CS 579: Online Social Network Analysis

Network Models

Reading: Chapter 4

Spring 2022

Kai Shu

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

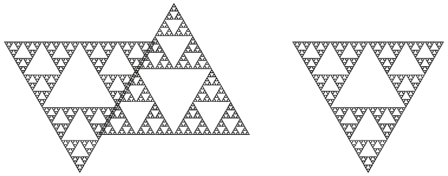
1. What are the principal underlying processes that help initiate these friendships?
2. How can these seemingly independent friendships form this complex friendship network?
3. In social media there are many networks with millions of nodes and billions of edges.
   - **They are complex and it is difficult to analyze them**

So, what do we do?

Design models that generate graphs

- Those graphs should be similar to real-world networks

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



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

What are properties of real-world networks we should accurately model?

Basic Intuition:

Hopefully, our complex output [social network] is generated by a simple process

Properties of Real-World Networks

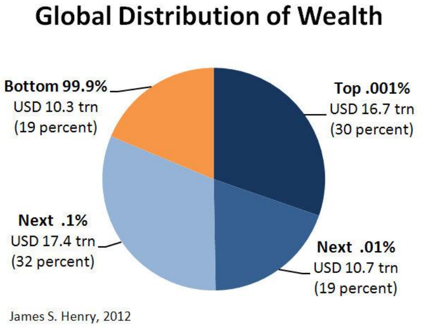
Power-law Distribution
High Clustering Coefficient
Small Average Path Length

Degree Distribution

Degree Distribution

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

Social Media:

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

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

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

Degree Distribution

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.

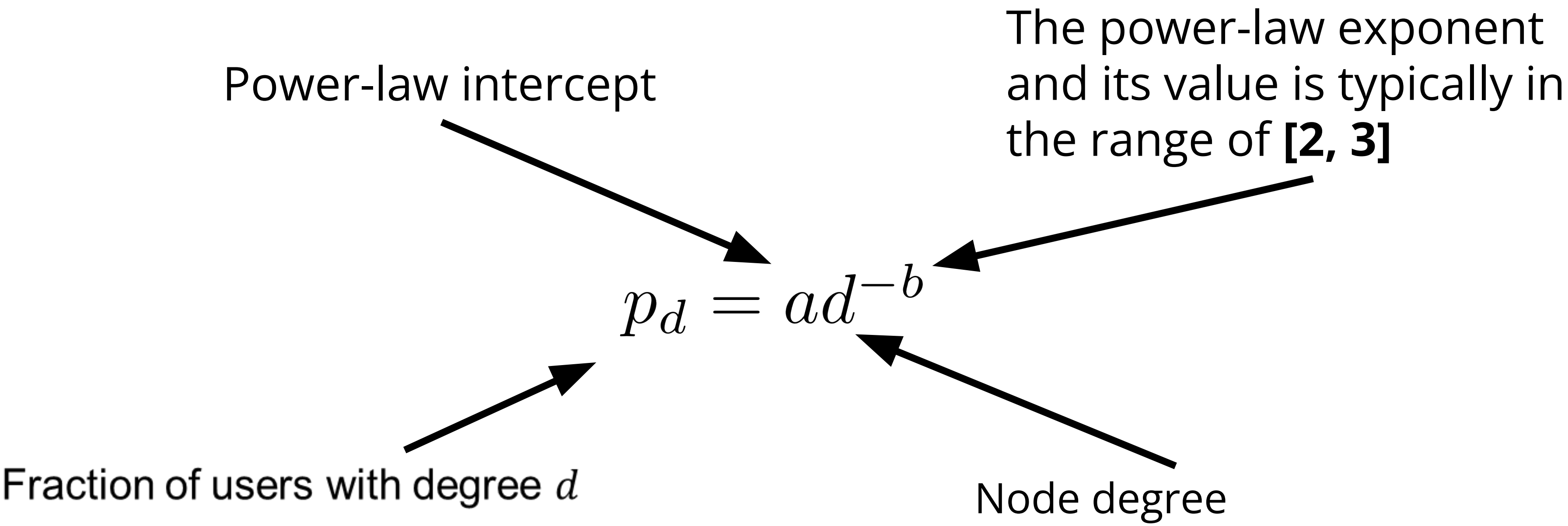
Friendships:

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

(Degree Distribution)

Power Law Distribution

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



Power-law intercept

The power-law exponent and its value is typically in the range of **[2, 3]**

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

Fraction of users with degree d

Node degree

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

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 $\mathbf{1/k^2}$

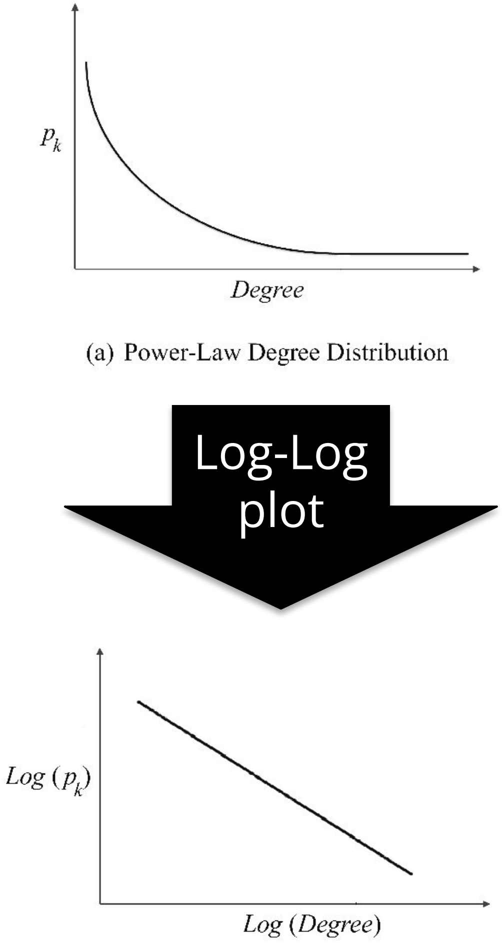
Power Law Distribution

- 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

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

Power-law Distribution: A 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!

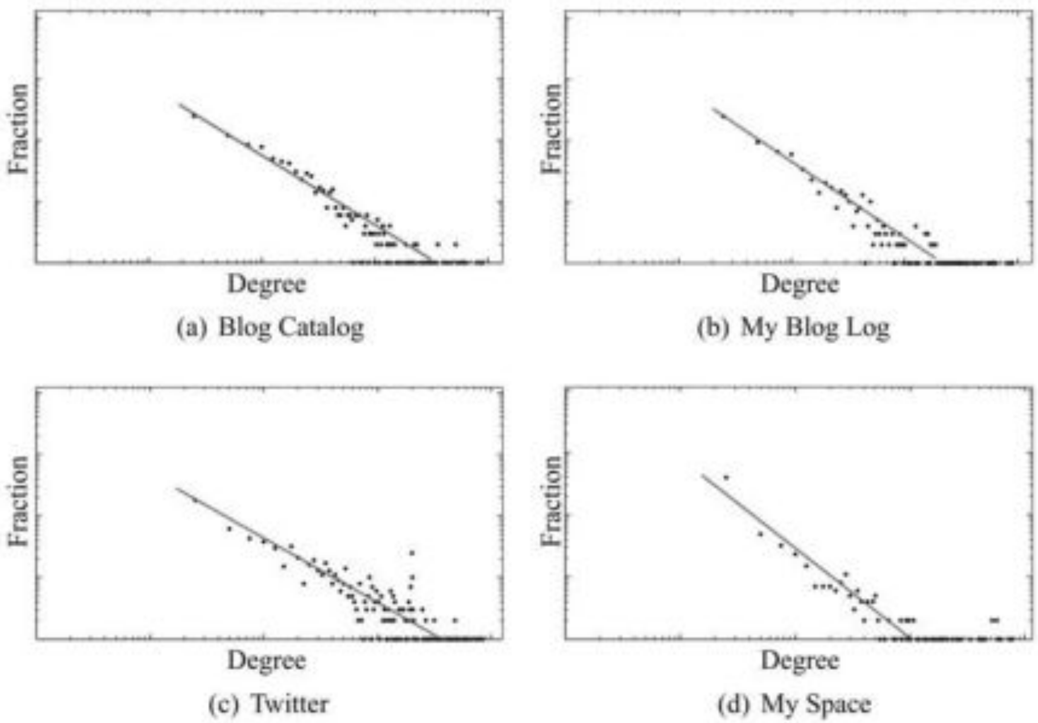
1. Other distributions could also exhibit this pattern
2. The results [estimations for parameters] can be biased and incorrect

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 power-law degree distribution are often called **scale-free** networks



(a) Blog Catalog

(b) My Blog Log

(c) Twitter

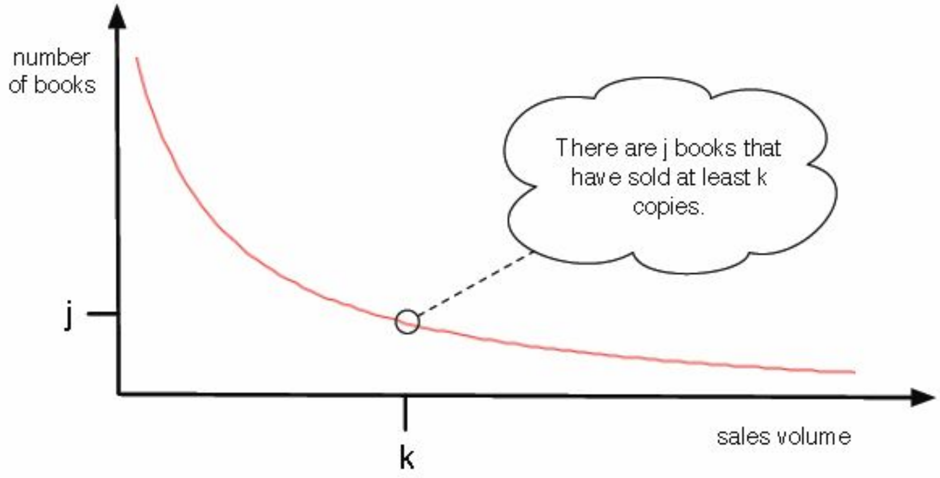
(d) My Space

The Long Tail

- In a company, 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, can be very significant.

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

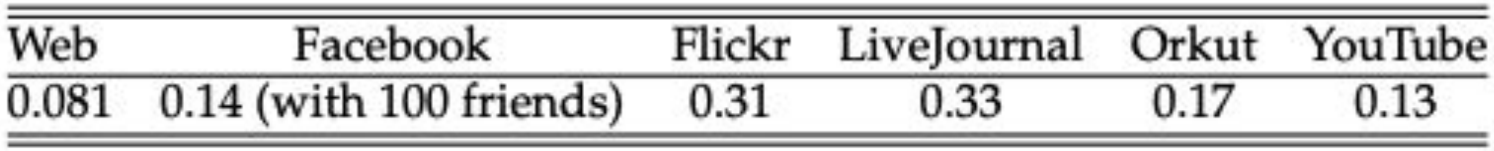


Clustering Coefficient

Clustering Coefficient

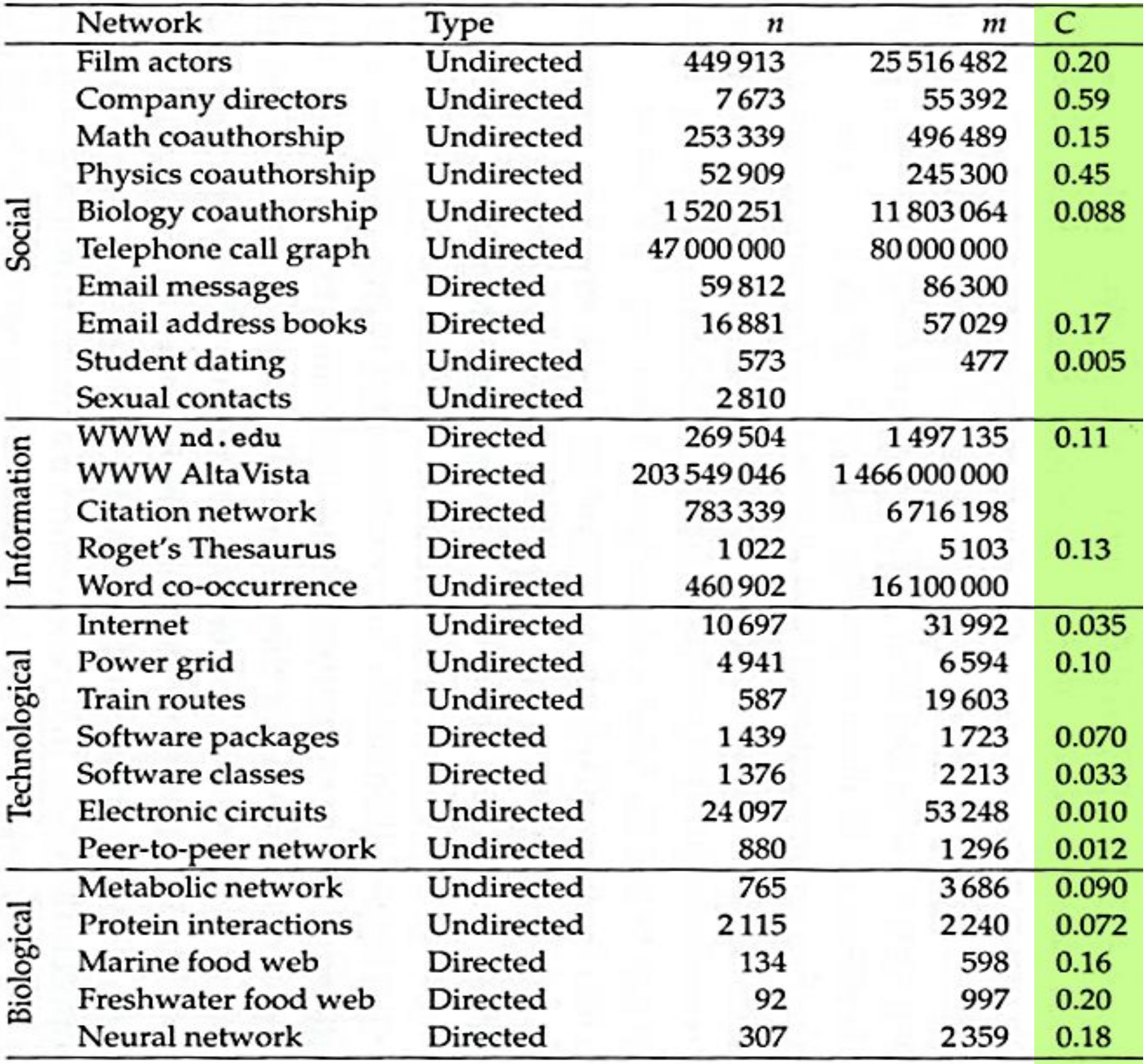
- In real-world networks, friendships are highly transitive, i.e., friends of an individual are often friends with one another
  - These friendships form triads -> high average [local] clustering coefficient

- In May 2011, Facebook had an average clustering coefficient of 0.5 for individuals who had 2 friends.



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



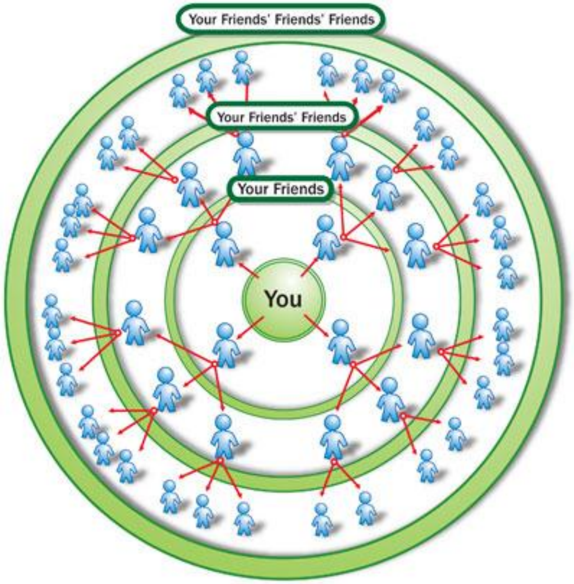
	Network	Type	n	m	C
Social	Film actors	Undirected	449 913	25 516 482	0.20
	Company directors	Undirected	7 673	55 392	0.59
	Math coauthorship	Undirected	253 339	496 489	0.15
	Physics coauthorship	Undirected	52 909	245 300	0.45
	Biology coauthorship	Undirected	1 520 251	11 803 064	0.088
	Telephone call 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?

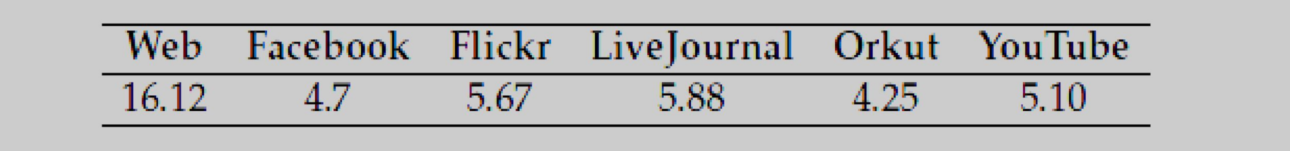
- A rumor (or a disease) spreads over a social network
  - assuming all nodes pass it immediately to all of their neighbors



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?

The Average Shortest Path

- In real-world networks, any two members of the network are usually connected via short paths. In other words, the average path length is small
  - Six degrees of separation:
    - **Stanley Milgram** in the well-known small-world experiment conducted in the 1960's conjectured that people around the world are connected to one another via a path of at most 6 individuals
  - Four degrees of separation:
    - **Lars Backstrom et al.** in May 2011, the average path length between individuals in the Facebook graph was 4.7. (4.3 for individuals in the US)



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

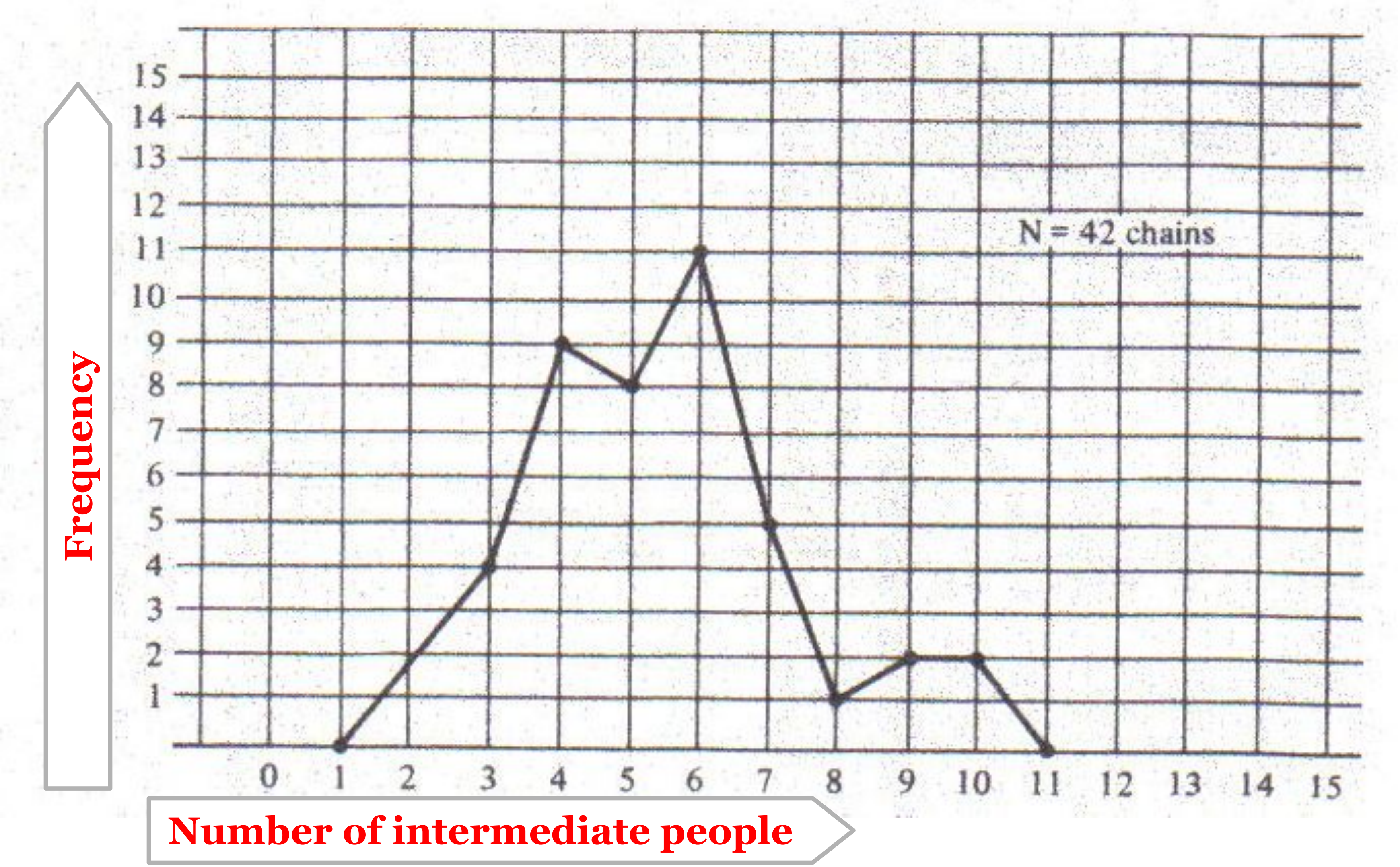
Stanley Milgram’s Experiments

- 296 random people from Nebraska (196) and Boston (100) were asked to send a letter (via intermediaries) to a stock broker in Boston
- S/he could only send to someone with whom they were on a first-name basis

Stanley Milgram (1933-1984)

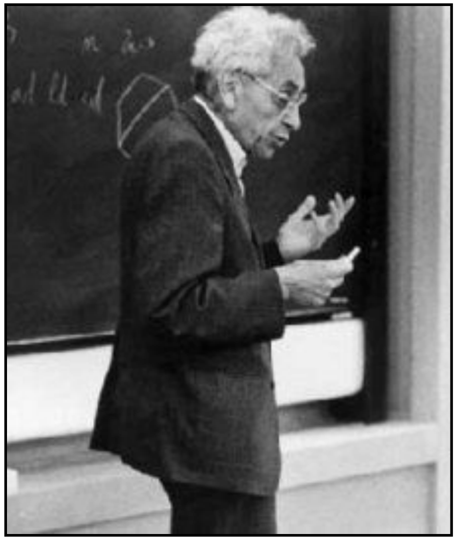
Among the letters that found the target (64), the average number of links was around **six**.

Small-world Effect, Milgram’s Experiment



Average Number of Intermediate people is 5.595

Erdős Number (or Bacon Number?)



Paul Erdős (1913-1996)

Number of links required to connect scholars to Erdős, via co-authorship of papers

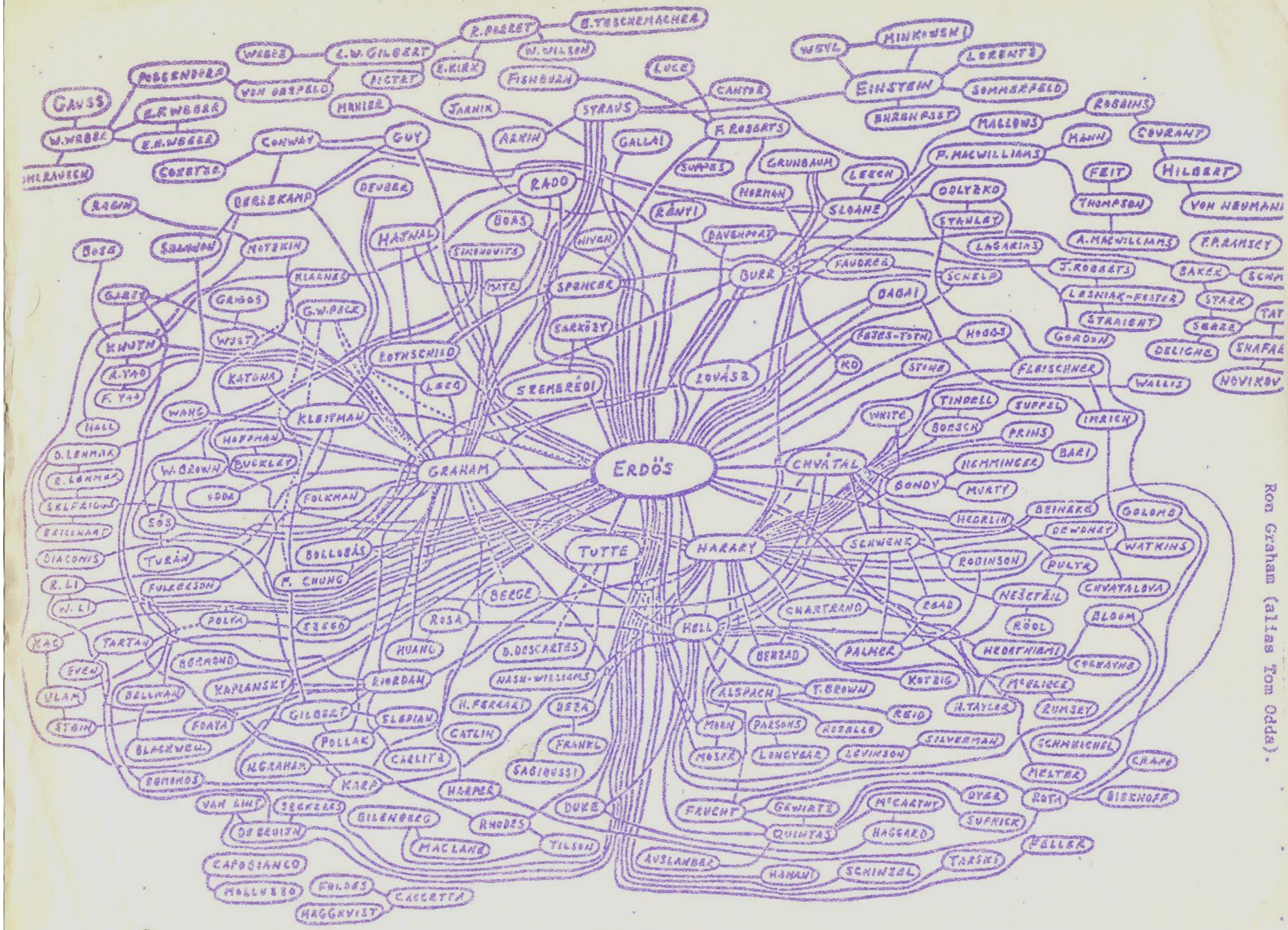
Erdős wrote 1500+ papers with 507 co-authors.

Jerry Grossman's (Oakland Univ.) website allows mathematicians to compute their Erdos numbers: http://www.oakland.edu/enp/

Connecting path lengths, among mathematicians only:

- Avg is *4.65 and* Maximum is *13*

Watch his documentary "N is a number" on YouTube

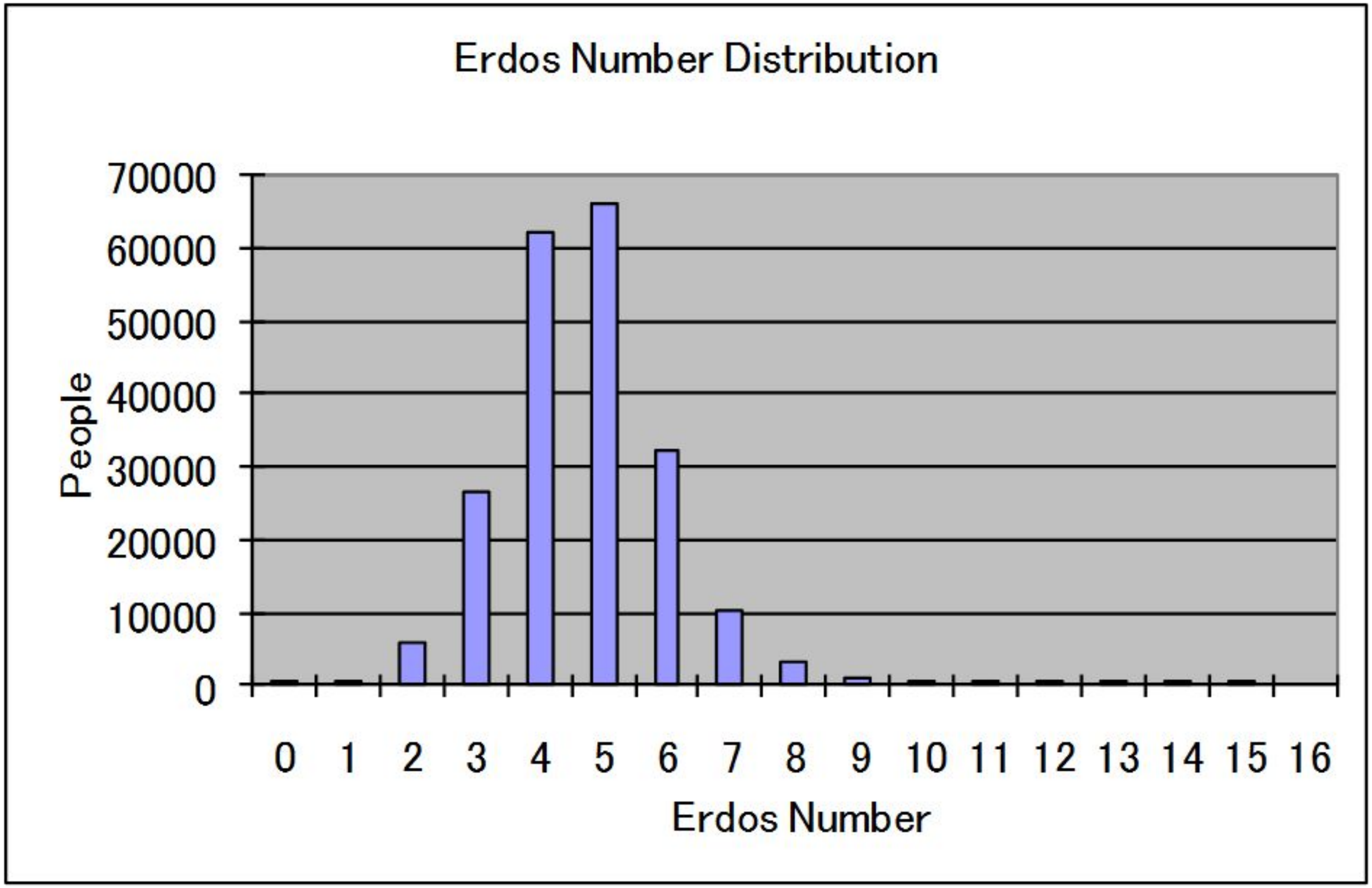


Ron Graham (alias Tom Odda).

Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

Erdos Number Project: http://www.oakland.edu/enp/index.html

Erdos Number Distribution



The median Erdös number is 5;
the mean is 4.69, and the standard deviation is 1.27.

The Average Shortest Path

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

Facebook

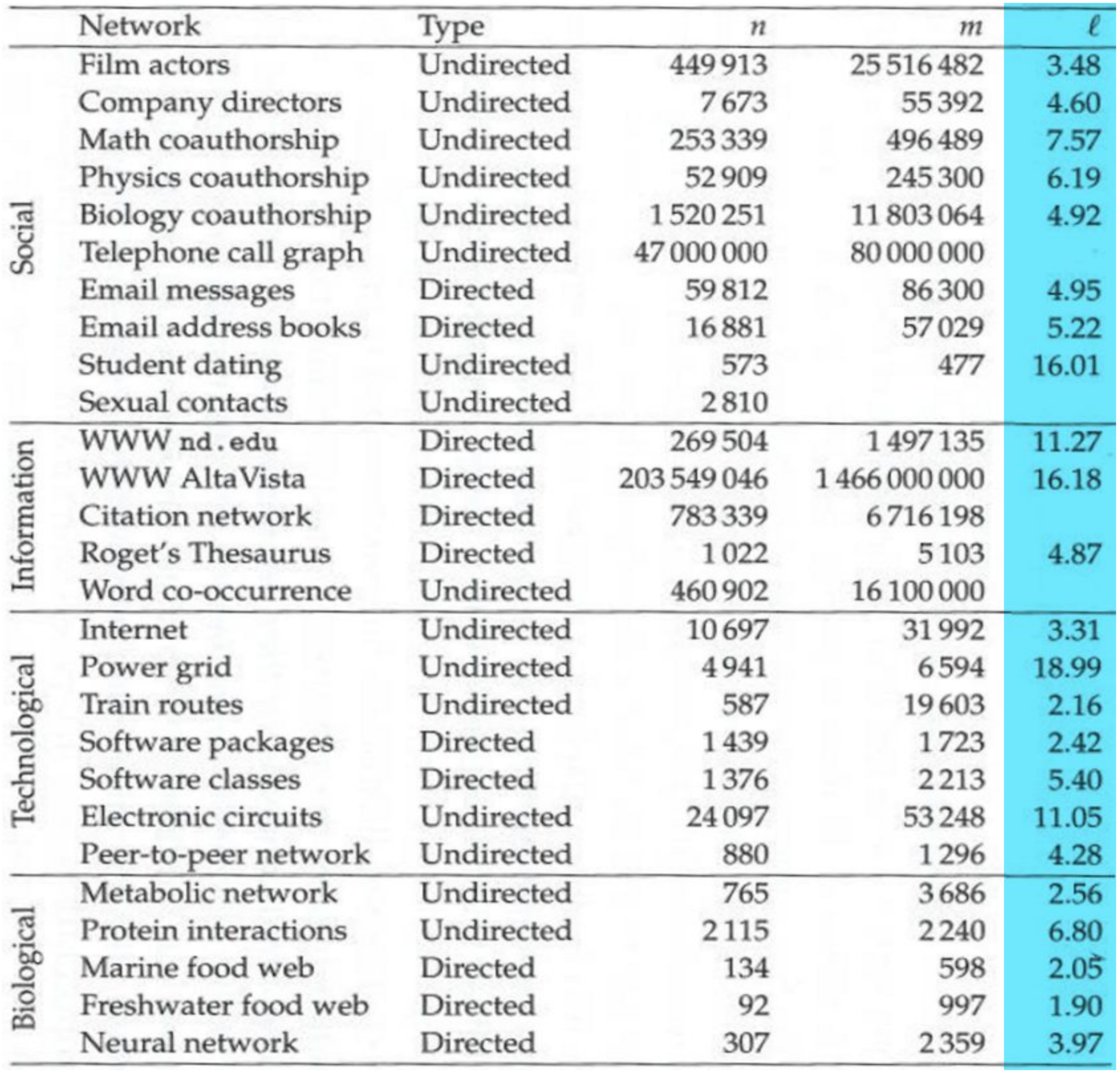
May 2011:

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

[Four degrees of separation]

The average path length is small

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

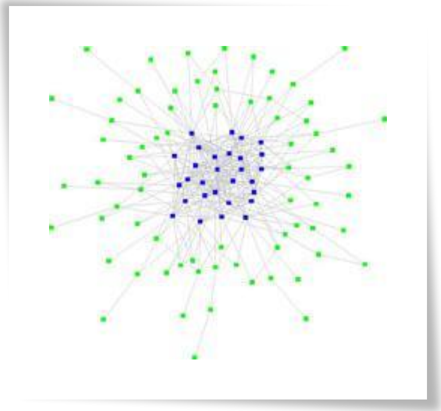
More Properties of Real-World Networks

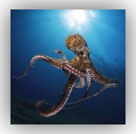
Friendship Paradox [Feld 1991]

- 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 start with the most basic assumption on how friendships are formed.

A Random Graph’s main 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 - $G(n, m)$

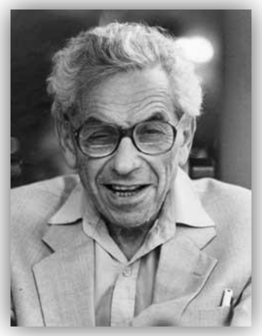
- Assume both number of nodes n and number of edges m are fixed.

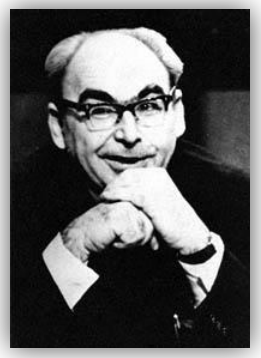
- Determine which m edges are selected from the set of possible edges

- 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 possible 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,

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

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

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