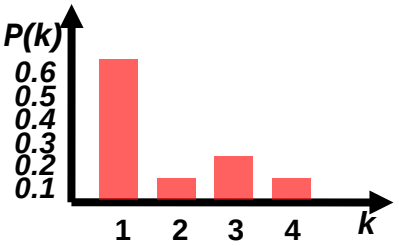
Diameter of G_{np} and the Small-World Phenomena

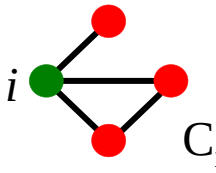
CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu

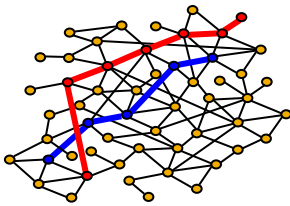
Recap: Network Properties & G_{np}



How to characterize networks?

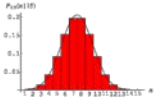
- Degree distribution *P(k)*
- Clustering Coefficient *C*
- Diameter (avg. shortest path length) *h*



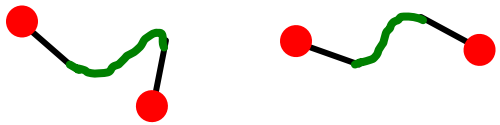
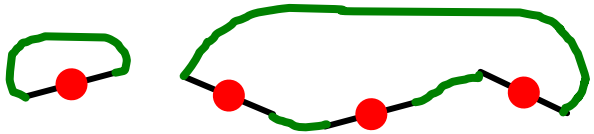
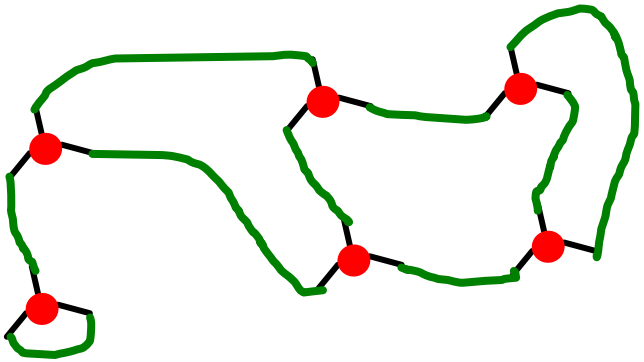


How to model networks?

- **Erdös-Renyi Random Graph** [Erdös-Renyi, '60]
  - $G_{n,p}$: undirected graph on ***n*** nodes where each edge ***(u,v)*** appears independently with prob. ***p***
    - **Degree distribution:** Binomial(n, p)
    - **Clustering coefficient:** $C \approx p \approx \frac{k}{n}$
    - **Diameter:** (now)



Def: Random k-Regular Graphs

- To prove the diameter of a G_{np} we define few concepts
- **Define: Random k-Regular graph**
  - **Assume each node has k spokes (half-edges)**
    - **k=1:** 
      Graph is a set of pairs
    - **k=2:** 
      Graph is a set of cycles
    - **k=3:** 
      Arbitrarily complicated graphs
  - **Randomly pair them up!**

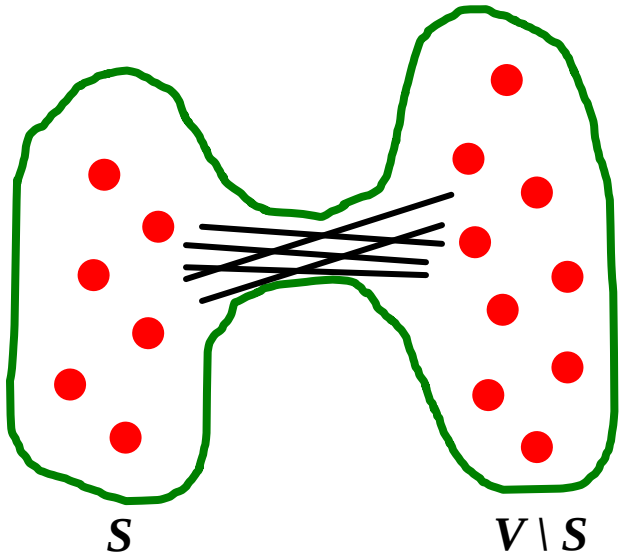
Def: Expansion

- Graph $G(V, E)$ has **expansion α**: if $\forall S \subseteq V$:

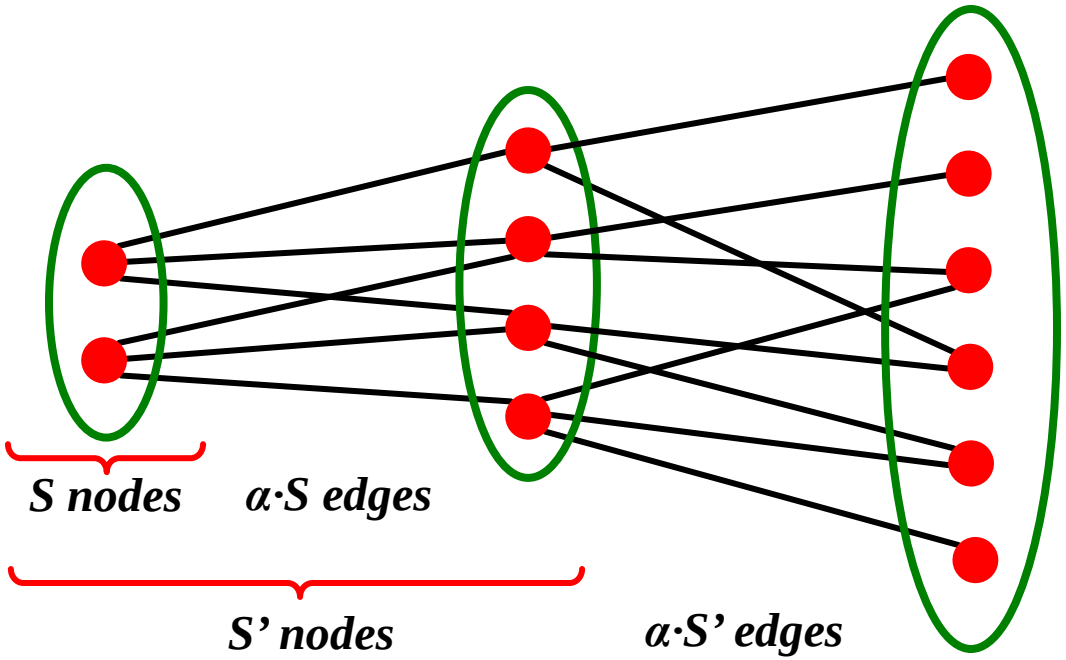
  # of edges leaving $S \geq \alpha \cdot \min(|S|, |V \setminus S|)$

- **Or equivalently:**

$$\alpha = \min_{S \subseteq V} \frac{\#edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$



Expansion: Intuition



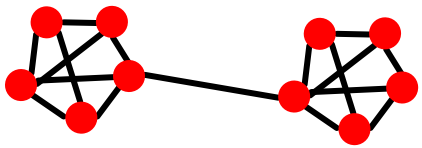
$$\alpha = \min_{S \subseteq V} \frac{\#edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$

(A big) graph with "good" expansion

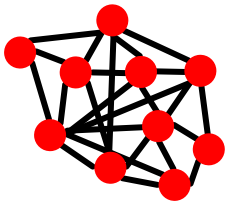
Expansion: Measures Robustness

$$\alpha = \min_{S \subseteq V} \frac{\#\,edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$

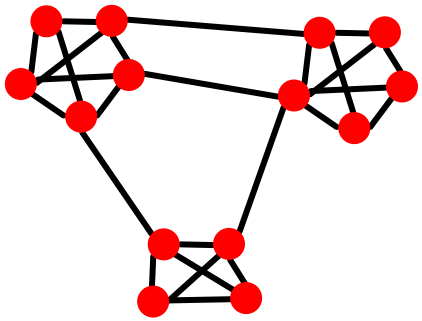
- Expansion is **measure of robustness:**
  - To disconnect l nodes, we need to cut $\geq \alpha \cdot l$ edges
- Low expansion:



- High expansion:



- Social networks:
  - "Communities"



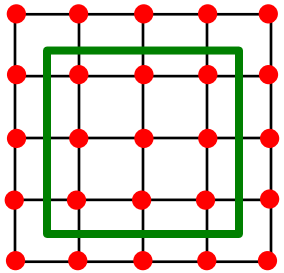
Expansion: k-Regular Graphs

$$\alpha = \min_{S \subseteq V} \frac{\#\,edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$

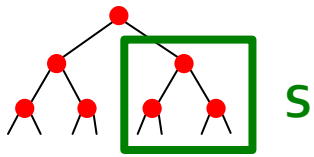
- ***k*-regular graph** (every node has degree k):
  - Expansion is at most k (when S is a single node)
- **Is there a graph on n nodes ($n \to \infty$), of fixed max deg. k, so that expansion α remains const?**

Examples:

- **n×n grid:** $k=4$: $\alpha = 2n/(n^2/4) \to 0$
  (S=n/2 × n/2 square in the center)



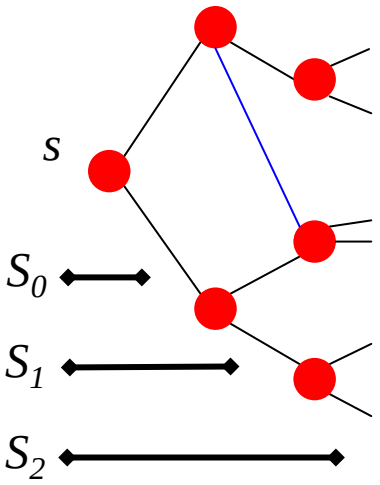
- **Complete binary tree:**
  $\alpha \to 0$ for $|S|=(n/2)-1$



- **Fact:** For a random **3-regular graph** on n nodes, there is some const α ($\alpha > 0$, independent. of n) such that w.h.p. the expansion of the graph is $\geq \alpha$

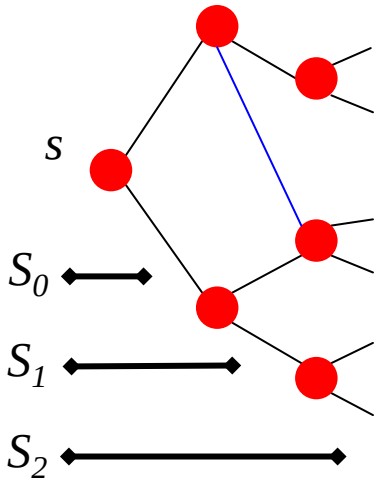
Diameter of 3-Regular Rnd. Graph

- **Fact:** In a graph on n nodes with expansion α for all pairs of nodes s and t there is a path of $O((\log n) / \alpha)$ edges connecting them.
- Proof:
  - Proof strategy:
    - We want to show that from any node s there is a path of length $O((\log n)/\alpha)$ to any other node t
  - Let S_j be a set of all nodes found within j steps of BFS from s.
  - **How does S_j increase as a function of j?**

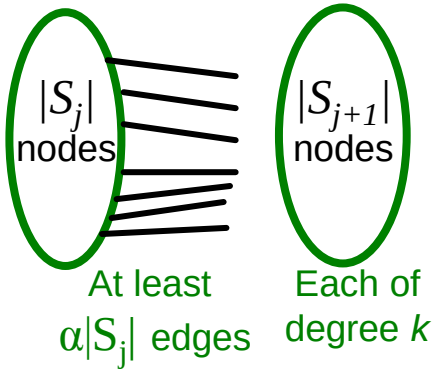


Diameter of 3-Regular Rnd. Graph

- Proof (continued):
  - Let S_j be a set of all nodes found within j steps of BFS from s.
  - **We want to relate S_j and S_{j+1}**



$$|S_{j+1}| \geq |S_j| + \underbrace{\frac{\overbrace{\alpha |S_j|}^{\text{Expansion}}}{k}}_{\substack{\text{At most } k \text{ edges} \\ \text{“collide” at a node}}} =$$



$$|S_{j+1}| \geq |S_j| \left(1 + \frac{\alpha}{k}\right) = \left(1 + \frac{\alpha}{k}\right)^{j+1}$$

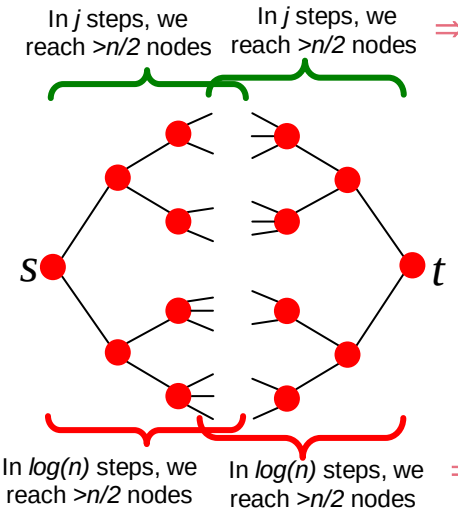
Diameter of 3-Regular Rnd. Graph

$$e = \lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x$$

- Proof (continued):
  - **In how many steps of BFS we reach >*n/2* nodes?**
  - Need j so that: $S_j = \left(1+\frac{\alpha}{k}\right)^j \geq \frac{n}{2}$
  - Let's set: $j = \frac{k\log_2 n}{\alpha}$
  - Then:

$$\left(1+\frac{\alpha}{k}\right)^{\frac{k\log_2 n}{\alpha}} \geq 2^{\log_2 n} = n > \frac{n}{2}$$

  - In $2k/\alpha \cdot \log n$ steps $|S_j|$ grows to $\Theta(n)$. So, **the diameter of *G* is $O(\log(n)/\alpha)$**



Claim: $\left(1+\frac{\alpha}{k}\right)^{\frac{k\log_2 n}{\alpha}} \geq 2^{\log_2 n}$

Remember $n>0$, $\alpha \leq k$ then:

if $\alpha = k$: $(1+1)^{\frac{1}{1}\log_2 n} = 2^{\log_2 n}$

if $\alpha \to 0$ then $\frac{k}{\alpha} = x \to \infty$:

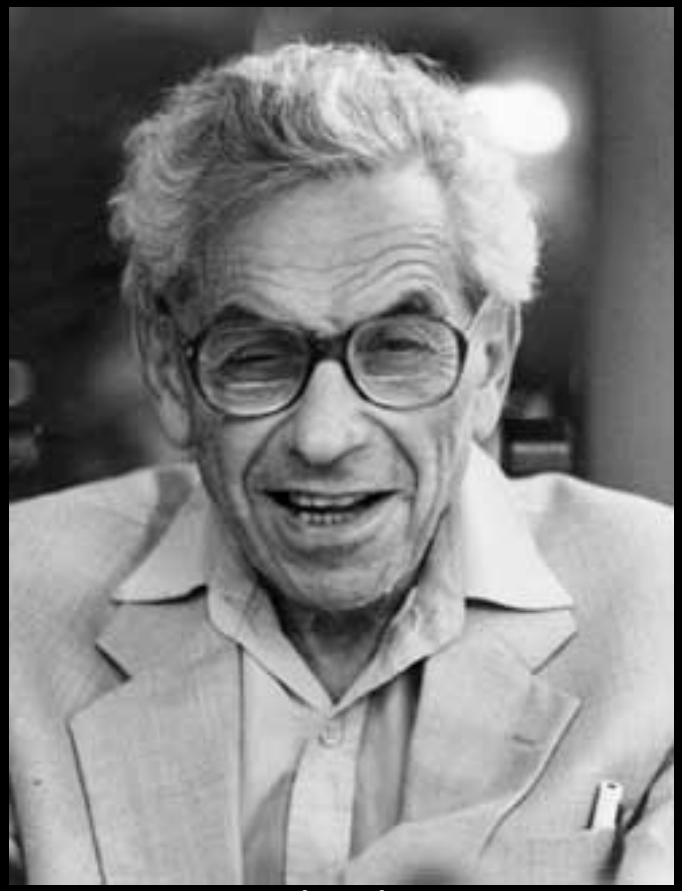
and $\left(1+\frac{1}{x}\right)^{x\log_2 n} = e^{\log_2 n} > 2^{\log_2 n}$

Network Properties of G_{np}

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

Path length: $O(\log n)$

Clustering coefficient: $C = p = \bar{k} / n$



Paul Erdös

G_{np} is so cool!
Let's also look at the connectivity

Back to Node Degrees of G_{np}

- **Remember, expected degree:** $E[X_v] = (n-1)p$
- **We want $E[X_v]$ be independent of n**

  So let: $p=k/(n-1)$
- **Observation:** If we build random graph G_{np} with $p=k/(n-1)$ we have many isolated nodes
- **Why?**

$$P[v \text{ has degree } 0] = (1-p)^{n-1} = \left(1-\frac{k}{n-1}\right)^{n-1} \xrightarrow[n\to\infty]{} e^{-k}$$

$$\lim_{n\to\infty}\left(1-\frac{k}{n-1}\right)^{n-1} = \left(1-\frac{1}{x}\right)^{-x\cdot k} = \left[\underbrace{\lim_{x\to\infty}\left(1-\frac{1}{x}\right)^{-x}}_{e}\right]^{-k} = e^{-k}$$

Use substitution $\frac{1}{x} = \frac{k}{n-1}$

By definition: $e = \lim_{x\to\infty}\left(1+\frac{1}{x}\right)^{x}$

No Isolated Nodes

- **How big do we have to make p before we are likely to have no isolated nodes?**
- We know: $P[v \text{ has degree } 0] = e^{-k}$
- **Event we are asking about is:**
  - I = some node is isolated
  - $I = \bigcup_{v \in N} I_v$ where I_v is the event that v is isolated
- **We have:**

$$P(I) = P\left(\bigcup_{v \in N} I_v\right) \leq \sum_{v \in N} P(I_v) = ne^{-k}$$

Union bound

$$\left|\bigcup_i A_i\right| \leq \sum_i |A_i|$$

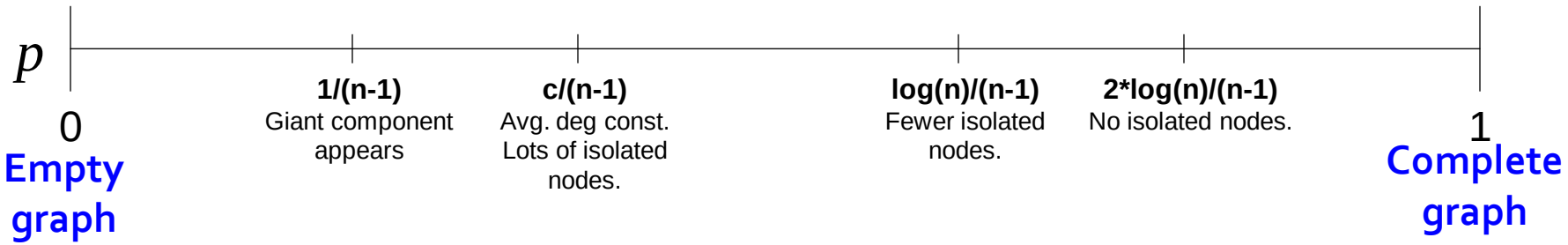
No Isolated Nodes

- **We just learned:** $P(I) \leq n\, e^{-k}$
- **Let's try:**
  - $k = \ln n$ then: $n\, e^{-k} = n\, e^{-\ln n} = n \cdot 1/n = 1$
  - $k = 2 \ln n$ then: $n\, e^{-2 \ln n} = n \cdot 1/n^2 = 1/n$
- **So if:**
  - $k = \ln n$ then: $P(I) \leq 1$
  - $k = 2 \ln n$ then: $P(I) \leq 1/n \rightarrow 0$ as $n \rightarrow \infty$

So, for ***p=2ln(n)*** we get no isolated nodes (as $n \rightarrow \infty$)

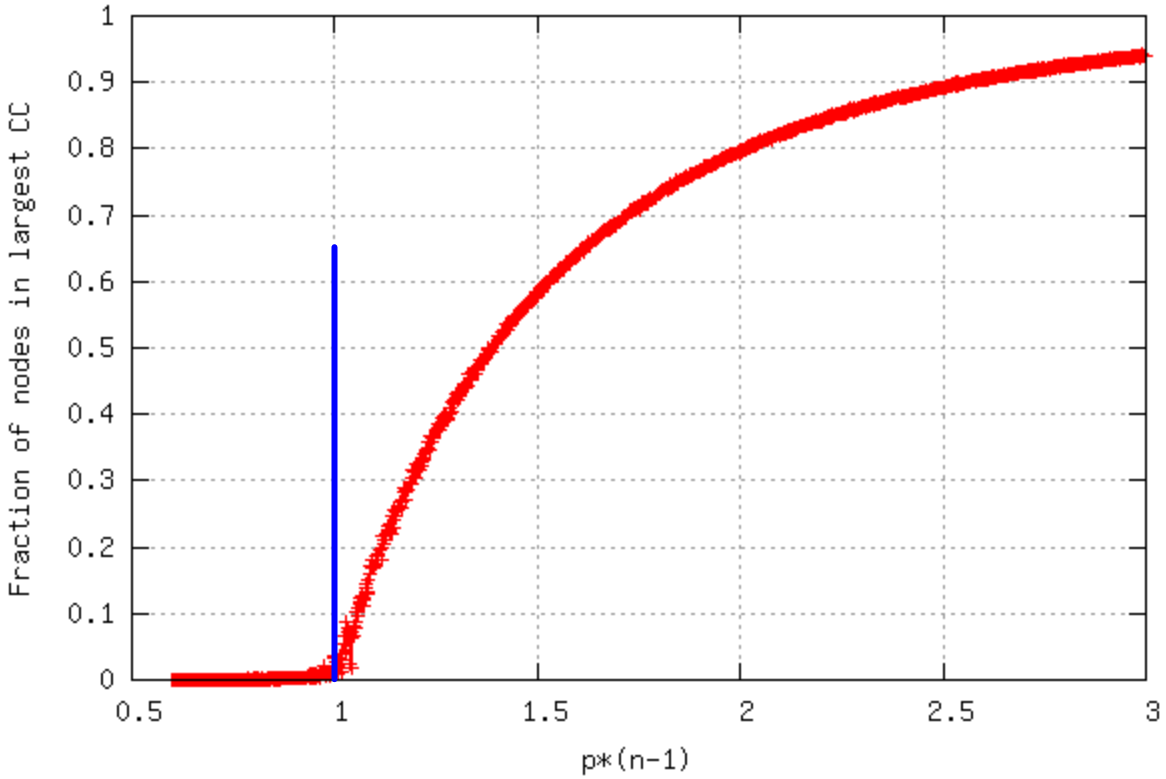
"Evolution" of a Random Graph

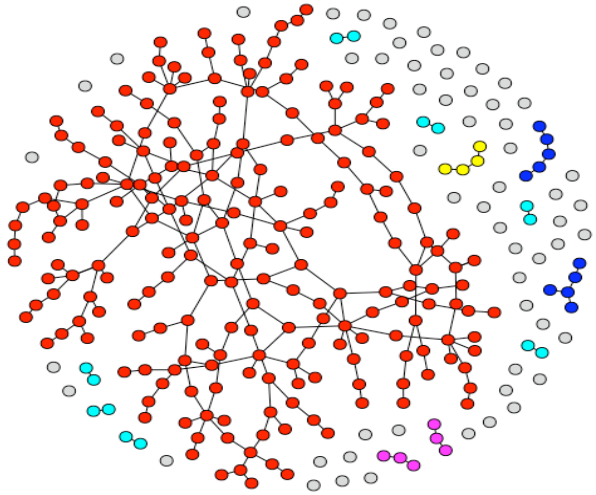
- **Graph structure of G_{np} as p changes:**



- **Emergence of a Giant Component:**
  avg. degree $k=2E/n$ or $p=k/(n-1)$
  - $k=1-\varepsilon$: all components are of size $\Omega(\log n)$
  - $k=1+\varepsilon$: 1 component of size $\Omega(n)$, others have size $\Omega(\log n)$

G_{np} Simulation Experiment



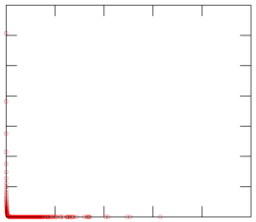
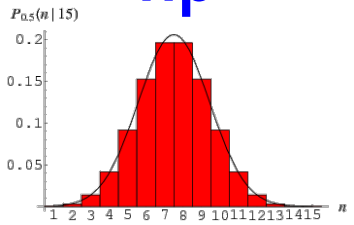


Fraction of nodes in the largest component

- G_{np}, n=100,000, $k=p(n-1)$ = 0.5 … 3


DIRECT FROM
★RINGSIDE!★
THE FIGHT EVERYONE WANTS TO SEE...
MSN
BOMBER
VS
Gnp
ROCKER
ROCK'EM
SOCK'EM
ROBOTS
★ ★ ★ MAIN EVENT ★ ★ ★

Back to MSN vs. G_{np}

	MSN	G_{np}
Degree distribution:		
Path length:	**6.6**	$O(\log n)$ $h \approx 8.2$
Clustering coefficient:	0.11	$\bar{k} / n$ $C \approx 8 \cdot 10^{-8}$
Connected component:	99%	$\bar{k} \approx 14$ $\log_{10}(180M) \approx 8$ So, GCC should kind of be there.

Real Networks vs. G_{np}

- **Are real networks like random graphs?**
  - Average path length: ☺
  - Giant connected component: ☺
  - Clustering Coefficient: ☹
  - Degree Distribution: ☹
- **Problems with the random network model:**
  - Degreed distribution differs from that of real networks
  - Giant component in most real networks 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!**

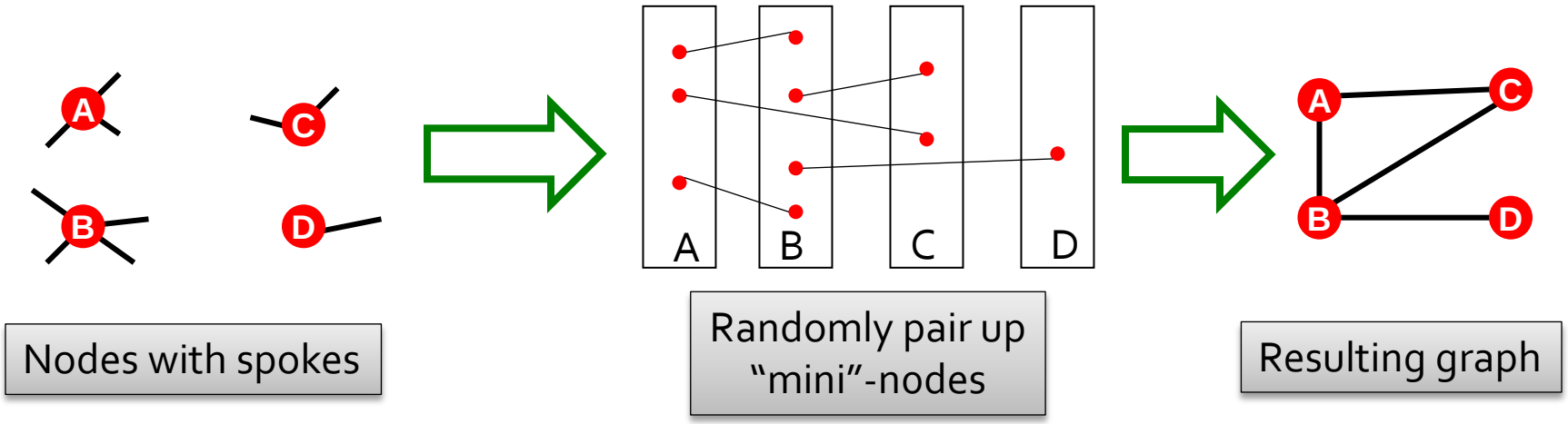
Real Networks vs. G_{np}

- **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!

Intermezzo: Configuration Model

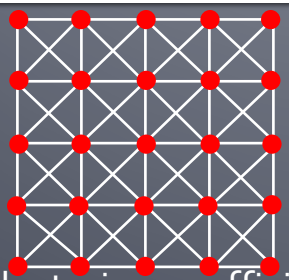
- **Goal:** Generate a random graph with a given degree sequence $k_1, k_2, \ldots k_N$
- **Configuration model:**



- **Useful for as a "null" model of networks**
  - We can compare the real network G and a "random" G' which has the same degree sequence as G

The Small-World Model

Can we have high clustering while also having short paths?



High clustering coefficient,
High diameter

Vs.

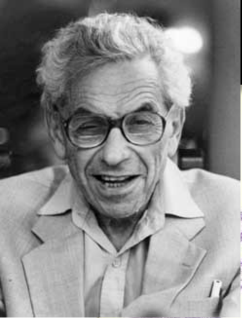
Low clustering coefficient
Low diameter

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!



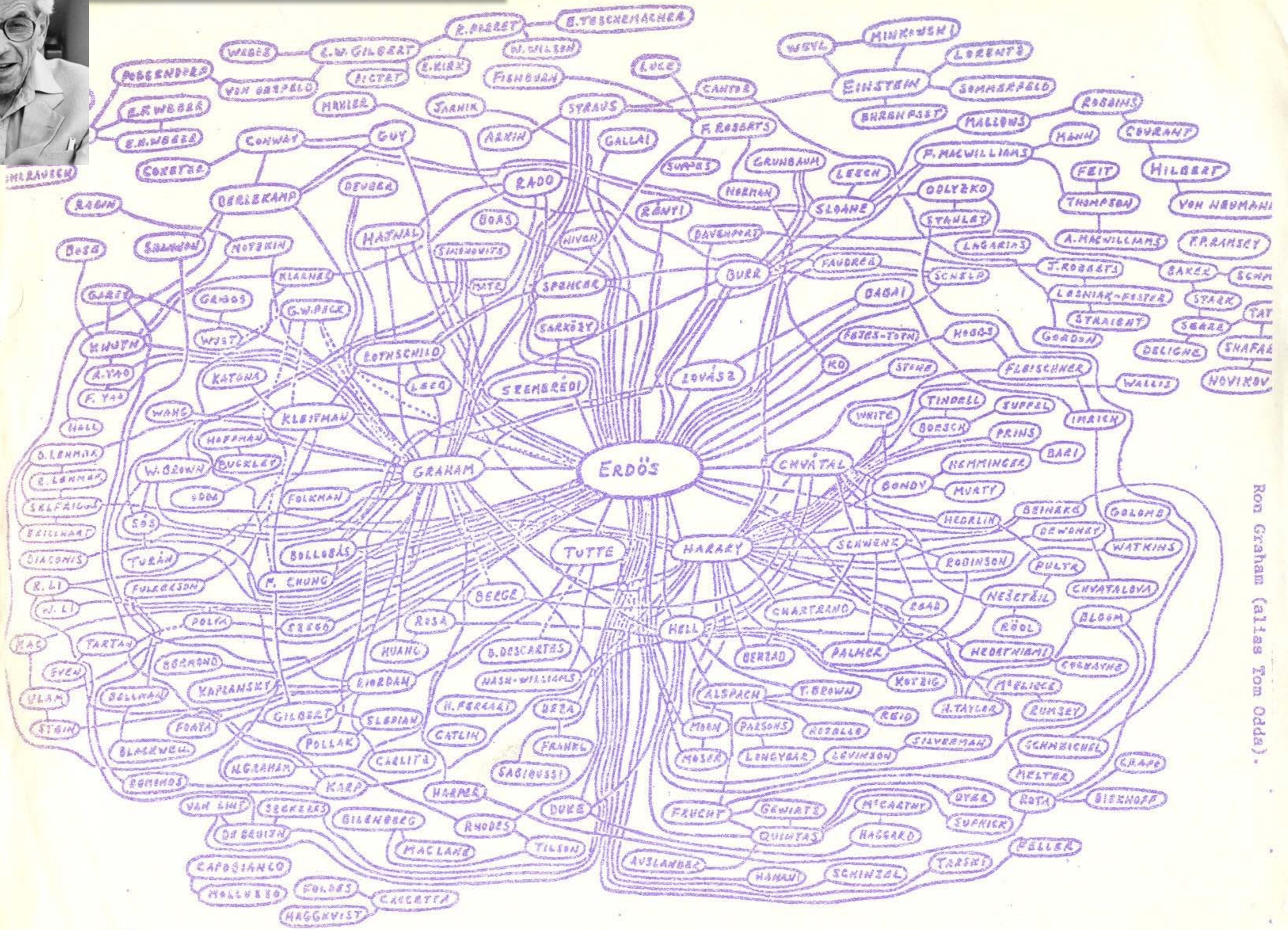
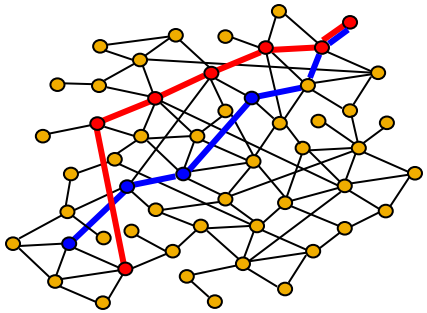


Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

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?**

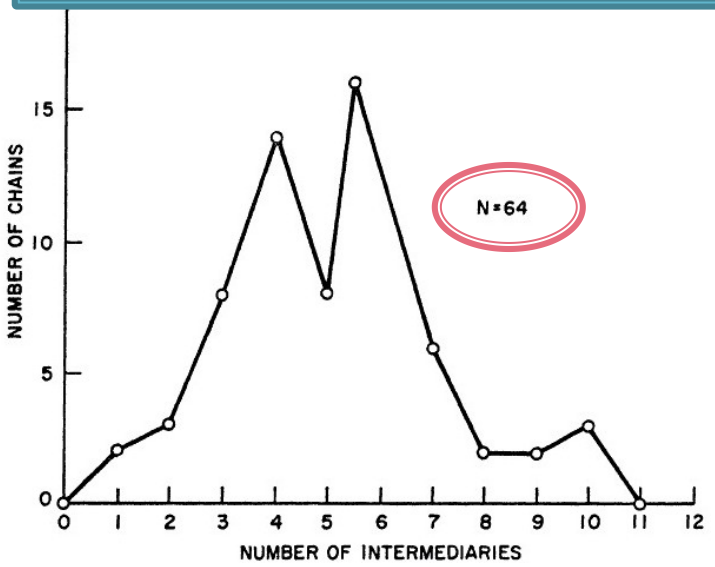


[Milgram, '67]

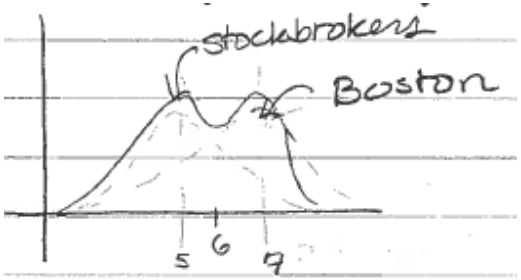
The Small-World Experiment

- **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

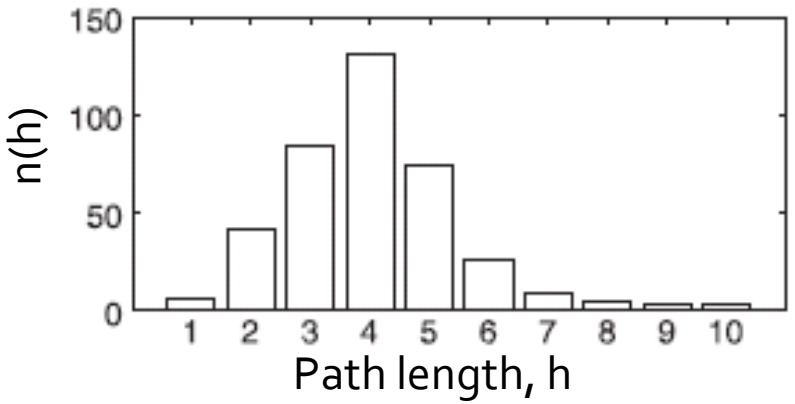


- **Boston vs. occupation networks:**
- **Criticism:**
  - **Funneling:**
    - 31 of 64 chains passed through 1 of 3 people ass their final step → Not all links/nodes are equal
  - Starting points and the target were non-random
  - People refused to participate (25% for Milgram)
  - **Some sort of social 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

[Dodds-Muhamad-Watts, '03]

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

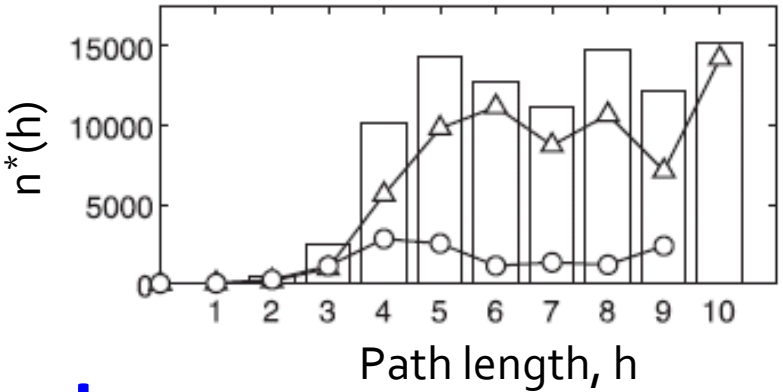
Correction factor: $n^*(h) = \frac{n(h)}{\prod_{i=0}^{h-1}(1-r_i)}$

r_i .... drop-out rate at hop i

[Dodds-Muhamad-Watts, '03]

Small-World in Email Study

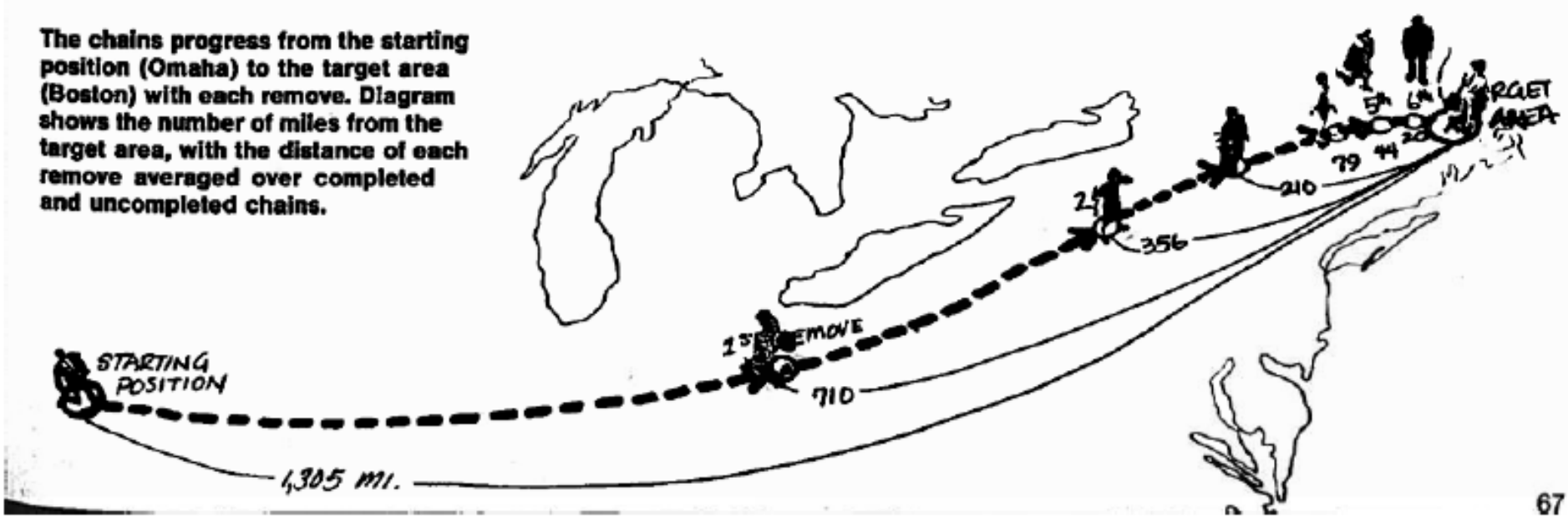
- **After the correction:**
  - **Typical path length h = 7**



- **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 network structure

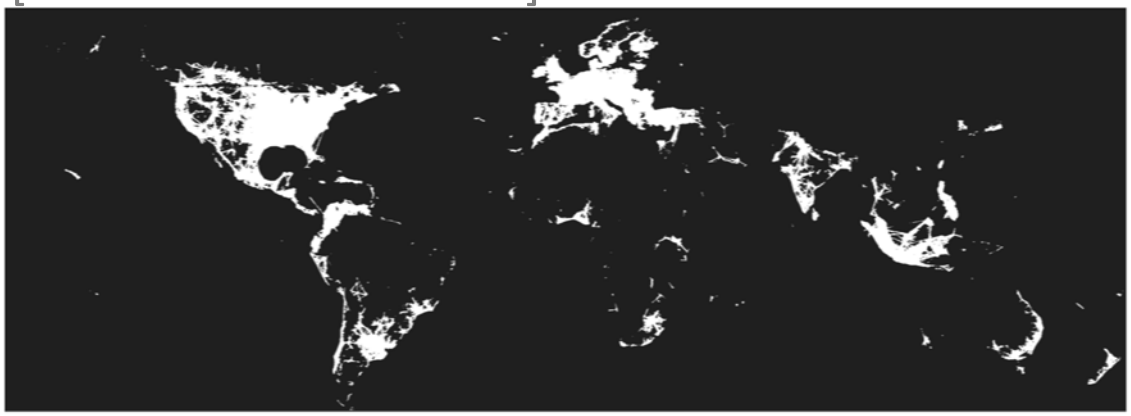
Two Questions

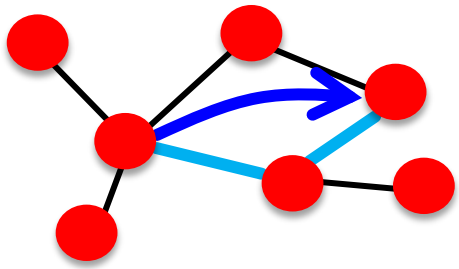
- (Today) What is the structure of a social network?
- (Next class) What kind of mechanisms do people use to route and find the target?



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]





Clustering Implies Edge Locality

- **MSN network has 7 orders of magnitude larger clustering than the corresponding G_{np}!**
- **Other examples:**

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

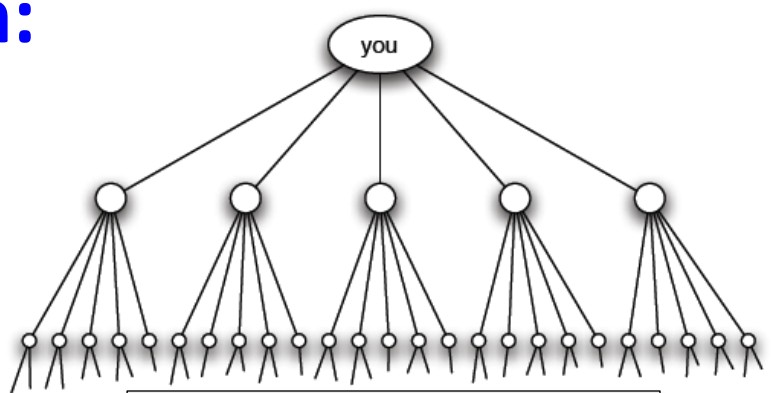
Network	h_{actual}	h_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power Grid	18.70	12.40	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

h ... Average shortest path length
C ... Average clustering coefficient
"actual" … real network
"random" … random graph with same avg. degree

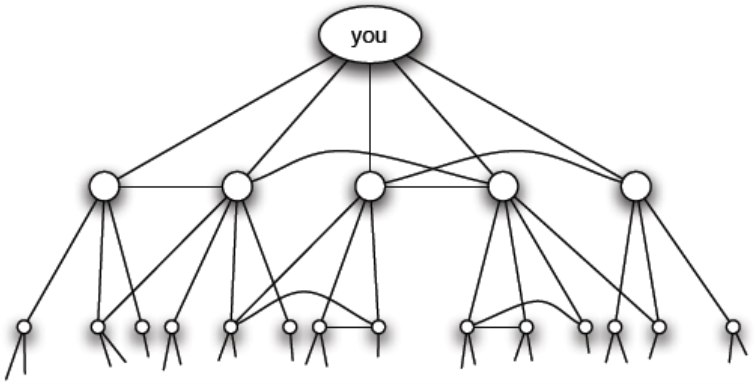
Back to the Small-World

- **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?**



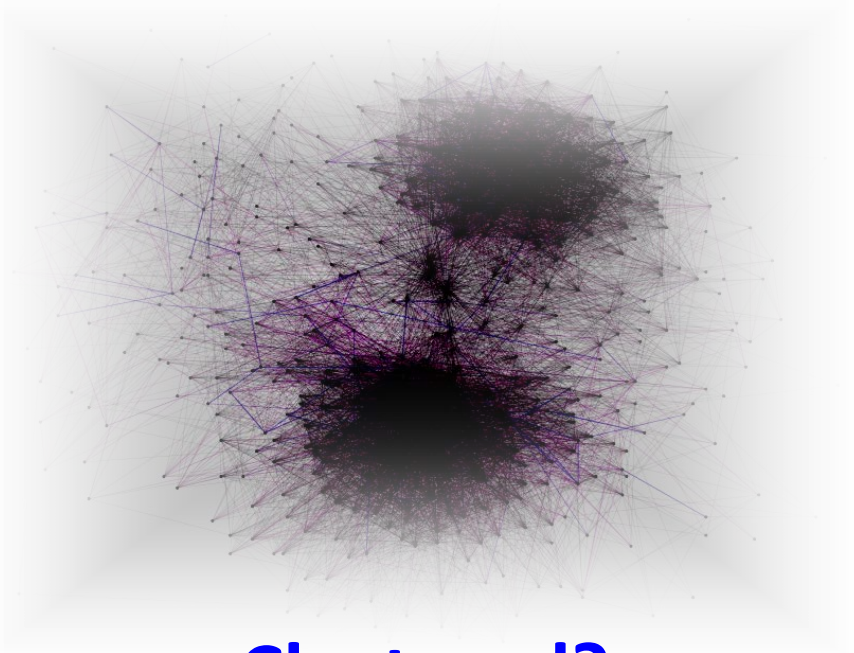
Pure exponential growth



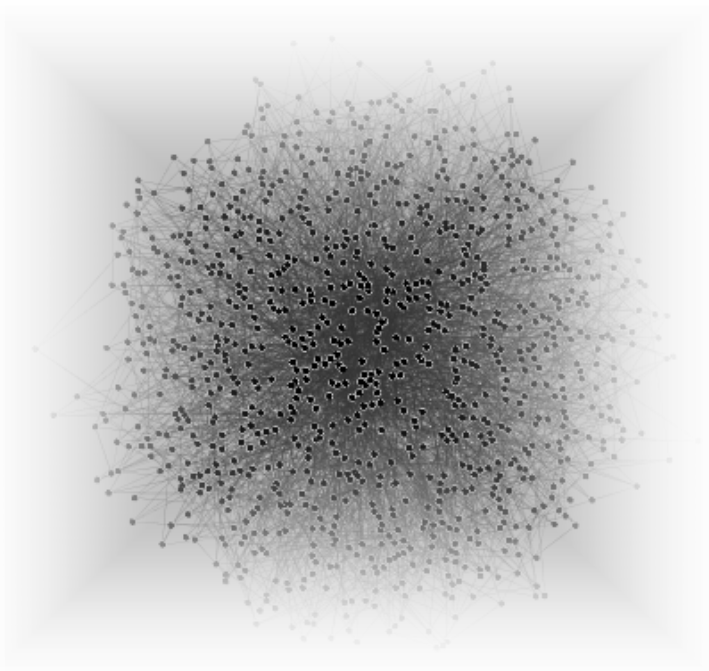
Triadic closure reduces growth rate

Clustering vs. Randomness

Where should we place social networks?



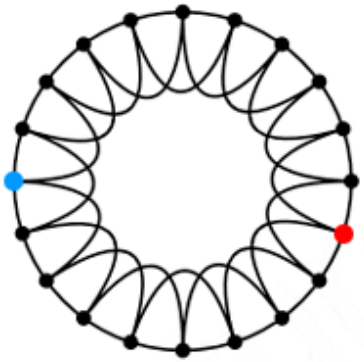
Clustered?



Random?

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"

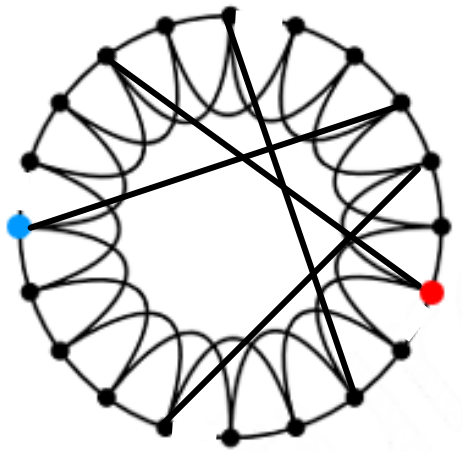
[Watts-Strogatz, '98]

Solution: The Small-World Model

Small-world Model [Watts-Strogatz '98]

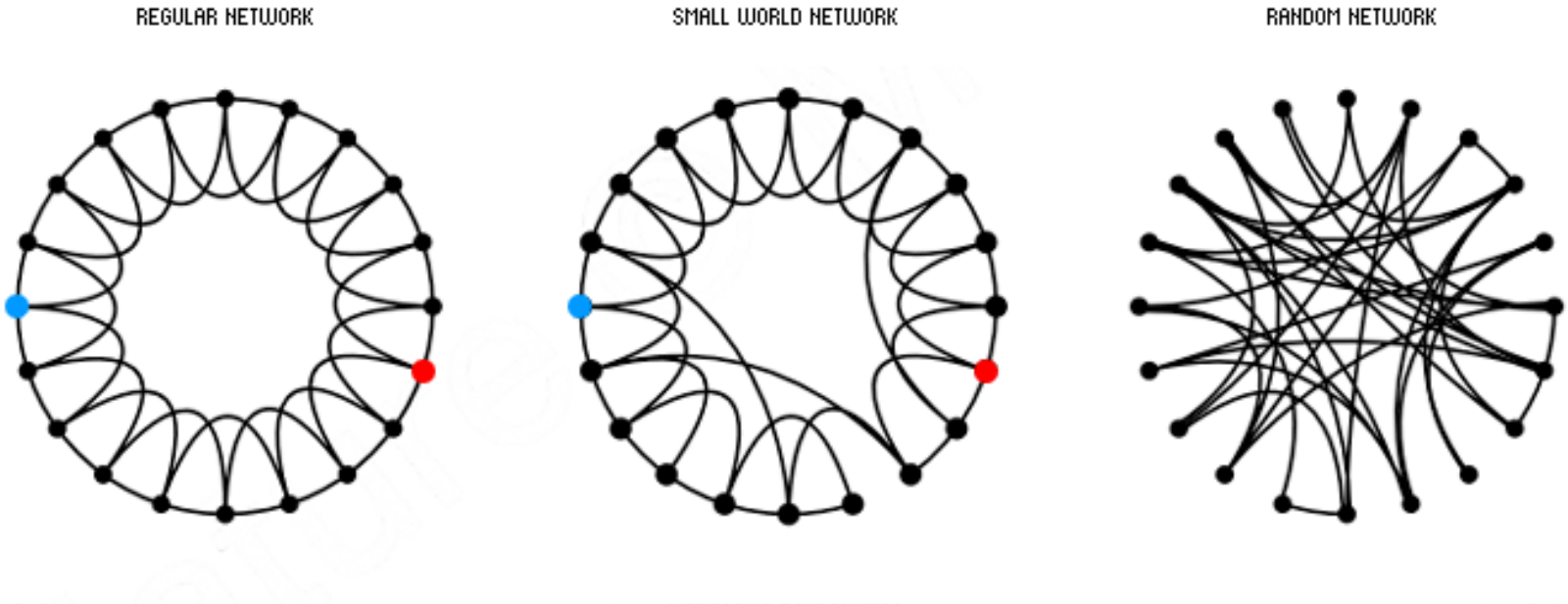
2 components to the model:

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



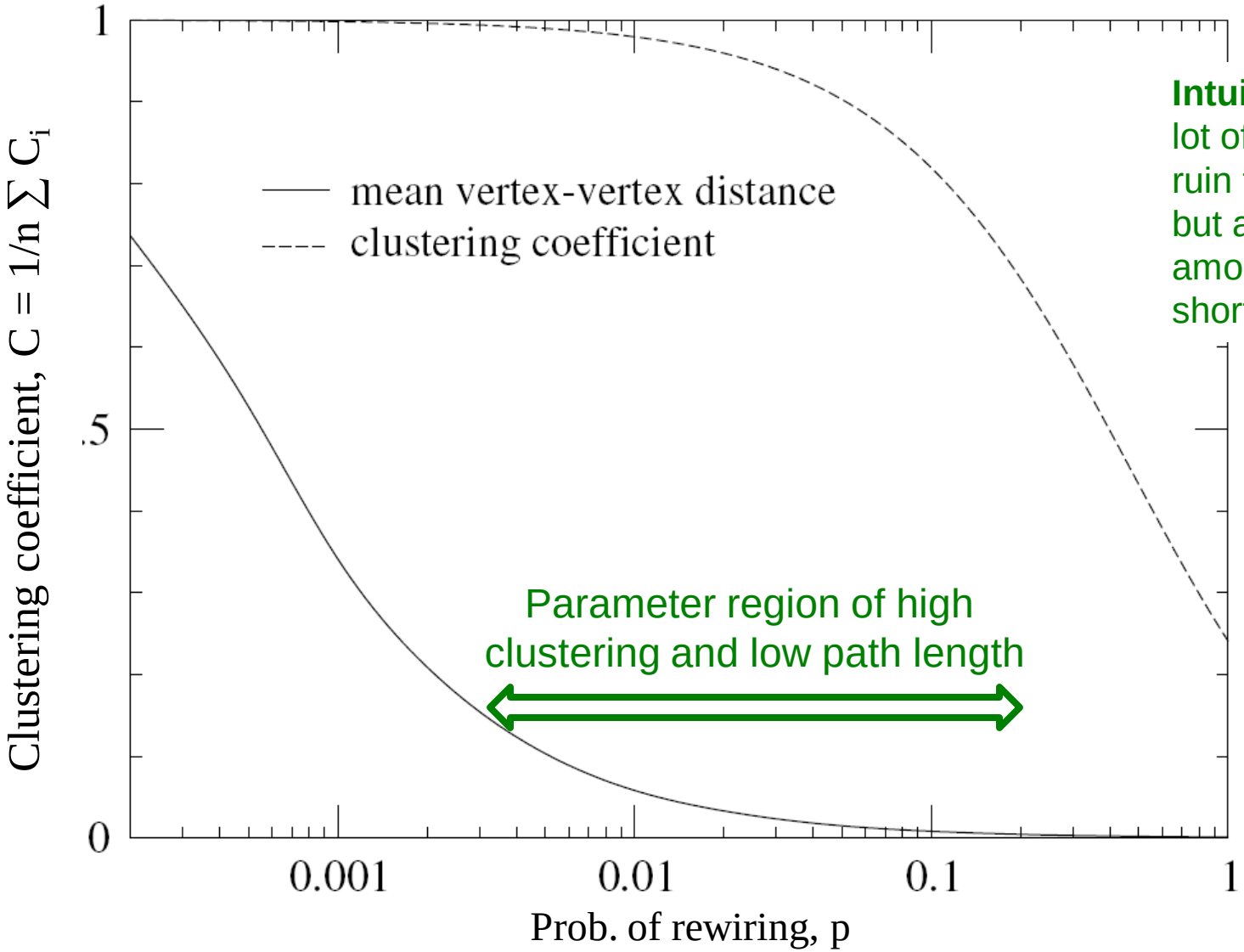
[Watts-Strogatz, '98]

The Small-World Model



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

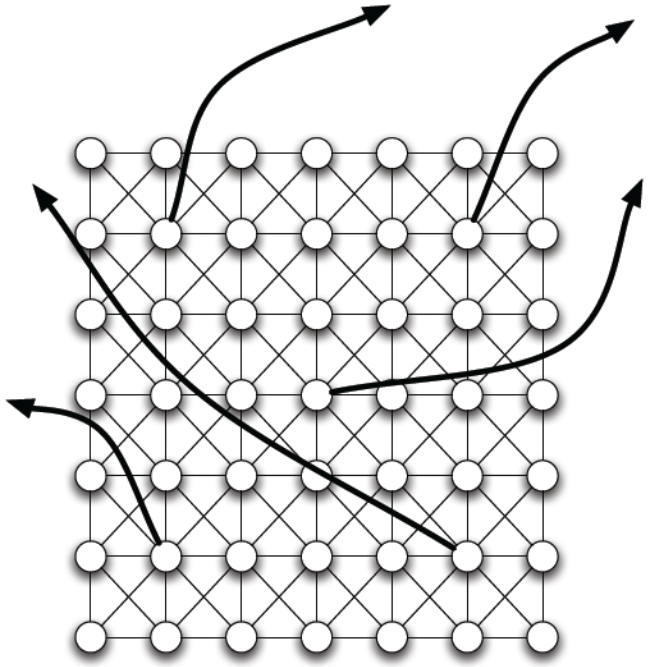
The Small-World Model



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-range edge can only close more.

What's the diameter?

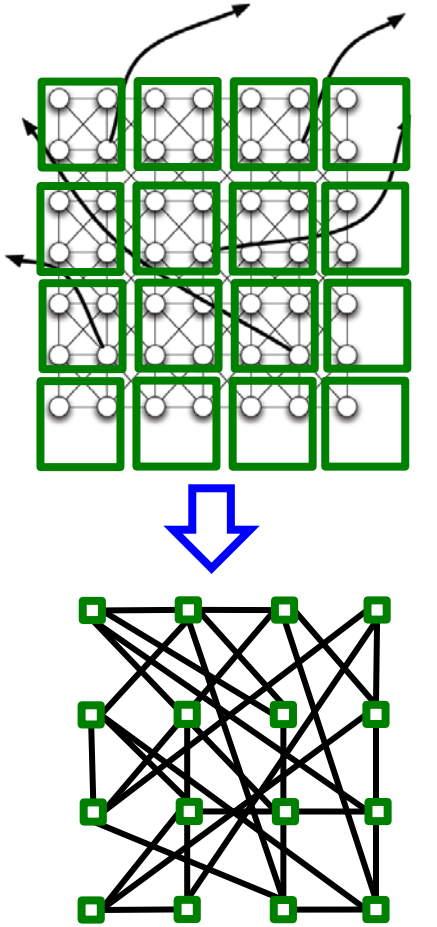
It is *O(log(n))*

Why?

Diameter of the Watts-Strogatz

- **Proof:**
  - 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

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



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 a Network?

- (Next time) **What mechanisms do people use to navigate networks and find the target?**