LECTURE 4: SMALL-WORLD PHENOMENA

COMP4641: Social Information Network Analysis and Engineering

Friday February 13rd 2015

Small world: a simplistic argument

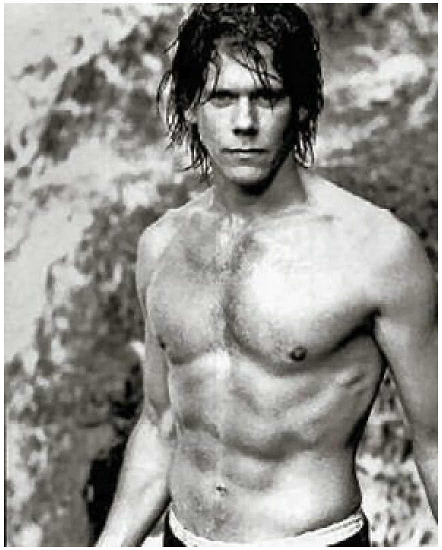
- How many people would you recognize by name?
  - '67 M. Gurevitch (MIT): about 500
- Roughly, how many are socially related to you?

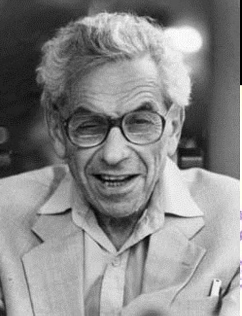
	how close to you?	Compares to	%. US pop.
500	direct acquaintance	C.S. dept	0.00017%
250,000	share an acquaintance with you	Harlem district	0.083%
125m	share an acquaintance with a friend of yours	Northeast + Midwest	42%

Six Degrees of Kevin Bacon

Origins of a small-world idea:

- **The Bacon number:**
  - Create a network of Hollywood actors
  - Connect two actors if they co-appeared in the movie
  - Bacon number: number of steps to Kevin Bacon
- As of Dec 2007, the highest (finite) Bacon number reported is 8
- Only approx. 12% of all actors cannot be linked to Bacon





Erdös numbers are small!

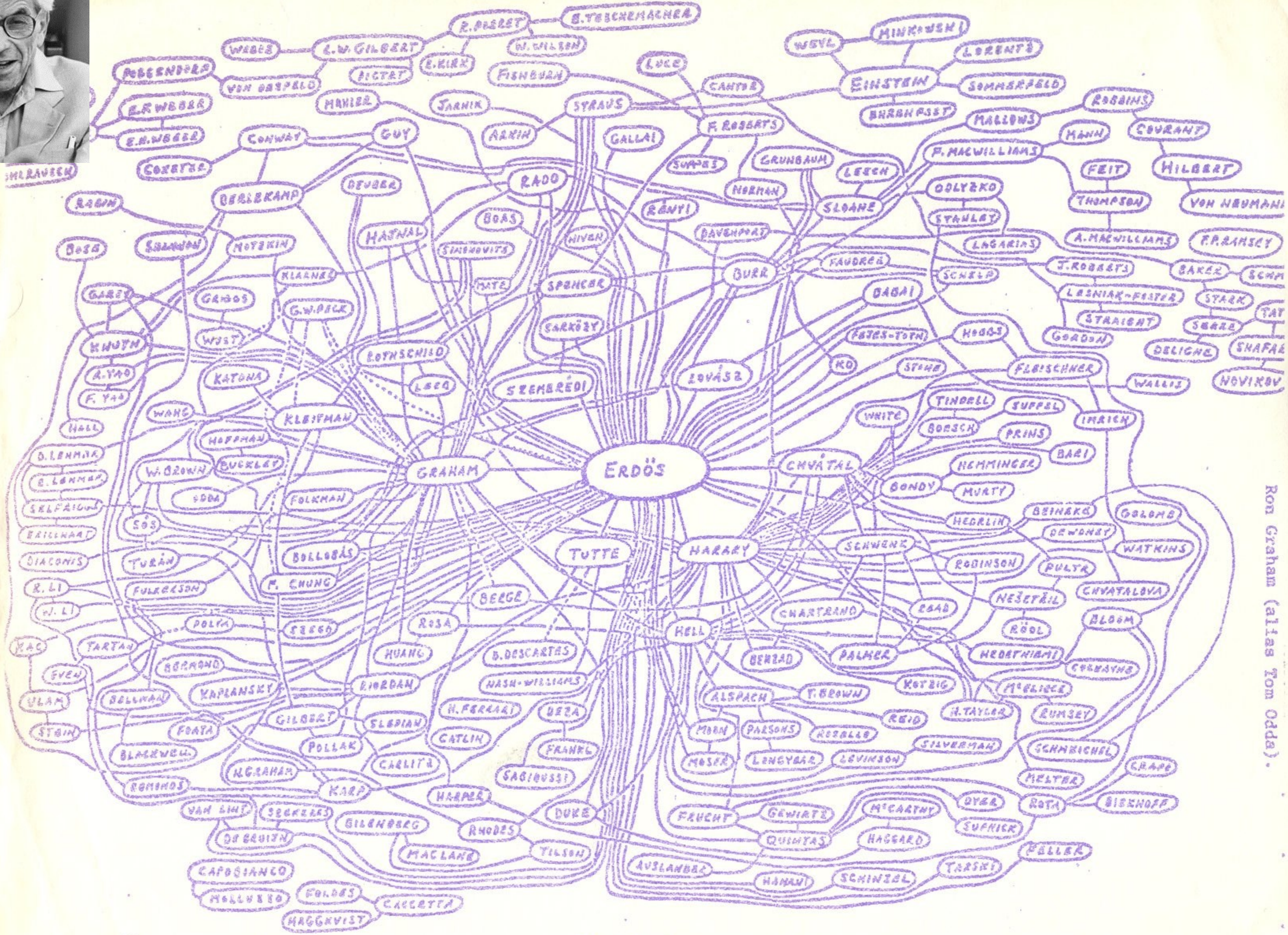
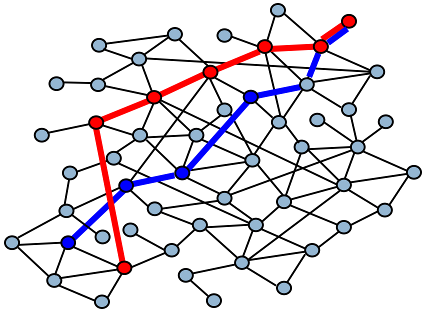


Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

The Small-World Experiment

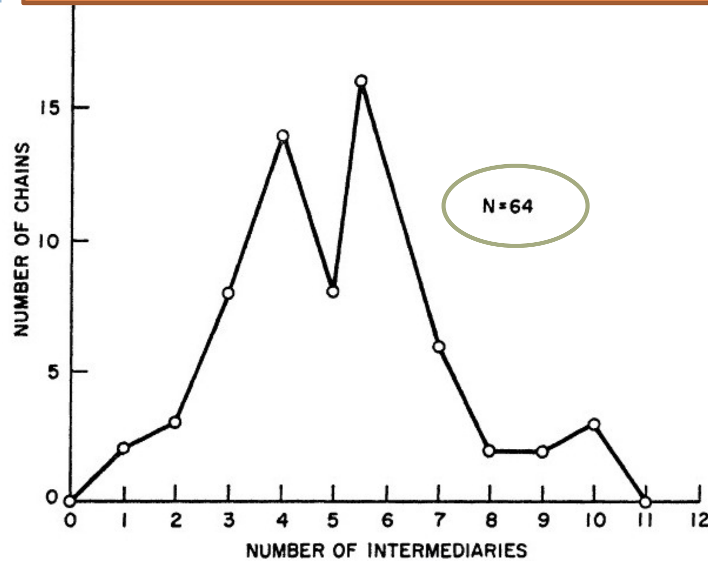
- **What is the typical shortest path length between any two people?**
  - Experiment on the global friendship network
    - Can't measure, need to probe explicitly
- **Small-world experiment** [Milgram '67]
  - Picked 300 people in Omaha, Nebraska and Wichita, Kansas
  - Ask them to get a letter to a stock-broker in Boston by passing it through friends
- **How many steps did it take?**



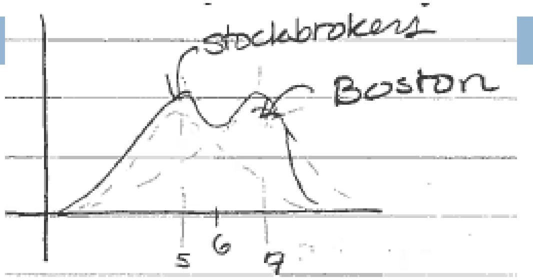
The Small-World Experiment

- 64 chains completed:
  (i.e., 64 letters reached the target)
  - It took 6.2 steps on the average, thus **"6 degrees of separation"**
- Further observations:
  - People what owned stock had shortest paths to the stockbroker than random people: 5.4 vs. 5.7
  - People from the Boston area have even closer paths: 4.4

Milgram's small world experiment



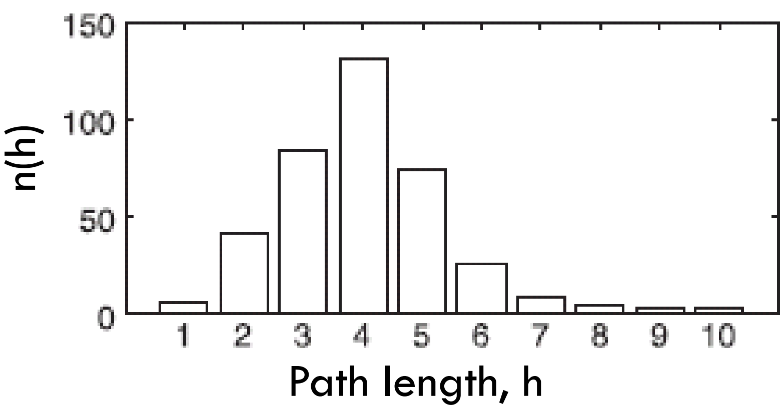
Milgram: Further Observations



- Boston vs. occupation networks:
- Criticism:
  - **Funneling:**
    - 31 of 64 chains passed through 1 of 3 people ass their final step → 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

Columbia Small-World Study

- **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

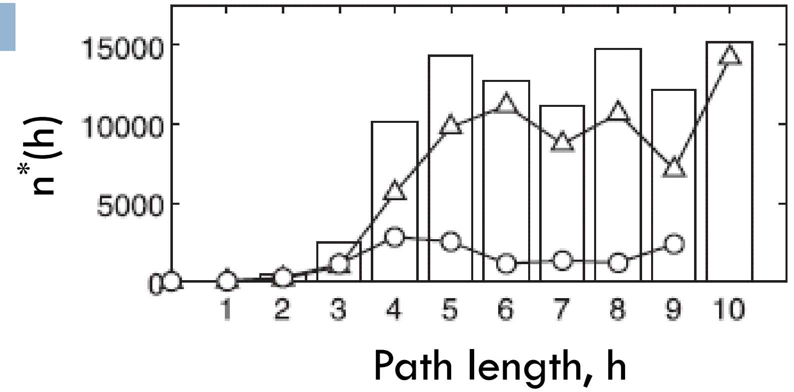
Correction factor:

$$n^*(h) = \frac{n(h)}{\prod_{i=0}^{h-1}(1-r_i)}$$

r_i …. drop-out rate at hop i

Small-World in Email Study

- After the correction:
  - **Typical path length h = 7**

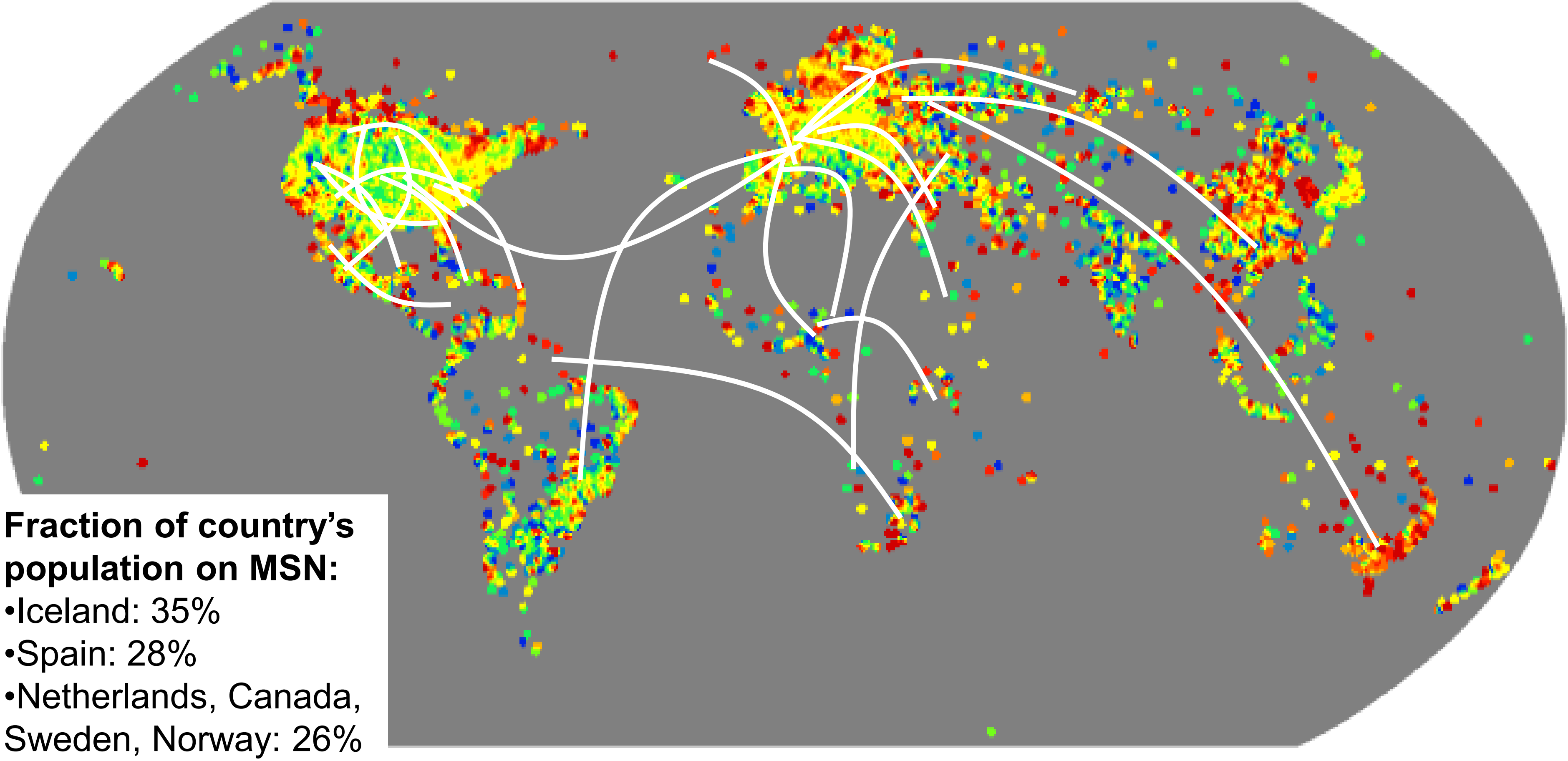


- Some not well understood phenomena in social networks:
  - **Funneling effect:** Some target's friends are more likely to be the final step.
    - Conjecture: High reputation/authority
  - **Effects of target's characteristics:** Structurally why are high-status target easier to find
    - Conjecture: Core-periphery net structure

The MSN Messenger

- **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

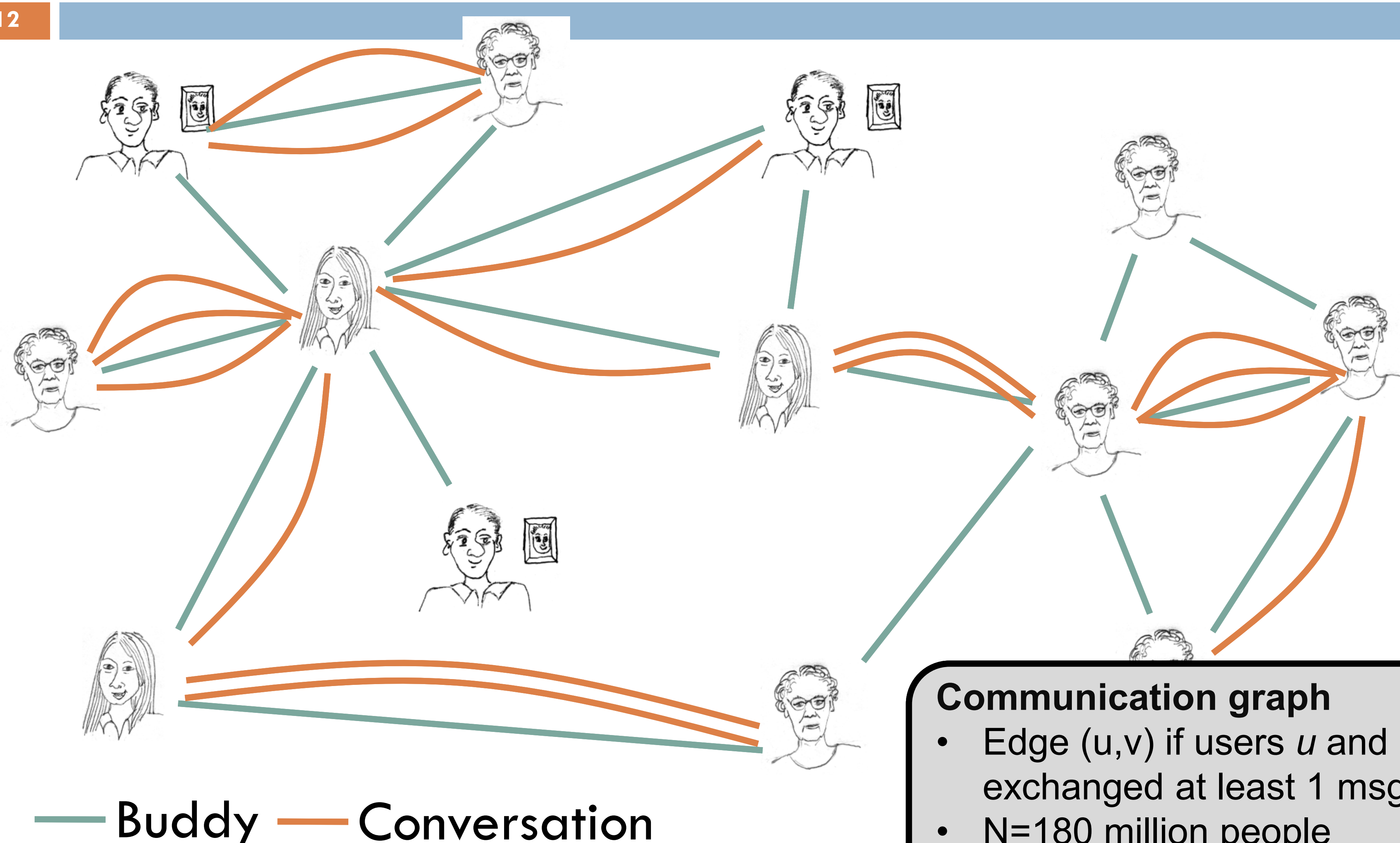
Messaging as a Network



Fraction of country's population on MSN:

- Iceland: 35%
- Spain: 28%
- Netherlands, Canada, Sweden, Norway: 26%
- France, UK: 18%
- USA, Brazil: 8%

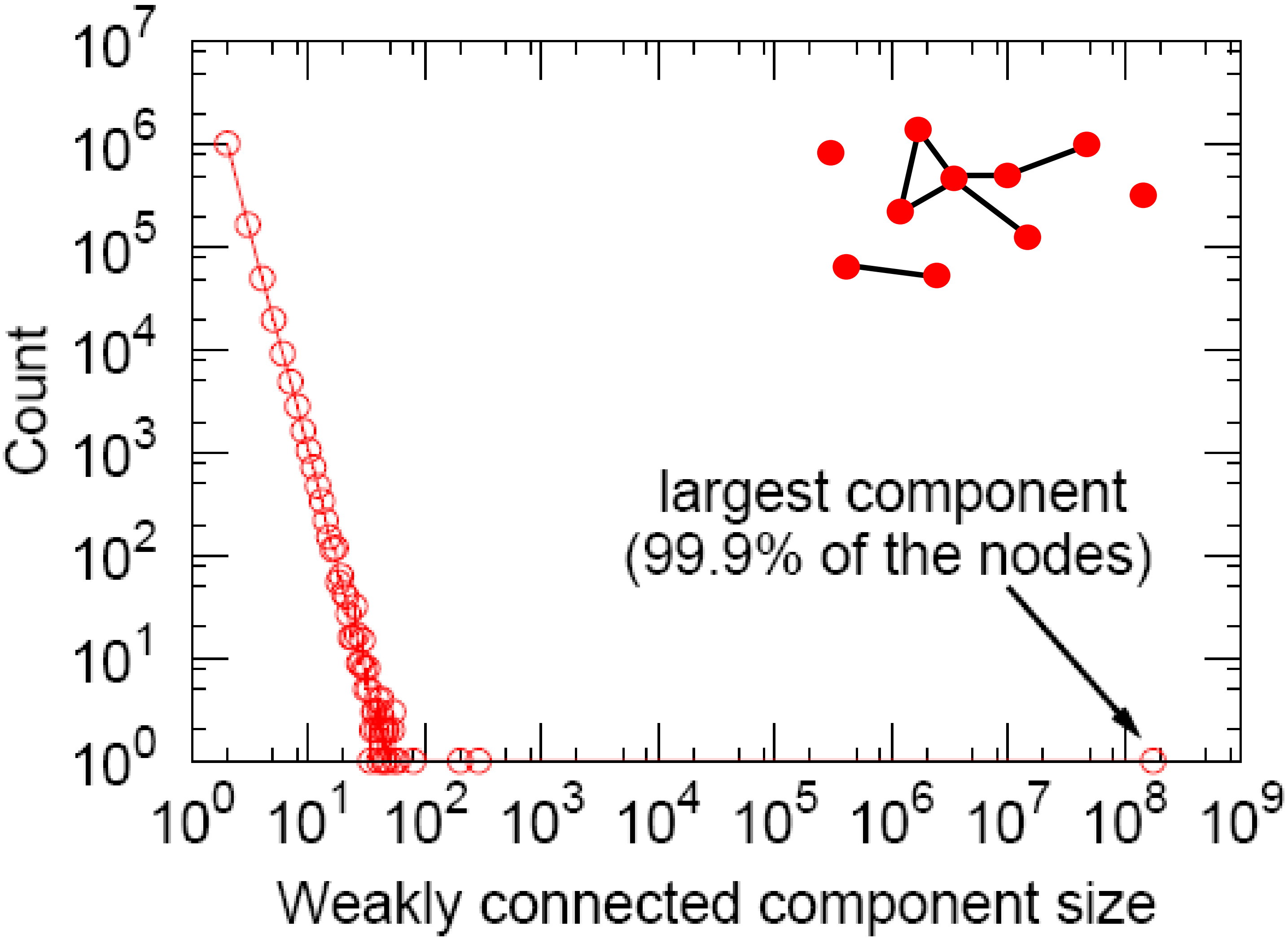
Messaging as a Network



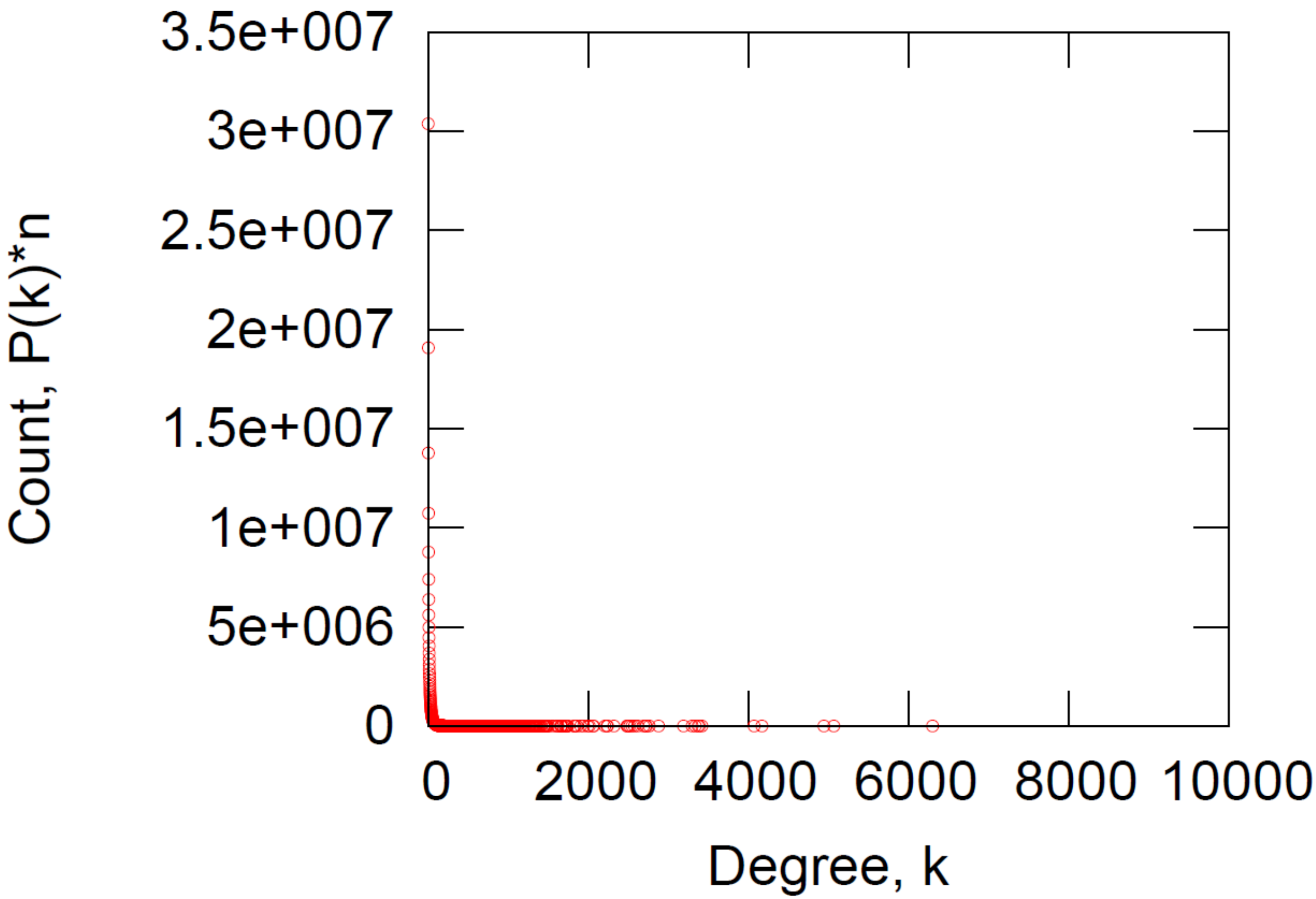
Communication graph

- Edge (u,v) if users *u* and *v* exchanged at least 1 msg
- N=180 million people
- E=1.3 billion edges

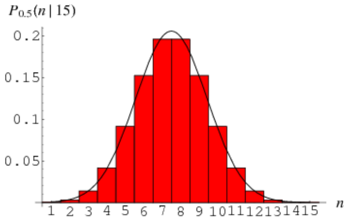
MSN Network: Connectivity



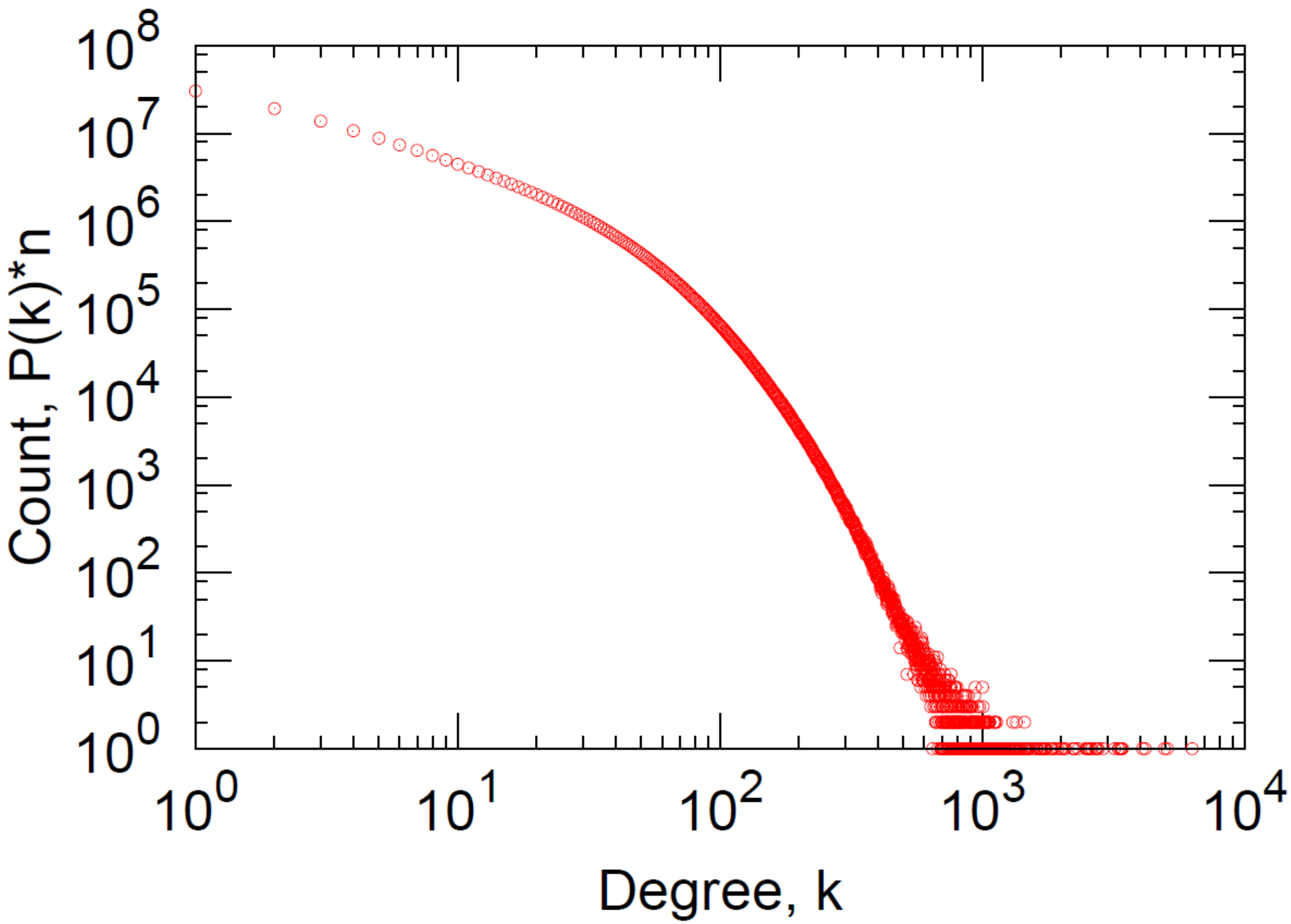
MSN: Degree Distribution



Note: Degree distribution of the MSN looks nothing like the G_{np}:

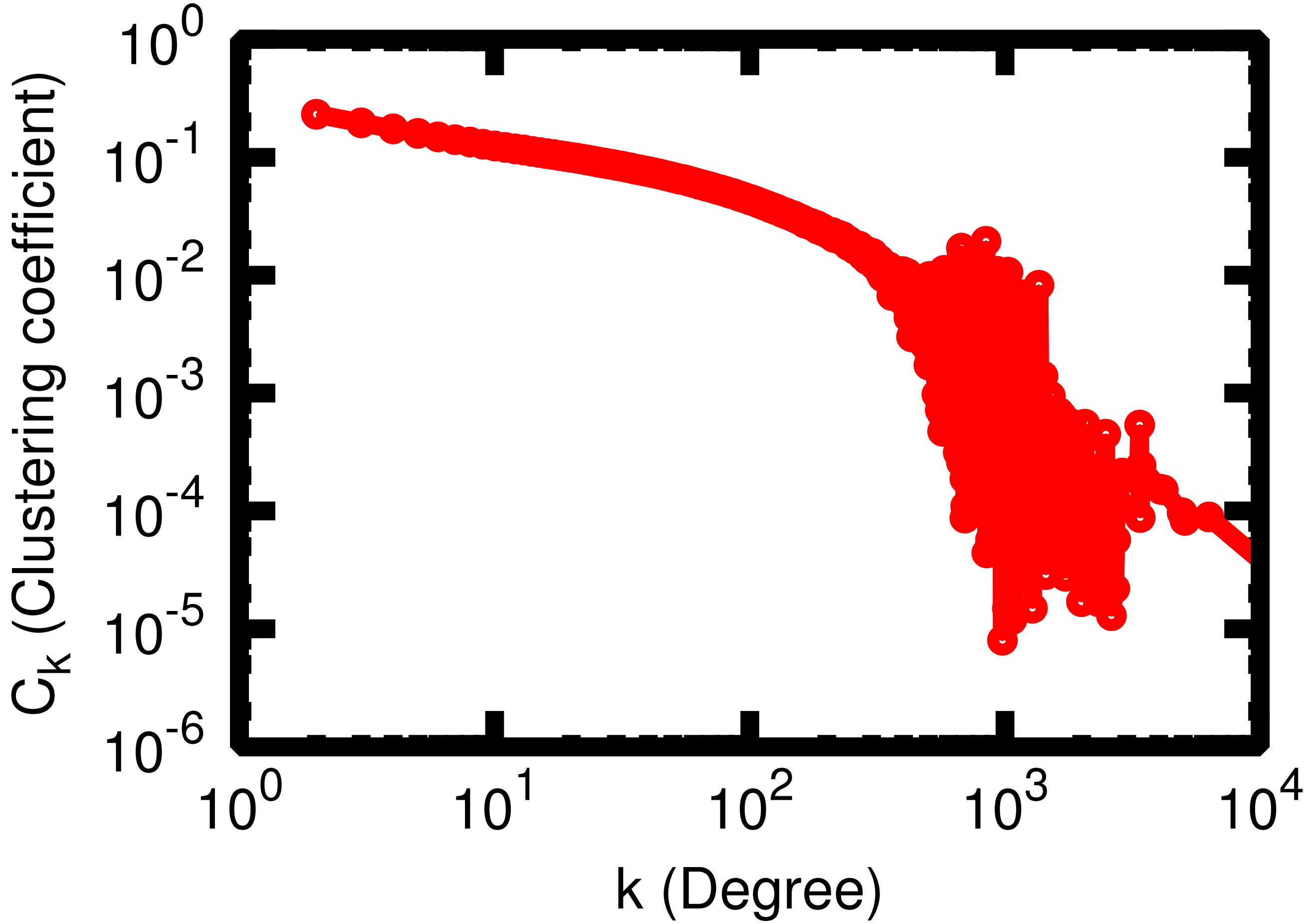


MSN: Log-Log Degree Distribution



We plot the same data as on the previous slide, just the axes are now logarithmic.

MSN: Clustering

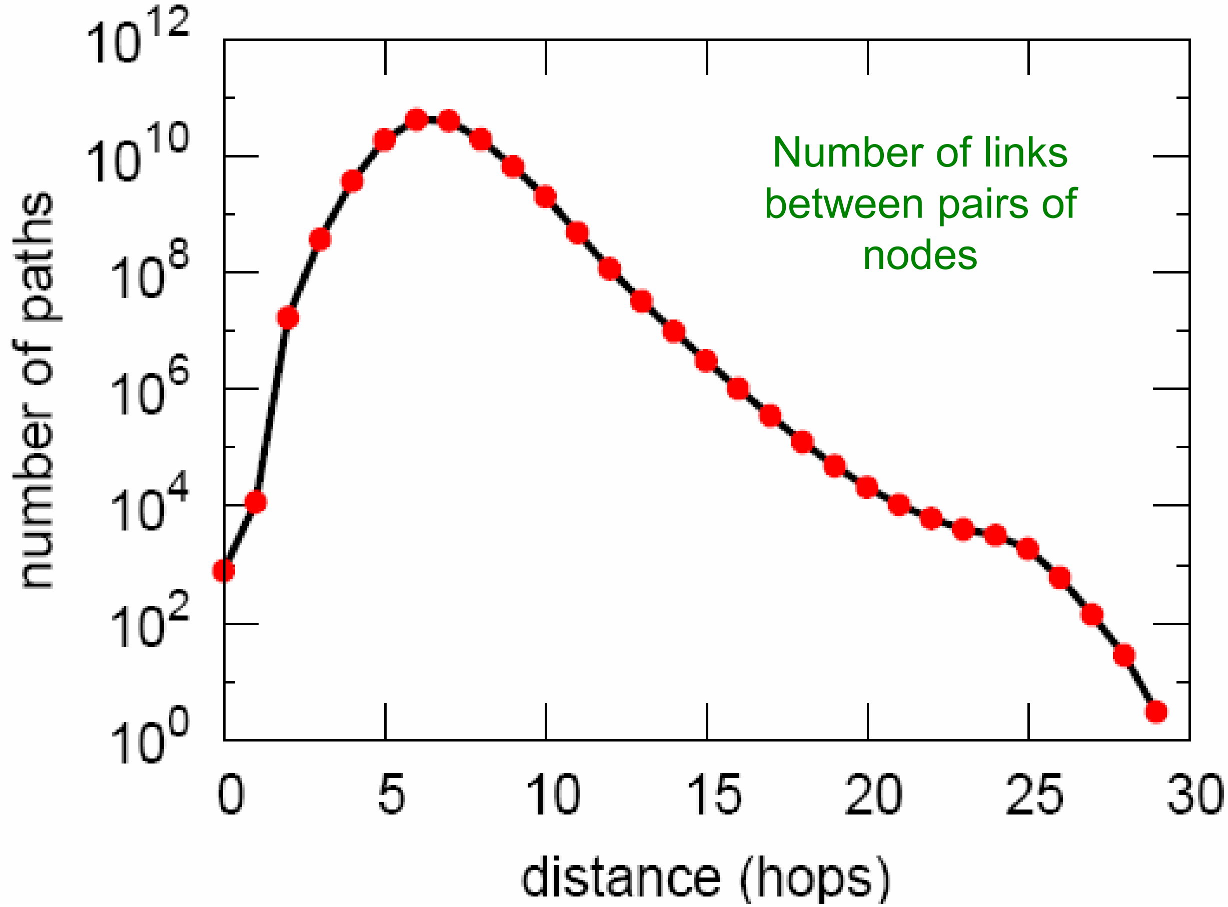


Avg. clustering of the MSN:
$C = 0.1140$

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

C_k: average C_i of nodes *i* of degree k: $$C_k = \frac{1}{N_k} \sum_{i:k_i=k} C_i$$

MSN: Diameter



Avg. path length 6.6

90% of the people can be reached in < 8 hops

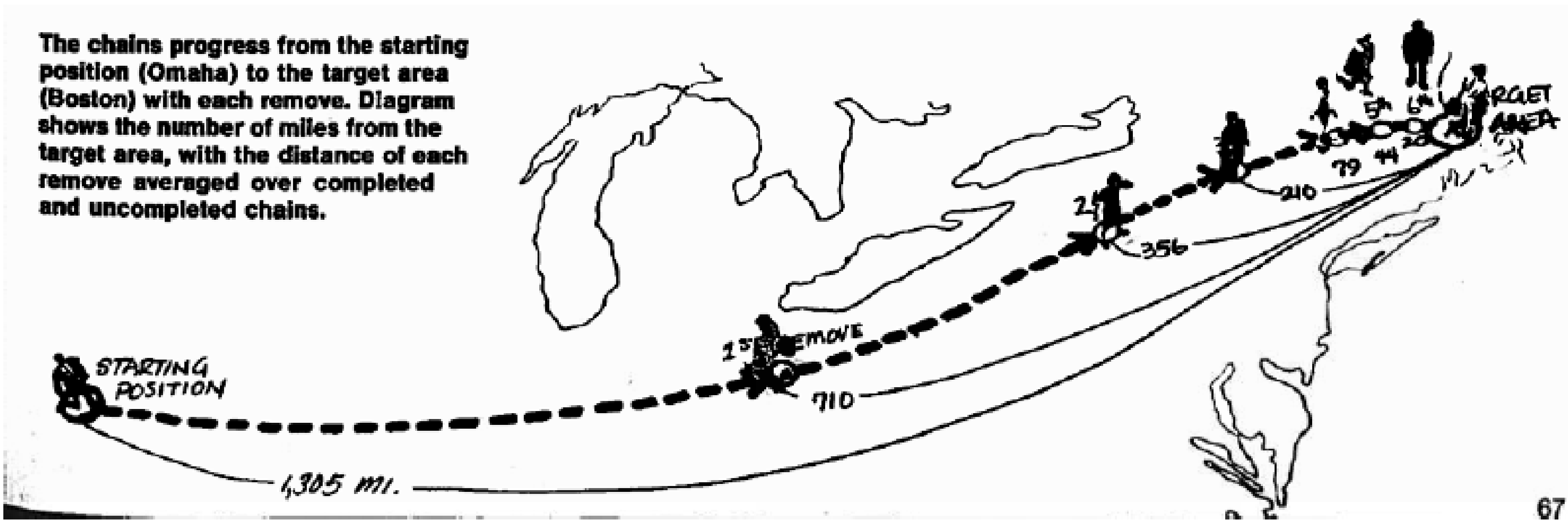
nodes as we do BFS out of a rand

Steps	#Nodes
0	1
1	10
2	78
3	3,96[illegible]
8	[illegible]
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

Table: use the expansion terminology: j and Sj.

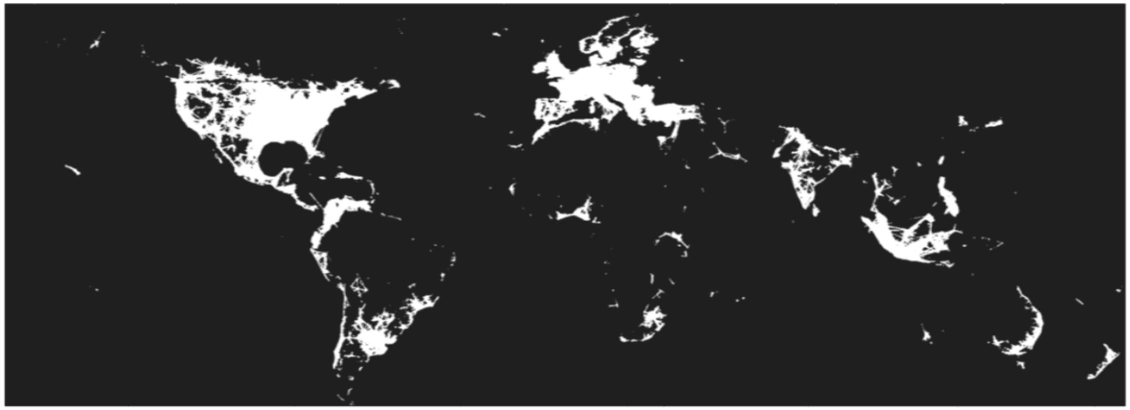
Two Questions

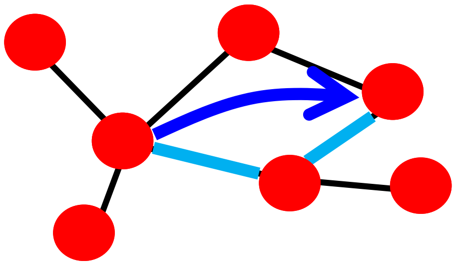
- (Today) What is the structure of a social network?
- (Later) Which mechanisms do people use to route and find the target?



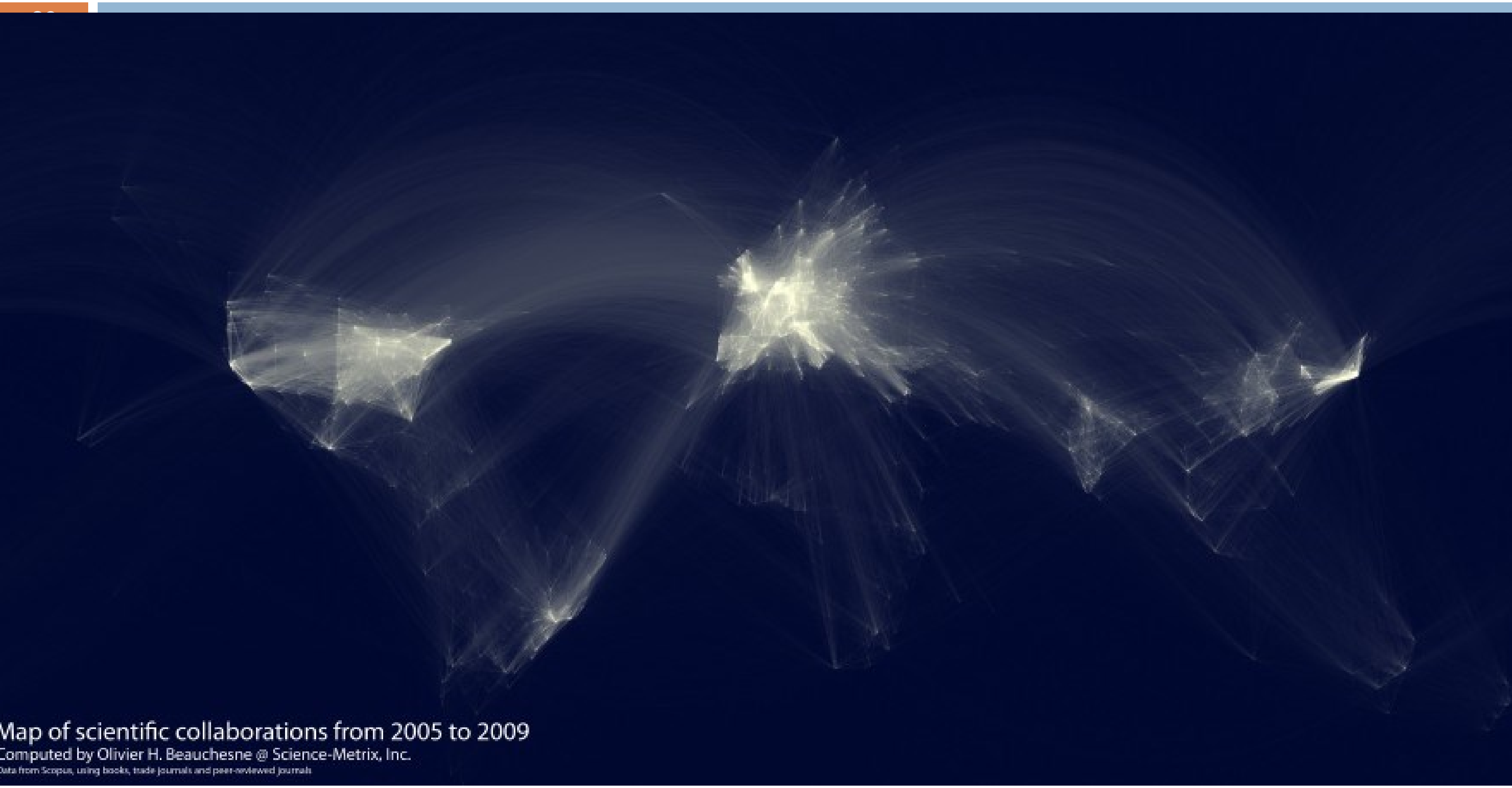
6-Degrees: Should We Be Surprised?

- Assume each human is connected to 100 other people. Then:
  - Step 1: reach 100 people
  - Step 2: reach 100*100 = 10,000 people
  - Step 3: reach 100*100*100 = 1,000,000 people
  - Step 4: reach 100*100*100*100 = 100M people
  - **In 5 steps we can reach 10 billion people**
- What's wrong here?
  - **92% of new FB friendships are to a friend-of-a-friend** [Backstom-Leskovec '11]





Scientific Collaborations



Clustering Implies Edge Locality

- **MSN network has 7 orders of magnitud**
clustering than the corresponding G_{np}!

The difference between “random” and “actual” – add green text

- **Other examples:**

Actor Collaborations (IMDB): N = 225,226 nodes, avg. degree $\bar{k}$ = 61
Electrical power grid: N = 4,941 nodes, $\bar{k}$ = 2.67
Network of neurons: N = 282 nodes, $\bar{k}$ = 14

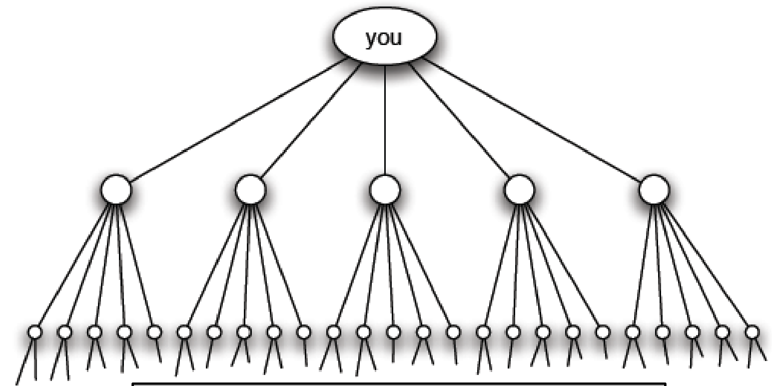
Network	h_{actual}	h_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power Grid	18.70	12.40	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

h ... Average shortest path length
C ... Average clustering coefficient

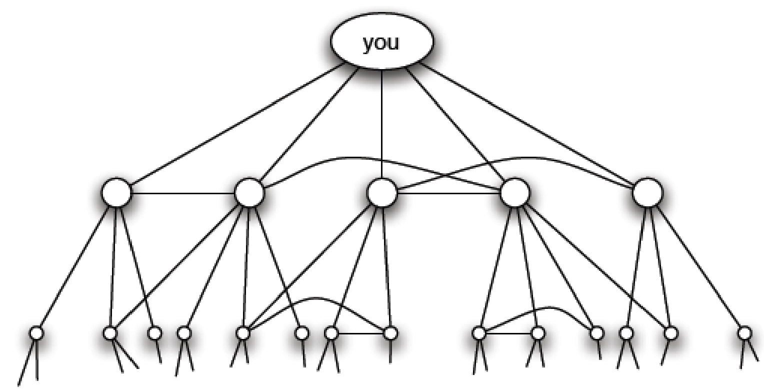
Back to the Small-World

- **Consequence of expansion:**
  - Short paths: $O(\log n)$
    - This is "best" we can do if the have a constant degree
    - and there are *n* nodes
- **But networks have local structure:**
  - **Triadic closure:**

    Friend of a friend is my friend
- **How can we have both?**



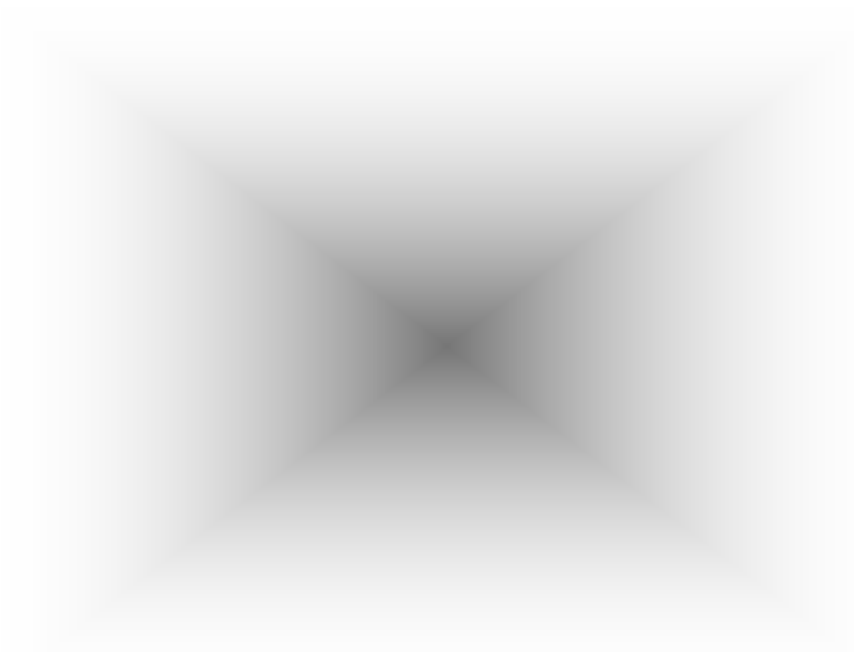
Pure exponential growth



Triadic closure reduces growth rate

Clustering vs. Randomness

Where should we place social networks?



Clustered?

Random?

Simplest Model of Graphs

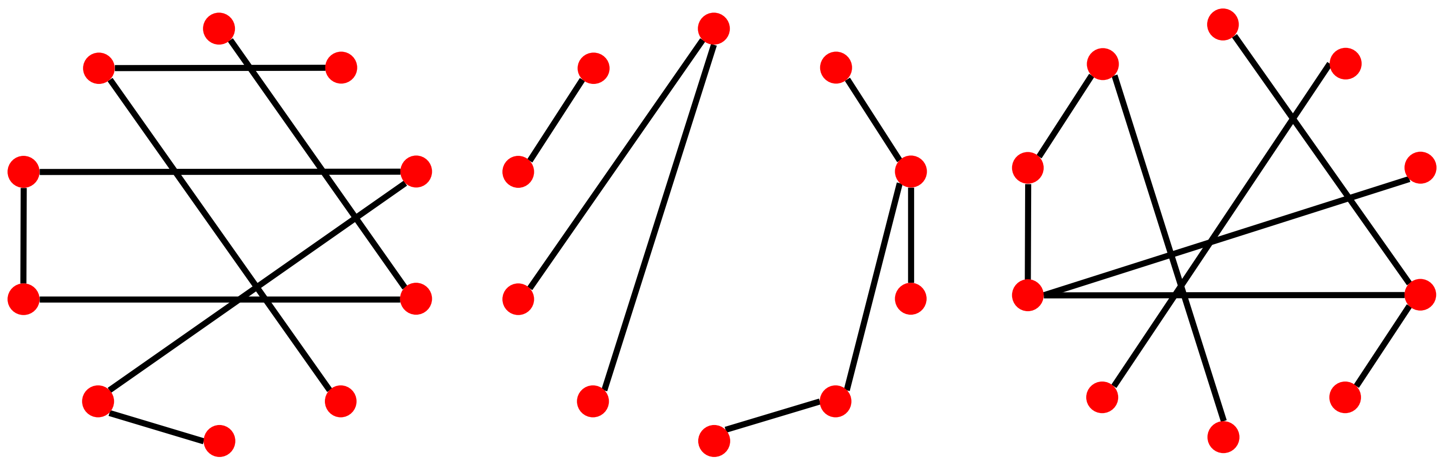
- **Erdös-Renyi Random Graphs** [Erdös-Renyi, '60]

- Two variants:
  - $G_{n,p}$: undirected graph on n nodes and each edge (u,v) appears i.i.d. with probability p
  - $G_{n,m}$: undirected graph with n nodes, and m uniformly at random picked edges

What kinds of networks does such model produce?

Random Graph Model

- n **and** p **do not uniquely determine the graph!**
  - The graph is a result of a random process
- We can have many different realizations



n = 10
p= 1/6

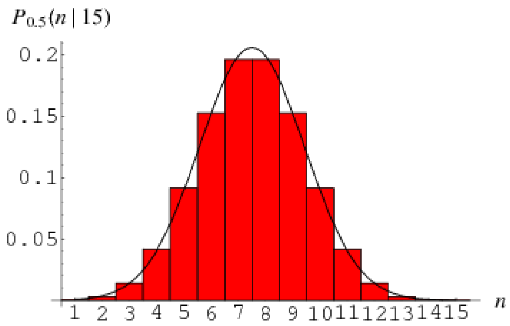
Random Graph Model: Edges

- **How likely is a graph on E edges?**
- $P(E)$: the probability that a given G_{np} generates a graph on exactly E edges:

$$P(E) = \binom{E_{\max}}{E} p^{E} (1-p)^{E_{\max}-E}$$

where $E_{max}=n(n-1)/2$ is the maximum possible number of edges in an undirected graph of n nodes

Binomial distribution >>>



Node Degrees in a Random Graph

- **What is expected degree of a node?**
  - Let X_v be a rnd. var. measuring the degree of node v
  - **We want to know:** $E[X_v] = \sum_{j=0}^{n-1} j\, P(X_v = j)$
    - **For the calculation we will need: Linearity of expectation**
      - For any random variables $Y_1, Y_2, \dots, Y_k$
      - If $Y = Y_1 + Y_2 + \dots Y_k$, then $E[Y] = \sum_i E[Y_i]$
- **Easier way:**
  - Decompose X_v to $X_v = X_{v,1} + X_{v,2} + \dots + X_{v,n-1}$
    - where $X_{v,u}$ is a $\{0,1\}$-random variable which tells if edge (v,u) exists or not

$$E[X_v] = \sum_{u=1}^{n-1} E[X_{vu}] = (n-1)p$$

How to think about this?

- Prob. of node u linking to node v is p
- u can link (flips a coin) to all other $(n-1)$ nodes
- Thus, the expected degree of node u is: $p(n-1)$

Properties of G_{np}

Degree distribution: $P(k)$

Path length: h

Clustering coefficient: C

What are values of these properties for G_{np}?

Degree Distribution

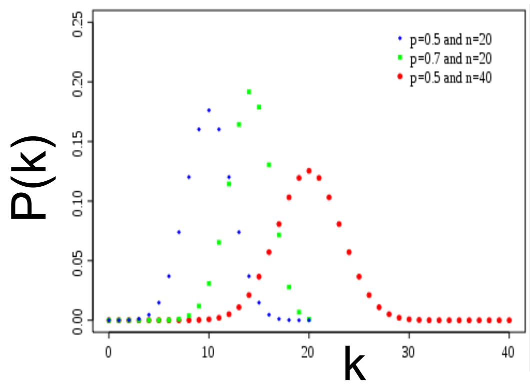
- **Fact:** **Degree distribution of G_{np} is Binomial.**
- Let $P(k)$ denote a fraction of nodes with degree k:

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

Select k nodes out of n-1

Probability of having k edges

Probability of missing the rest of the n-1-k edges



Mean, variance of a binomial distribution

$$\bar{k} = p(n-1)$$

$$\sigma^2 = p(1-p)(n-1)$$

$$\frac{\sigma}{\bar{k}} = \left[\frac{1-p}{p}\frac{1}{(n-1)}\right]^{1/2} \approx \frac{1}{(n-1)^{1/2}}$$

As the network size increases, the distribution becomes increasingly narrow—we are increasingly confident that the degree of a node is in the vicinity of $\bar{k}$.

Clustering Coefficient of G_{np}

- **Remember:** $$C_i = \frac{2e_i}{k_i(k_i-1)}$$
- Edges in G_{np} appear i.i.d with prob. p
- **So:** $$e_i = p\frac{k_i(k_i-1)}{2}$$

Each pair is connected with prob. p

No. of distinct pairs of neighbors of node i of degree k_i

- **Then:** $$C = \frac{p \cdot k_i(k_i-1)}{k_i(k_i-1)} = p = \frac{\bar{k}}{N}$$

Clustering coefficient of a random graph is small.
For a fixed avg. degree, C decreases with the graph size N.

Real Networks vs. G_{np}

- **Are real networks like random graphs?**
  - Giant connected component: ☺
  - Average path length: ☺
  - Clustering Coefficient: ☹
  - Degree Distribution: ☹
- **Problems with the random network model:**
  - Degreed distribution differs from that of real networks
  - Giant component in most real network does NOT emerge through a phase transition
  - No local structure – clustering coefficient is too low
- **Most important: Are real networks random?**
  - The answer is simply: **NO!**

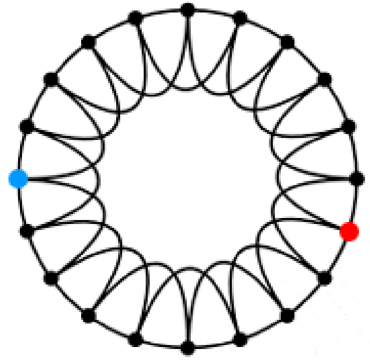
Real Networks vs. G_{np}

- **If G_{np} is wrong, why did we spend time on it?**
  - It is the reference model for the rest of the class.
  - It will help us calculate many quantities, that can then be compared to the real data
  - It will help us understand to what degree is a particular property the result of some random process

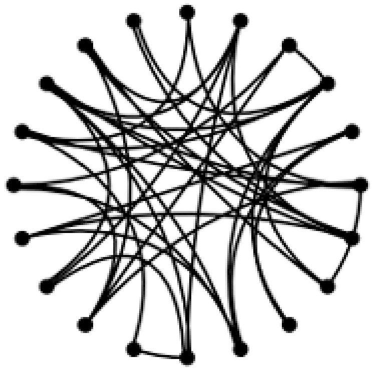
So, while G_{np} is WRONG, it will turn out to be extremly USEFUL!

Small-World: How?

- **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

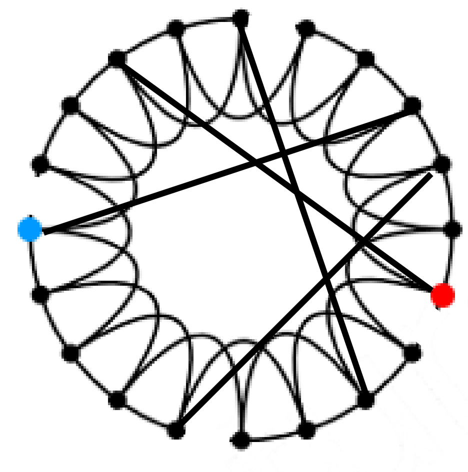
  - Clustering implies edge "locality"
  - Randomness enables "shortcuts"

Solution: The Small-World Model

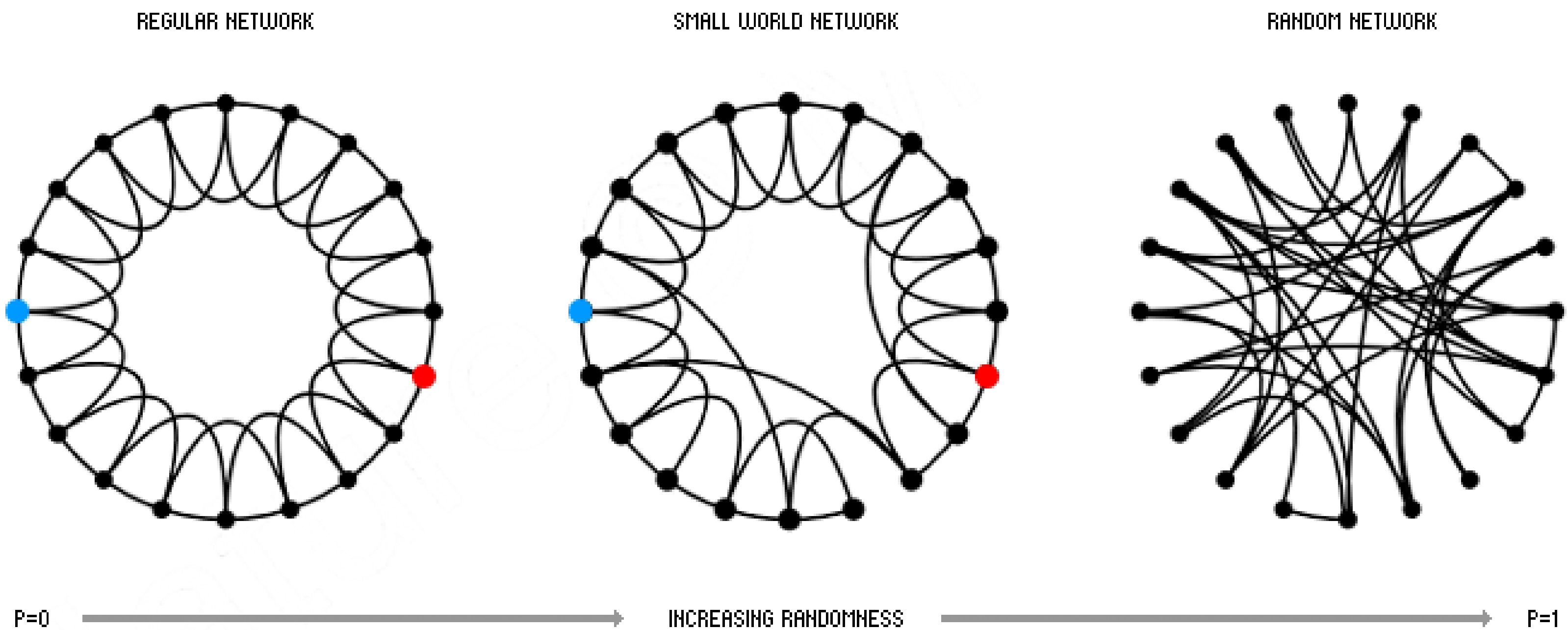
Small-world Model [Watts-Strogatz '98]:

2 components to the model:

- (1) Start with a **low-dimensional regular lattice**
  - Has high clustering coefficient
- Now introduce randomness ("shortcuts")
- (2) **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



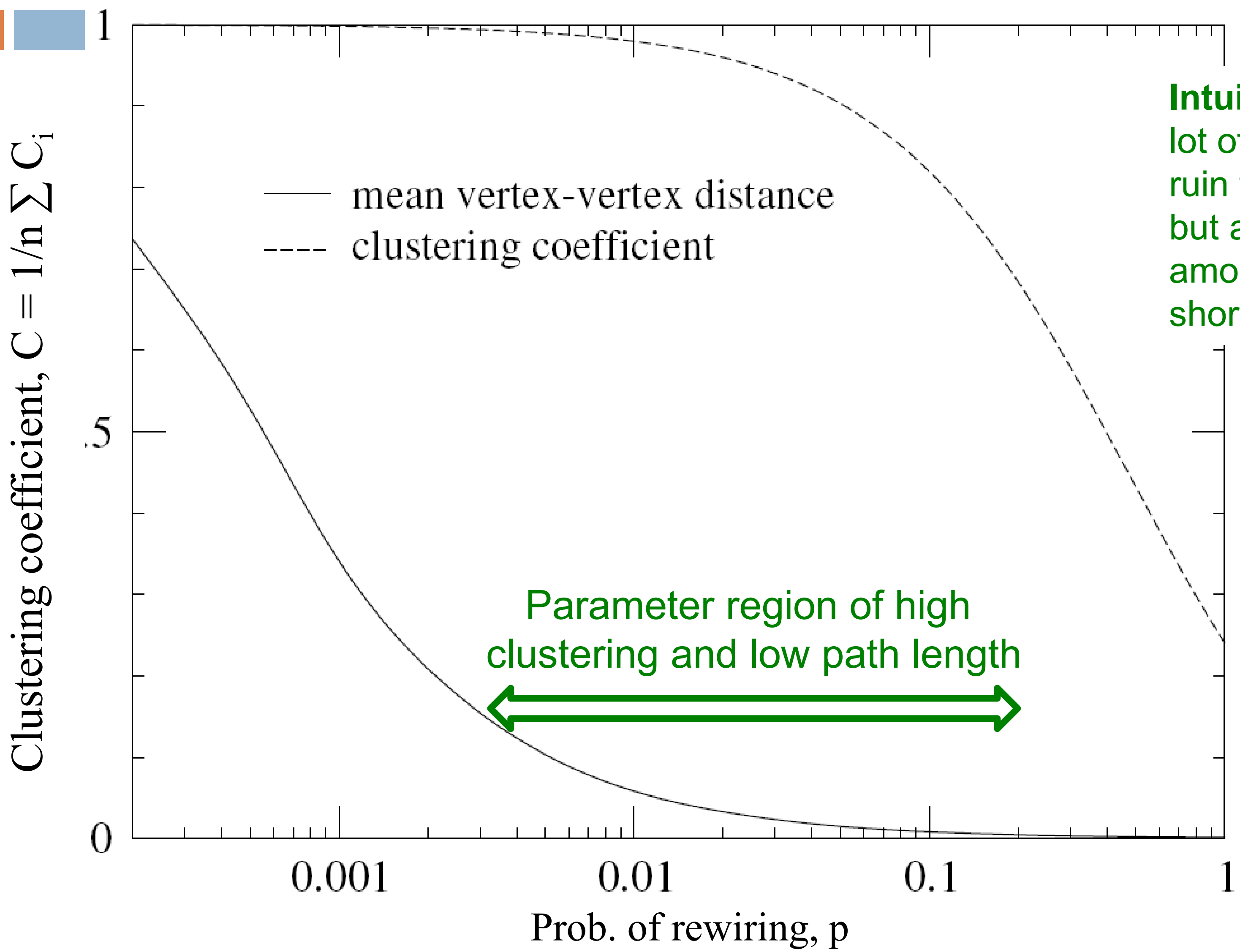
The Small-World Model



High clustering High diameter	High clustering Low diameter	Low clustering Low diameter
$h = \frac{N}{2\bar{k}}$ $C = \frac{3}{4}$		$h = \frac{\log N}{\log \alpha}$ $C = \frac{\bar{k}}{N}$

Rewiring allows us to “interpolate” between
a regular lattice and a random graph

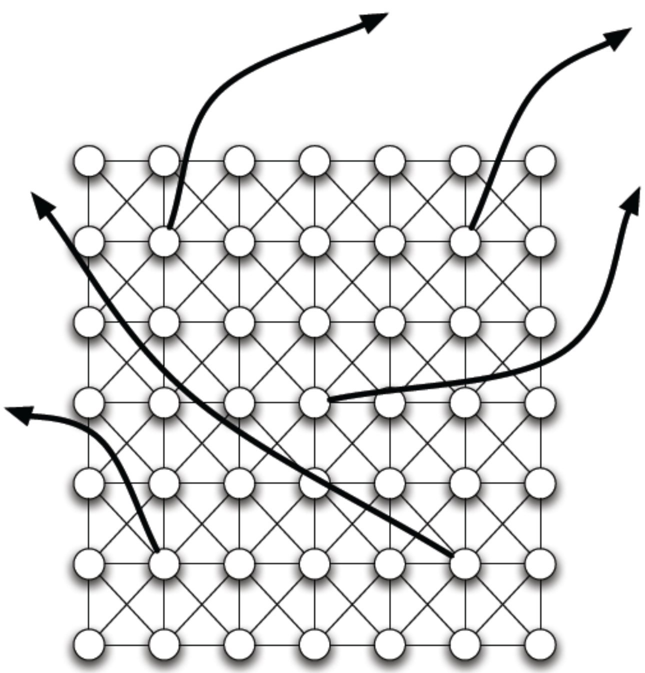
The Small-World Model



Intuition: It takes a lot of randomness to ruin the clustering, but a very small amount to create shortcuts.

Diameter of the Watts-Strogatz

- **Alternative formulation of the model:**
  - Start with a square grid
  - Each node has 1 random long-range edge
    - Each node has 1 spoke. Then randomly connect them.



$$C_i = \frac{2 \cdot e_i}{k_i(k_i - 1)} = \frac{2 \cdot 12}{9 \cdot 8} \geq 0.33$$

There are already 12 triangles in the grid and the long-rage edge can only close more.

What's the diameter?

It is *log(n)*

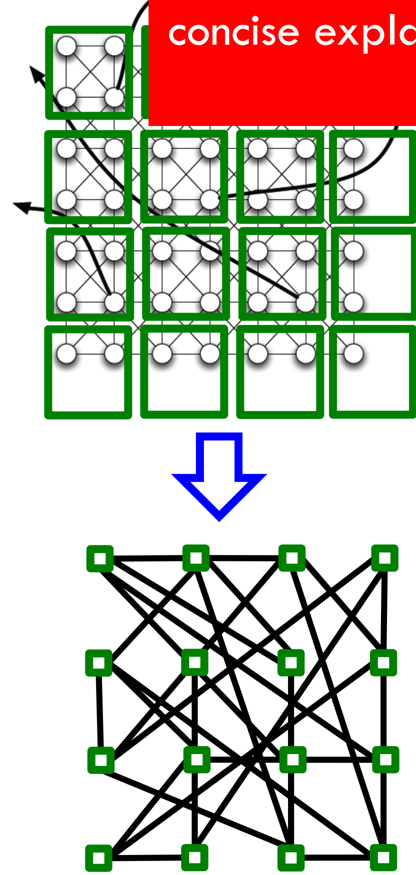
Why?

Diameter of the Watts-Strogatz

- <u>Proof:</u>
  - Consider a graph where we contract 2x2 subgraphs into supernodes
  - Now we have 4 edges sticking out of each supernode
    - **4-regular random graph!**
  - From Thm. we have short paths between super nodes
  - We can turn this into a path in a real graph by adding at most 2 steps per hop

$\Rightarrow$ **Diameter of the model is** $O(2 \log n)$

Need a good concise explanation



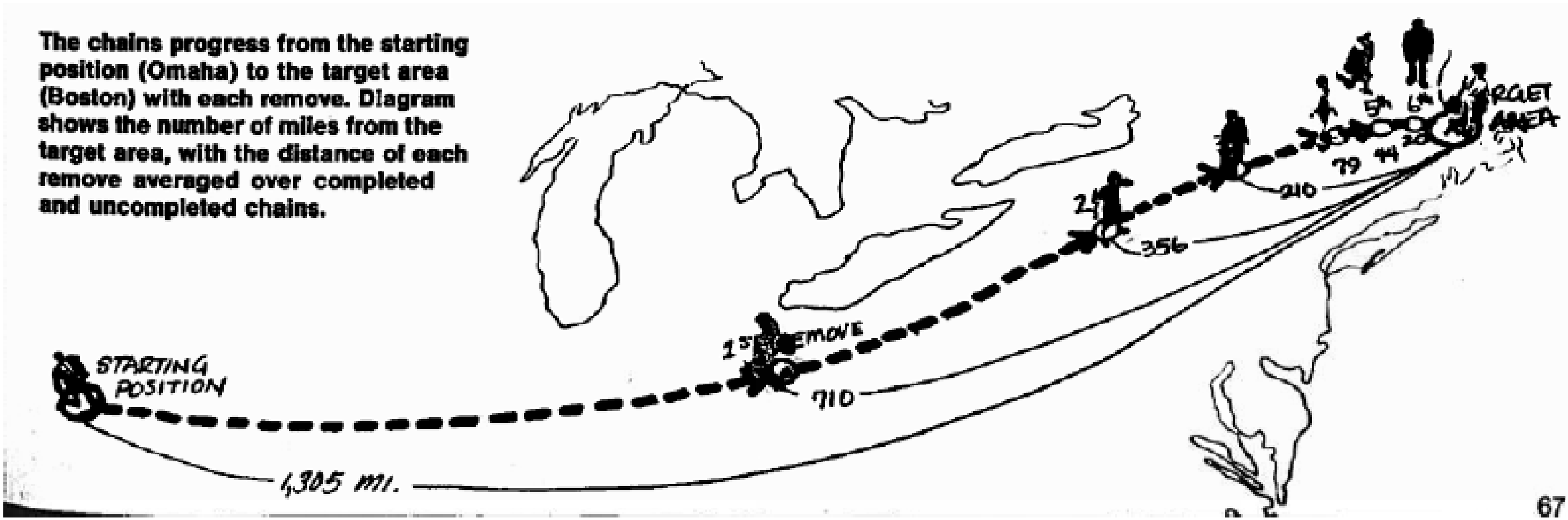
4-regular random graph

Small-World: Summary

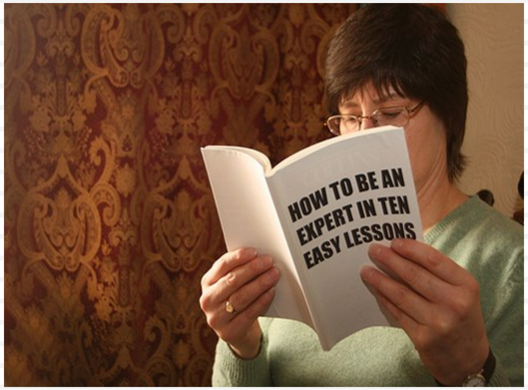
- **Could a network with high clustering be at the same time a small world?**
  - Yes. You don't need more than a few random links.
- **The Watts Strogatz Model:**
  - Provides insight on the interplay between clustering and the small-world
  - Captures the structure of many realistic networks
  - Accounts for the high clustering of real networks
  - Does not lead to the correct degree distribution
  - Does not enable **navigation** **(next lecture)**

How to Navigate the Network?

- (1) What is the structure of a social network?
- (Next) Which mechanisms do people use to route and find the target?



The 10 papers that will make you a social expert



10 sociological must-reads

1. S.Milgram, "The small world problem," *Psychology today*, 1967.
2. M. Granovetter, "The strength of weak ties: A network theory revisited," *Sociological theory*, vol. 1, pp. 201–233, 1983.
3. M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual review of sociology*, vol. 27, pp. 415–444, Jan. 2001.
4. M. O. Lorenz, "Methods of measuring the concentration of wealth," *Publications of the American Statistical Association*, vol. 9, no. 70, pp. 209–219, 1905.
   + H. Simon, "On a Class of Skew Distribution Functions," Biometrika, vol. 42, no. 3, pp. 425–440, 1955.
5. R. I. M. Dunbar, "Coevolution of Neocortical Size, Group-Size and Language in Humans," *Behav Brain Sci*, vol. 16, no. 4, pp. 681–694, 1993.
6. D. Cartwright and F. Harary, "Structural balance: a generalization of Heider's theory.," *Psychological Review*, vol. 63, no. 5, pp. 277–293, 1956.
7. M. Granovetter, "Threshold Models of Collective Behavior," *The American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, May 1978.
8. B. Ryan and N. C. Gross, "The diffusion of hybrid seed corn in two Iowa communities," *Rural sociology*, vol. 8, no. 1, pp. 15–24, 1943.
   + S. Asch, "Opinions and social pressure," *Scientific American*, 1955.
9. R. S. Burt, *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.
10. F. Galton, "Vox Populi," *Nature*, vol. 75, no. 1949, pp. 450–451, Mar. 1907.