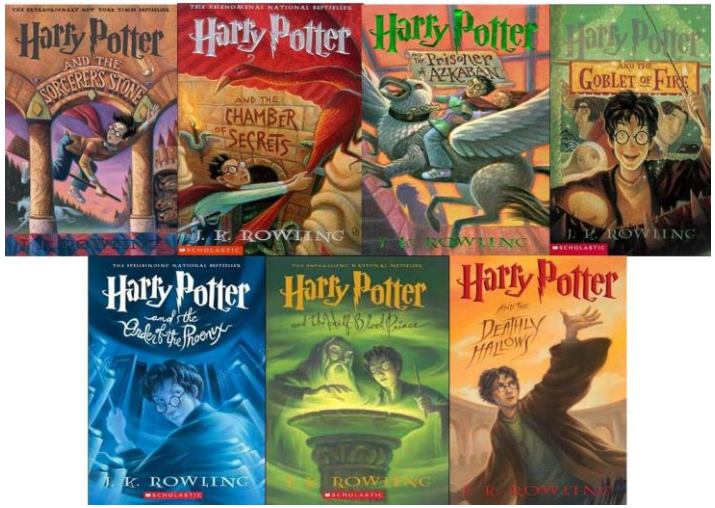
- Similar to random graphs, we can simulate real-world networks by generating a preferential attachment model by setting the expected degree m
  - See Algorithm 4.2 in the book

Real-World Networks and Simulated Graphs

	Original Network				Simulated Graph	
Network	*Size*	*Average Degree*	*Average Path Length*	*C*	*Average Path Length*	*C*
Film Actors	225,226	61	3.65	0.79	4.90	≈ 0.005
Medline Coauthorship	1,520,251	18.1	4.6	0.56	5.36	≈ 0.0002
E.Coli	282	7.35	2.9	0.32	2.37	0.03
C.Elegans	282	14	2.65	0.28	1.99	0.05

Unpredictability of the Rich-Get-Richer Effects

- The initial stages of one's rise to popularity are fragile
- Once a user is well established, the rich-get-richer dynamics of popularity is likely to push the user even higher
- **But** getting the rich-get-richer process started in the first place is full of potential accidents and near-misses



If we could roll time back to 1997, and then run history forward again, would the Harry Potter books again sell hundreds of millions of copies?

See more: Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. "Experimental study of inequality and unpredictability in an artificial cultural market." *science* 311.5762 (2006): 854-856.

Network Models Extended

- **Model-Driven Models!**

Configuration Model
Kleinberg Small-World Model
Vertex Copying Model

Configuration Model

Configuration Model

Problem: the degree distribution in random graphs is now power law

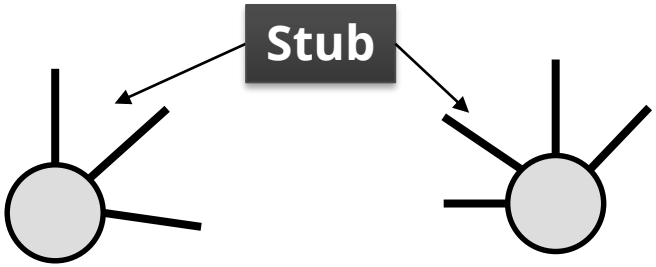
- Binomial for small n and Poisson for large n

Solution:

- Given a degree sequence, ensure that the graph generated has that degree sequence.
- The degree sequence can
  1. Come from a real-world graph
  2. Be sampled from a degree distribution, i.e., power law

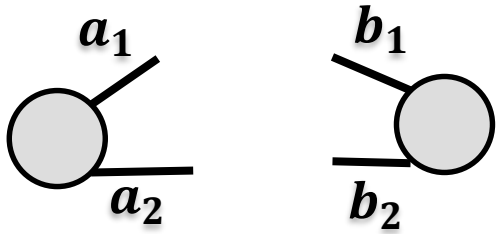
Configuration Model

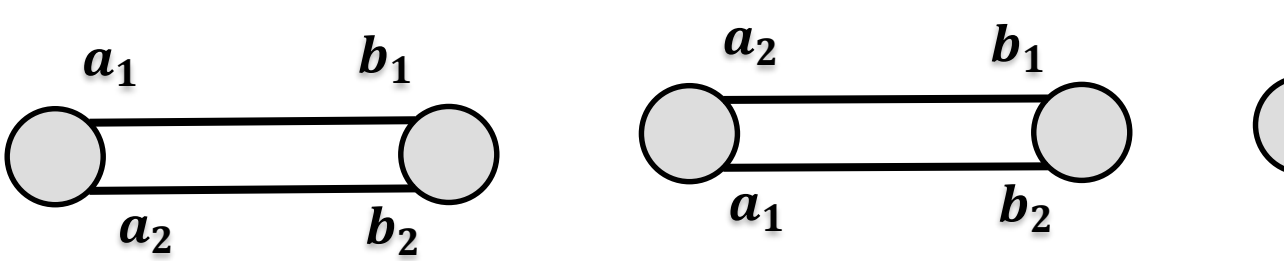
I. Given the degree sequence $d_1, d_2, \ldots, d_n$ generate n nodes, such that node i has d_i ***stubs*** (also known as ***half-edges***)



II. Randomly match stubs until no edges there are no more stubs

- You can have loops and multiple edges
- Each **configuration** (**not each graph!**) appears with equal probability
- You can get the same graph





How to generate the Configuration model

1. Create a list where the **node id** for node v_i with degree d_i is repeated d_i times
2. Shuffle the list
3. Starting from the first index, join adjacent nodes

Example: Degree sequence (2,2,2)

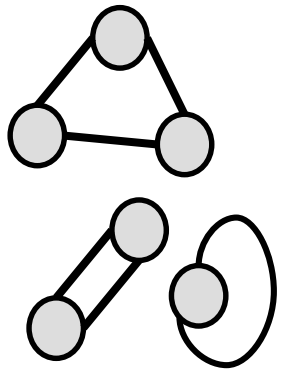
v_1	v_1	v_2	v_2	v_3	v_3

Random Shuffle 1:

v_1	v_2	v_2	v_3	v_3	v_1

Random Shuffle 2:

v_1	v_1	v_2	v_3	v_3	v_2



Properties of the Configuration Model

The probability that node v_i gets connected to node v_i is approximately

$$\frac{d_i d_j}{2m}$$

Proof:

In the shuffled list, for each v_i that appears:

- There are d_j instances of v_j that it could be next to
- The probability of being next to one is $d_j/(2m-1)$
- There are d_i instances of v_j ; therefore, the total probability is $(d_i d_j)/(2m-1) \approx (d_i d_j)/2m$

Properties of the Configuration Model

Expected Number of **Multi-Edges**

$$\frac{1}{2}\left[\frac{(\bar{d^2})-(\bar{d})}{\bar{d}}\right]^2$$

- It depends on the first and second moment
  - Most likely a very small number
- **How to prove?**

$$\sum_{i\neq j}\left(\frac{d_i d_j}{2m}\right)\left(\frac{(d_i-1)(d_j-1)}{2m}\right)$$

Expected Number of **Self-Loops**

$$\frac{(\bar{d^2})-(\bar{d})}{2\bar{d}}$$

- Note that probability of a node getting connected to itself is $(\ d_i(d_i-1)\ )/4m$

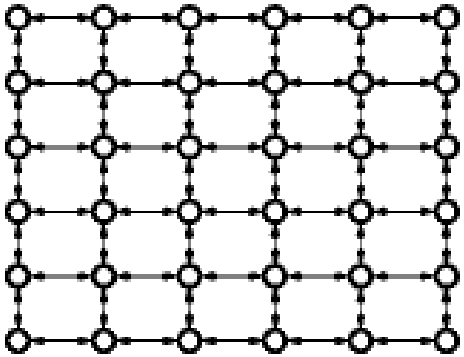
Kleinberg Small-World Model

Kleinberg Small-World Model

- The Watts-Strogatz Model cannot explain how navigation in the Milgram's Small-World experiment works
- Nodes only have their local information, but seem to find global short paths.

Construction:

- $n \times n$ lattice (n^2 nodes)
- Lattice distance is *Manhattan* distance (L_1- norm)



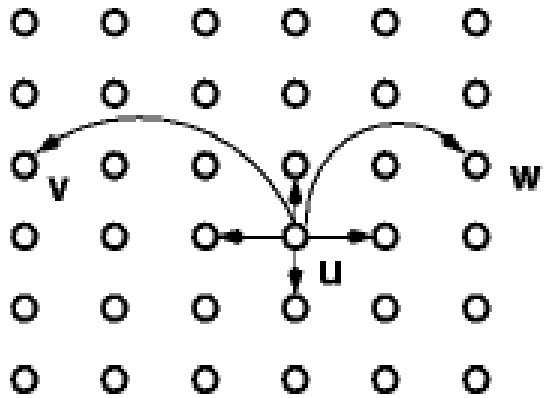
$$D((i,j),(k,l)) = |k-i| + |j-l|$$

Short- and Long-Range Contacts

- Nodes have short- and long-range contacts

Short-Range Contacts:

- Nodes are connected via **undirected** edges to all nodes within distance p



$p = 1$ and $q = 2$

Long-Range Contacts

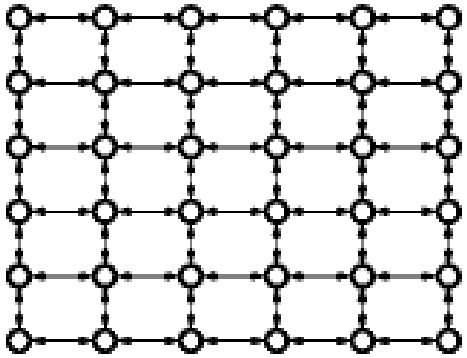
- For $q \geq 0$ and $r \geq 0$, q **directed** edges are made using independent random trials
- Connection probability for long-range contacts follows power-law

$$p(u \to v) = \frac{d(u,v)^{-r}}{\sum_{u \neq v} d(u,v)^{-r}}$$

Navigation Algorithm

Consider $r = 2$, $q = 1$, $p = 1$, and the lattice to be 2-dimensional

- Works as long as r = dimension of the lattice



The following greedy algorithm will reach the target in $O(log^2 n)$

1. Examine all of your neighbors (4 short range and the 1 long-range contact)
2. Find the closest to the target
3. Jump to that neighbor

Getting Lower Bounds on Long-Range Contacts

$$\sum_{u \neq v} d(u,v)^{-2} \leq \sum_{i=1}^{2n-2} (4i) i^{-2}$$

Max distance in the lattice

Number of nodes That are i away

$$= 4 \sum_{i=1}^{2n-2} i^{-1}$$

$$\leq 4(1 + \ln(2n-2))$$

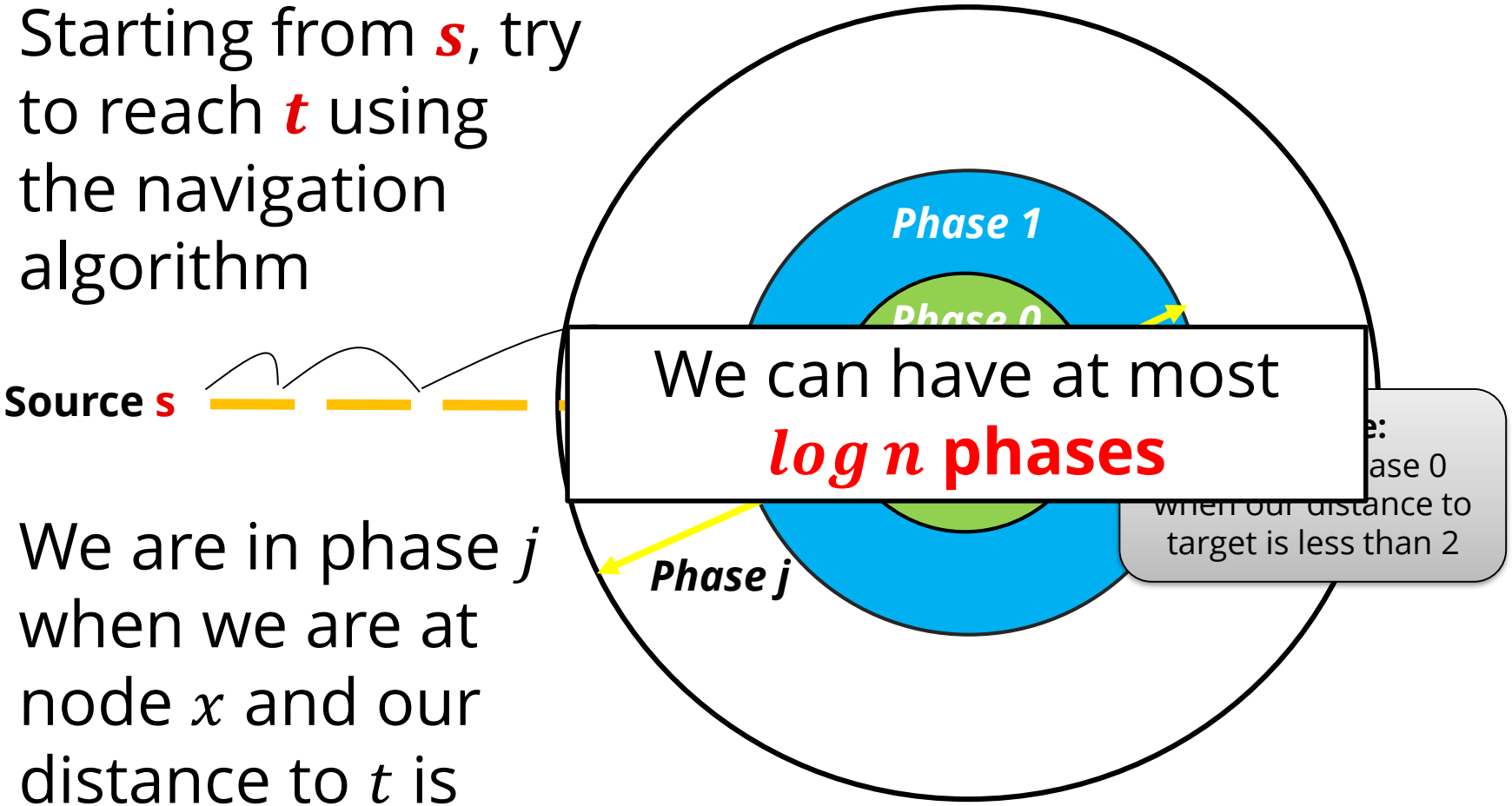
$$\leq 4(\ln 3 + \ln(2n))$$

$$\leq 4(\ln 6n)$$

$$p(u \rightarrow v) \geq \frac{d(u,v)^{-r}}{4 \ln(6n)}$$

Phases

Starting from s, try to reach t using the navigation algorithm



We are in phase j when we are at node x and our distance to t is

$$2^j < d(x,t) \leq 2^{j+1}$$

Phases

- If we can prove that we need to take $\log n$ jumps within each phase then we are done!
- Phase j ends when $d(x,t) < 2^j$
- To end phase j, a node x should connect to some long-range contact l with distance $d(l,t) < 2^j$

1. How many nodes are within distance 2^j?
2. What is the probability of connecting to one such node?

How many nodes are within distance 2^j

- At least
$$1 + \sum_{i=1}^{2^j} i = 1 + 2^{2j-1} + 2^j \geq 2^{2j-1}$$
- For all such nodes, there distance is at most
$$2^{j+1} + 2^j \leq 2^{j+2}$$
- We know that $p(u \to v) \geq \frac{d(u,v)^{-r}}{4 \ln(6n)}$
- So, the probability of ending the phase is at least
$$\frac{2^{2j-1}}{(2^{2j+4})4 \ln(6n)} = \frac{1}{128 \ln 6n}$$
- So, we need at most $128 \ln 6n = O(\ln n)$ jumps

Vertex Copying Model

Not all power-laws are created the same model

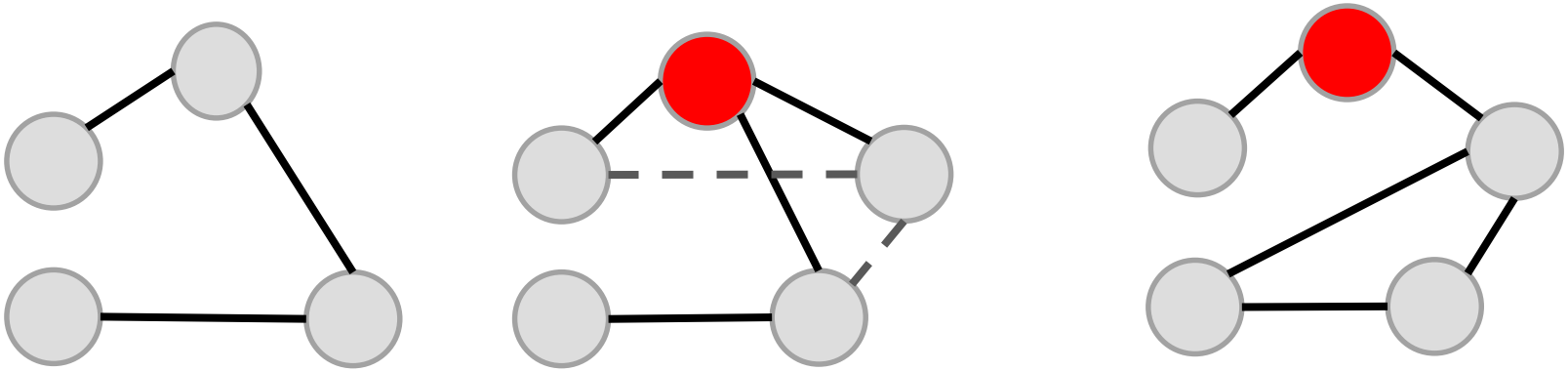
Basic idea:

- A new node comes in
- We select some other node at random
- We connect to that node (optional)
- We copy all of its connections

Problem: What if doesn't have any edges!

Solution:

- Instead of copying every connection, we flip coins for each
  - With prob. p we copy, with prob. (1-p) we select some other node uniformly at random and connect to



Vertex Copying Model: Properties

- Known to generate a power law degree distribution
- Known to preserve the community structure of large graphs