Network Models

Thanks to Jure Leskovec, Stanford and Panayiotis Tsaparas, Univ. of Ioannina for slides

Agenda

- Erdös-Renyi Random Graph Model
- The Small-World Model
- The Configuration Model

Erdös-Renyi
Random Graph Model

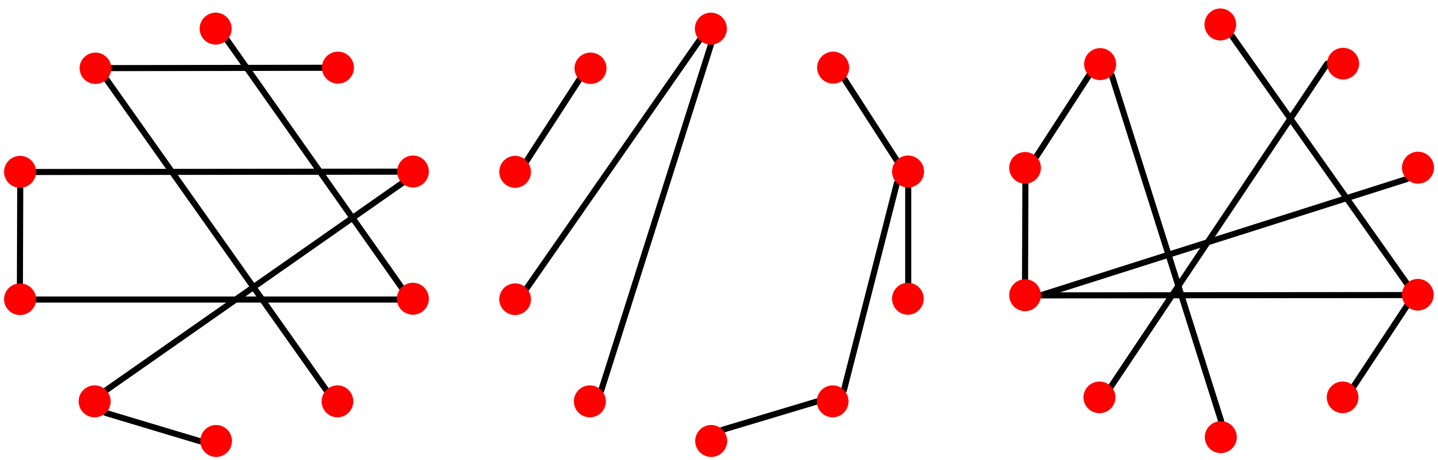
Simplest Model of Graphs

- **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 given the same *n* and *p*



n = 10
p= 1/6

Random Graph Model: Edges

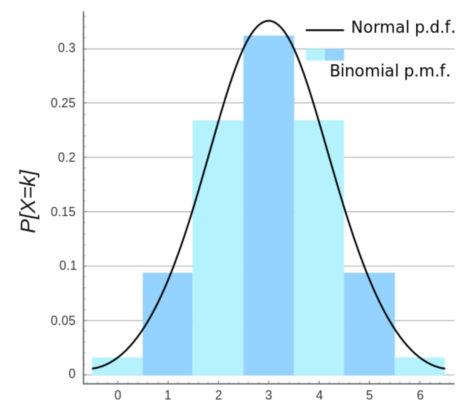
- **How likely is a graph on E edges?**
- $P(E)$: the probability that a given G_{np} generates a graph on exactly E edges:

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

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

P(E) is exactly the
Binomial distribution >>>

Number of successes in a sequence of E_{max} independent yes/no experiments



Properties of G_{np}

Degree distribution: $P(k)$

Path length: h

Clustering coefficient: C

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

Node Degrees in a Random Graph

- **What is expected degree of a node?**
  - Let $\boldsymbol{X_v}$ 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 $Y_1, Y_2, \ldots, Y_k$
      - If $Y=Y_1+Y_2+\ldots Y_k$, then $E[Y]= \sum_i E[Y_i]$
- **An easier way:**
  - Decompose X_v to $X_v= X_{v,1}+X_{v,2}+\ldots+X_{v,n-1}$
    - where $\boldsymbol{X_{v,u}}$ 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: $p(n-1)$

Degree Distribution

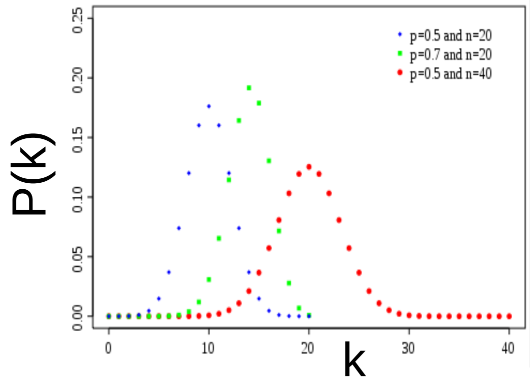
- **Fact:** **Degree distribution of G_{np} is <u>Binomial</u>.**
- Let $P(k)$ denote a fraction of nodes with degree k:

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

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

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

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

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

By the law of large numbers, 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 k.

Clustering Coefficient of G_{np}

- **Remember:** $C_i = \dfrac{2e_i}{k_i(k_i - 1)}$

Where e_i is the number of edges between i's neighbors

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

Each pair is connected with prob. p

Number of distinct pairs of neighbors of node i of degree k_i

- **Then:** $C_i = \dfrac{p \cdot k_i(k_i - 1)}{k_i(k_i - 1)} = p = \dfrac{\bar{k}}{n-1} \approx \dfrac{\bar{k}}{n}$

Clustering coefficient of a random graph is small.
For a fixed avg. degree (that is p=1/n), *C* decreases with the graph size *n*.

Network Properties of G_{np}

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

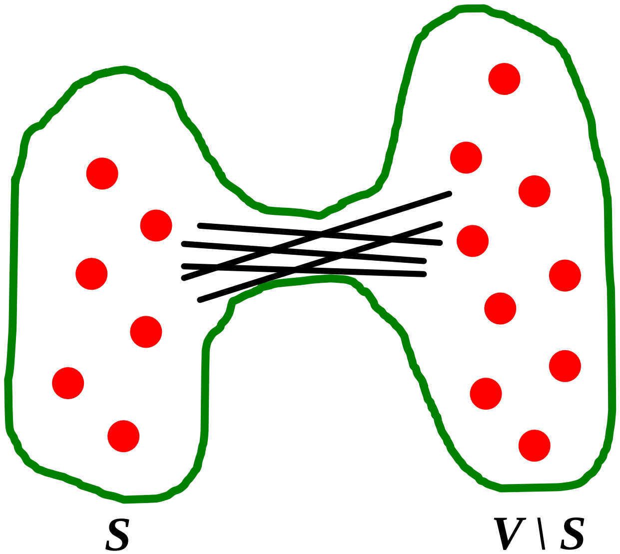
Clustering coefficient: $C = p = \bar{k}/n$

Path length: ***next!***

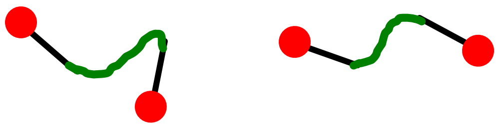
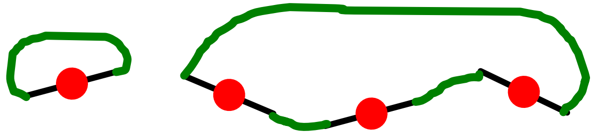
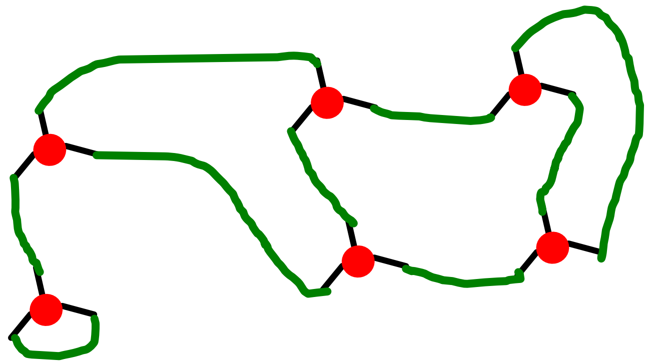
Def: Random k-Regular Graphs

- **We need to define two concepts**
- **1) Define: Random k-Regular graph**
  - Assume each node has k spokes (half-edges)
  - Randomly pair them up!
- **2) Define: Expansion**
  - Graph $G(V, E)$ has **expansion α**: if $\forall S \subseteq V$: #edges leaving S $\geq \alpha \cdot \min(|S|, |V \setminus S|)$
  - **Or equivalently:**

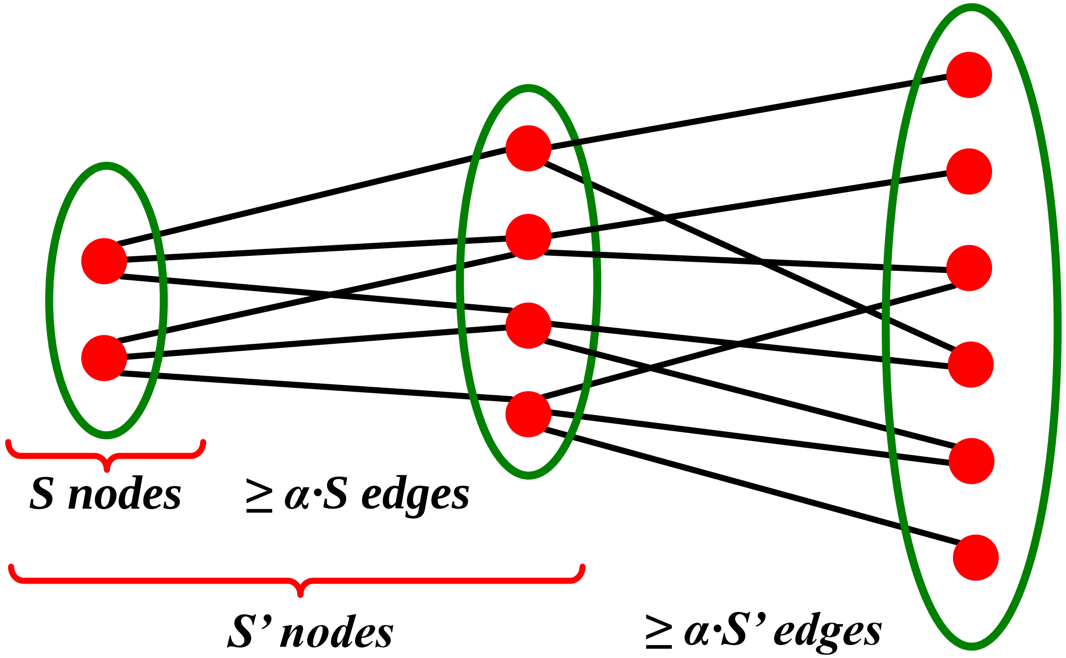
$$\alpha = \min_{S \subseteq V} \frac{\#\, edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$



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!**

Expansion: Intuition



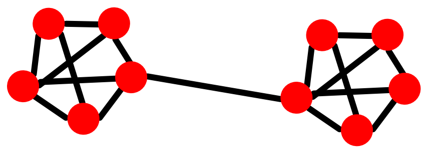
$$\alpha = \min_{S \subseteq V} \frac{\# edges\ leaving\ S}{\min(|S|, |V \setminus S|)}$$

(A big) graph with "good" expansion

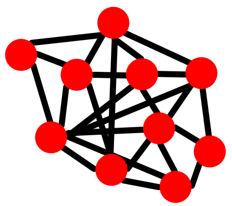
Expansion: Measures Robustness

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

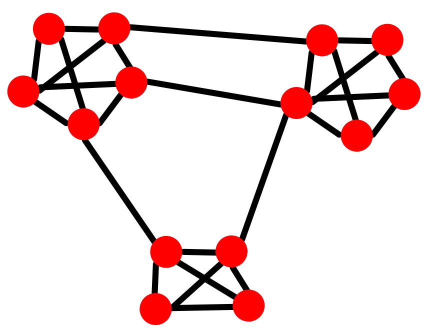
- Expansion is **measure of robustness:**
  - To disconnect l nodes, we need to cut $\geq \alpha \cdot l$ edges
- Low expansion:



- High expansion:



- Social networks:
  - "Communities"



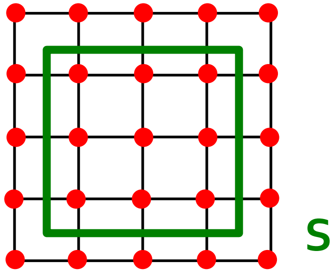
Expansion: k-Regular Graphs

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

- ***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?**

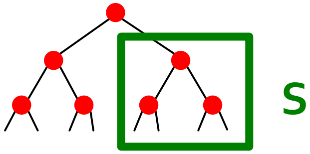
Examples:

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



Make this into 6x6 grid!

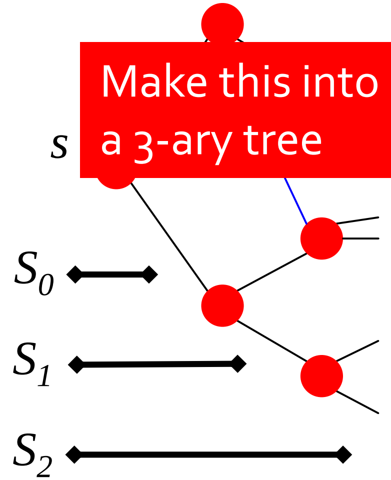
- **Complete binary tree:**
  $\alpha \to 0$ for $|S| = (n/2)-1$



- **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 $\geq \alpha$ (In fact, $\alpha = d/2$ as $d \to \infty$)

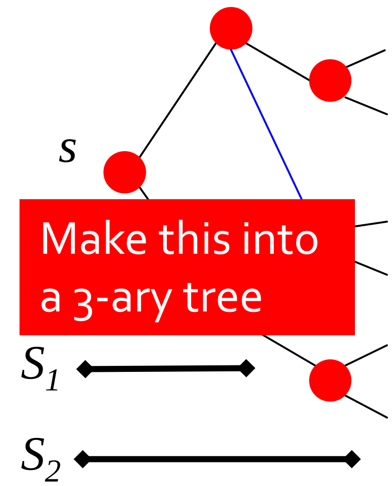
Diameter of 3-Regular Rnd. Graph

- **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

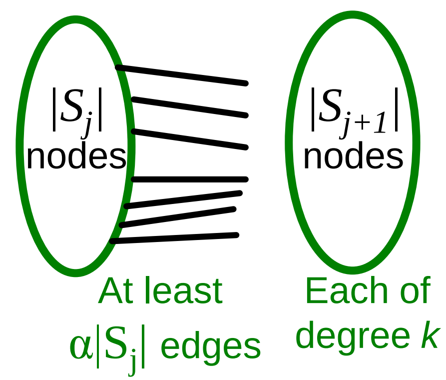
- <u>Proof (continued):</u>
  - Let S_j be a set of all nodes found within j steps of BFS from s.
  - **We want to relate S_j and S_{j+1}**



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

$$|S_{j+1}| \geq |S_j|\left(1+\frac{\alpha}{k}\right) = S_0\left(1+\frac{\alpha}{k}\right)^{j+1}$$

where $S_0=1$

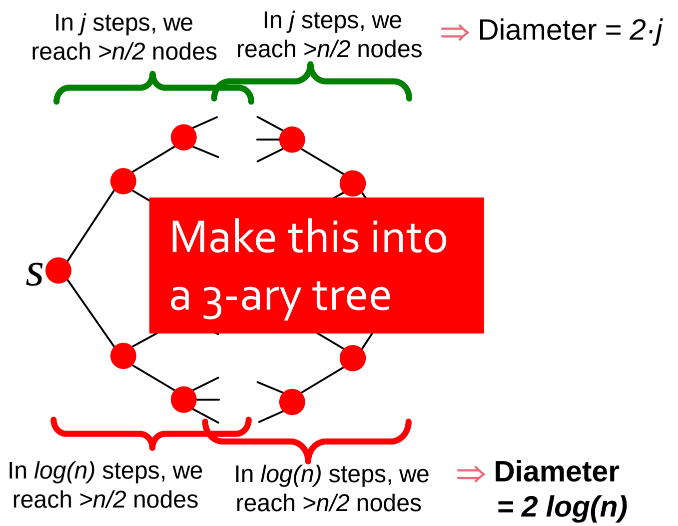


Diameter of 3-Regular Rnd. Graph

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

- <u>Proof (continued):</u>
  - **In how many steps of BFS do 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:

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



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

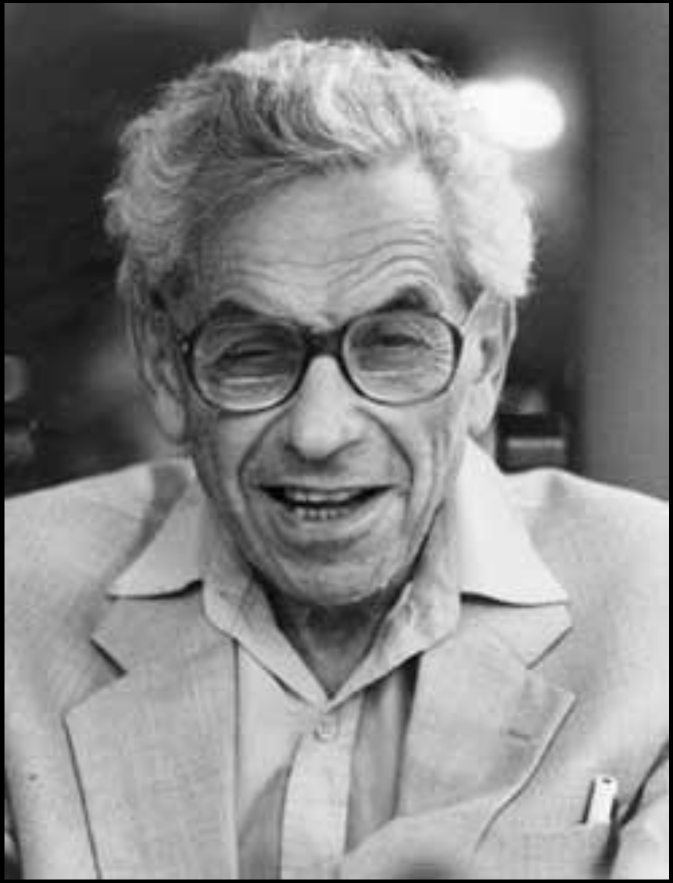
  - In $2k/\alpha \cdot \log n$ steps $|S_j|$ grows to $\Theta(n)$. So, **the diameter of *G* is $O(\log(n)/\alpha)$**

Network Properties of G_{np}

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

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

Path length: $$O(\log n)$$



Paul Erdös

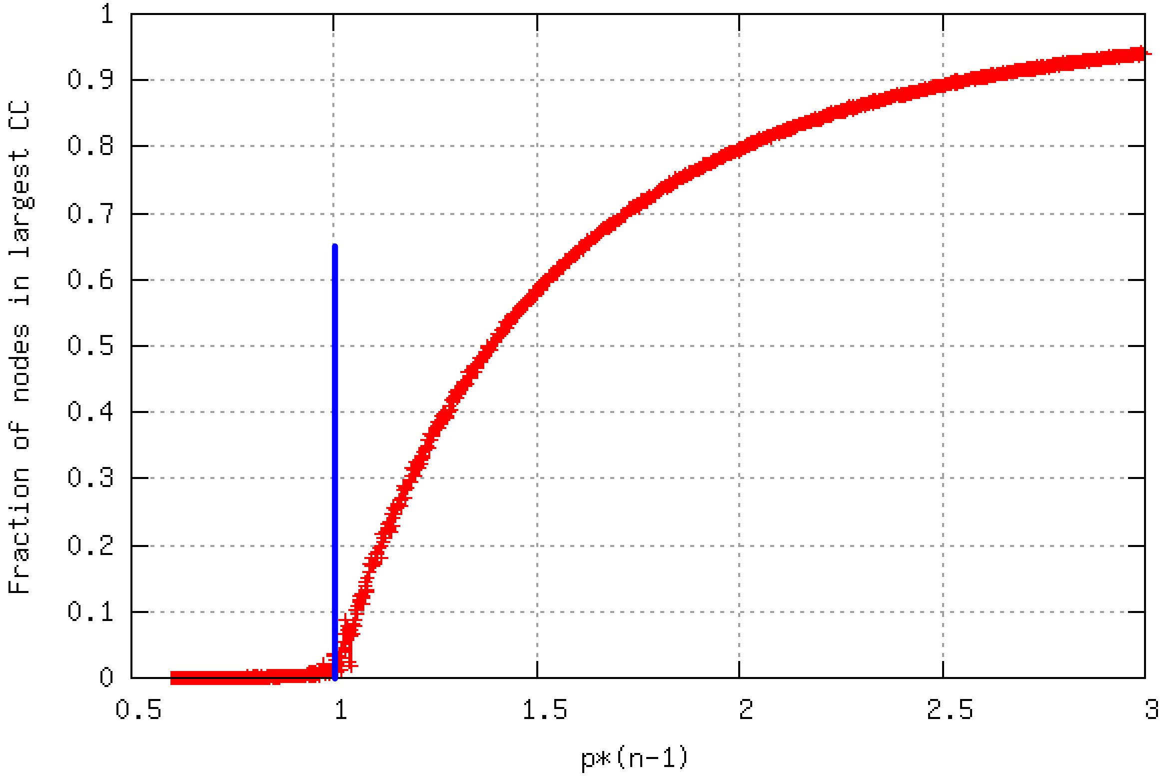
G_{np} is so cool!
Let's also look at its evolution

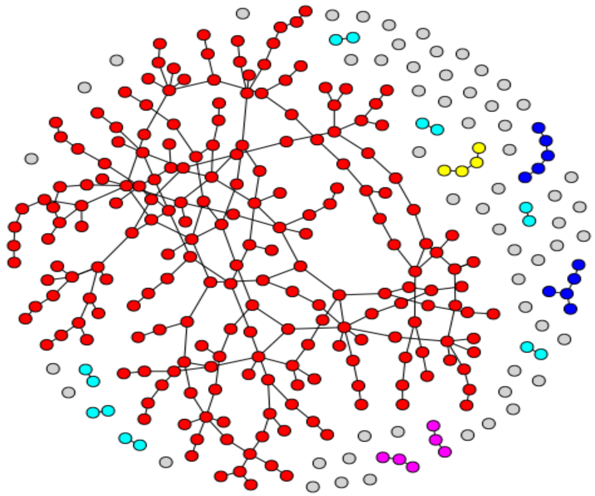
"Evolution" of a Random Graph

- **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

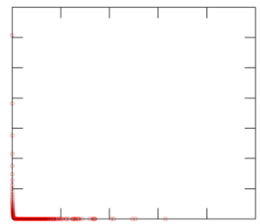
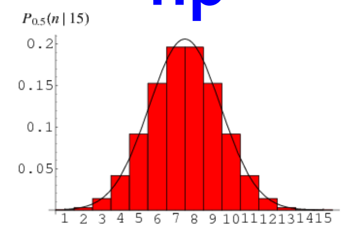




Fraction of nodes in the largest component

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

Back to MSN vs. G_{np}

	MSN	G_{np}	n=180M
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%*	GCC exists when $\bar{k}$>1. $\bar{k} \approx 14$.	✓

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!**

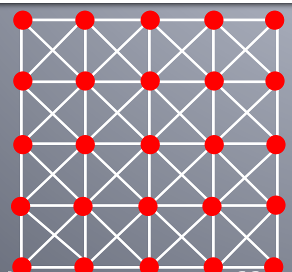
Real Networks vs. G_{np}

- **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!

The Small-World Model

Can we have high clustering while also having short paths?



High clustering coefficient,
High diameter

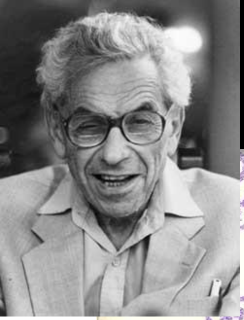
Vs.

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

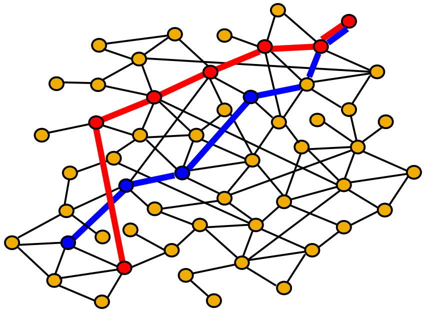


Erdös numbers are small!
(Jure’s number is 3.)

Find out your Erdos number: http://www.ams.org/mathscinet/collaborationDistance.html

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?**

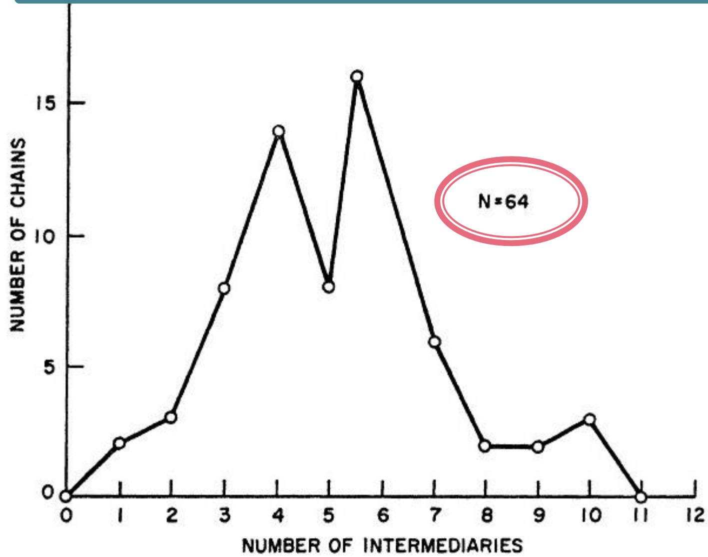


The Small-World Experiment

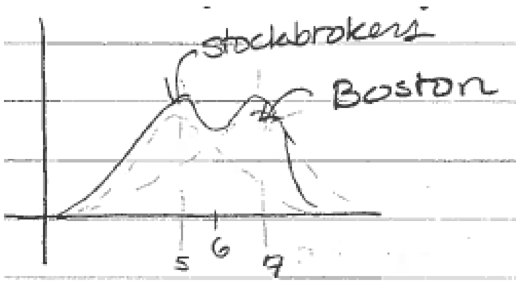
[Milgram, '67]

- **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 who owned stock had shorter paths to the stockbroker than random people: 5.4 vs. 6.7
  - People from the Boston area have even closer paths: 4.4

Milgram's small world experiment



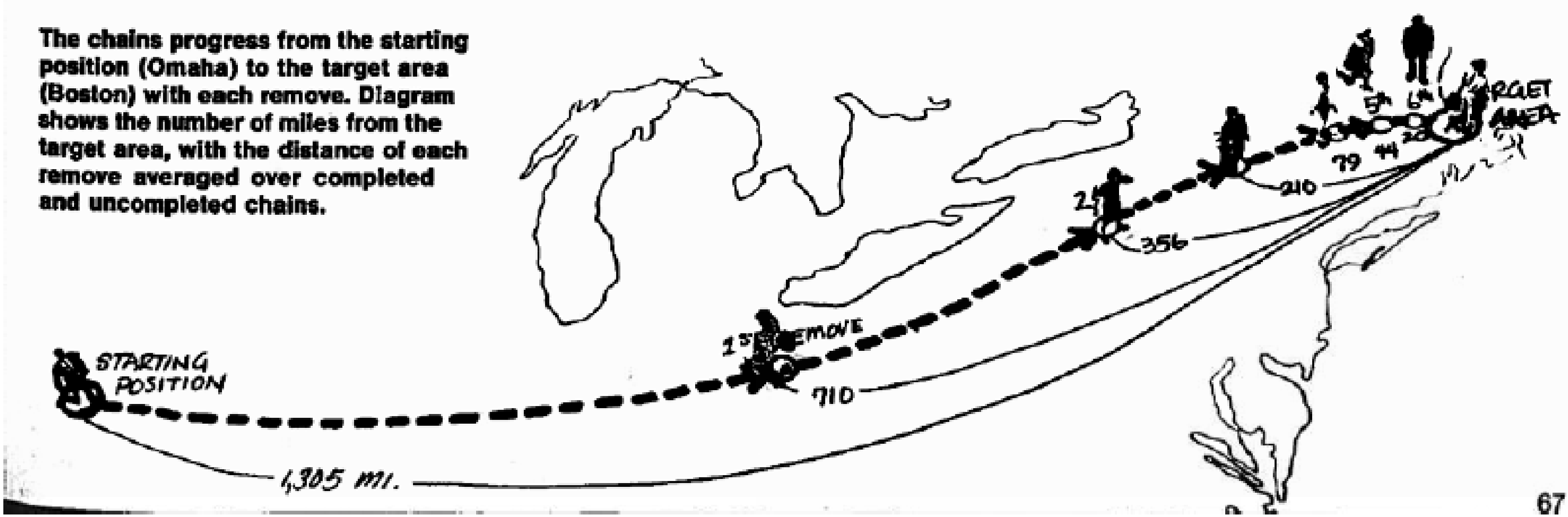
Milgram: Further Observations



- **Boston vs. occupation networks:**
- **Criticism:**
  - **Funneling:**
    - 31 of 64 chains passed through 1 of 3 people as their final step → Not all links/nodes are equal
  - Starting points and the target were non-random
  - There are not many samples (only 64)
  - People refused to participate (25% for Milgram)
    - Not all searches finished (only 64 out of 300)
  - **Some sort of social search:** People in the experiment follow some strategy instead of forwarding the letter to everyone. **They are not finding the shortest path!**
  - People might have used extra information resources

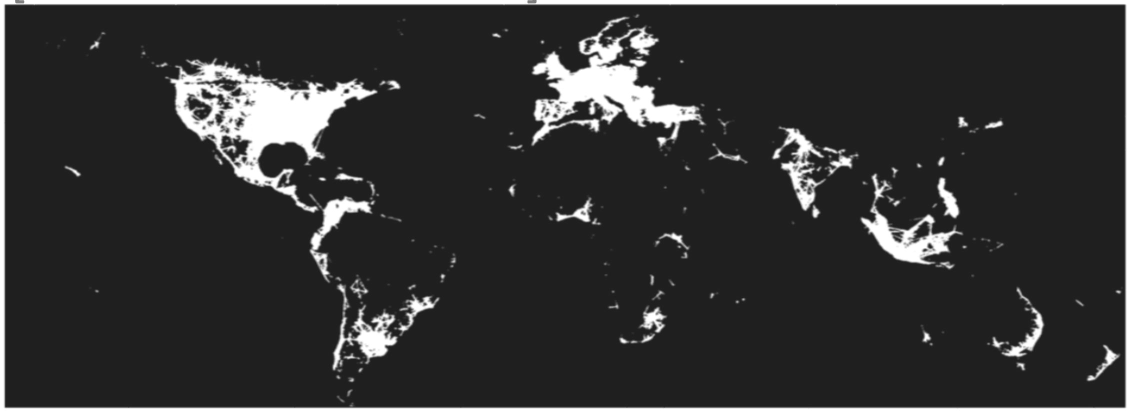
Two Questions

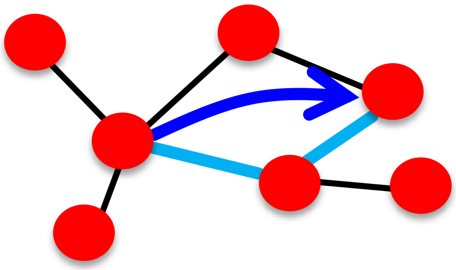
- (Today) What is the structure of a social network?
- (offline) What kind of mechanisms do people use to route and find the target?



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:**

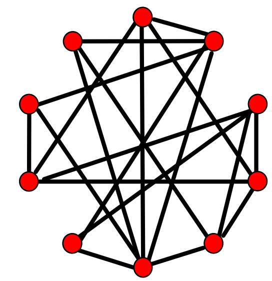
Actor Collaborations (IMDB): N = 225,226 nodes, avg. degree $\bar{k}$ = 61
Electrical power grid: N = 4,941 nodes, $\bar{k}$ = 2.67
Network of neurons: N = 282 nodes, $\bar{k}$ = 14

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

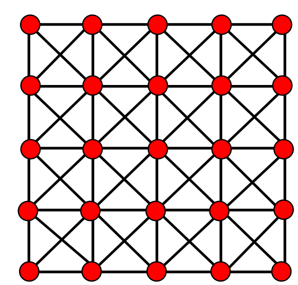
h ... Average shortest path length
C ... Average clustering coefficient
"actual" … real network
"random" … random graph with same avg. degree

The "Controversy"

- **Consequence of expansion:**
  - **Short paths: *O(log n)***
    - This is "best" we can do if we have a constant degree
  - But clustering is low!
- **But networks have "local" structure:**
  - **Triadic closure:**
    Friend of a friend is my friend
  - High clustering but diameter is also high
- **How can we have both?**



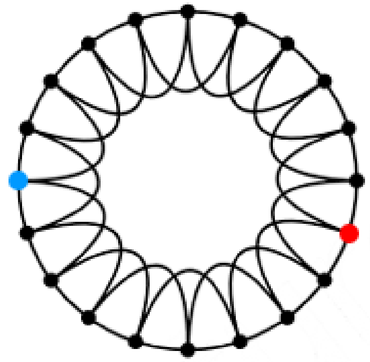
Low diameter
Low clustering coefficient



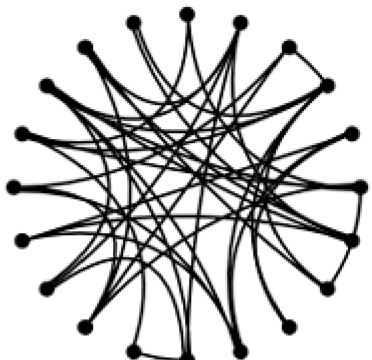
High clustering coefficient
High diameter

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



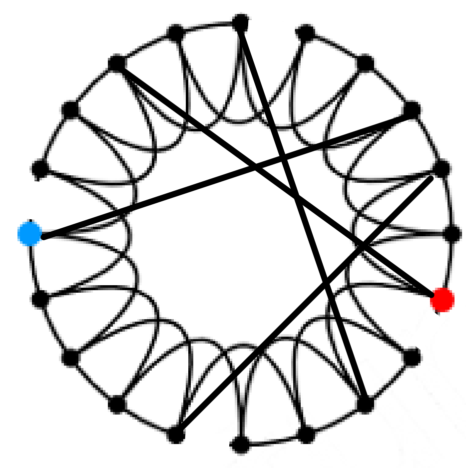
Low clustering
Low diameter

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

Solution: The Small-World Model

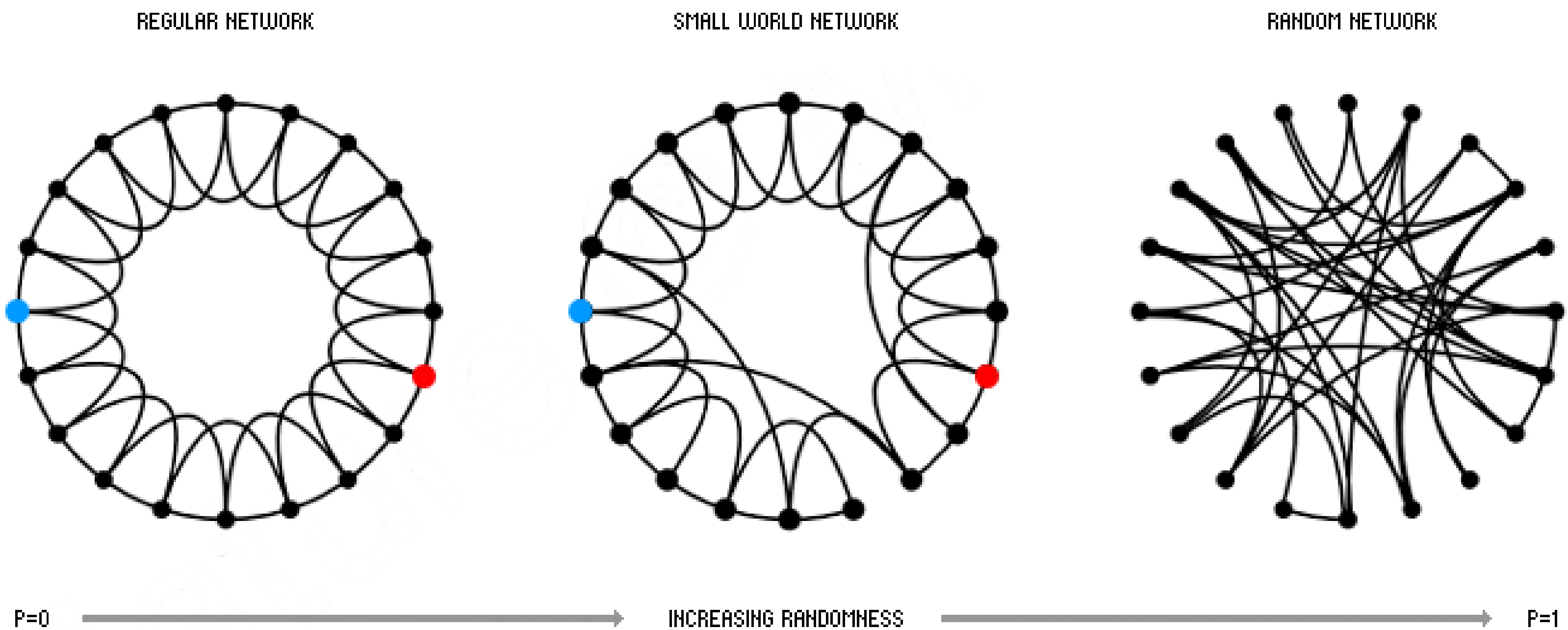
Small-world Model [Watts-Strogatz '98]
Two components to the model:

- **(1)** Start with a **low-dimensional regular lattice**
  - (In our case we using a ring as a 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



[Watts-Strogatz, ‘98]

The Small-World Model



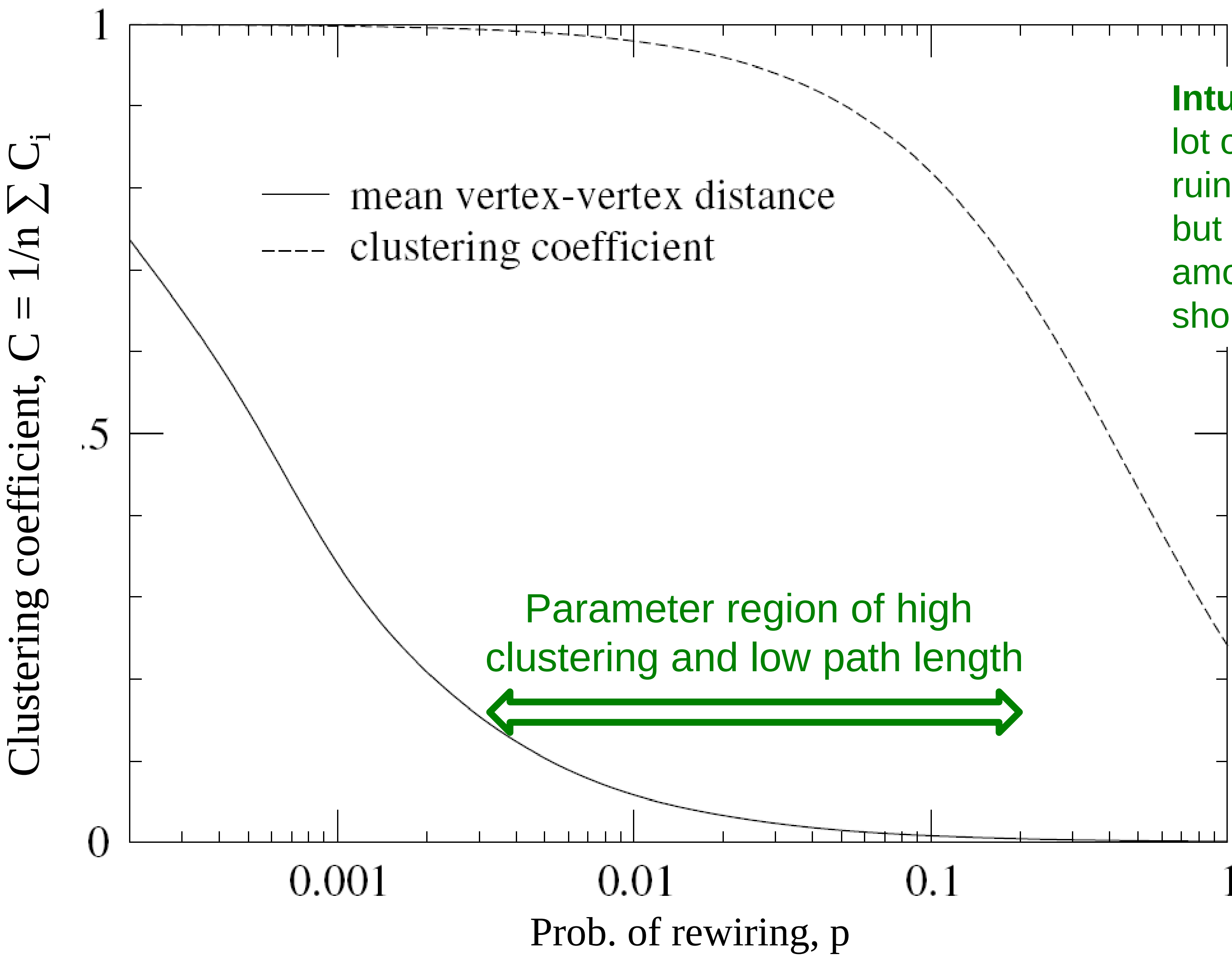
Regular network	Small world network	Random network
High clustering High diameter	High clustering Low diameter	Low clustering Low diameter

$$h = \frac{N}{2\bar{k}} \qquad C = \frac{3}{4}$$

$$h = \frac{\log N}{\log \alpha} \qquad C = \frac{\bar{k}}{N}$$

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

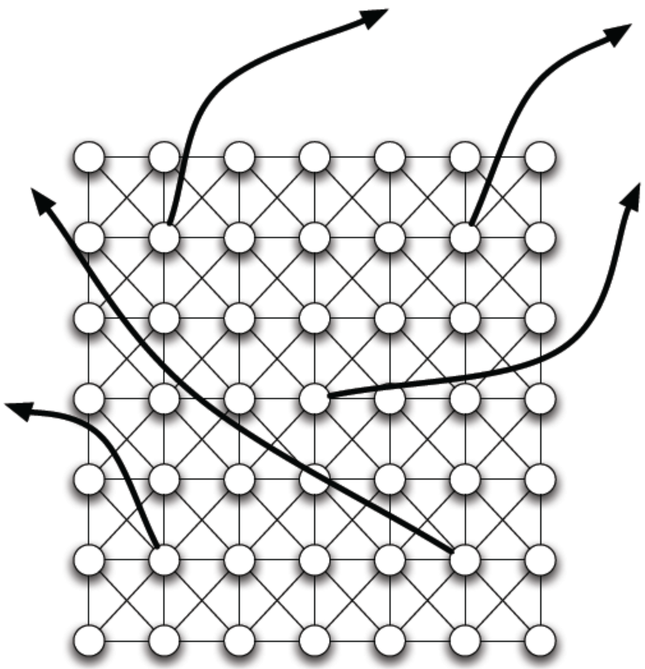
The Small-World Model



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.



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

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

What's the diameter?

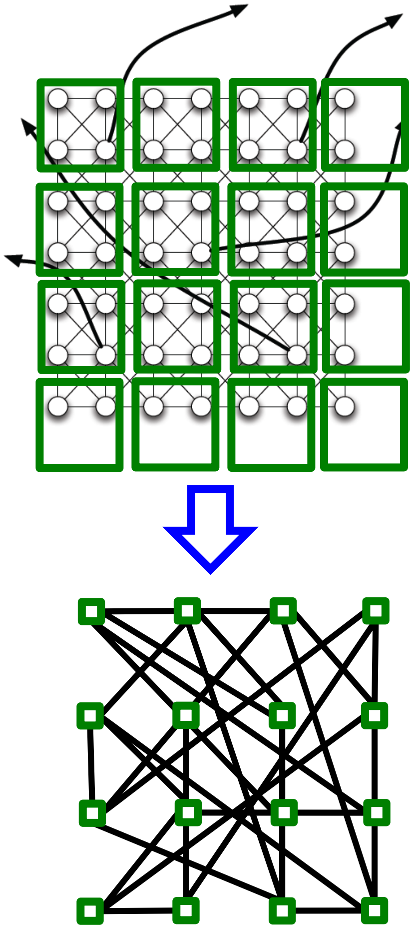
It is $O(\log(n))$

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 long range edge (by having to traverse internal nodes)

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



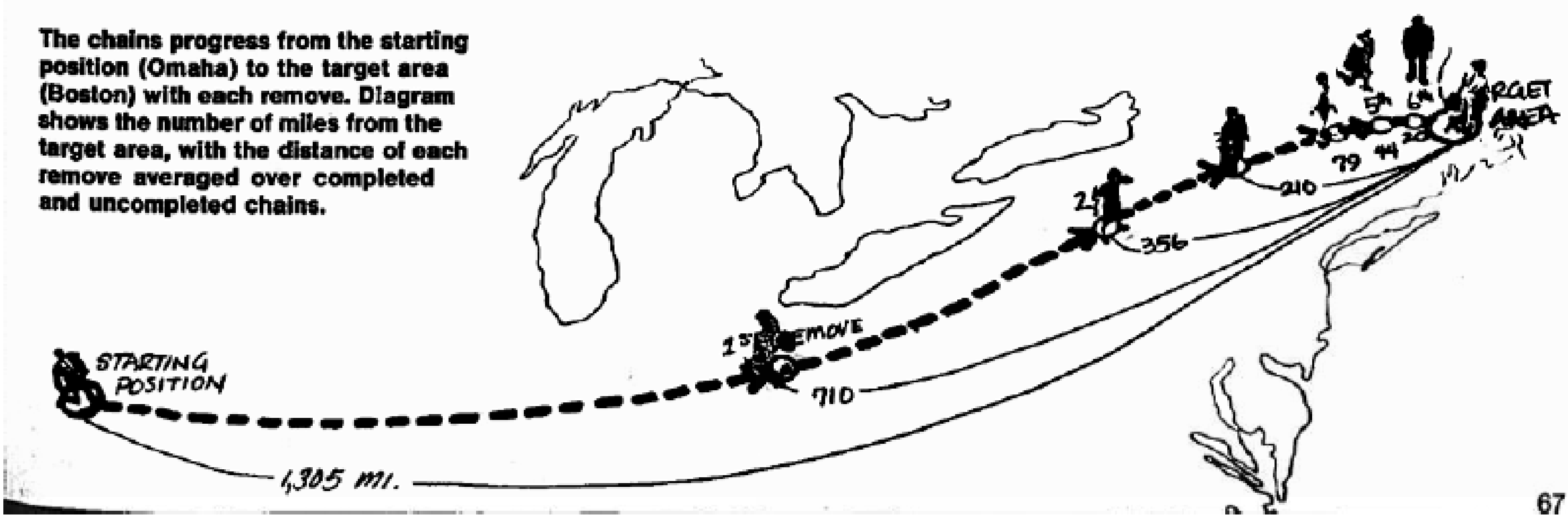
4-regular random graph

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** **(offline lecture)**

How to Navigate a Network?

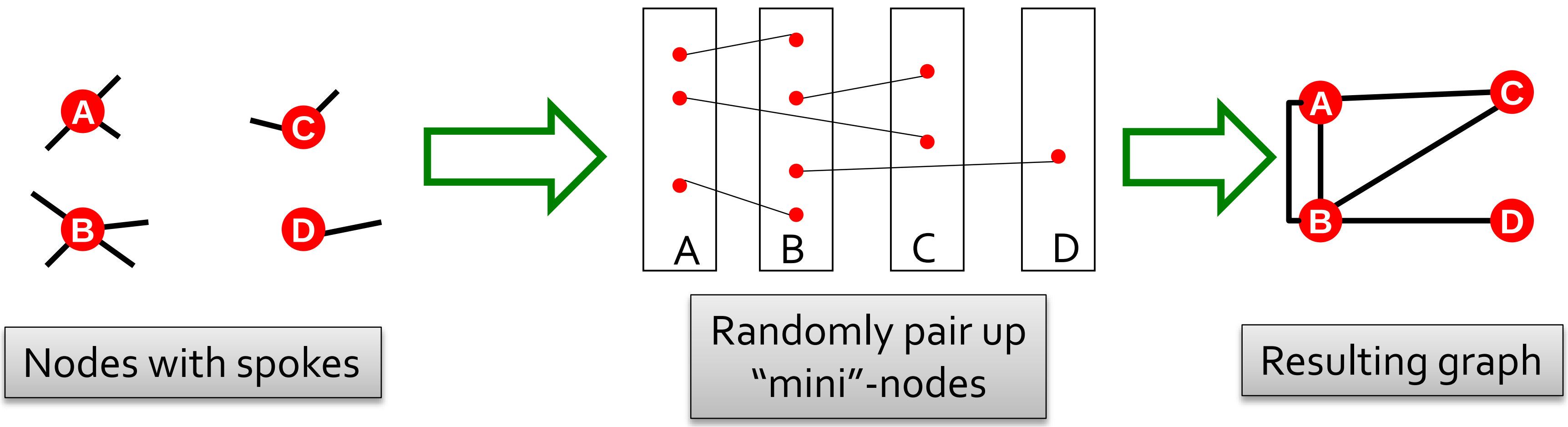
- (offline) **What mechanisms do people use to navigate networks and find the target?**



The Configuration Model

Intermezzo: Configuration Model

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



- **Useful 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