Network Models

CS 579 Online Social Network Analysis

Dr. Cindy Hood
11/18/25

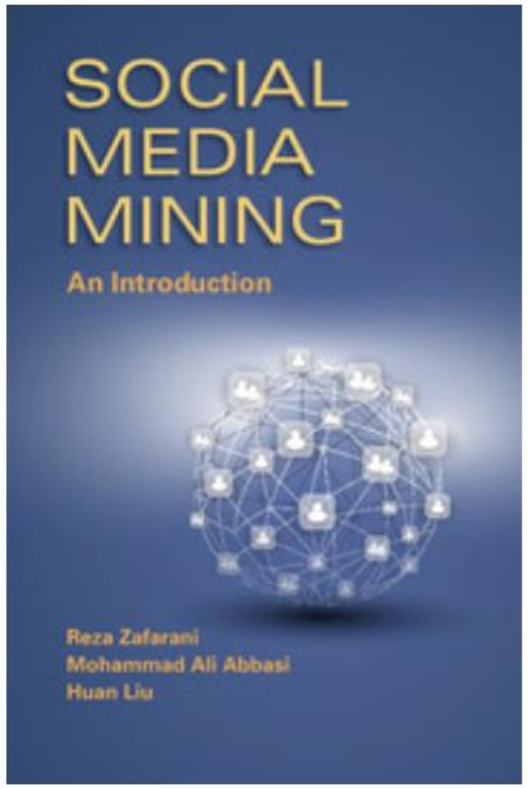
Remaining Exam and Deliverables

- Final project progress report ~~and~~ video posted
  - Due 11/21
- HW #5 posted
  - Problems that will help you prepare for Exam 2
  - Due 11/24 (No late days)
  - Social Media Mining book posted on Canvas
- Exam 2
  - Cumulative
  - 12/2
- Final project poster presentation/video (online students)
  - 12/4
- Final project report
  - Week of 12/8
    - Specific date tbd

Final Project Progress Report

- You will create a presentation about your progress on the final project to date and then submit a video of you/your team presenting the slides. Each student should speak and the speaker should be shown in the video while they are presenting. The video should be 2-4 minutes long.
- The presentation should include:
- Intro slide with the title of the project and student name(s)
- A summary of the project components clearly illustrating pieces that are being reused from HW 4 along with other components and how the components fit together.
- A plan for completing the project highlighting the starting point (what was done for HW4) and the steps to be completed along with discussion on progress made to date.
- You will submit a link to your video presentation. Please be sure that it is accessible to Prof Hood and the TAs. One submission per team.

References



http://www.socialmediamining.info

Some additional resources

- Myatt and Johnson (2014), *Making Sense of Data I*, 2nd Edition, Wiley, ISBN: 978-1-118-40741-7
  - https://onlinelibrary.wiley.com/doi/epdf/10.1002/9781118422007
- *Speech and Language Processing*, Dan Jurafsky and James H. Martin, https://web.stanford.edu/~jurafsky/slp3/
- *An Introduction to Statistical Learning*, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor (Python version only)
  - https://www.statlearning.com
- Good YouTube channel for getting intuition
  - https://www.youtube.com/@statquest/videos
- Networks, Crowds, and Markets: Reasoning About a Highly Connected World by David Easley and Jon Kleinberg.
  - http://www.cs.cornell.edu/home/kleinber/networks-book/

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?

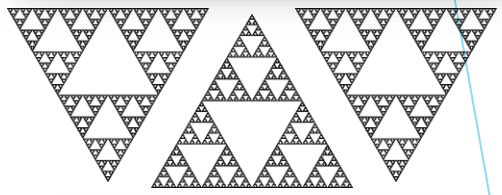
Design models that generate graphs

- 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

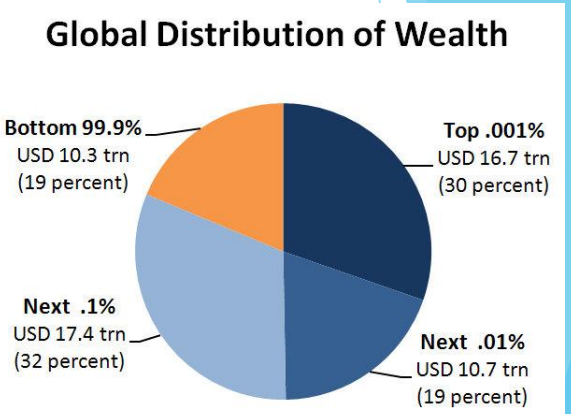
Small Average Path Length

Degree Distribution

Distributions

Wealth Distribution:

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



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.

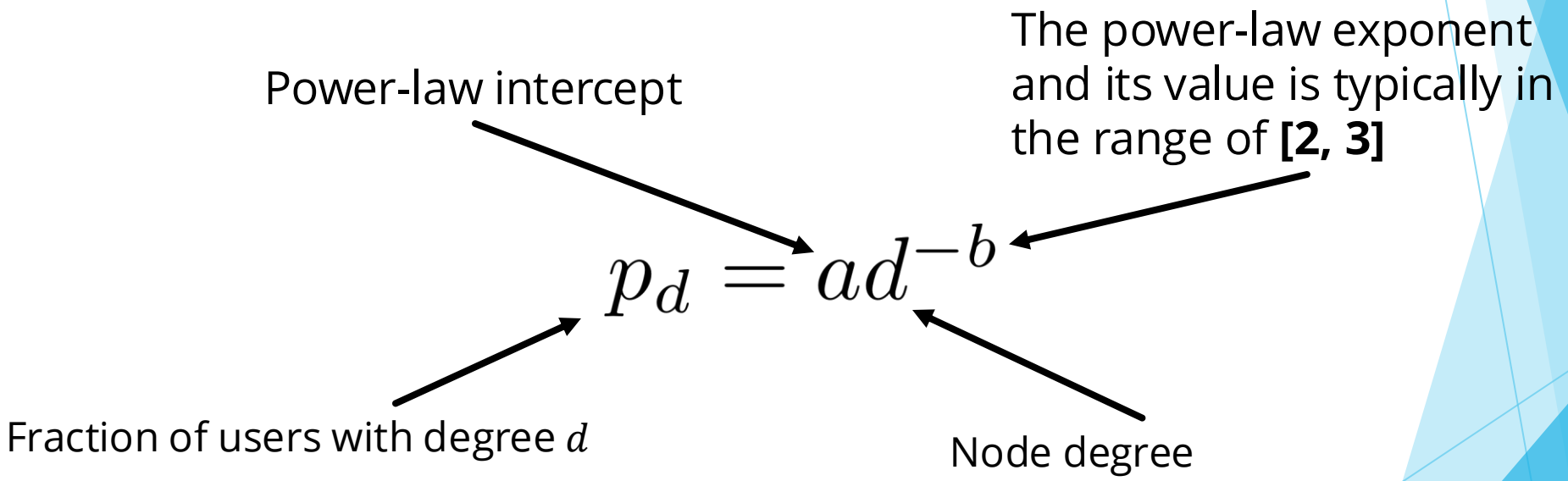
Friendships:

- 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**



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

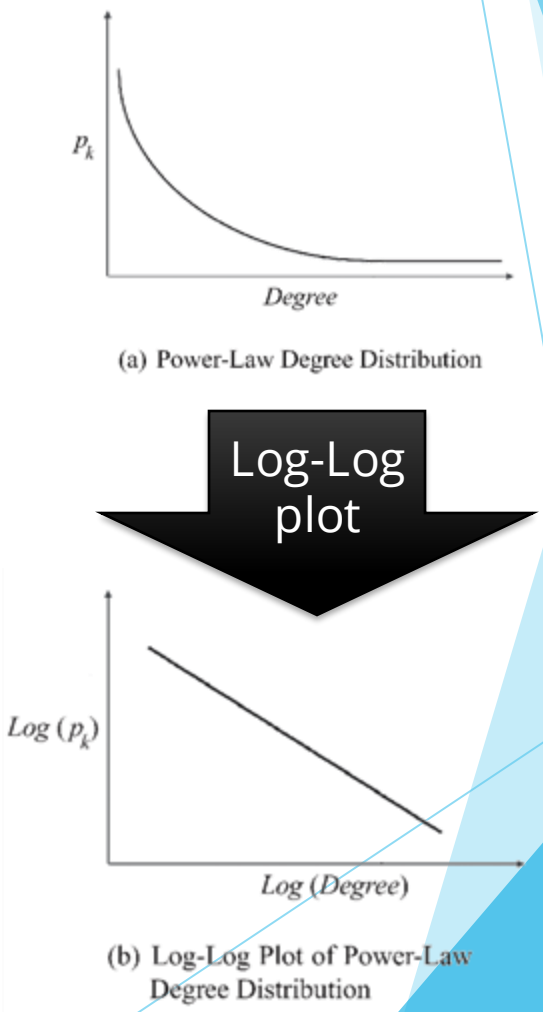
Power-Law Distribution: Examples

- **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



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!

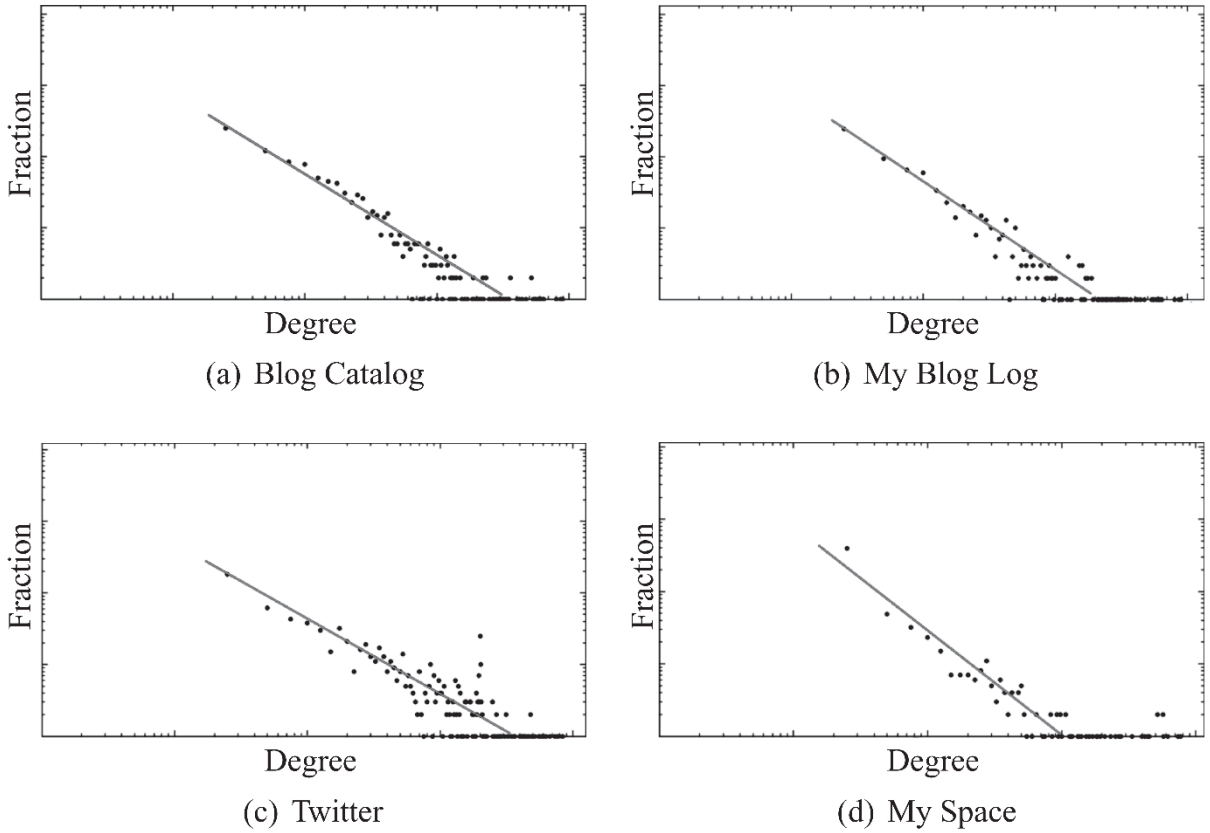
1. Other distributions could also exhibit this pattern
2. The results [estimations for parameters] can be biased and incorrect

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



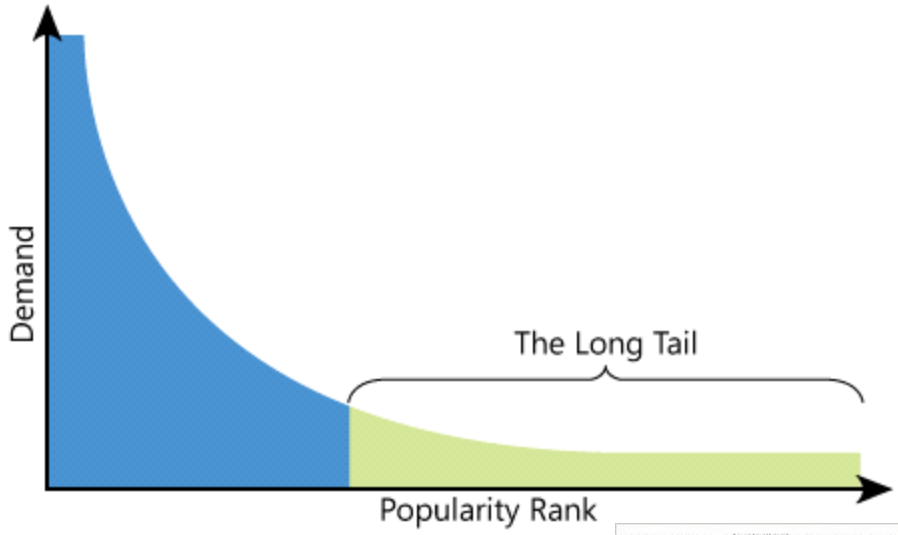
The tail of the power-law distribution is long!

The Looooooong Tail

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

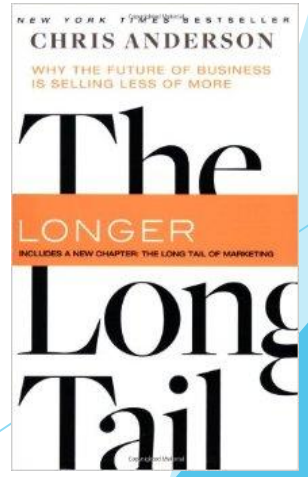
OR

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



Article | Open access | Published: 04 March 2019

Scale-free networks are rare

Anna D. Broido & Aaron Clauset

Nature Communications **10**, Article number: 1017 (2019) | Cite this article

Abstract

Real-world networks are often claimed to be scale free, meaning that the fraction of nodes with degree k follows a power law $k^{-\alpha}$, a pattern with broad implications for the structure and dynamics of complex systems. However, the universality of scale-free networks remains controversial. Here, we organize different definitions of scale-free networks and construct a severe test of their empirical prevalence using state-of-the-art statistical tools applied to nearly 1000 social, biological, technological, transportation, and information networks. Across these networks, we find robust evidence that strongly scale-free structure is empirically rare, while for most networks, log-normal distributions fit the data as well or better than power laws. Furthermore, social networks are at best weakly scale free, while a handful of technological and biological networks appear strongly scale free. These findings highlight the structural diversity of real-world networks and the need for new theoretical explanations of these non-scale-free patterns.

https://www.nature.com/articles/s41467-019-08746-5#data-availability

https://www.nature.com/articles/s41467-019-08746-5#data-availability

Clustering Coefficient

Clustering Coefficient

- Captures how connected your friends are
- For each vertex, it is a measure of the density of the 1.5-degree egocentric network
- Value between 0 and 1
- The higher the value, the more an individual's friends know each other

Clustering Coefficient

- In real-world networks, friendships are highly transitive
  - Friends of a user are often friends with one another
  - These friendships form triads
  - High average [local] clustering coefficient

Facebook

May 2011:

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

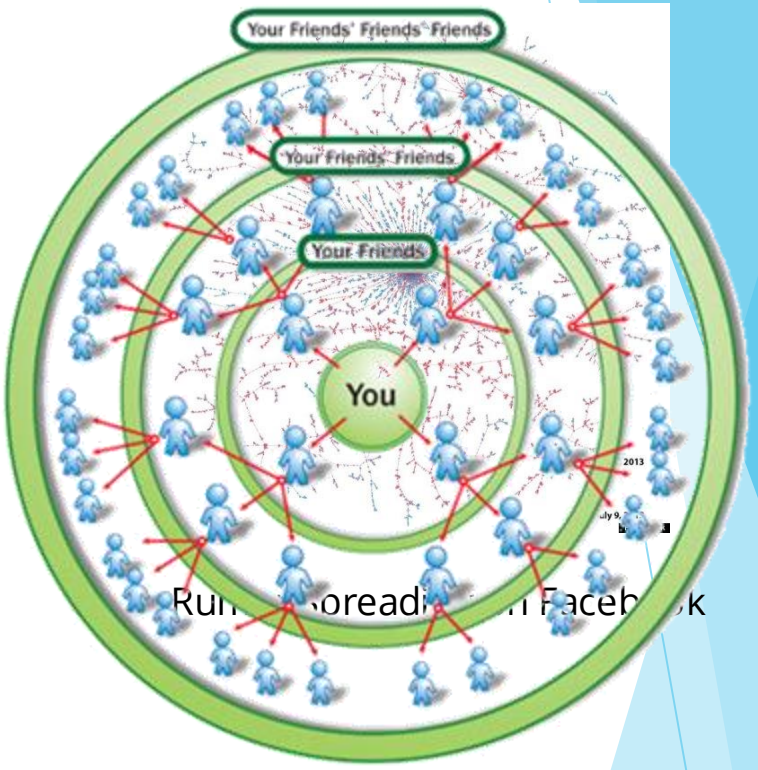
Source: M. E. J Newman

Average Path Length

How Small is the World?

A rumor 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?

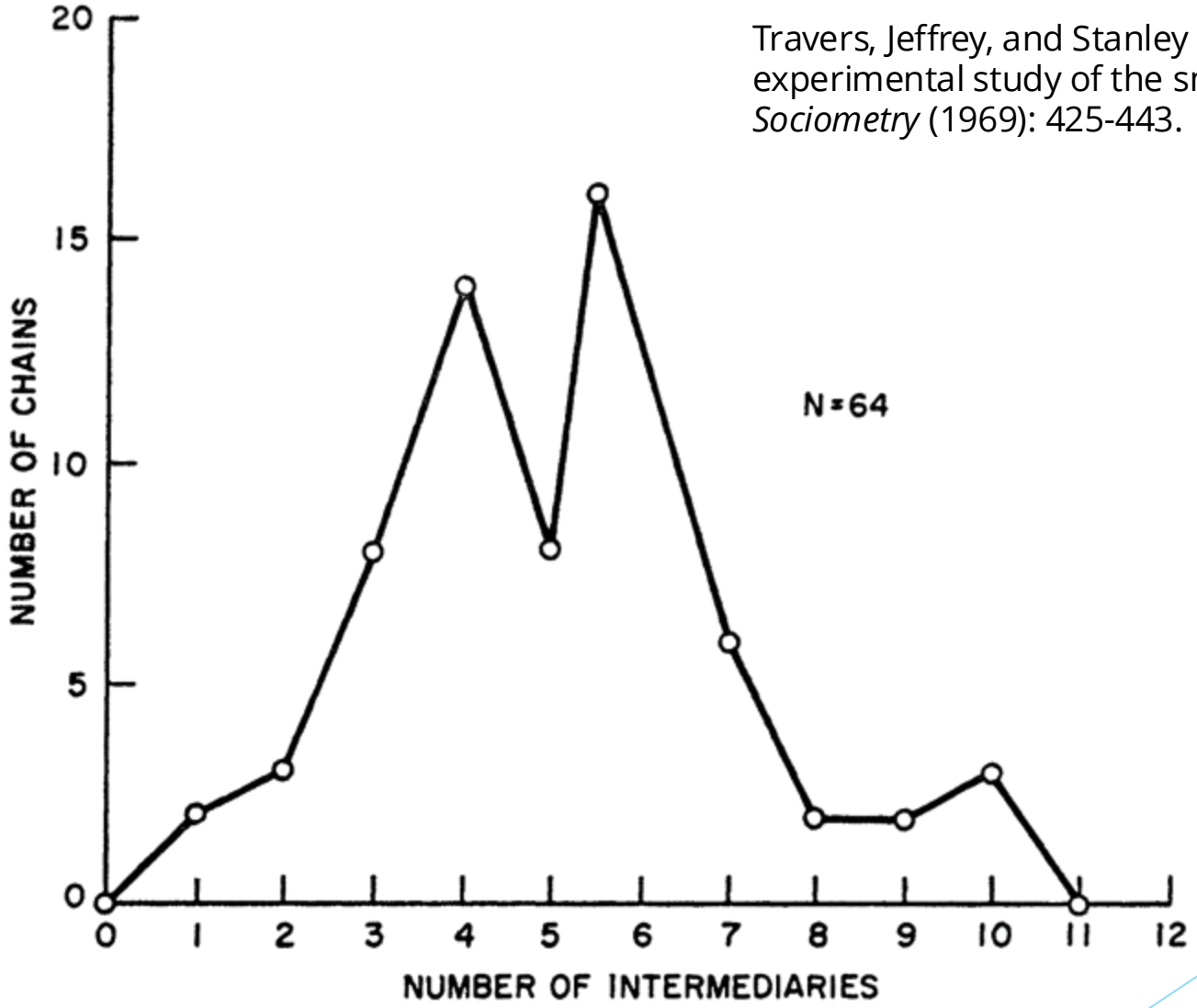
Milgram's Experiment

- 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

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

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