# Graph Models

Sarunas Girdzijauskas

ID2211

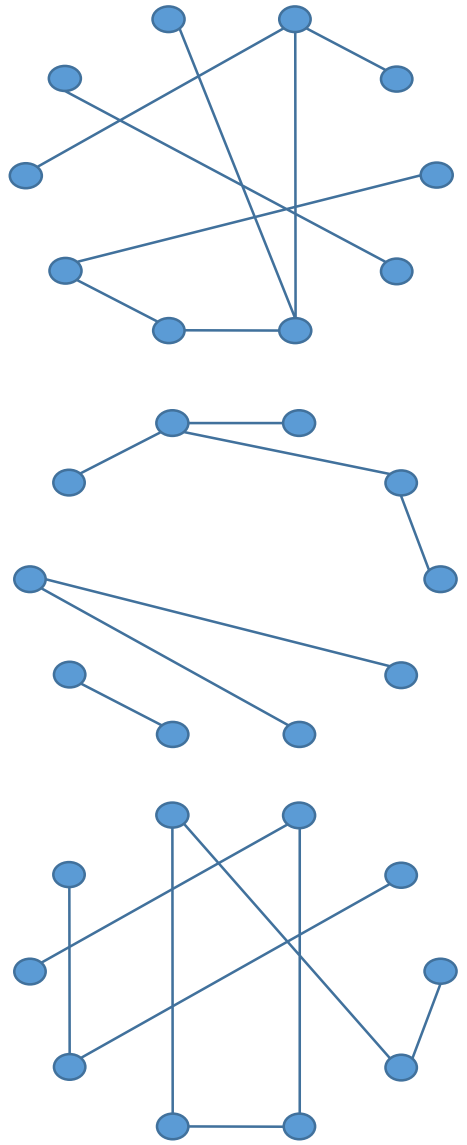
March 2019

# Recap

- Basic Notations
- Type of graphs
- Paths/cycles/Connectivity/Giant component
- Centrality measures
- Metrics Comparison
- Clustering Coefficient
- Degree Distributions
- Real Networks
- Graph Models: Random Graph

# Models of Graphs: Random Graphs

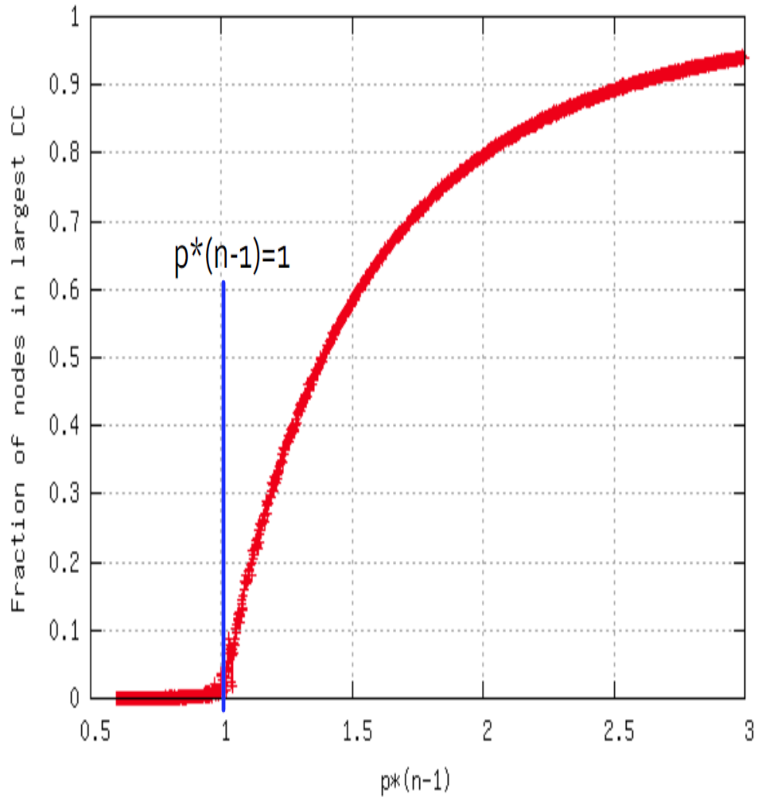
- **G(n,m) model**
  - Start with n isolated vertices
  - Place m edges among them at random.
  - G(n,m) defines a family of graphs (not a particular graph)
- **G(n, p) model (Erdos-Renyi random graph)**
  - Start with n isolated vertices
  - We place and edge between each distinct vertex pair with probability p.
  - n and p do not uniquely determine the graph! It is stochastic!

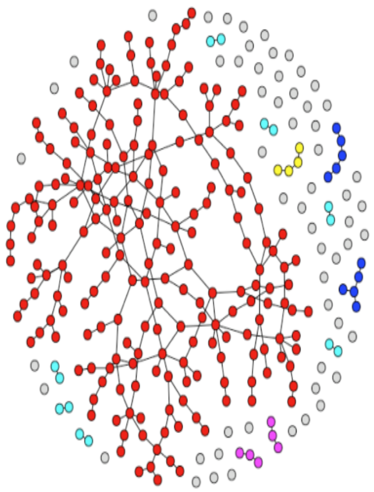


n = 10
p= 1/6

# Erdos-Renyi random graph

- What can you say about the graph when we move p from 0 to 1?
  - Diameter?
  - How big is a giant component when p=0 and when p=1?
  - How does the giant component grow inbetween those p values?
    - Network undergoes “***phase transition***”



Fraction of nodes in the largest component

# Erdös-Renyi random graph (cont.)

as N becomes large:

If $p < 1/N$, probability of a giant component goes to 0

If $p > 1/N$, probability of a giant component goes to 1, and all other components will have size at most log(N)

- at $p \sim 1/N$, average degree is ~ 1 (sparse graph)
  - When $p \sim 2*\log(N)/N$ – no isolated nodes
  - If we force each node to have at least 3 neighbors then the graph is connected a.a.s.
- Any monotone property exhibits **Threshold phenomena** in Erdos-Renyi with respect to p.
- E.g., network has a cycle of at least K vertices.
- Diameter: approx log(N)/log(d)
- Average degree?: ?
- Clustering coefficient?: ?
- Degree distribution?: ?
- **Giant Component Example**

# Degree Distribution

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

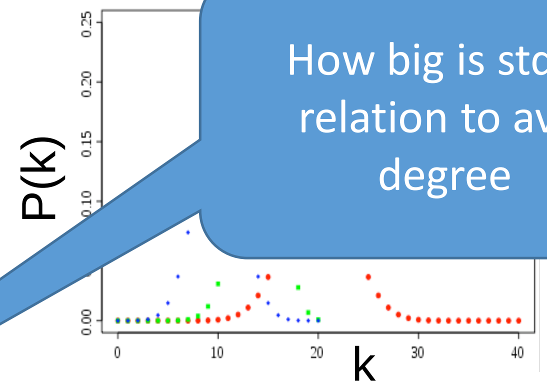
This is a probability mass function of a binomial distribution

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

Select $k$ nodes out of $n-1$

Probability of having $k$ edges

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




How big is std in relation to avg. degree

**Mean, variance of a binomial distribution**

Average degree

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

Variance

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

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

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 $k$.

We can assume that for very large N and fixed p each node eds up with almost the same degree

# Clustering Coefficient of $G_{np}$

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

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:** $$e_i = p\frac{k_i(k_i - 1)}{2}$$

Each pair is connected with prob. $p$

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

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

For very large graphs with fixed degree the clustering coefficient goes to zero

Clustering coefficient of a random graph is small.
If we generate bigger and bigger graphs with fixed avg. degree k (that is we set $p = k \cdot 1/n$), then $C$ decreases with the graph size $n$.

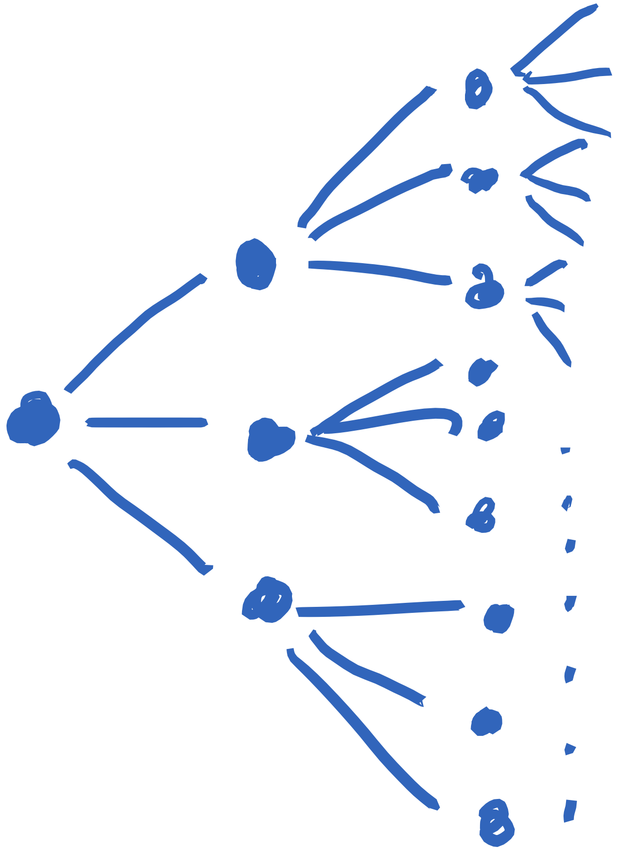
# Network Properties of $G_{np}$

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

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

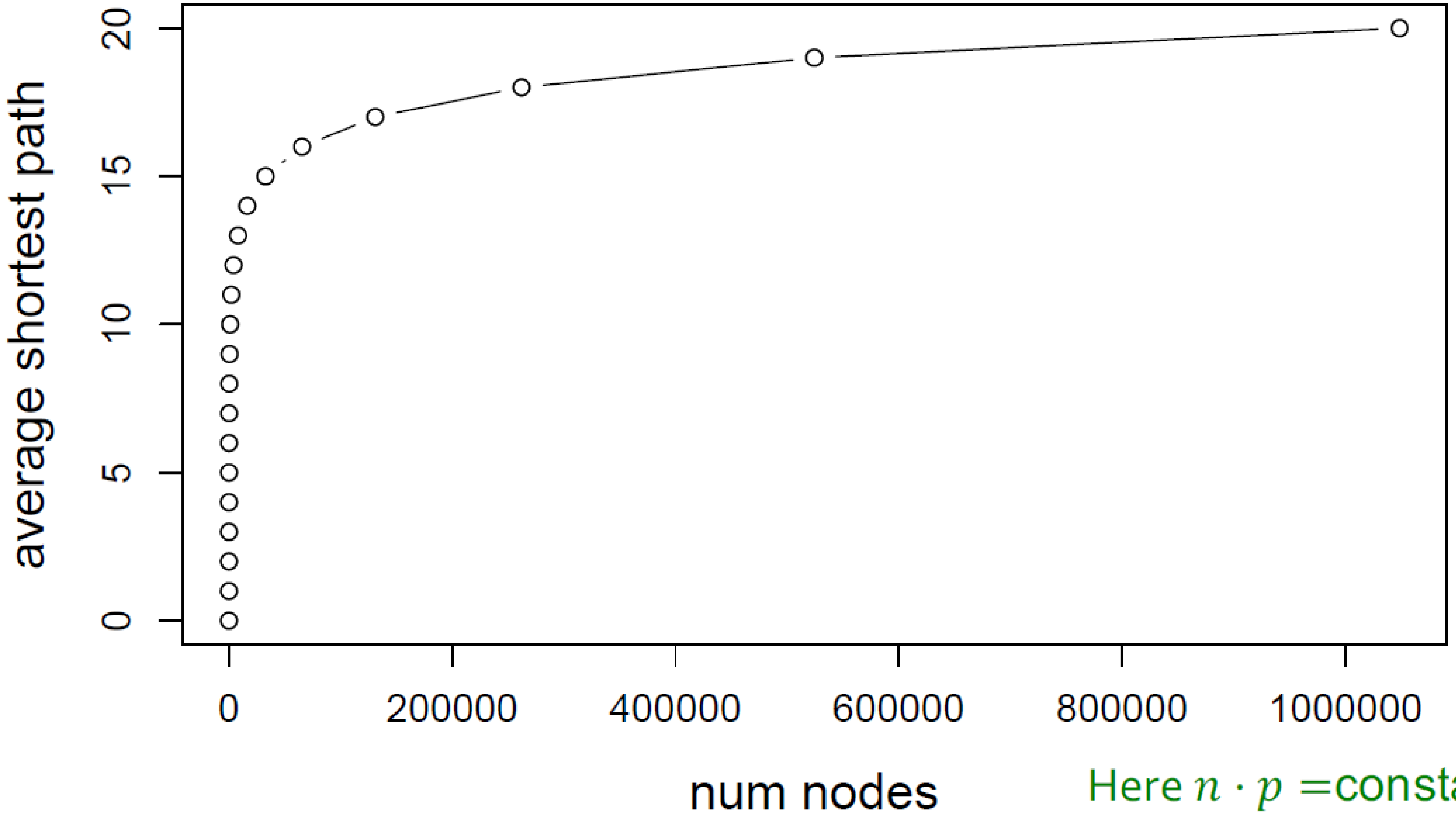
**Path length:** ***next!***

# Intuition on Diameter Calculation



- Take a tree with degree d
- Nodes reached in
  - $1^{st}$ step 1+d
  - $2^{nd}$ step $1+d+d^2$
  - $3^{rd}$ step $1+d+d^2+d^3$
  - $k^{th}$ step $1+d+d^2+d^3+\ldots+d^k \sim d^k$
- When do we reach N nodes?
  - $N=d^k$
  - $k=\log_d N$
- Intuition for Random graphs
  - Most of the degrees in random graph are between 3d and d/3 (chernoff bounds), so log(3d) almost equal to log(d/3) and equal to log(d)
  - Very few neighbors are neighbors themselves, so we do not hit many "covered" nodes.
    - Random Graphs are good expanders (later about that)
    - Most of the nodes are hit on the last step.

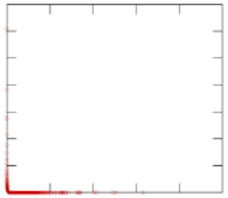
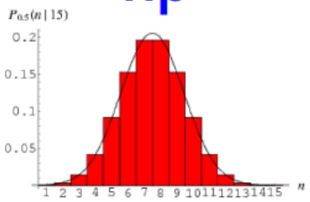
# Diameter in large ER graphs




Here $n \cdot p =$constant
That is, avg deg $k$ is const

# Can Erdos-Renyi explain real-world networks?

- MSN network vs. Erdos Renyi
  - Degree distribution?
  - Avg. Path length?
  - Avg. Clustering coef.?
  - Largest connected component?

| | MSN | $G_{np}$ | n=180M |
|---|---|---|---|
| Degree distribution? |  |  | ✗ |
| Avg. Path length? | **6.6** | $O(\log n)$ <br> $h \approx 8.2$ | ✓ |
| Avg. Clustering coef.? | $0.11$ | $\bar{k}/n$ <br> $C \approx 8\cdot 10^{-8}$ | ✗ |
| Largest connected component? | $99\%$ | GCC exists when $\bar{k}>1$. <br> $\bar{k} \approx 14$. | ✓ |

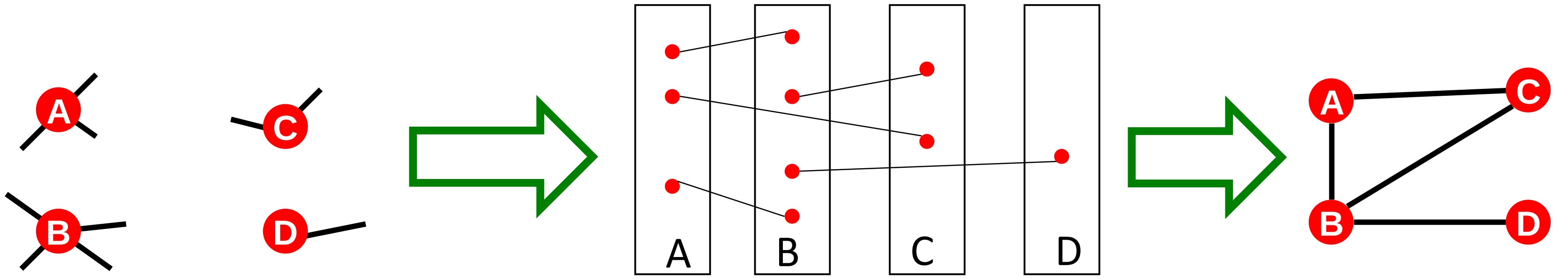
# Preferential attachment Model

- Start with Two connected nodes
  - Add a new node v
  - Create a link between v and one of the existing nodes with **probability proportional to the degree** of the that node
    - P(u,v) = d(u)/Total_network_degree
- **Rich get richer** phenomenon!
- Exhibits power-law distributions
- Can be extended to m links (Barabasi-Albert model)
  - Start with **connected network of $m_0$** nodes
  - Each new node connects to m nodes ($m \leq m_0$) with aforedescribed pref. attachment principle.

- [Example]

# Configuration Model

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




Nodes with spokes

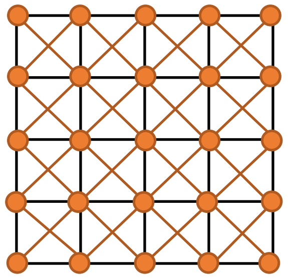
Randomly pair up "mini"-nodes

Resulting graph

- **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 $G$

# Clustering coefficient problem

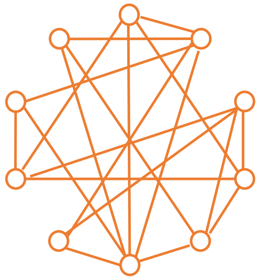
- What model can you think of that gives us a large clustering coefficient?
  - Grid?



Our Society seems to be like this!

High clustering coefficient,
High diameter

**Vs.**



Low clustering coefficient
Low diameter

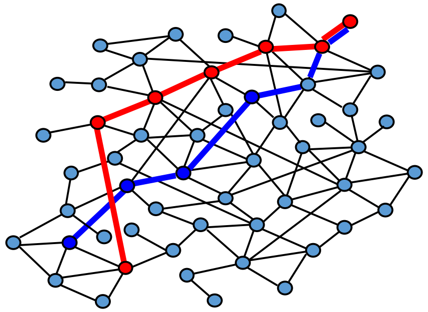
But there are many anectotal evidence of Small-World

# Bacon number

- **Kevin Bacon number**
  - Number of steps to Kevin Bacon in a Hollywood actor move co-appearance network
    - Actors are connected if co-appeared in the same movie.
- Origins?
  - Milgram’s Small World Network.
    - Six-degrees of separation
    - Surprising result at the time!


# The Small-World Experiment

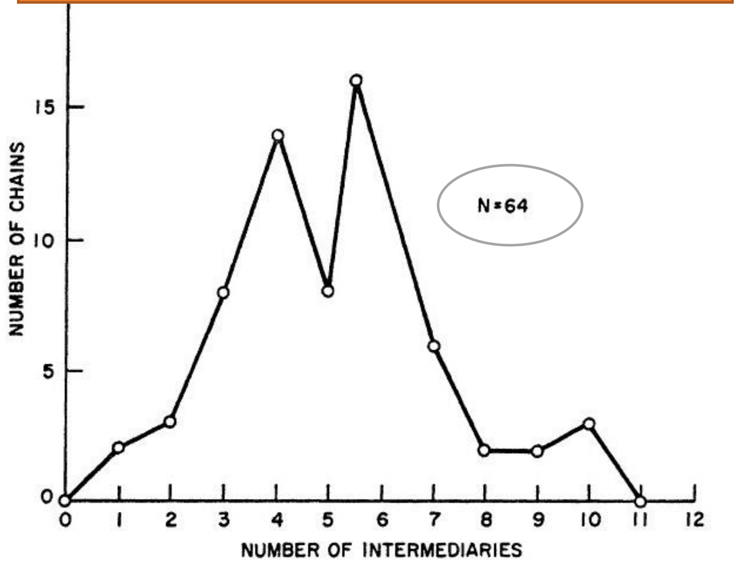
- **What is the typical shortest path length between any two people?**
  - **Experiment on the global friendship network**
    - Can’t measure, need to probe explicitly
- **Small-world experiment** [Milgram ’67]
  - Picked 300 people in Omaha, Nebraska and Wichita, Kansas
  - Ask them to get a letter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



# The Small-World Experiment

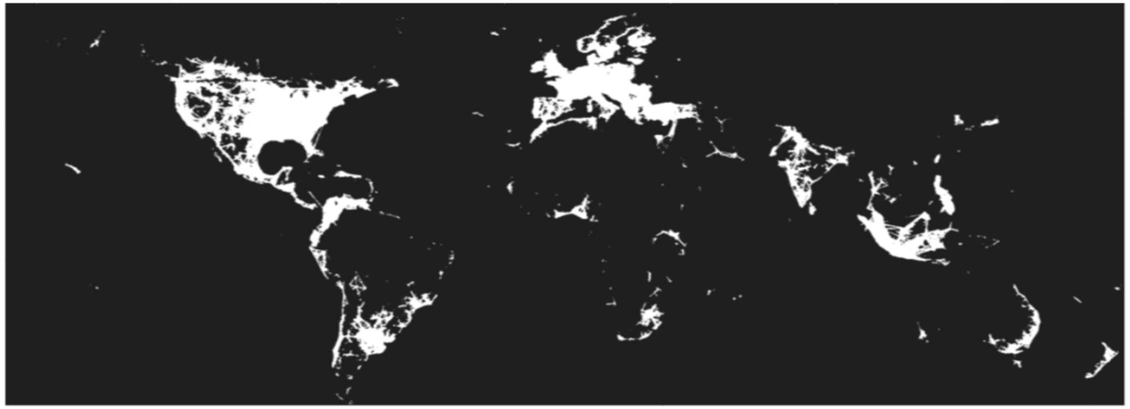
- **64 chains completed:**
  (i.e., 64 letters reached the target)
  - It took 6.2 steps on the average, thus **"6 degrees of separation"**
- **Further observations:**
  - People who owned stock had shorter paths to the stockbroker than random people: 5.4 vs. 6.7
  - People from the Boston area have even closer paths: 4.4
- Replicated Study in 2003
  - **In 2003 Dodds, Muhamad and Watts performed the experiment using e-mail**

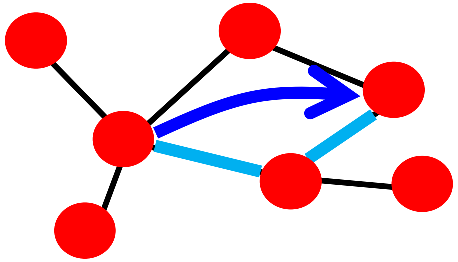
Milgram's small world experiment



# 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:**

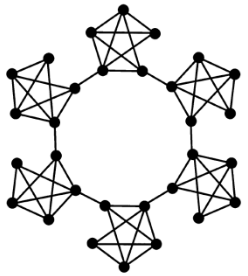
Actor Collaborations (IMDB): $N$ = 225,226 nodes, avg. degree $\overline{k}$ = 61
Electrical power grid: $N$ = 4,941 nodes, $\overline{k}$ = 2.67
Network of neurons: $N$ = 282 nodes, $\overline{k}$ = 14

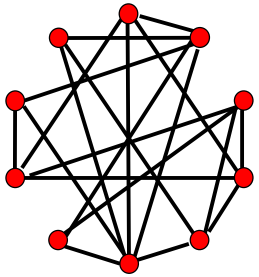
| Network | $h_{actual}$ | $h_{random}$ | $C_{actual}$ | $C_{random}$ |
|---|---|---|---|---|
| **Film actors** | 3.65 | 2.99 | 0.79 | 0.00027 |
| **Power Grid** | 18.70 | 12.40 | 0.080 | 0.005 |
| **C. elegans** | 2.65 | 2.25 | 0.28 | 0.05 |

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

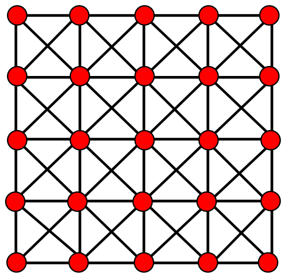
# The “Controversy”

- **Random Graphs:**
  - **Short paths:** $O(\log n)$
    - This is the smallest diameter we can get 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?**
  - **Ideas?**
  - **Caveman model**





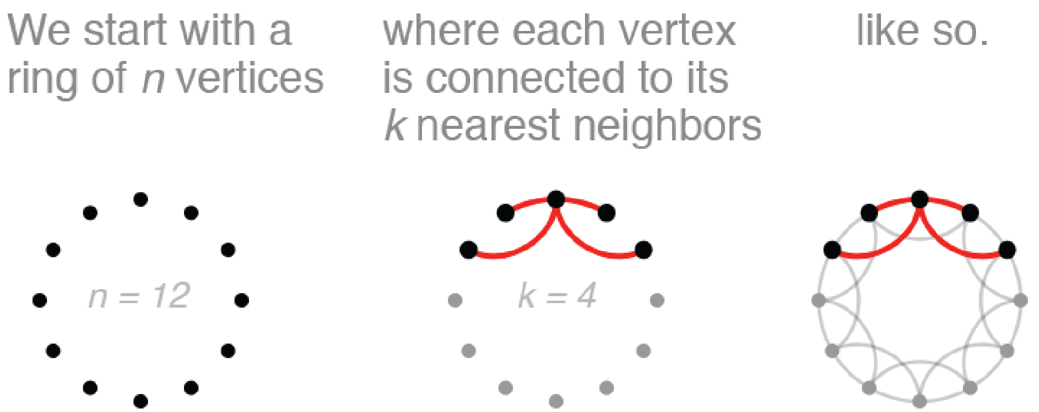
**Low diameter**
**Low clustering coefficient**



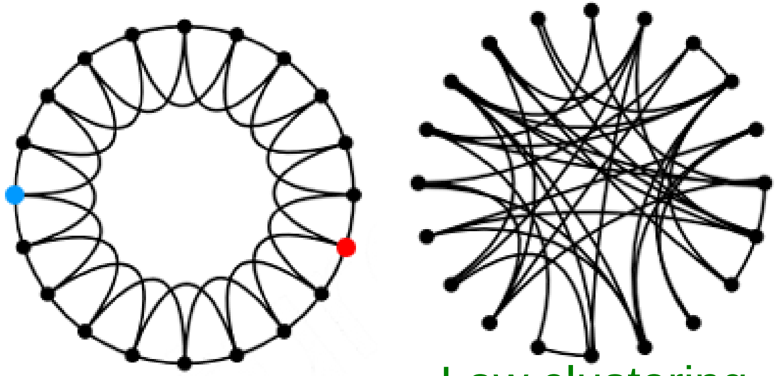
**High clustering coefficient**
**High diameter**

# Towards Small-World Model

- Regular graph with degree k connected to nearest neighbors




- Can be also a grid, torus, or any other “geographical” structure which has high clusterisation and high diameter
- ***With probability p rewire each edge in the network to a random node.***
  - ***Q: What happens when p=1?***
  - ***Q: What happens when p=0?***

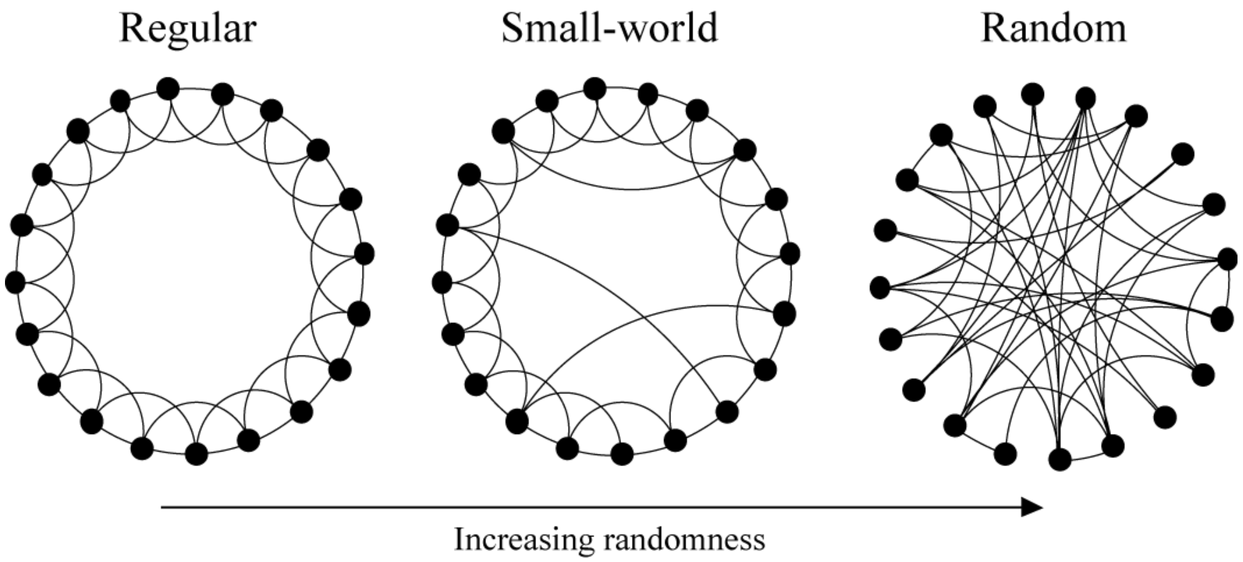


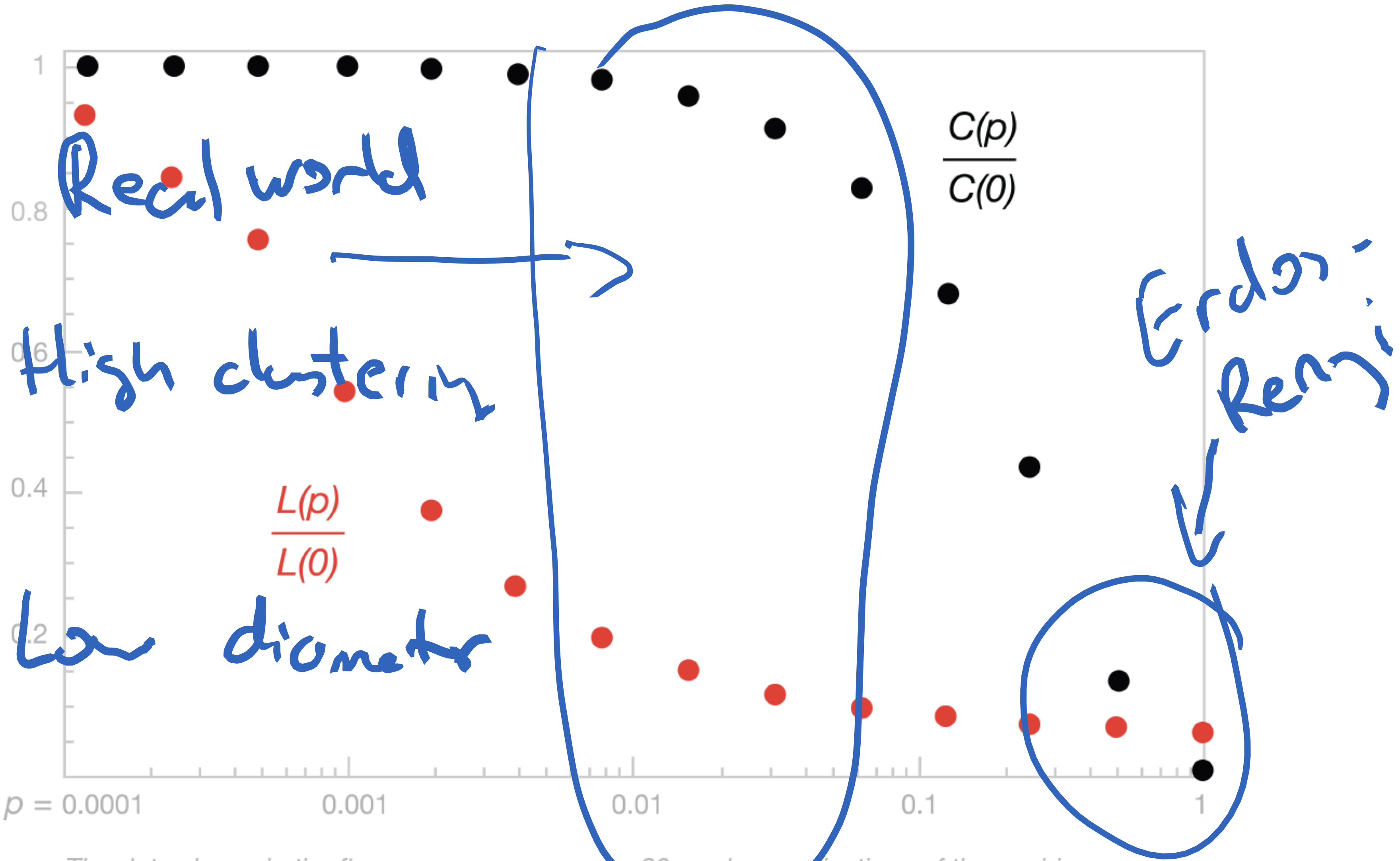
High clustering
High diameter

Low clustering
Low diameter

# Watts-Strogatz Model (cont.)

- When p=1 we have ~Erdos-Renyi network
- There is a range of p values where the network exhibits properties of both: random and regular graphs:
  - High clusterisation;
  - Short path length.



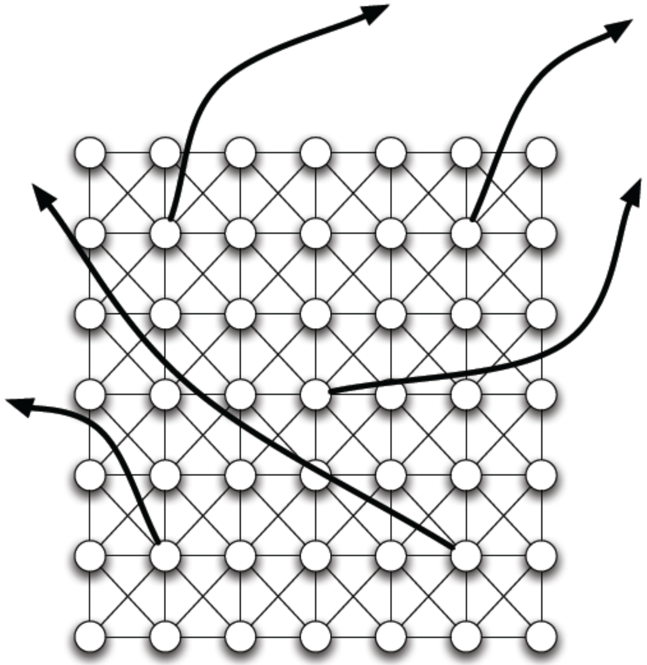



*The data shown in the figure are averages over 20 random realizations of the rewiring process, and have been normalized by the values L(0), C(0) for a regular lattice. All the graphs have n = 1000 vertices and an average degree of k = 10 edges per vertex. We note that a logarithmic horizontal scale has been used to resolve the rapid drop in L(p), corresponding to the onset of the small-world phenomenon. During this drop, C(p) remains almost constant at its value for the regular lattice, indicating that the transition to a small world is almost undetectable at the local level.*

- [Small World Example]

# 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?**

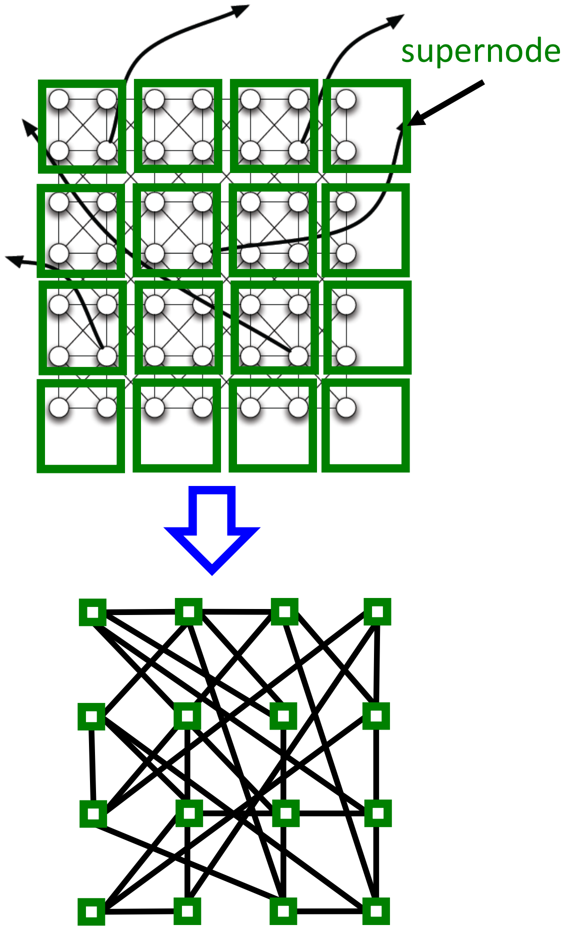
**It is *O(log(n))***

**Why?**

# Diameter of the Watts-Strogatz

- **Proof:**
  - Consider a graph where we contract 2x2 subgraphs into supernodes
  - Now we have 4 long-range edges sticking out of each supernode
    - **4-regular random graph – always connected!**
  - We can turn this into a path in the original graph by adding at most 2 steps per long range edge (by having to traverse internal nodes)

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

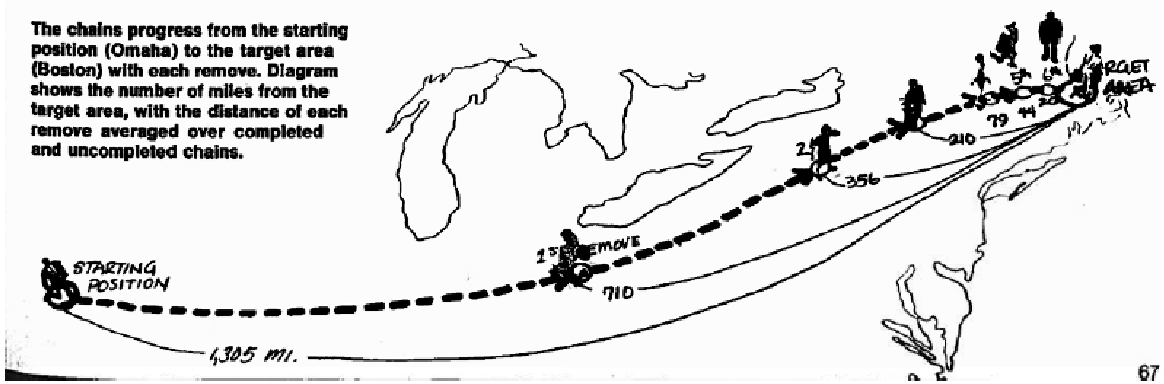



**4-regular random graph on supernodes**

Note that this analysis ignores edges between neighbors of super-nodes, but this does not matter since those edges would make the diameter only go further down.

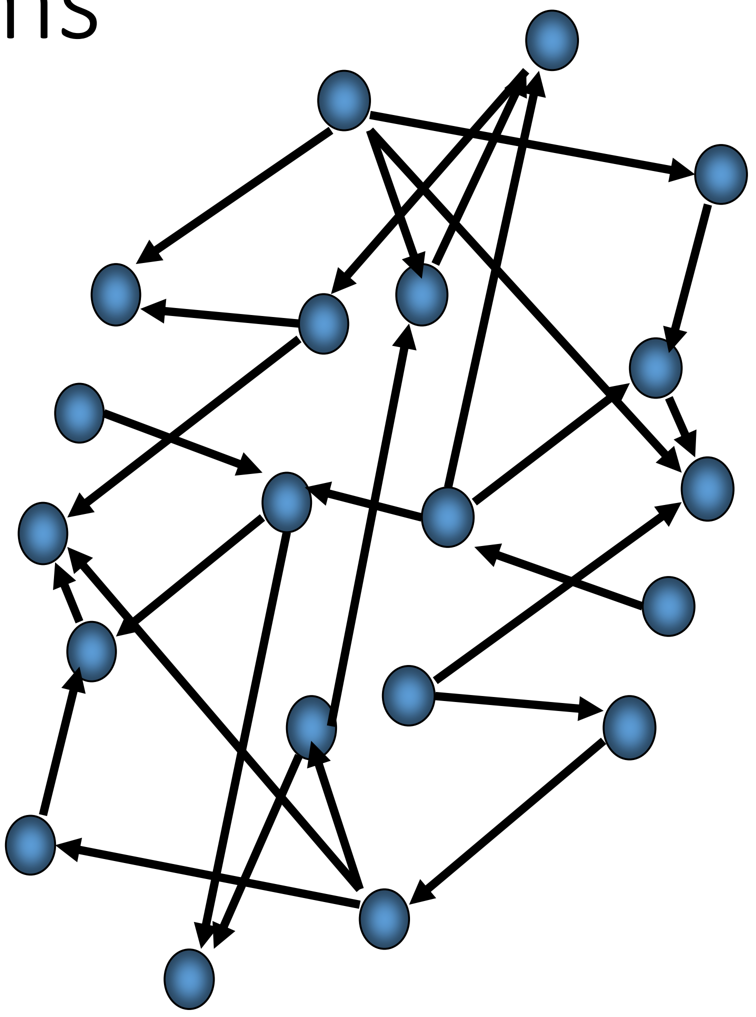
# Small-World: remaining questions

- Is it enough to explain Milgram's experiment?
  - If there exist shortest path between any two nodes - **where is the global knowledge** that we can find this path?!
  - Why should **arbitrary pairs of strangers be able to find short chains** of acquaintances that link them together???
  - **Why decentralized "search algorithm" works?**
    - Very few nodes were involved in the discovery of the "shortest path"



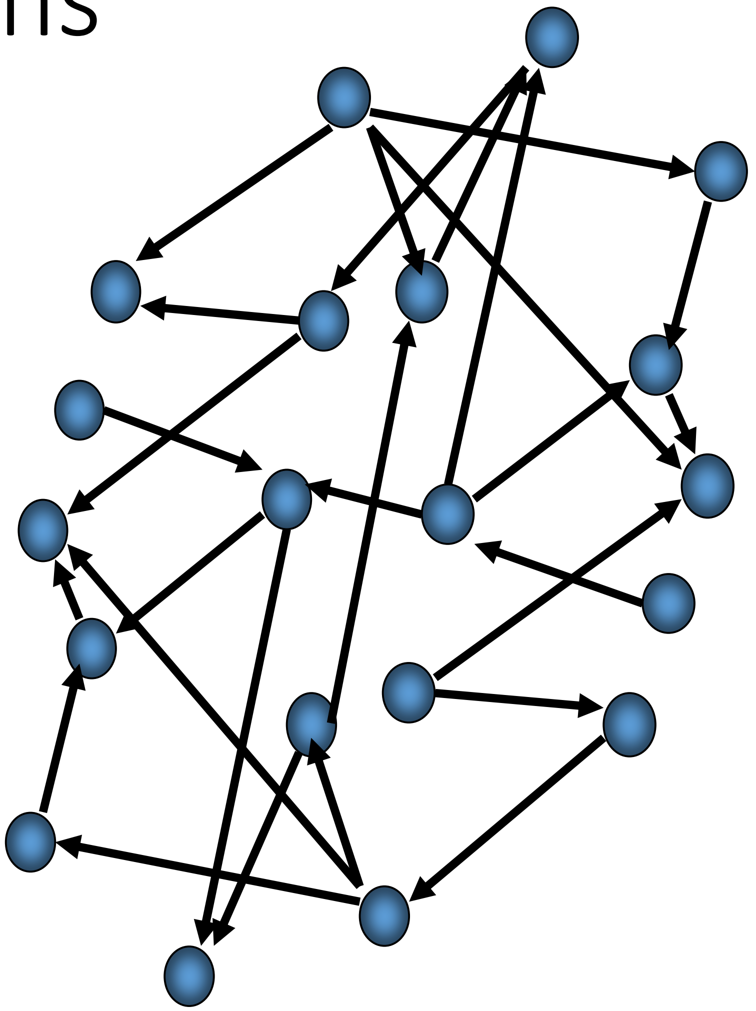
# Implications for P2P systems

- Each P2P system can be interpreted as a directed graph where peers correspond to the nodes and their routing table entries as directed links to the other nodes



# Implications for P2P systems

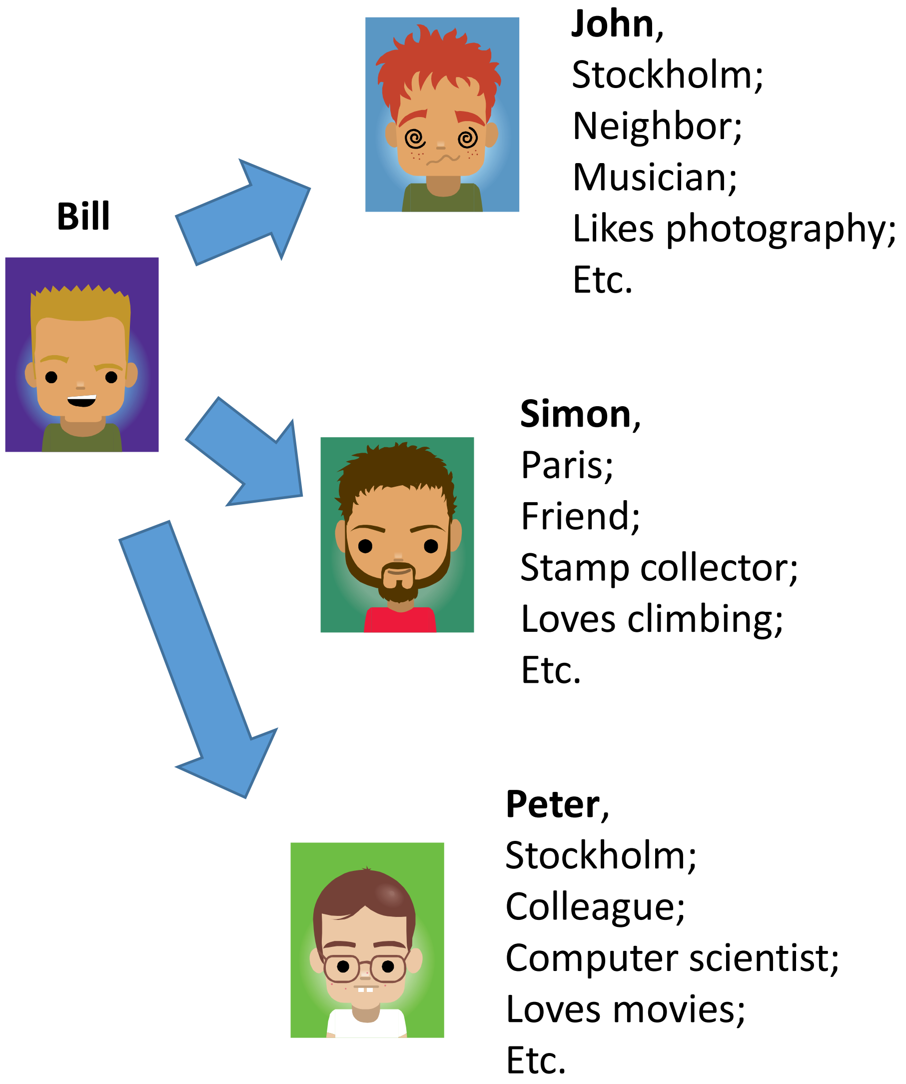
- Task for P2P:
  - Design a **completely decentralized algorithm** that would route message from any node A to any other node B with relatively few hops compared with the size of the graph



- **Is it possible?**
  - Milgram experiment suggests YES!

# So why Milgram's experiment worked?

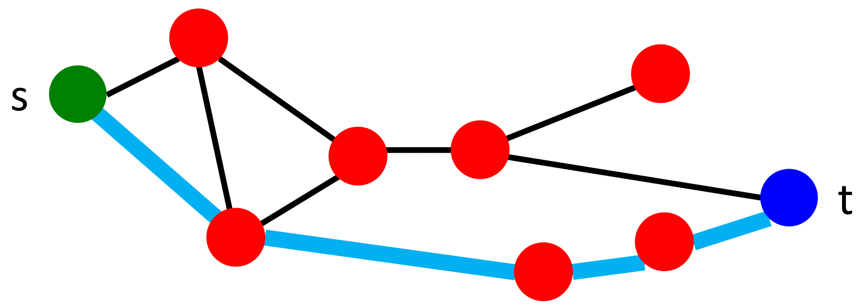
- Social network is not a bare graph of vertices and edges, but a graph with certain "labels"
- The "labels" representing various dimensions of our life
  - Hobbies, work, geographical distribution etc.
- There is (are multiple) "**labeling space(s)**" with a **distance metric**!!!
- We can greedily minimize the distance!!
  - Decentralized search: a greedy-routing algorithm
  - We need to build right graph where decentralized algorithm might perform the best



# Decentralized Navigation (Search)

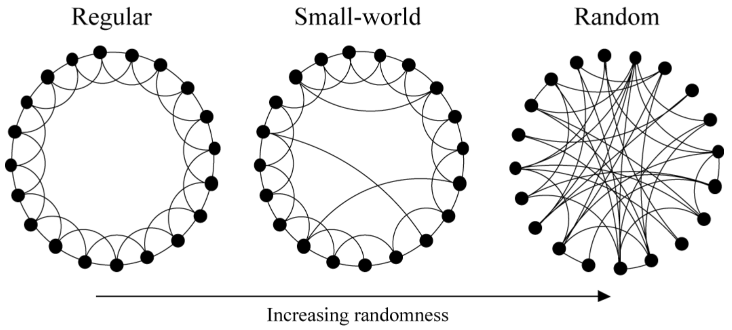
**The setting:**

- $s$ only knows locations of its friends and location of the target $t$
- $s$ does not know links of anyone else but itself
- **ID-space (e.g., geographic) Navigation:** $s$ "navigates" to a node geographically closest to $t$
- **Search time T:** Number of steps to reach $t$



# Navigation in Watts-Strogatz Small-Worlds

- Watts-Strogatz model
  - High clusterisation;
  - **Short path length**.




- Construction involves a notion of "**ID space**" and a "**distance**" function.
  - Think how to connect to k "closest neighbors" in the initial step…
- **Short Paths exist** in Watts-Strogatz model, but decentralized greedy routing **can not find** them!

# Overview of the Results

**Navigable**

Search time T:

$$O((\log n)^{\beta})$$

**Kleinberg's model**

$$O((\log n)^{2})$$

**Not navigable**

Search time T:

$$O(n^{\alpha})$$

**Watts-Strogatz**

$$O(n^{\frac{2}{3}})$$

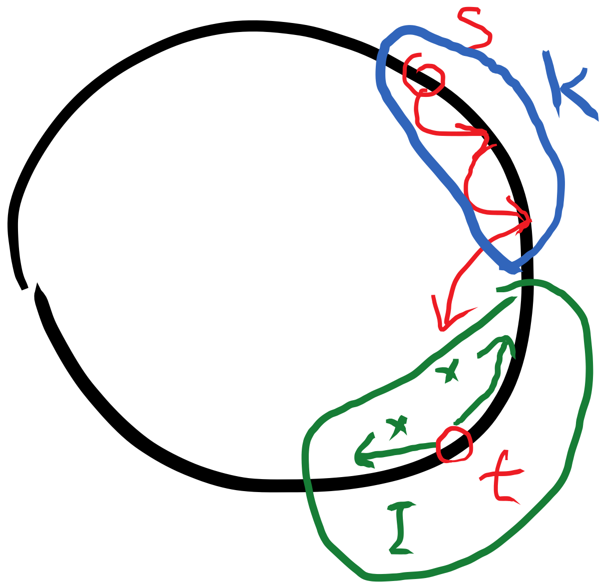
**Erdős–Rényi**

$$O(n)$$

**Note:** We know these graphs have diameter *O(log n)*.
So in Kleinberg's model search time is polynomial in *log n*,
while in Watts-Strogatz it is exponential (in *log n*).

# Why Watts-Strogatz is not navigable?

- Assume **1-d ring** structure + **1 random link** from each node
- Assume any **decentralized distance minimizing** (on the ring) search algorithm
- $\boldsymbol{E}$ = event that any of the first $\boldsymbol{k}$ nodes search algorithm visits has a link to $\boldsymbol{I}$



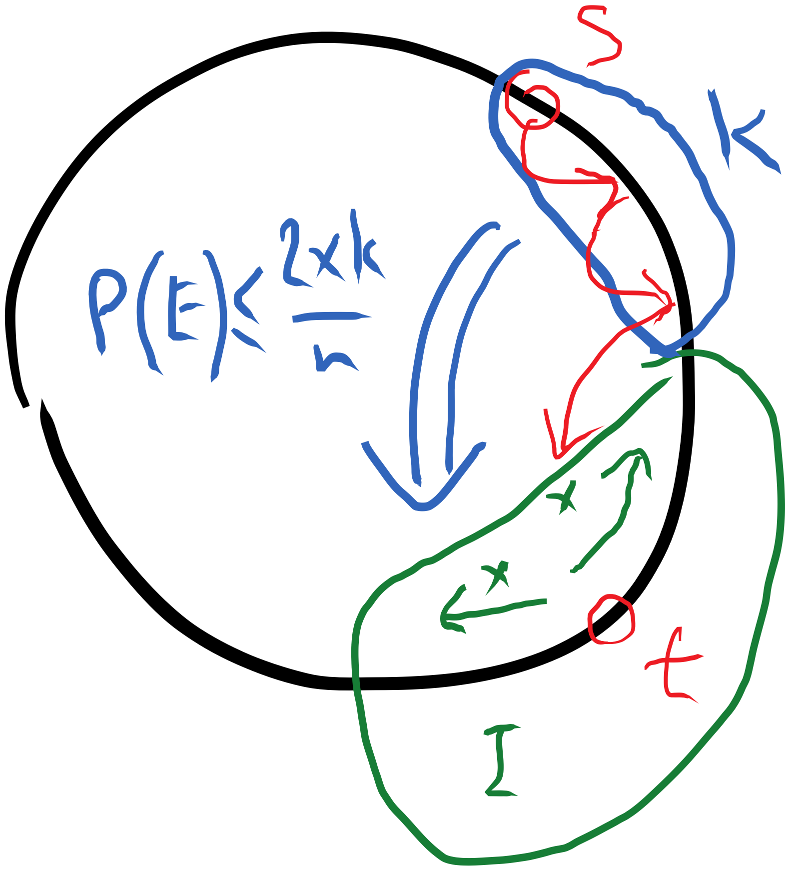

- **Let:** $\boldsymbol{E_i}$= event that long link out of node $\boldsymbol{i}$ points to some node in interval $\boldsymbol{I}$ of width $\boldsymbol{2 \cdot x}$ nodes (for some $\boldsymbol{x}$) around target $\boldsymbol{t}$
- **Then:** $P(E_i) = \frac{2x}{n-1} \approx \frac{2x}{n}$

# Why Watts-Strogatz is not navigable? (cont.)

**Then:** $$P(E) = P\left(\bigcup_i^k E_i\right) \leq \sum_i^k P(E_i) = k\frac{2x}{n}$$

**Let's choose** $$k = x = \tfrac{1}{2}\sqrt{n}$$

**Then:** $$P(E) \le 2\frac{\left(\frac{1}{2}\sqrt{n}\right)^2}{n} = \frac{1}{2}$$

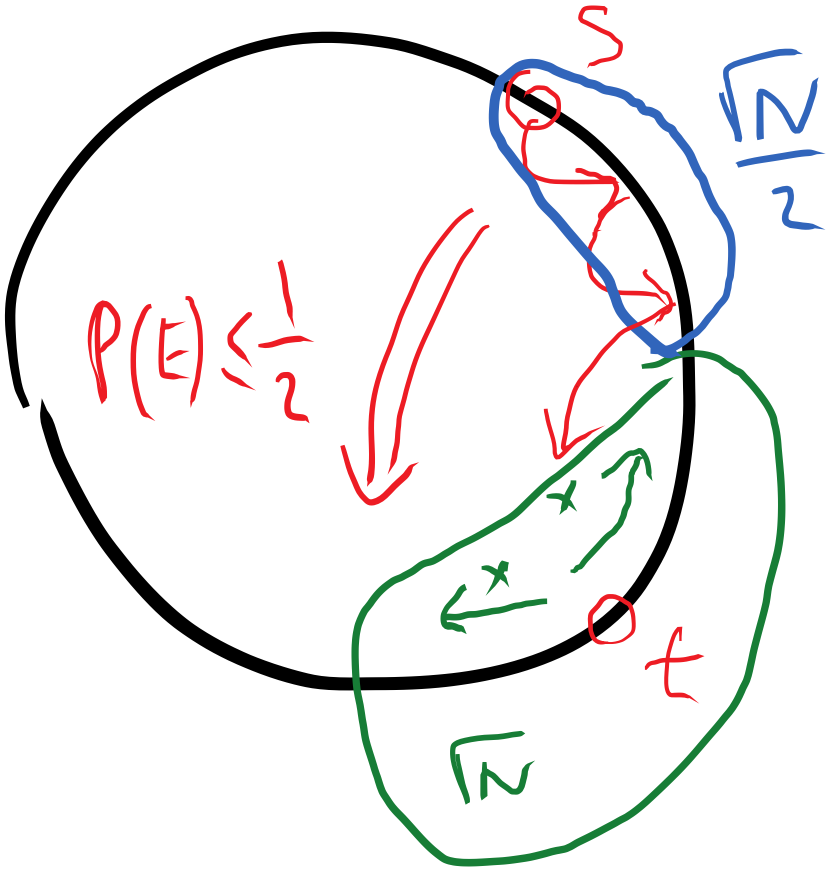


# Why Watts-Strogatz is not navigable? (cont.)

**Then:** $$P(E) = P\left(\bigcup_{i}^{k} E_i\right) \le \sum_{i}^{k} P(E_i) = k\,\frac{2x}{n}$$

**Let's choose** $$k = x = \tfrac{1}{2}\sqrt{n}$$

**Then:** $$P(E)\,\text{£}\,2\,\frac{\left(\frac{1}{2}\sqrt{n}\right)^2}{n} = \frac{1}{2}$$



# Why Watts-Strogatz is not na

It is BAD!
How did Milgram's experiment showed such short routes???!!!
Ideas?

- P(E) = P(in $\frac{1}{2}\sqrt{n}$ steps we jump inside $\frac{1}{2}\sqrt{n}$ of $t$) ≤ $\frac{1}{2}$
- E does not happen with prob. at least $\frac{1}{2}$, i.e. **P(not E) $\geq \frac{1}{2}$**
- If E does not happen, we must traverse $\frac{1}{2}\sqrt{n}$ steps to get to t
- Expected number of steps in the **"blue area"?** at least $\frac{1}{2}*\frac{1}{2}\sqrt{n}$

- **Thus expected number** of search steps in such a network at least **O($\sqrt{n}$)**
  - For **d**-dim. lattice: $T \geq O(n^{d/(d+1)})$