Lecture 3:
Random Graphs

01204456 Social Networks Data Mining

Department of Computer Engineering
Faculty of Engineering, Kasetsart University
Bangkok, Thailand.

Department of Computer Engineering Kasetsart University

MIKE LABORATORY

Outline

- Definition
- Degree Distribution
- Evolution of G_{np}
- Giant Component

Network Models

- Why model?
  - Be a simple representation of complex network
  - Can derive properties mathematically
  - Predict properties and outcomes

Downloading NetLogo

- NetLogo website:
  http://ccl.northwestern.edu/netlogo/

- Models specified in the class website

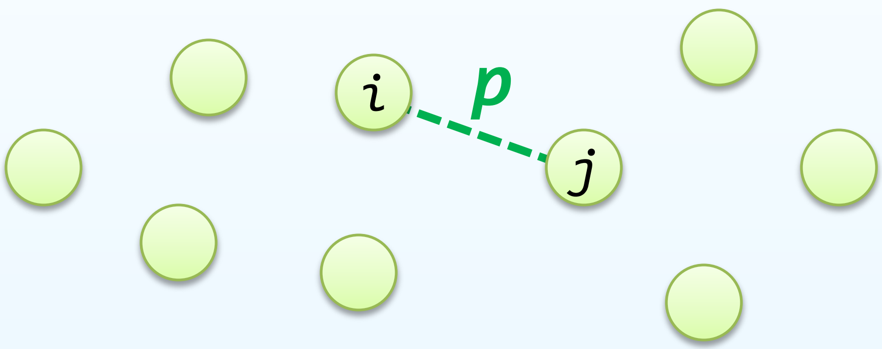
Random Graph

- Assumptions:
  - Nodes connect at random
  - Network is undirected

- Key parameters:
  - n = number of nodes in the graph
  - m = total number of edges in the graph
  - p = probability that any two nodes are connected

Random Graph

- In the $G(n,p)$ model,
  - a graph is constructed by connecting nodes randomly.
  - each edge is included in the graph with probability p independent from every other edge.

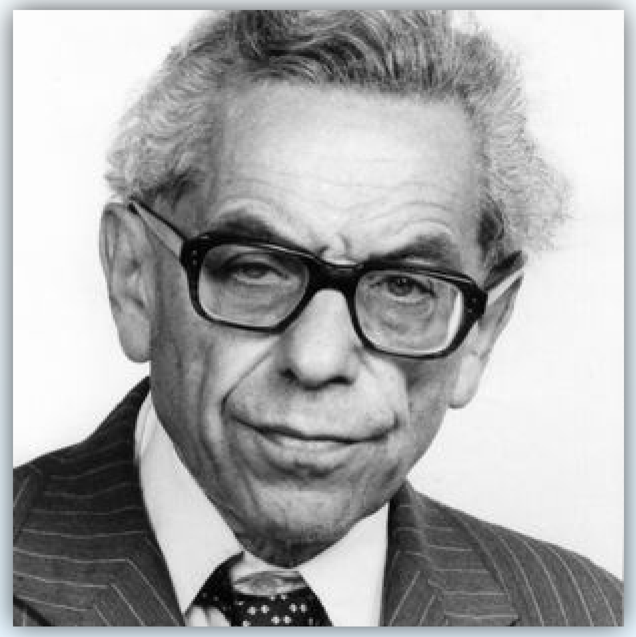


- So, a specific graph with m edges will appear with

$$\boxed{\Pr[G_{np}] = p^m(1-p)^{\binom{n}{2}-m}}$$

Paul Erdös and Alfréd Rényi

- Hungarian Mathematicians
- The Erdös-Rényi model of random graphs (1959)



The Erdös-Rényi Model

- A simplest network model
- In the $G(n,m)$ model,
  - a graph is chosen uniformly at random from the collection of all graphs which have n nodes and m edges.

- So, that specific graph will appear with

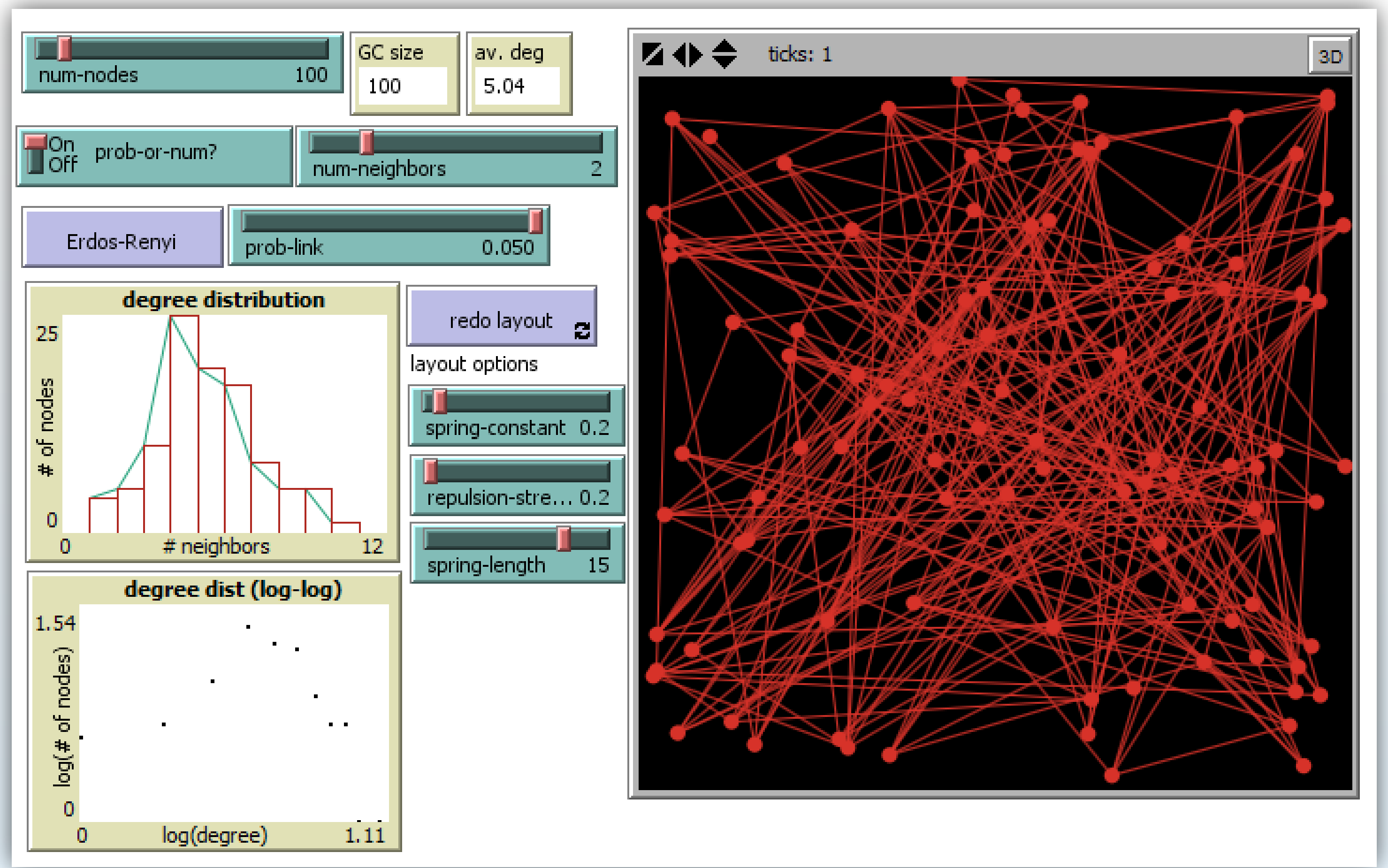
$$\Pr[G_{nm}] = 1/\binom{\binom{n}{2}}{m}$$

- Moreover, we expect $G(n,p) \approx G(n,m)$ if we choose p and m such that

$$m = \binom{n}{2} p$$

Question #1

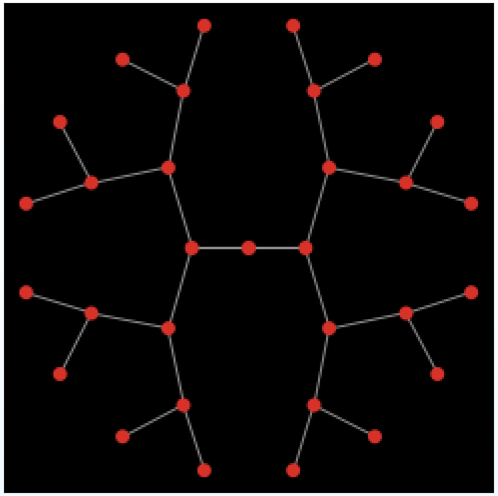
- As the size of the network increases, if you keep *p* (i.e., the probability of any two nodes being connected) the same, what happens to the average degree?
  a) stays the same
  b) increases
  c) decreases



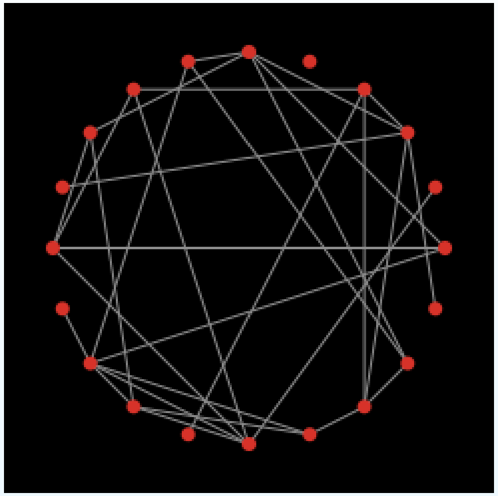
ErdosRenyiDegDist.nlogo

NetLogo: Layouts

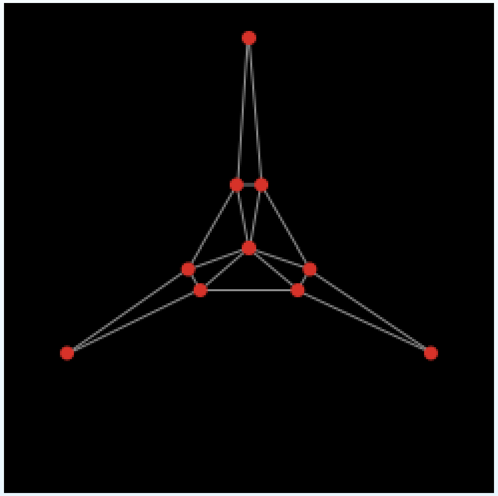
radial

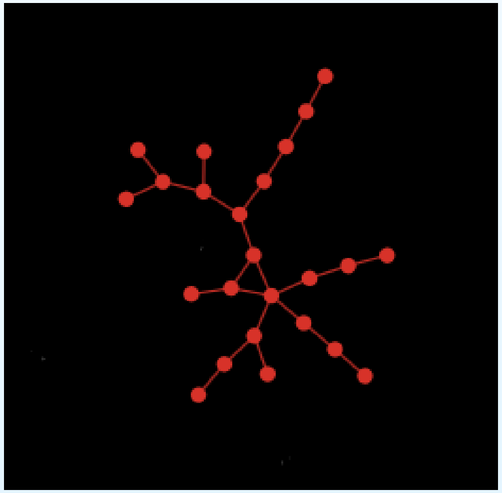


circle



tutte





spring

Degree Distribution

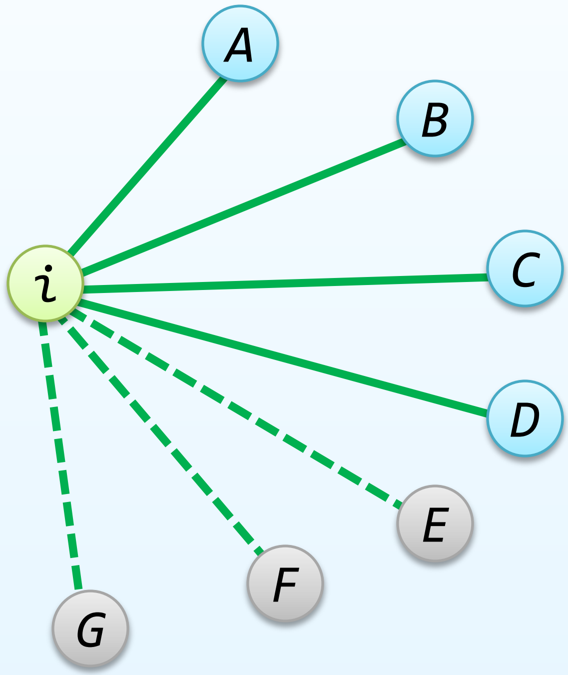
- $G(n,p)$: for each potential edge we flip a biased coin
  - with probability p we add the edge
  - with probability $1-p$ we don't
- What is the probability that a node has 0,1,2,3,… edges?
- Probabilities sum to 1

Question #2

- What is the maximum degree of a node in a simple (i.e., no multiple edges between the same two nodes) n-node graph?
  a) n
  b) $n-1$
  c) $n/2$

Question #3

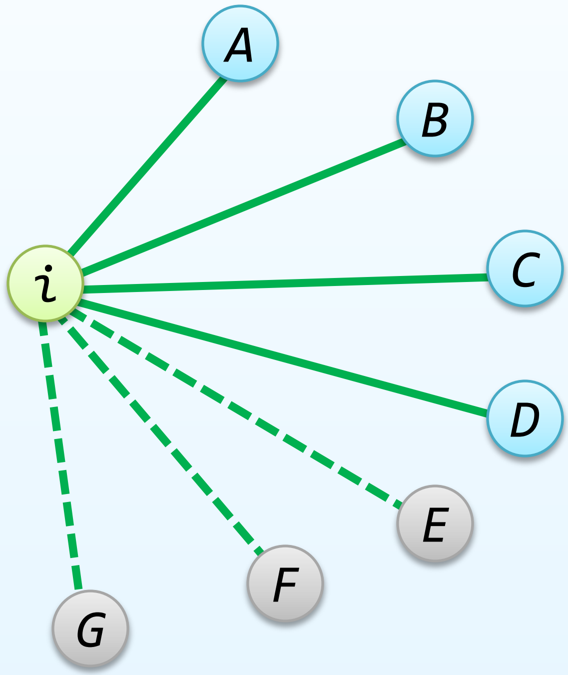
- Given a 8-node graph, how many ways can a node i be connected to its neighbors to get degree 4?



$$\binom{7}{4} = 35$$

Question #4

- Given a 8-node graph and probability p of any two nodes sharing an edge, what is the probability that a node i has degree 4?



$$p^4(1-p)^3$$

Degree Distribution

- Each node has $n-1$ tries to get edges
- Each try is a success with probability p
- The probability of a node to have degree k:

$$\boxed{\Pr[K=k] = \binom{n-1}{k} p^k (1-p)^{n-1-k}}$$

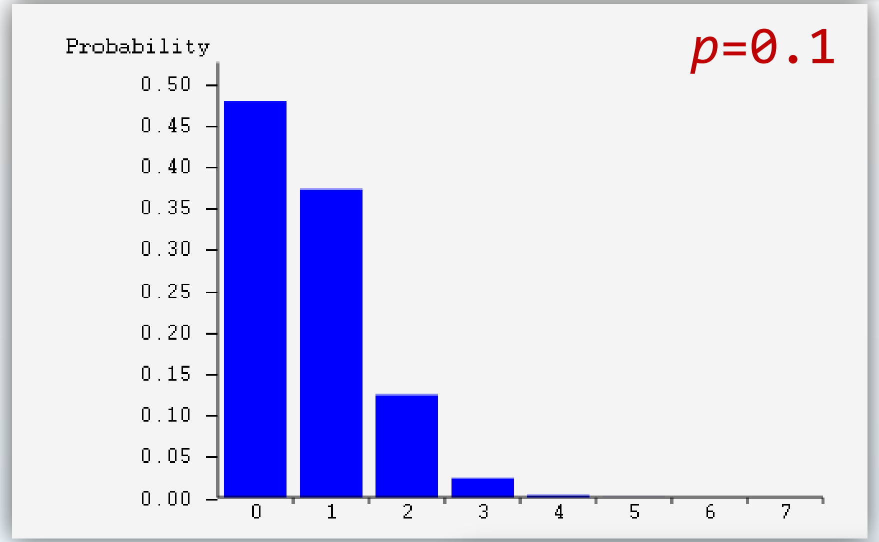
⇨ Binomial distribution, $B(n-1;k;p)$

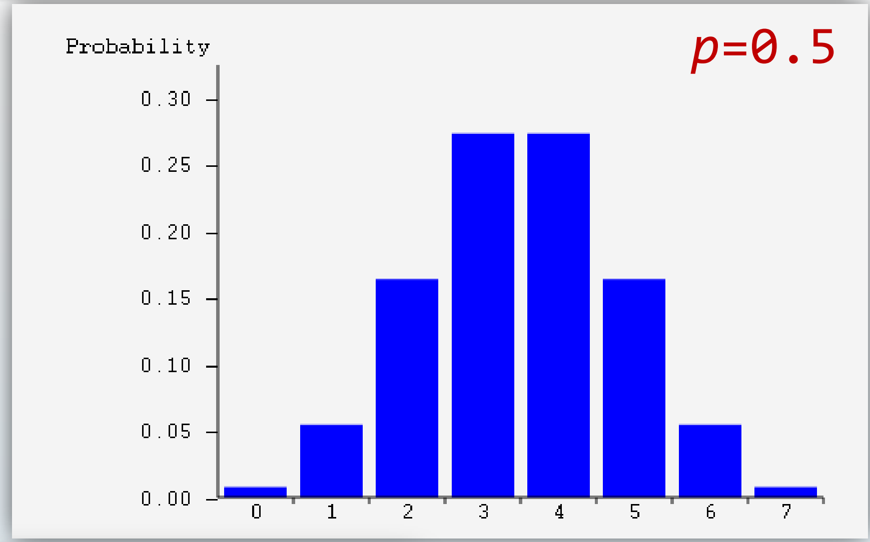
Example: Binomial Distribution

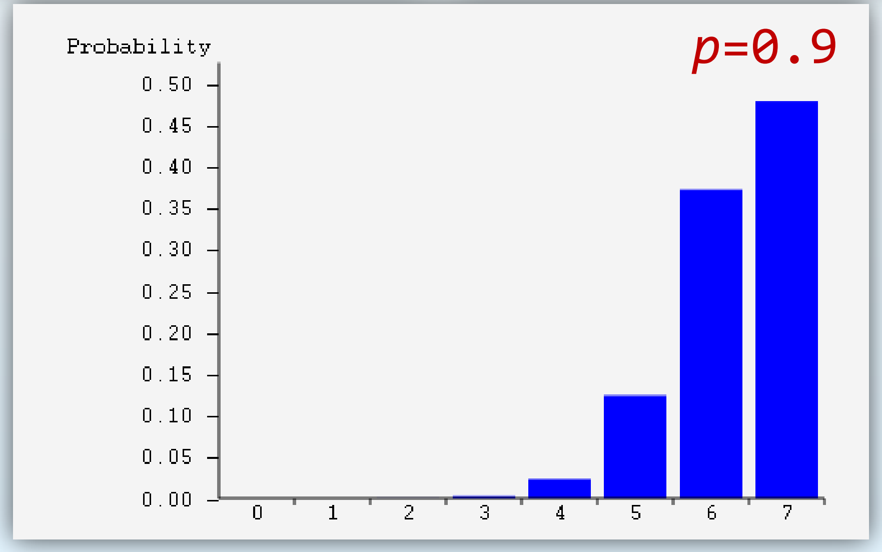
- Given a 8-node graph and probability *p* of any two nodes sharing an edge
- What is the probability that a given node has degree 4?

$$B(7;4;p) = \binom{7}{4} p^4 (1-p)^3$$

Example: if varying *p*

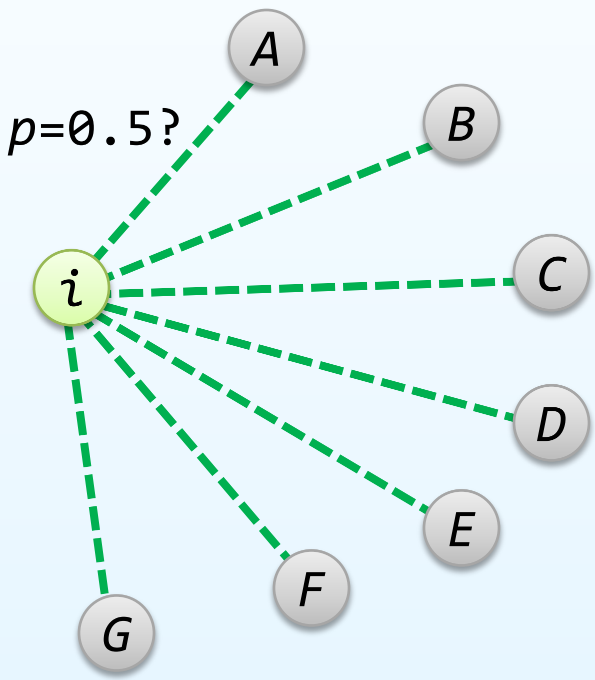


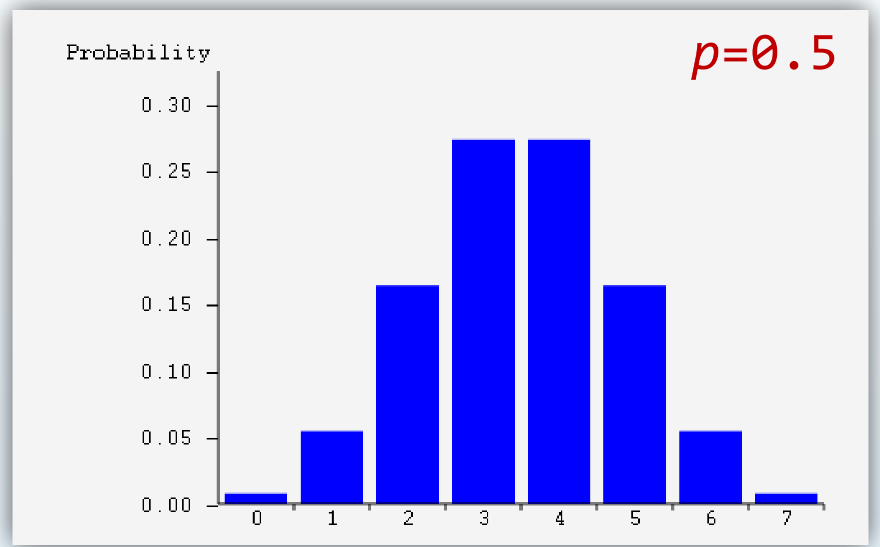




Question #5

- What is the average degree of a graph with 8 nodes and probability $p=0.5$ of an edge existing between any two nodes?





$$
\begin{aligned}
E[K] &= (0\times0.0078125)+(1\times0.0546875) \\
&\quad +(2\times0.1640625)+(3\times0.2734375) \\
&\quad +(4\times0.2734375)+(5\times0.1640625) \\
&\quad +(6\times0.0546875)+(7\times0.0078125) \\
&= 3.5
\end{aligned}
$$

What is the mean?

- In general, we know

$$\mu = \mathrm{E}[X] = \Sigma x P_X(x)$$

- For a graph with n nodes and probability p of an edge existing between any two nodes, the average degree is

$$\begin{aligned} \mathrm{E}[K] &= \Sigma k \binom{n-1}{k} p^k (1-p)^{n-1-k} \\ &= (n-1)p \\ &\approx np \end{aligned}$$

What is the variance?

- In general, we know

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \Sigma(x-\mu)^2 P_X(x)$$

- For a graph with *n* nodes and probability *p* of an edge existing between any two nodes, the variance in degree is

$$\begin{aligned}\sigma_K^2 &= \Sigma(k-\mu)^2\binom{n-1}{k}p^k(1-p)^{n-1-k} \\ &= (n-1)p(1-p) \\ &\approx np(1-p)\end{aligned}$$

Approximation

$$\Pr[K=k] = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

$$\Pr[K=k] \rightarrow \frac{(np)^k e^{-np}}{k!}$$

$$\Pr[K=k] \rightarrow \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-np)^2}{2\sigma^2}}$$

Binomial

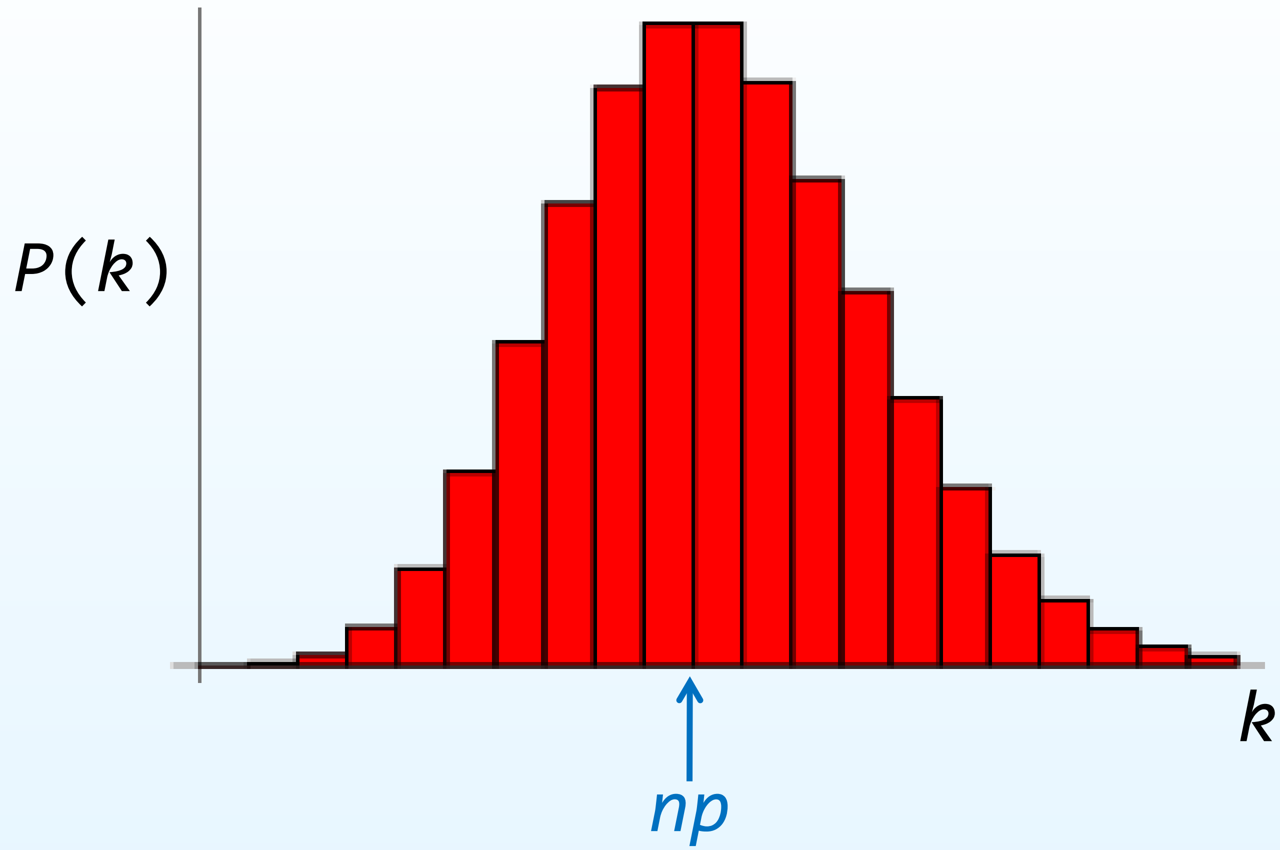
↓ large n and small p so, np is constant

Poisson

↓ large np

Normal

Poisson Distribution



Insights

- Previously: degree distribution
- Emergence of giant component
- Average shortest path

Keywords

- Connectivity
  - A graph is connected if for every pair of nodes, there is a path between them
- Component
  - If a graph is not connected, then it breaks apart naturally into a set of connected groups of nodes
- Giant component

Giant Component

- A connected component (SCC for directed network) is a large network, when its size is a constant fraction of the entire graph

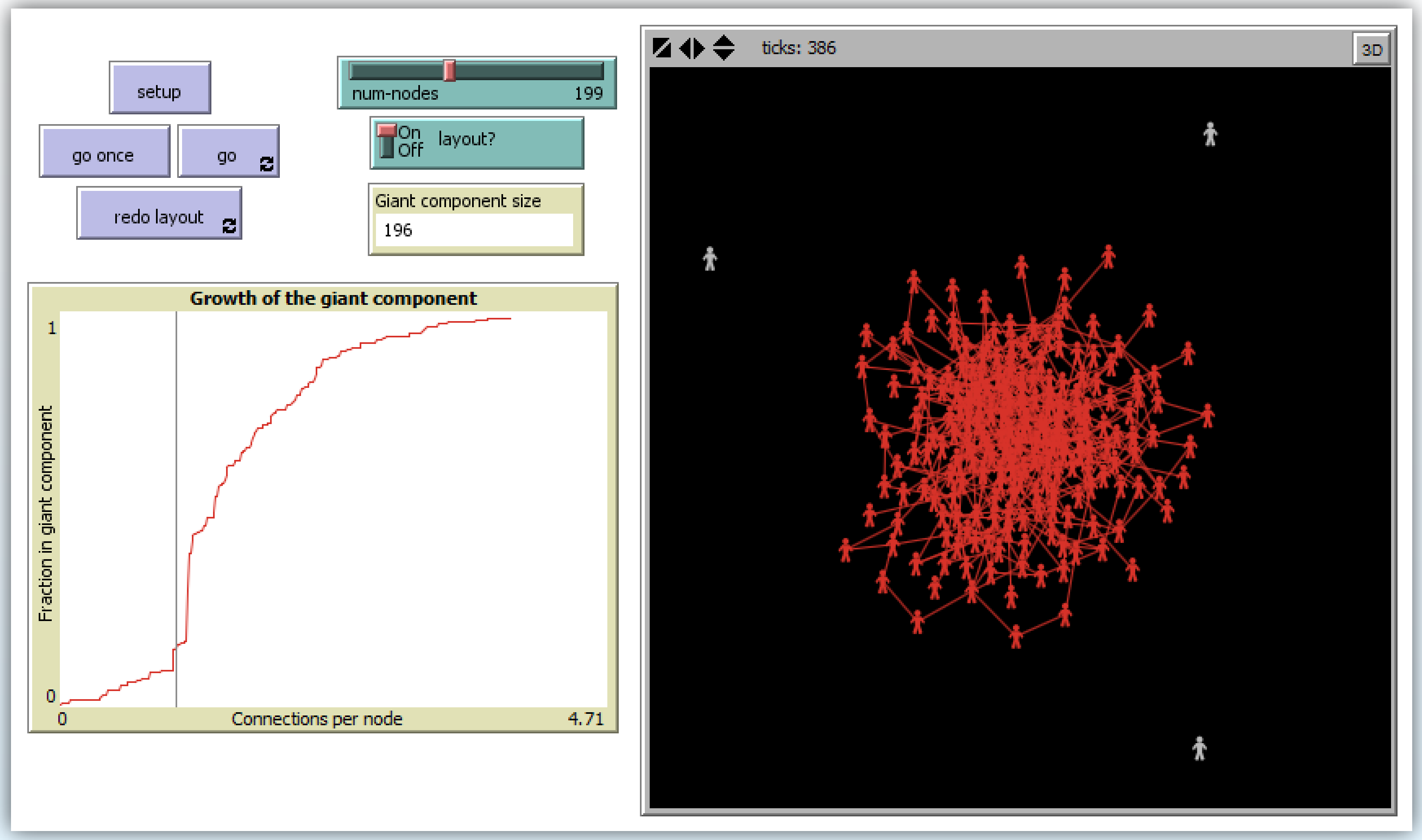
- Formally,
  - let n_X be the size of a connected component X in a network of size n
  - then, X is a giant component if

$$\lim_{n \to \infty} \frac{n_X}{n} = c > 0$$

Question #6

- What is the average degree $\langle k \rangle$ at which the giant component starts to emerge?
  a) 0
  b) 1
  c) 3/2
  d) 3

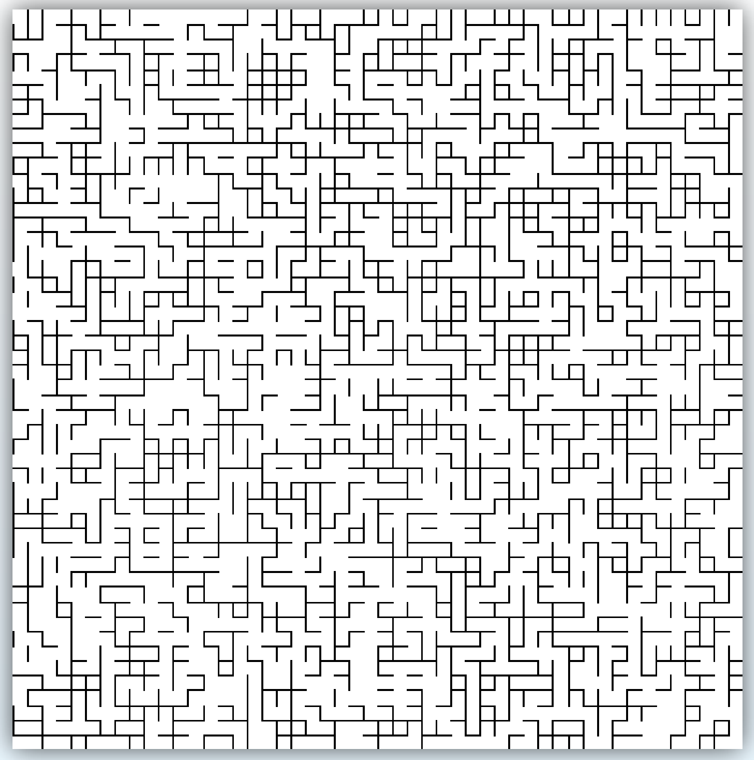
Emergence of a Giant Component



Giant Component.nlogo

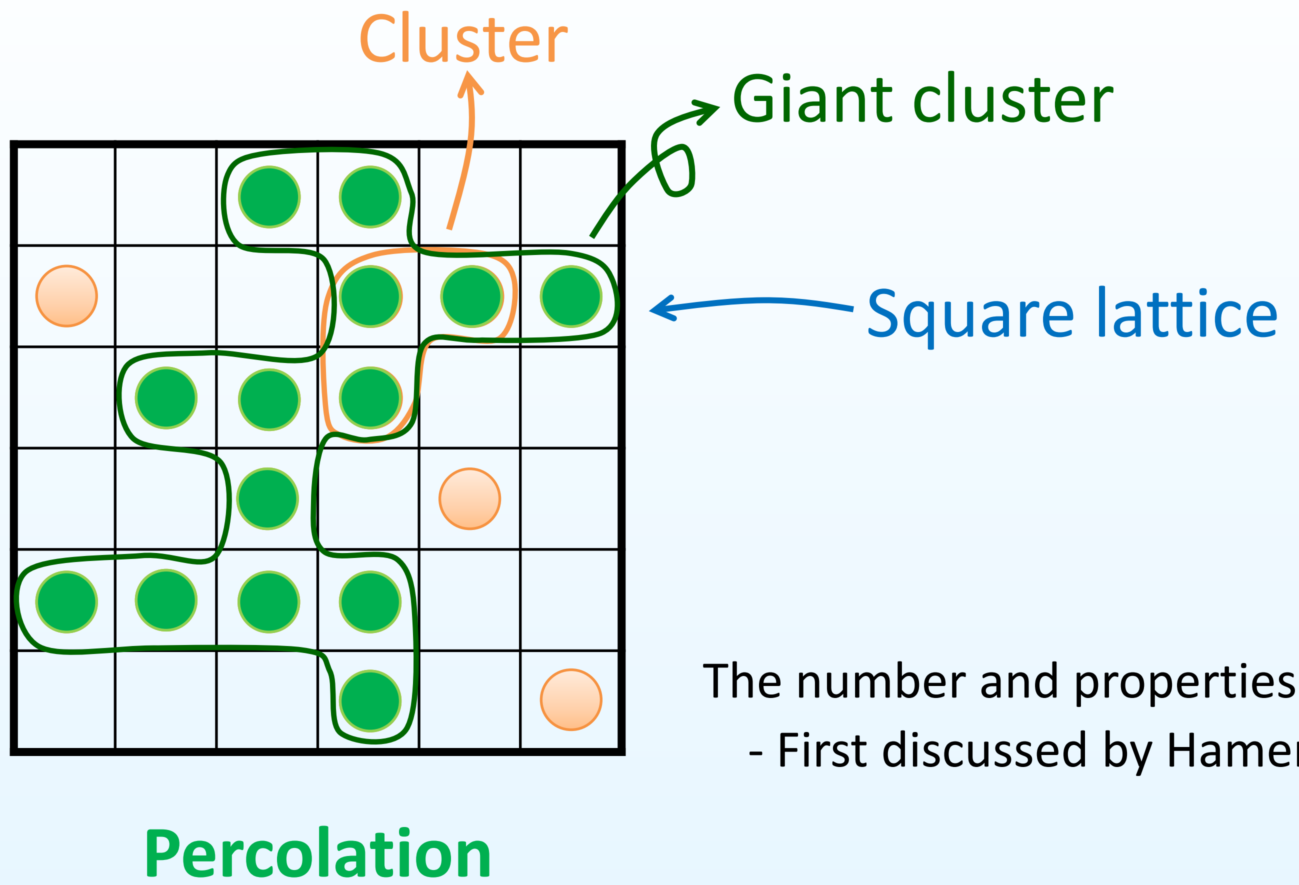
Percolation Theory

- Describe the behavior of connected clusters in a random graph
- Apply to material science and other domains
- Two slightly different mathematical models:
  - bond percolation
  - site percolation



a bond percolation on the square lattice in two dimensions

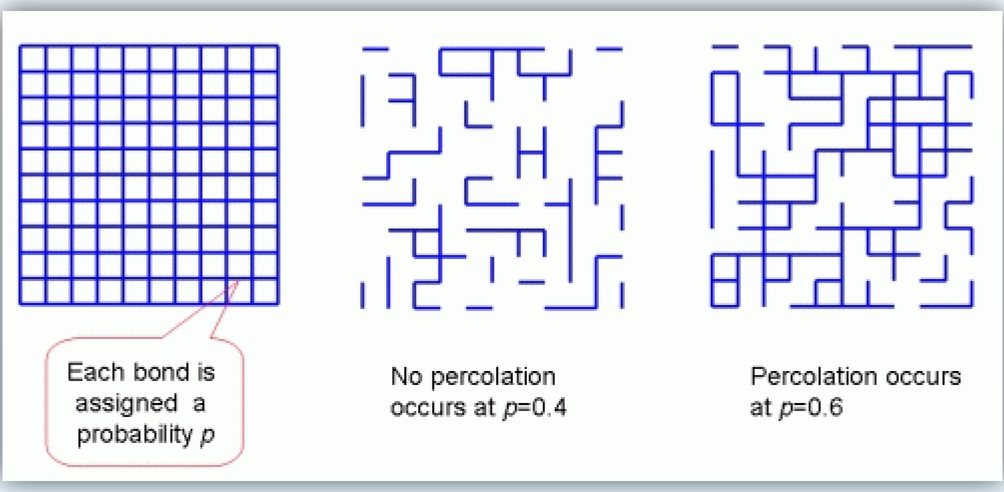
Example: Percolation (2D Lattice)



The number and properties of clusters?
- First discussed by Hamersley in 1957

Percolation Threshold

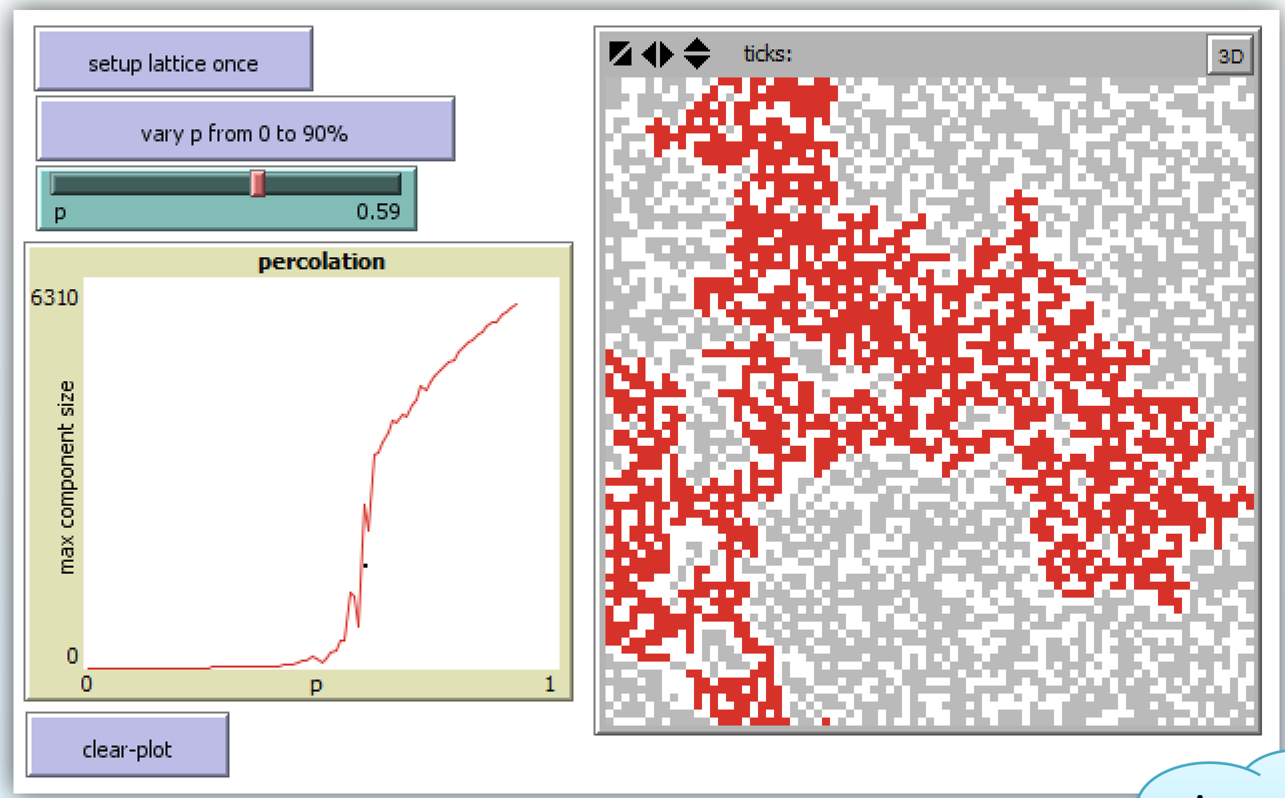
- A probability of bonds (or sites) needed to be made into a random network before the giant component appears
- Also, called critical threshold (p_c)



Question #7

- What is the percolation threshold of 2D lattice (i.e., fraction of sites that need to be occupied in order for a giant connected component to emerge)?
  a) 0
  b) ~1/4
  c) ~1/3
  d) ~1/2

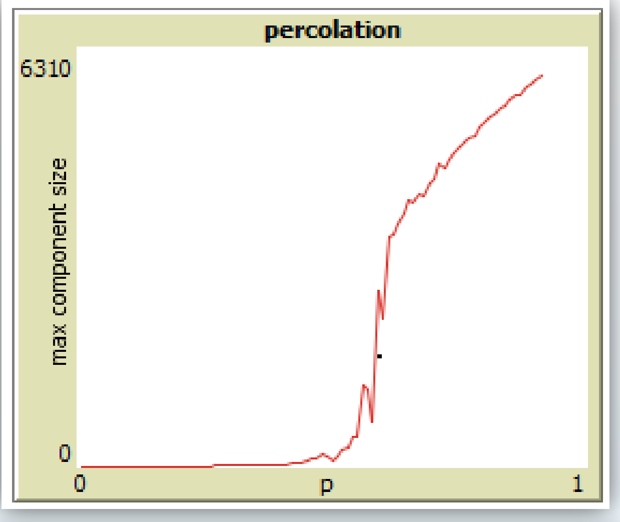
Percolation on a 2D Lattice

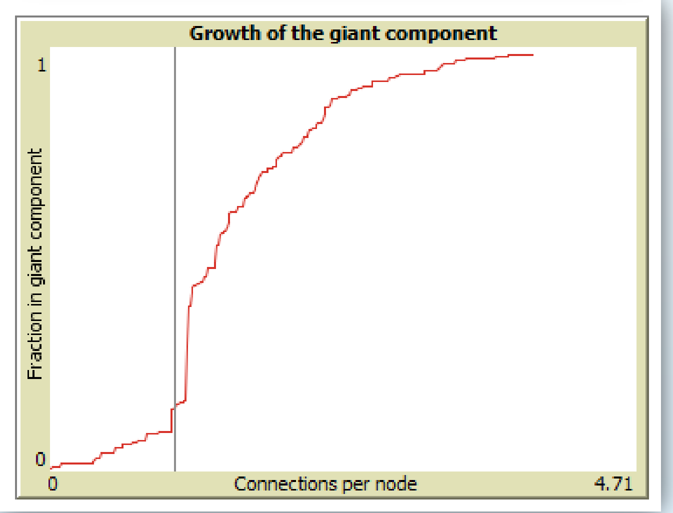


LatticePercolation.nlogo

site = 0.5927
bond = 0.5

Percolation on a Random Graph





- Percolation threshold:
  - A fraction of edges needed to be added before the giant component appears
- For Erdös-Rényi network, as the average degree increases to $\langle k \rangle = 1$, a giant component suddenly appears

What happens to G_{np} when we vary p?

Back to Node Degrees of G_{np}

- Remember, the expected degree is

$$\boxed{E[K] = np}$$

- So, if we want $E[K]$ being independent of n
  - Let

$$\boxed{p = c/n}$$

Probability of a Node Being Isolated

- **Observation:** If we build a random graph G_{np} with $p=c/n$, we will have many isolated nodes

- **Why?**

$$\Pr[v \text{ has degree } 0] = (1-p)^{n-1} = \left(1-\tfrac{c}{n-1}\right)^{n-1} \underset{n\to\infty}{\approx} e^{-c}$$

$$\lim_{n\to\infty}\left(1-\tfrac{c}{n-1}\right)^{n-1} = \lim_{x\to\infty}\left(1+\tfrac{1}{x}\right)^{-xc} \quad ; \tfrac{1}{x} = -\tfrac{c}{n-1}$$

$$= \Big(\underbrace{\lim_{x\to\infty}\left(1+\tfrac{1}{x}\right)^{x}}_{\text{e (by definition)}}\Big)^{-c}$$

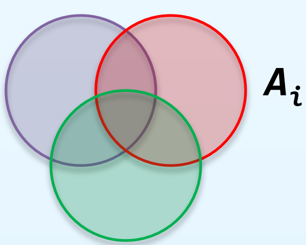
Case of No Isolated Nodes

"What *p* is before we are likely to have no isolated nodes?"

- We know $\Pr[v \text{ has degree } 0] = e^{-c}$
- Let I represent isolated nodes
  - Then, $I = \bigcup_{v \in V} I_v$ where I_v is the event that v is isolated
- So, we have

$$|I| = \left|\bigcup_{v \in V} I_v\right| \leq \sum_{v \in V} |I_v| = ne^{-c}$$

Union bound



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

Case of No Isolated Nodes

- We just learned: $|I| = ne^{-c}$
- Let's try:
  - $c = \ln(n)$ then $ne^{-c} = ne^{-\ln(n)} = 1$
  - $c = 2\ln(n)$ then $ne^{-c} = ne^{-2\ln(n)} = 1/n$
- So if:
  - $p = \ln(n)/n$ then $|I| \leq 1$
  - $p = 2\ln(n)/n$ then $|I| \leq 1/n \to 0$ as $n \to \infty$

Evolution of a Random Graph

Graph structure of $G(n,p)$ as p changes:

p	
0	empty graph
c/n	trees
$1/n$	giant component appears
c'/n	one giant component, others are trees
$\log(n)/n$	$G(n,p)$ is connected
$\Omega(\log(n)/n)$	No isolated nodes
$\Omega(n^{\varepsilon-1})$	finite diameter
$\Theta(1)$	dense graph, diameter is 2
1	complete graph

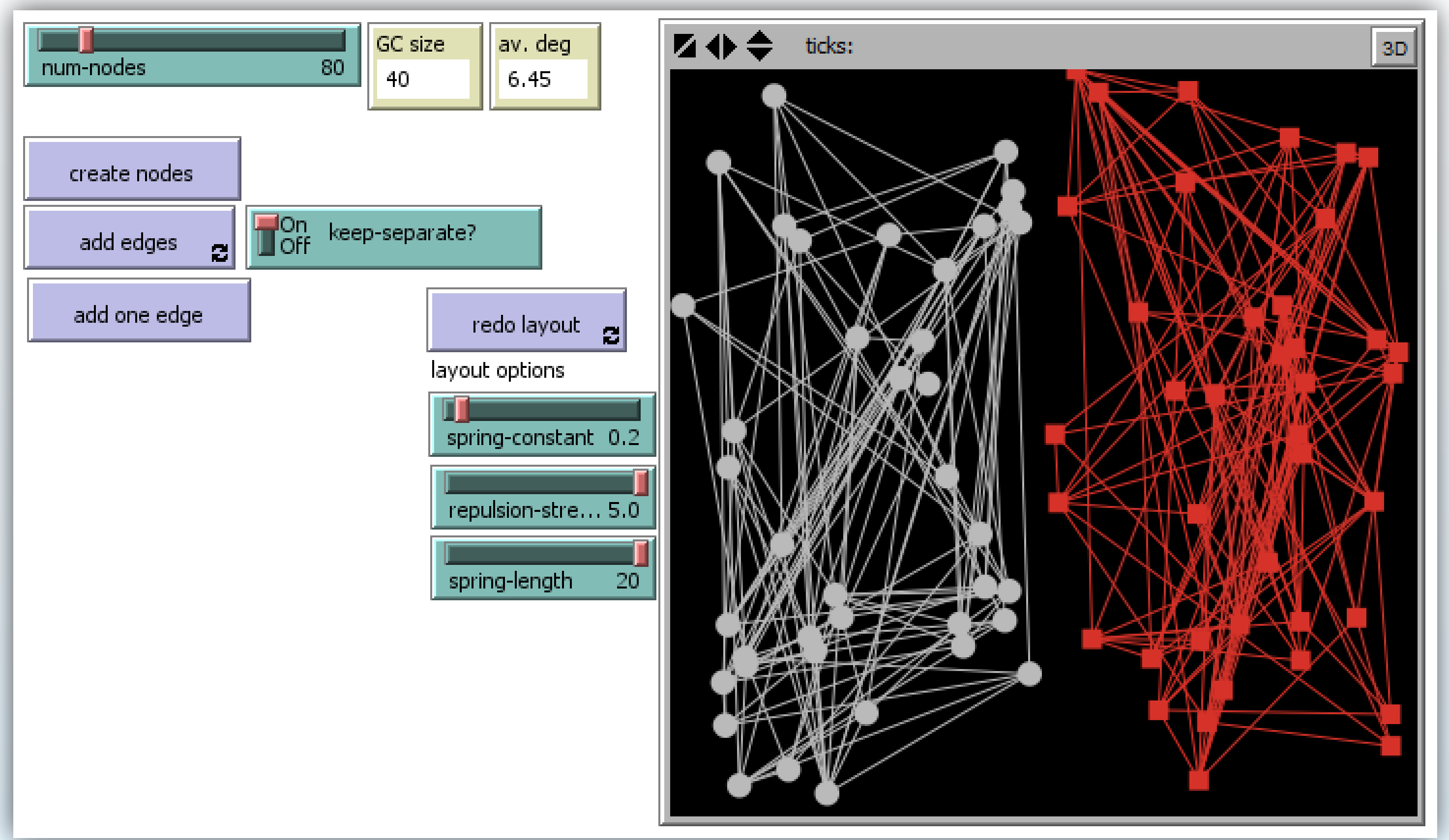
Emergence of a Giant Component

- If $p < {}^1\!/_n$
  - with high probability, there is no giant component, with all connected components of the graph having size $O(\log n)$
- If $p > {}^1\!/_n$
  - with high probability, there is a single giant component, with all other connected components having size $O(\log n)$
- If $p = {}^1\!/_n$
  - with high probability, the number of nodes in the largest component of the graph is proportion to $n^{2/3}$

Question #8

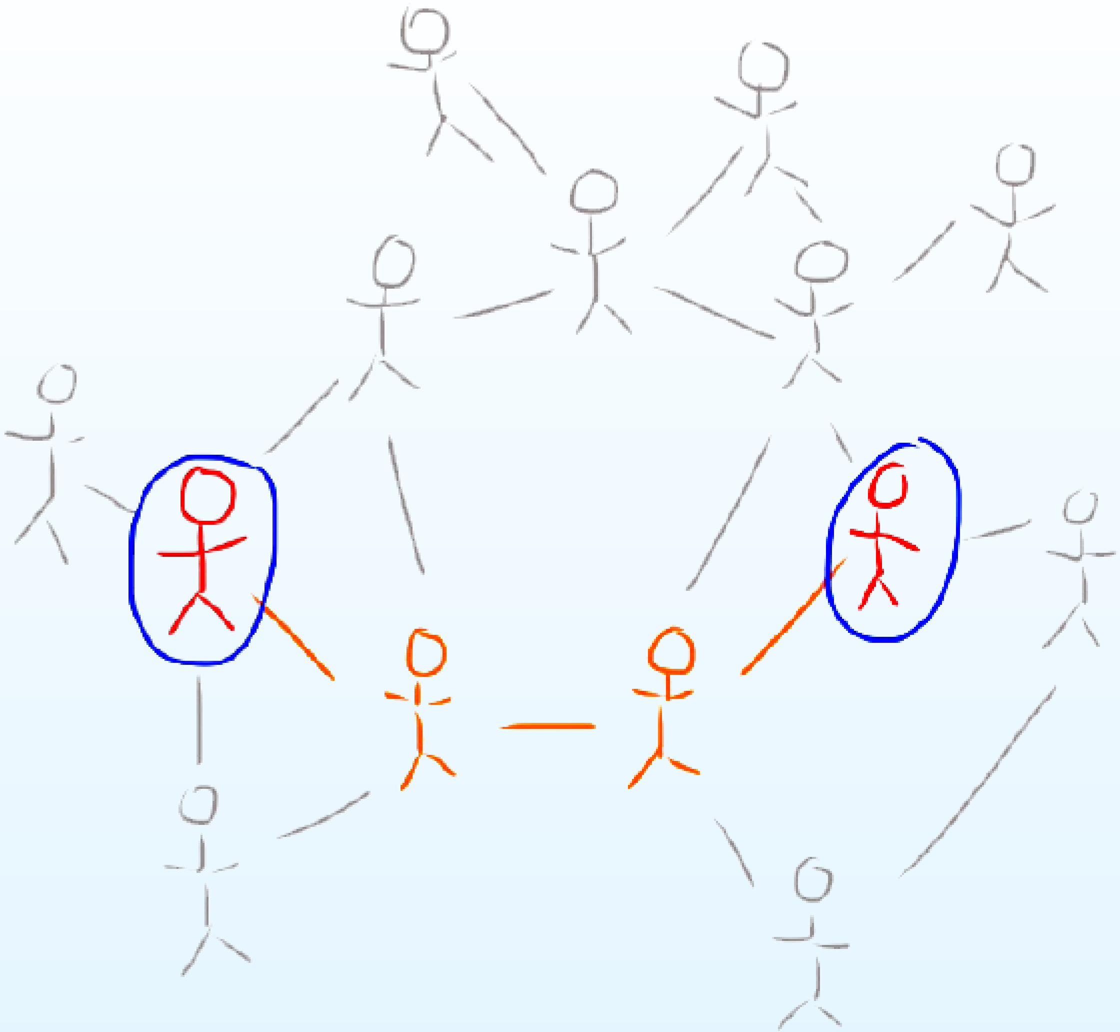
- If you have 2 large-components each occupying roughly ½ of the graph, how long does it typically take for the addition of random edges to join them into one giant component?
    a) 1-4 edge additions
    b) 5-20 edge additions
    c) over 20 edge additions

Why just One Giant Component?



ErdosRenyiTwoComponents.nlogo

Average Shortest Path

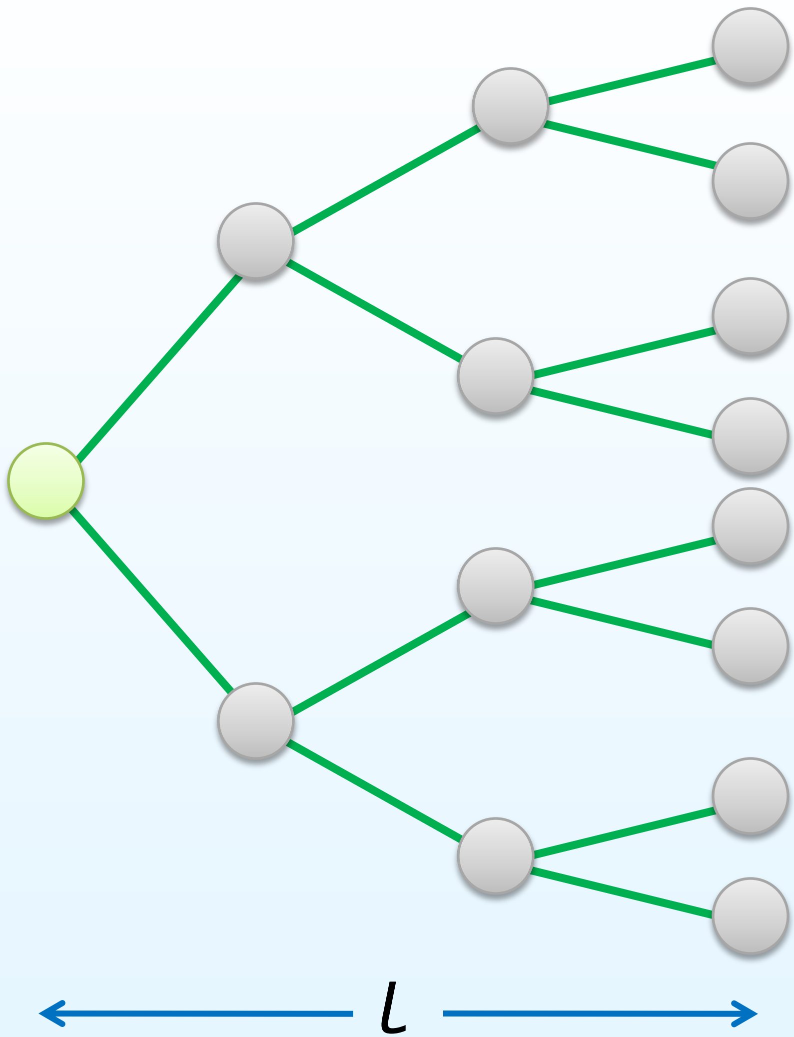


Average Shortest Path

"How many hops on average between each pair of nodes?"

- Again, each of your friends has $\langle k \rangle = avg\text{-}degree$ friends beside you
- Ignoring loops, the number of people you have at distance l is $\langle k \rangle^{l}$

Friends at Distance l



- Let N be the number of people you have at distance l

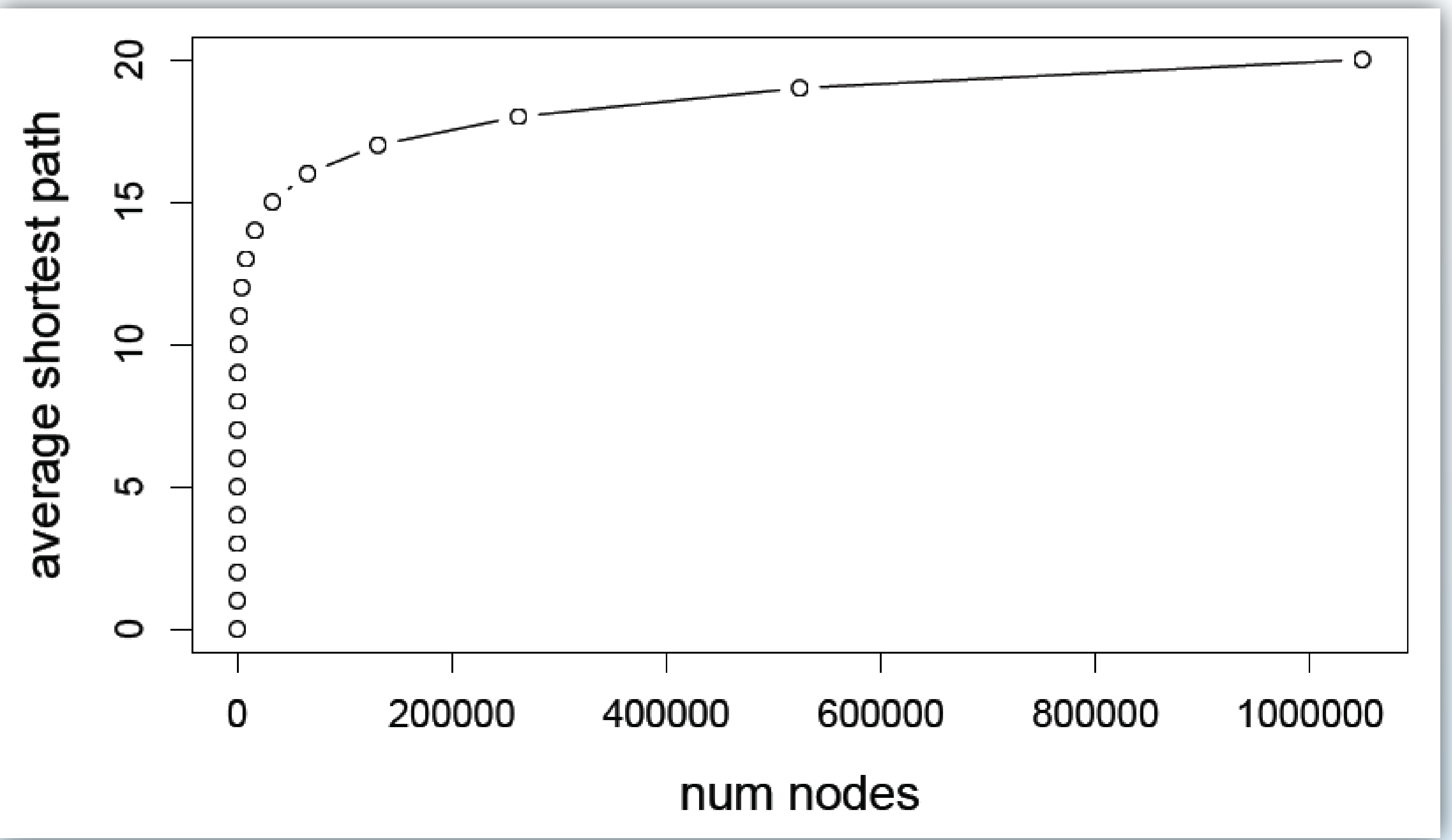
$$N = \langle k \rangle^{l}$$

- Scaling:
  - Average shortest path

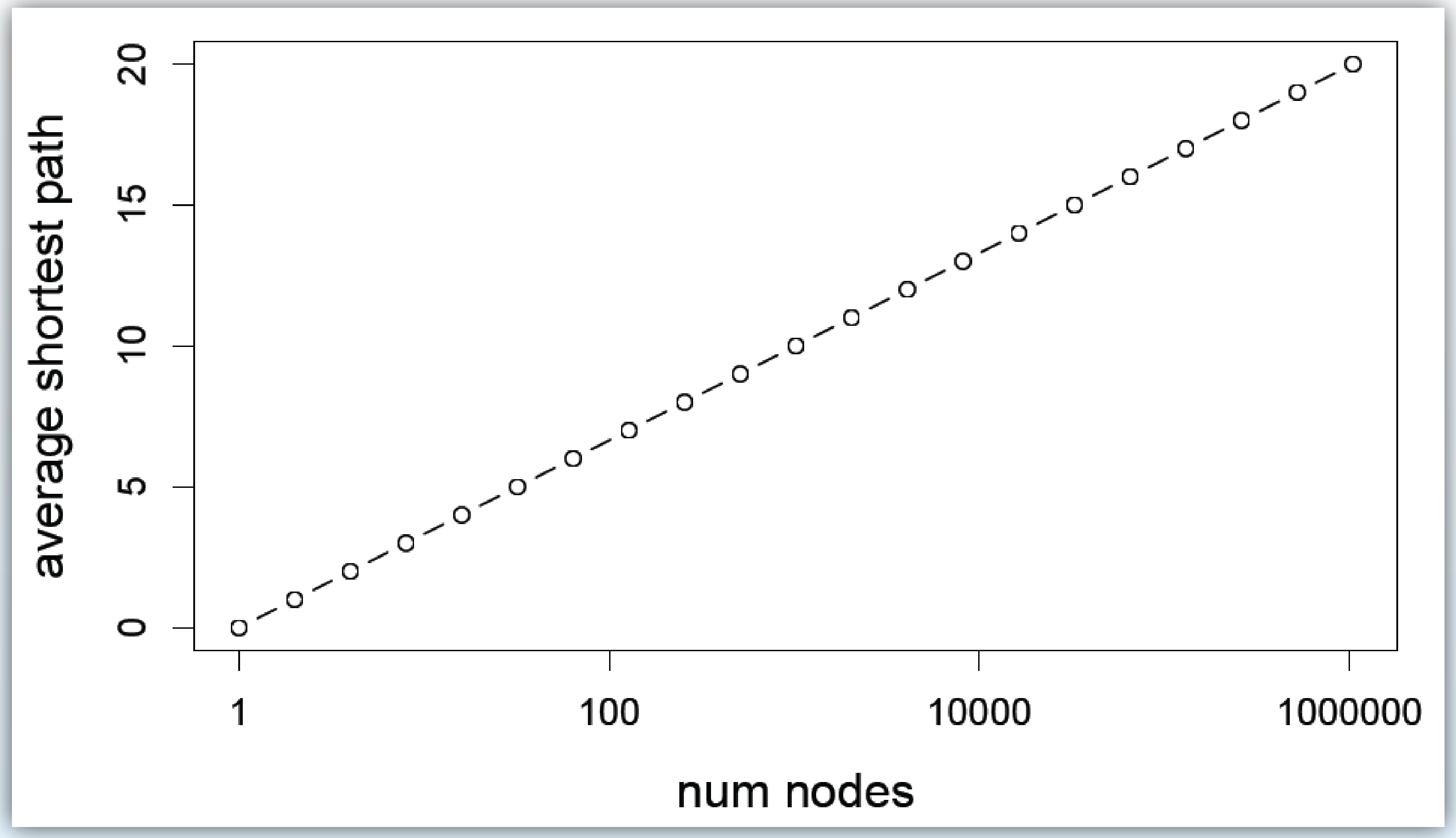
$$\boxed{l_{avg} \sim \frac{\log(N)}{\log(\langle k \rangle)}}$$

What This Means in Practice?

- Erdös-Rényi networks can grow to be very large but nodes will be just a few hops apart



Erdös-Rényi Avg. Shortest Path



Any Question?