Random graphs and Watts-Strogatz

CS 249B: Science of Networks
Week 05: Thursday, 02/28/08
Daniel Bilar
Wellesley College
Spring 2008

Goals this lecture

- How to model small world phenomenon
  - E-R "Random Graph" model
  - W-S "Small World" model
    - Regular Lattice – Random Lattice
    - Properties
- Parsing paper Watts-Strogatz (1998)

Review: Why should the world be anything other than 'small'?

It is <u>remarkable</u> because

1. The network is **numerically large** in the sense that the world contains $n >> 100$ people. In the real world, n is on the order of billions.
2. The network is **sparse** in the sense that each person is connected to an average of only k other people, which is, at most, on the order of thousands (Kochen 1989)—hundreds of thousands of times smaller than the population of the planet.
3. The network is **decentralized** in that there is **no dominant central vertex** to which most other vertices are directly connected. This implies a stronger condition than sparseness: not only must the average degree k be much less than n, but the *maximal* degree $kmax$ over all vertices must also be much less than n.
4. The network is **highly clustered**, in that most friendship circles are strongly overlapping. That is, we expect that many of our friends are friends also of each other

Historical developments

- Karinthy (1929):
  - **Hungarian novelist**: Short story "Chains", claim of '5 degrees'
- Solomonoff & Rapoport (1951)
  - **Mathematical Biology**: Introduces random graphs, weak connectivity, phase transition to giant component
- Erdos & Renyi (1960)
  - **Mathematics**: Fathers of modern random graph theory (8 papers (1950-1968)): Network property Q (giant components, subgraphs)
- Kochen and Pool (1978)
  - **Math and Poli Sci**: Social contacts, inspired Milgram
- Milgram and Travers (1967)
  - **Sociology and Psychology**: Acquaintance networks
  - 'Six degrees'
- Price (1965)
  - **Information Science:** Citation network between scientific papers
  - Power laws, 'cumulative advantage' mechanism
- Watts-Strogatz (1998)

.... and more to come

Review: Network Model Requirements

- (Good) models and reality should have properties of real-life networks in common!

- Features of real life networks that we have encountered so far
  - **Power law** distributions on degree $\boldsymbol{k_i}$
    - "Fat tail", "Megavalues do occur"
  - **Short** average path length $\boldsymbol{l}$
    - "Small world", "Six degrees of separation"
  - **High** average clustering $\boldsymbol{C}$
    - "My friends are friends with each other"

  - … and a few other properties we'll see in time

➔ Through what processes can the emergence of these features be explained?
  - ➔ One is the classic "random graph" model: Erdös-Renyi

Random Graphs

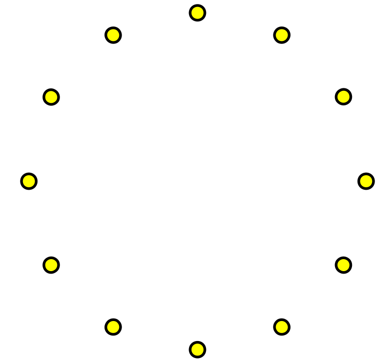
Erdős and Renyi (1959)

- *N* nodes .. Any two nodes has probability *p* of being connected
- Each of *n(n-1)/2* edges appear independently with probability *p*.
- Average degree <k> ≈ pN

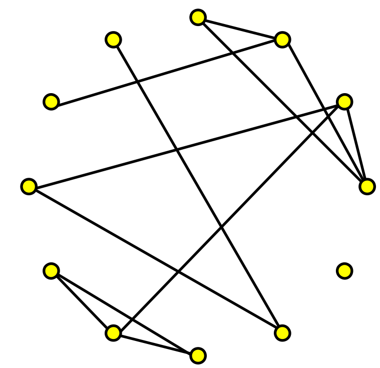
What interesting things can be said for different values of p or k ? (that are true as $N \rightarrow \infty$)

N = 12

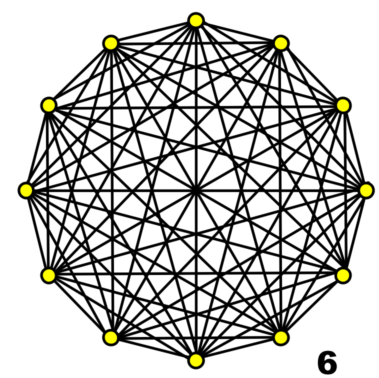
p = 0.0 ; *k* = 0



p = 0.09 ; *k* = 1



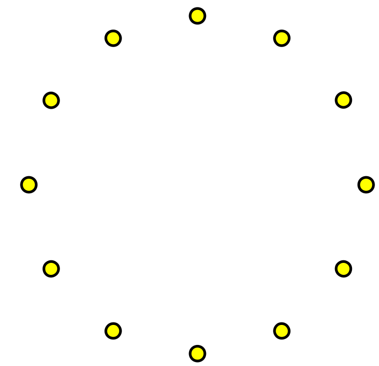
p = 1.0 ; *k* ≈ *N*

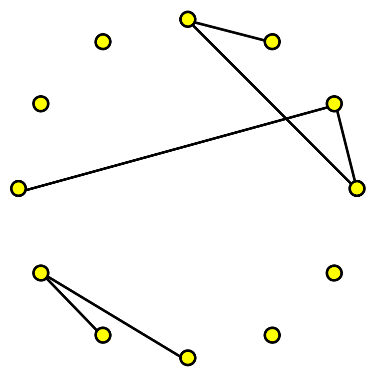


Random Graphs

Erdős and Renyi (1959)

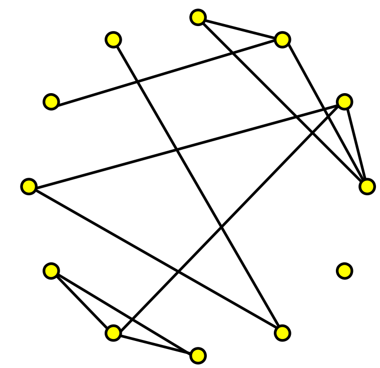
$p = 0.0$; $k = 0$





$p = 0.045$; $k = 0.5$

$p = 0.09$; $k = 1$



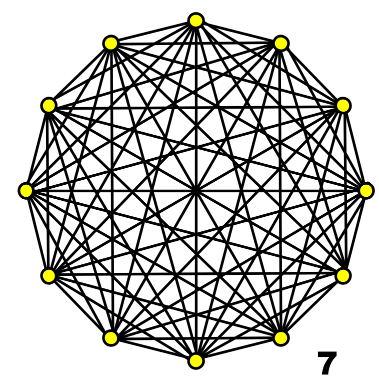
Let's look at…

Size of the largest connected cluster

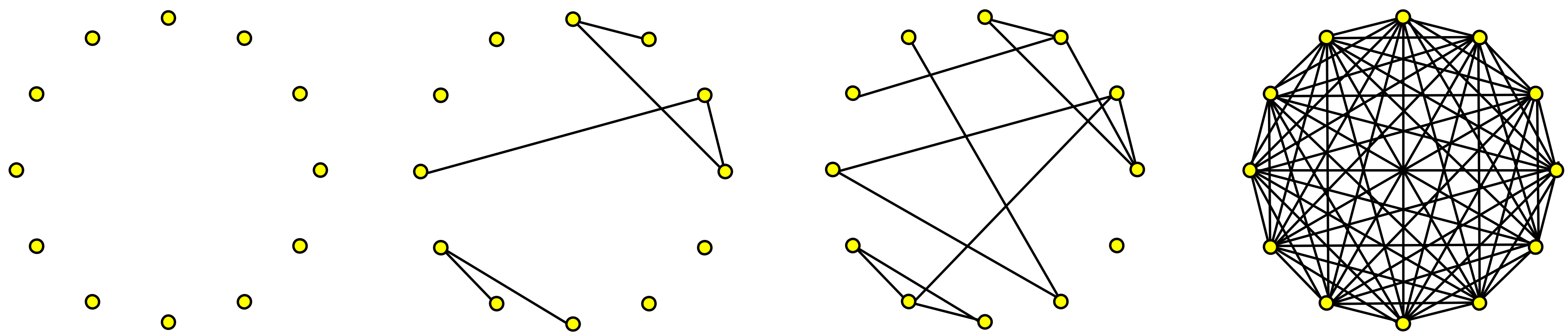
Diameter (maximum path length between nodes) of the largest cluster

Average path length between nodes (if a path exists)

$p = 1.0$; $k \approx N$



Random Graphs



$p = 0.0$; $k = 0$ — $p = 0.045$; $k = 0.5$ — $p = 0.09$; $k = 1$ — $p = 1.0$; $k \approx N$

	$p = 0.0$; $k = 0$	$p = 0.045$; $k = 0.5$	$p = 0.09$; $k = 1$	$p = 1.0$; $k \approx N$
Size of largest component	1	5	11	12
Diameter of largest component	0	4	7	1
Average path length between nodes	0.0	2.0	4.2	1.0

Effect of average degree <k>

If $k < 1$:

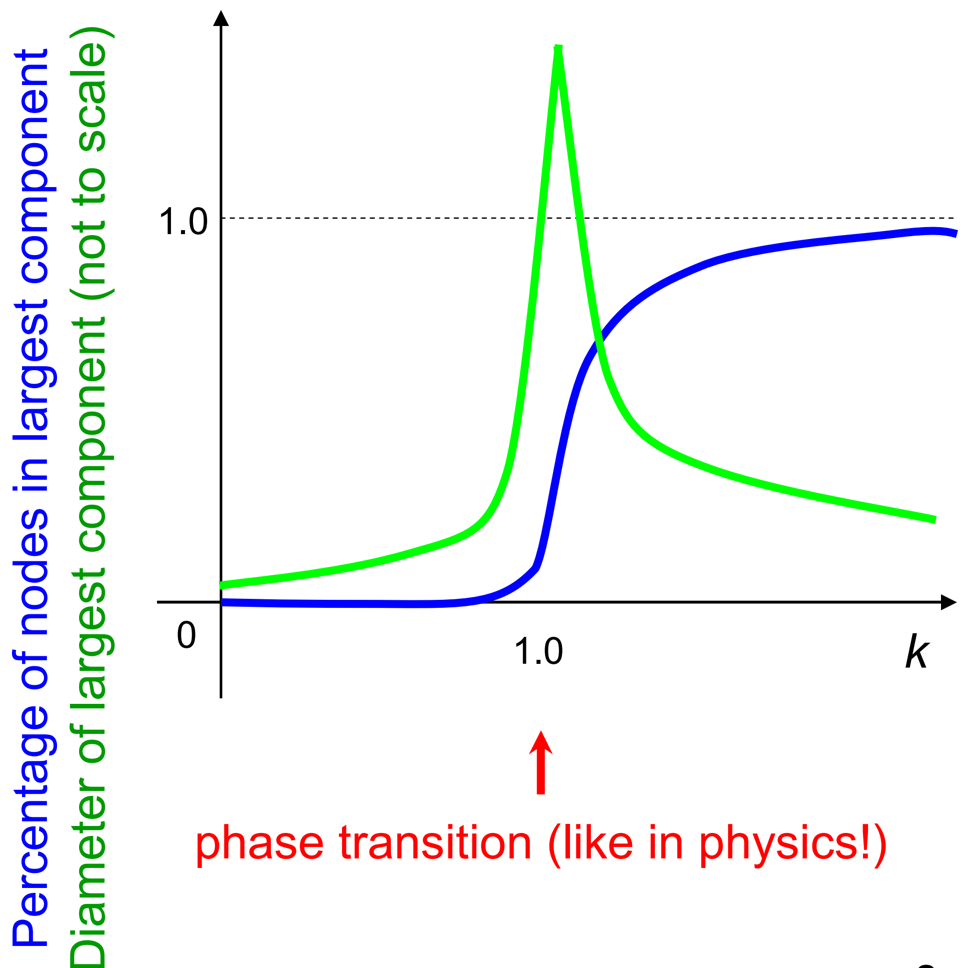
- small, isolated clusters
- small diameters
- short path lengths

At k = 1:

- a *giant component* appears
- diameter peaks
- path lengths are high

For k > 1:

- almost all nodes connected
- diameter shrinks
- path lengths shorten



Property "Giant component"

- In many/most (not all!) real-life networks, we see
  - *Small* diameter
  - A *high* degree of clustering
  - A *heavy-tailed* degree distribution

  … and
  - *Few* connected components:
    - often only 1 or a small number independent of network size

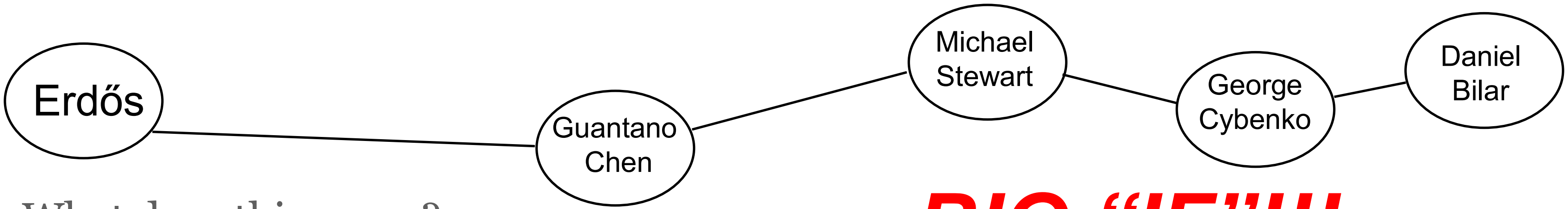
➔ The formation of a giant component is an example of a property Q (here the giant component) *emerging* at a threshold probability $P_c(Q)$

We'll encounter this sort of phenomenon ("Tipping points" (see Malcom Gladwell) and discuss the significance in coming lectures

Random Graphs

- Does E-R model explain the closeness of the world?
- If connections between people can be modeled as a random graph, then…
  - Because the average person easily knows more than one person (k >> 1), we live in a “small world” where within a few links, we are connected to anyone in the world.
  - Erdős and Renyi showed that average path length between connected nodes is $\dfrac{\ln N}{\ln k}$

Random Graphs



What does this mean?

BIG "IF"!!! What would indicate whether or not this could be plausible?

- If connections between people can be modeled as a random graph, then…
  - Because the average person easily knows more than one person ($k >> 1$),
  - We live in a "small world" where within a few links, we are connected to anyone in the world.
  - Erdős and Renyi computed average path length between connected nodes to be:

$$\frac{\ln N}{\ln k}$$

Random Graph: Clust. Coeff. <C>

- Compare real-life network's <C> to <C> in E-R model

→ In a random graph: $C_{\text{rand}} \sim 1/N$ (if the average degree <k> is held constant)

Network	N	ℓ	C	C_{rand}
movie actors	225 226	3.65	0.79	0.00027
neural network	282	2.65	0.28	0.05
power grid	4941	18.7	0.08	0.0005

→ E-R model is not a model that generates real life small-world graphs!

More comparative data on *l* and *C*

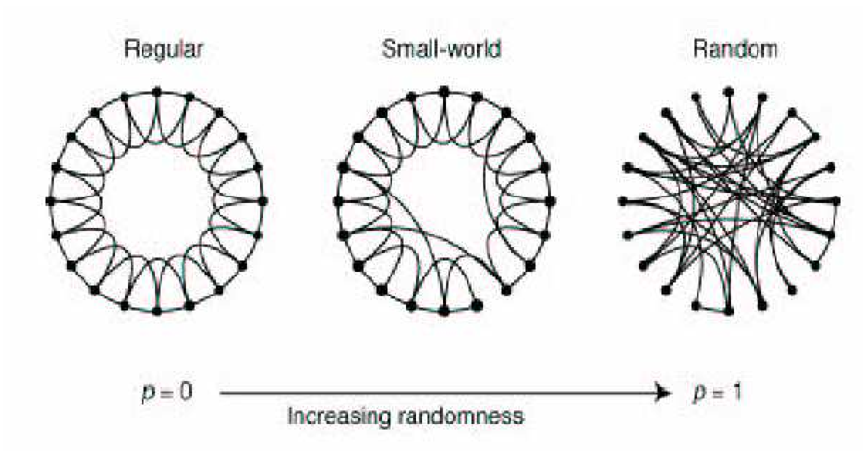
TABLE I. The general characteristics of several real networks. For each network we indicated the number of nodes, the average degree $\langle k \rangle$, the average path length ℓ and the clustering coefficient C. For a comparison we have included the average path length ℓ_{rand} and clustering coefficient C_{rand} of a random graph with the same size and average degree. The last column identifies the symbols in Figs. 8 and 9.

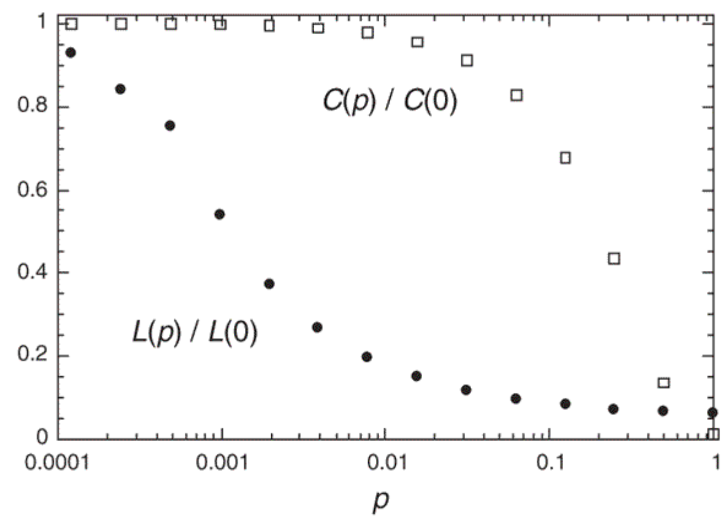
Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference	Nr.
WWW, site level, undir.	153,127	35.21	3.1	3.35	0.1078	0.00023	Adamic 1999	1
Internet, domain level	3015 - 6209	3.52 - 4.11	3.7 - 3.76	6.36 - 6.18	0.18 - 0.3	0.001	Yook *et al.* 2001a, Pastor-Satorras *et al.* 2001	2
Movie actors	225,226	61	3.65	2.99	0.79	0.00027	Watts, Strogatz 1998	3
LANL coauthorship	52,909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman 2001a,b	4
MEDLINE coauthorship	1,520,251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman 2001a,b	5
SPIRES coauthorship	56,627	173	4.0	2.12	0.726	0.003	Newman 2001a,b,c	6
NCSTRL coauthorship	11,994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman 2001a,b	7
Math coauthorship	70,975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.* 2001	8
Neurosci. coauthorship	209,293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.* 2001	9
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner, Fell 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner, Fell 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya, Solé 2000	12
Silwood park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya, Solé 2000	13
Words, cooccurence	460.902	70.13	2.67	3.03	0.437	0.0001	Cancho, Solé 2001	14
Words, synonyms	22,311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.* 2001	15
Power grid	4,941	2.67	18.7	12.4	0.08	0.005	Watts, Strogatz 1998	16
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts, Strogatz 1998	17

Albert-Barabasi (2002): “Statistical mechanics of complex networks”, p.8

Watts-Strogatz 'Small World' (1998)

- WS constructed a model in which networks can have both short path lengths ***l*** like ER … <u>and</u> high clustering ***C***
- Result of W-S model in English:
  - "a few random links in an otherwise clustered graph give an average shortest path close to that of a random graph"





D. J. Watts and S. H. Strogatz, *Collective dynamics of "small-world" networks*, Nature, 393 (1998), pp. 440–442.

W-S model investigation

- Effects of rewiring probability ***p***
  - Does small shortest path mean small clustering? And large shortest path mean large clustering?

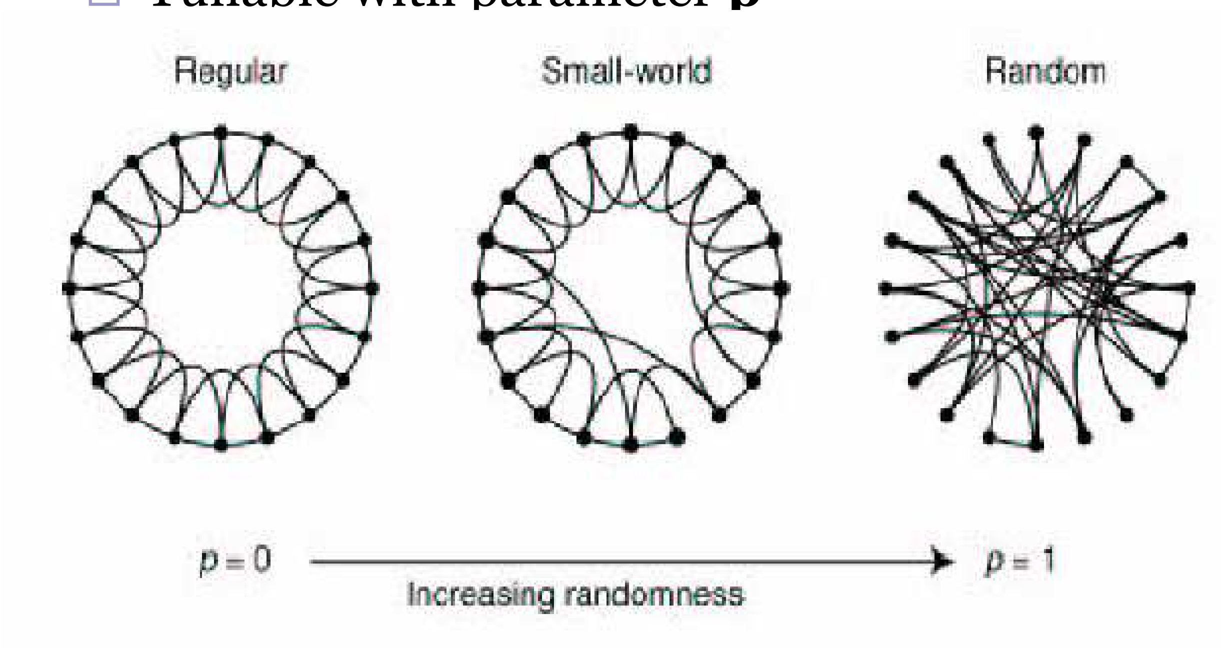
- Watts figured it out through numerical simulation

  As he increased p from 0 to 1 ….
  - **Fast** decrease of mean distance ***l(p)***
  - **Slow** decrease in clustering ***C(p)***

→ Exist range of ***p*** which generated graphs with short path lengths ***l(p)*** and high clustering ***C(p)***

Original W-S model

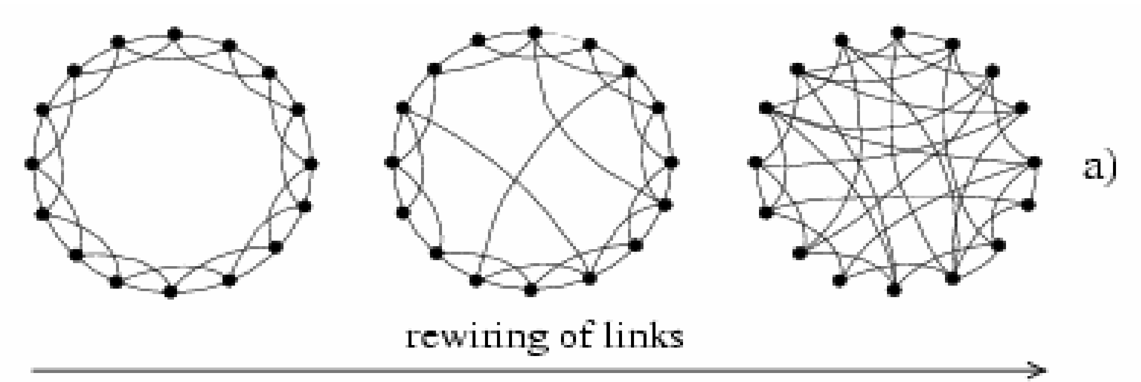
- Each node has K>=4 nearest neighbors (local)
- Probability ***p*** of rewiring to randomly chosen nodes
  - p small gives "**Regular lattice**" **(**also called **"Ordered")**
  - p large gives "**Random graph**"
  - Tunable with parameter **p**



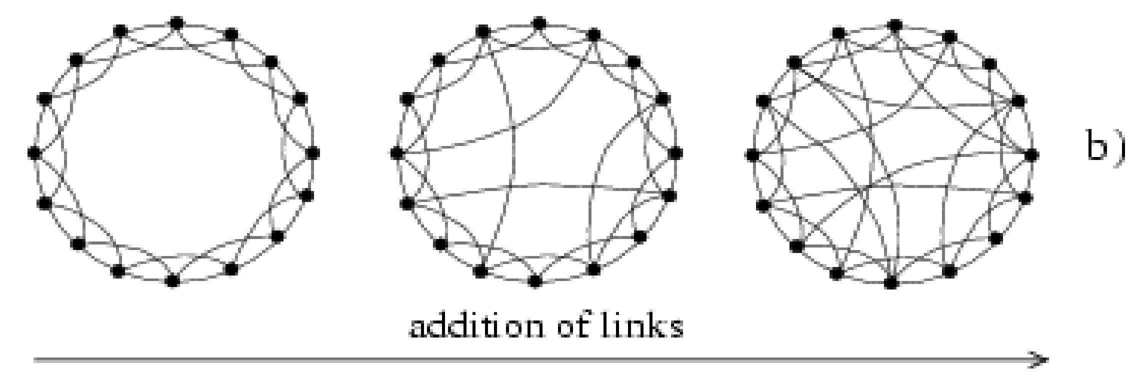
What is a

➔"Regular Lattice" ?

Constructing W-S



Select a fraction ***p*** of edges
Reposition on of their endpoints



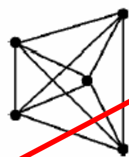
Add a fraction **p** of additional edges leaving underlying lattice intact

- No loops or multiple edges allowed

C(p) for Ordered lattice (p =0)

Cliques (completely connected subgraphs)

$$k = N - 1,\ n = \frac{N(N-1)}{2}$$



How close the neighborhood of a node is to a clique?

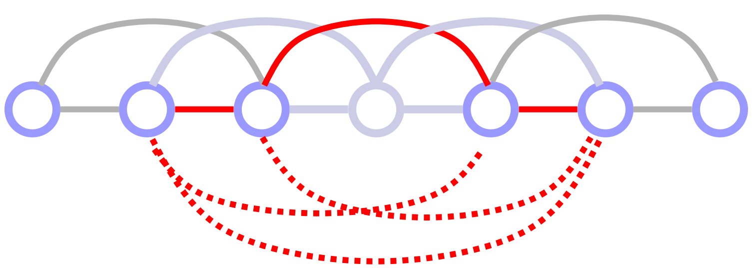
Edges among first neighbors of node *i*

$$C_i \equiv \frac{n_i}{k_i(k_i-1)/2},\quad k \neq 0,1$$

or

Recall Clustering Coefficient C

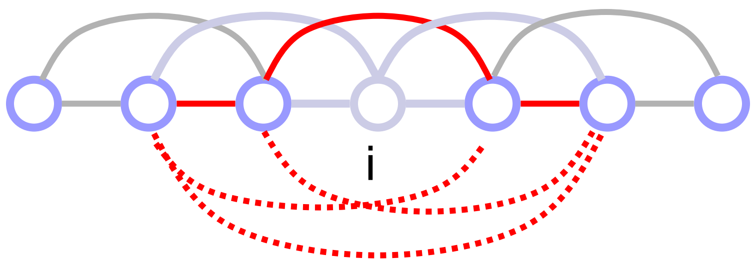
- k neighbors, who can have $k*(k-1)/2$ pairwise connections between them
- Some of the connections between them are present in the lattice
- Here, for K=4:



$C_i = 3/ ((4*3)/2) = ½$

$\langle C_i \rangle$ roughly ½ as well

C(p) for Ordered, K-connected lattice



Which nodes do *i* and *i's* neighbours have in common?

- **K/2** hops away from i
  - can connect to (K/2 – 1) of i's neighbors
- **K/2-1** hops away from i
  - can connect to (1 + K/2 – 1) neighbors
- **K/2 – 2** hops away from i
  - (2 + K/2 – 1) neighbors

....

- **1 hop** away from i
  - 2*(K/2 – 1)

➔ Sum this up

Note: We have to multiply by factor of 2 because *i* has neighbors on both sides but also have to divide by a factor of 2 because edges are undirected -> no net effect on summation

The number of connections between neighbors is given by

$$\sum_{j=0}^{\frac{K}{2}-1}\left(\frac{K}{2}+i-1\right)=\frac{3}{8}K(K-2)$$

C(p) for ordered, K-connected lattice (cont)

- The number of connections between neighbors is given by

$$\sum_{j=0}^{\frac{K}{2}-1}(\frac{K}{2}+i-1)=\frac{3}{8}K(K-2)$$

- The maximum number of connections is k*(k-1)/2

➔ Clustering coefficient is

$$\boxed{C=\frac{3(K-2)}{4(K-1)}}$$

l(p) for Ordered lattice

- Recall: Geodesic distance $\boldsymbol{l_{i,j}}$ between vertices *i* and *j is* the shortest path connecting *i* and *j*

- We are also interested in <***l***>, the *average* geodesic distance between vertex pairs in G
  - Also called sometimes ***average path***

- Average node is N/4 hops away (a quarter of the way around the ring), and you can hop over K/2 nodes at a time

$$l \approx \frac{N}{2K} >> 1$$

C(p), l(p) for Random Lattice (p >> 0)

- There are an average of K links per node.
- The probability that any two nodes are connected is p = K/N

$$l \approx \frac{\ln N}{\ln K} \qquad \text{small}$$

$$C \approx \frac{K}{N} \qquad \text{small}$$

- The probability that two nodes which share in a neighbor in common are connected themselves is the same as any two random nodes: K/N
  - (actually (K-1)/N because they have already expended one edge on their common neighbor.

➔ As p=1, W-S is (almost) like Erdos-Renyi random graphs

Summary: Regular vs Random Graph

- Ordered Graphs (p approx 0)
  - have a high clustering coefficient but high path lengths
- Random Graphs (p >> 0)
  - have low path length but a low clustering coefficient
- Each match the properties expected from real networks! Want BOTH!

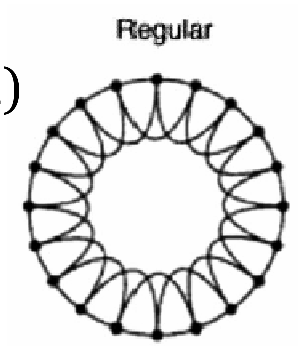
Ordered Graph (k=4)

Long paths

- $L \sim n/(2k)$

Highly clustered

- $C \sim 3/4$



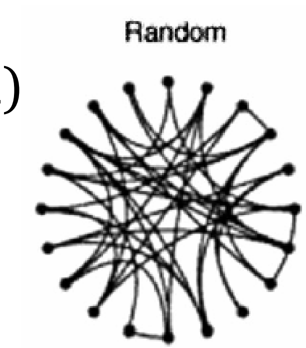
Random Graph (k=4)

Short path length

- $L \sim \log_k N$

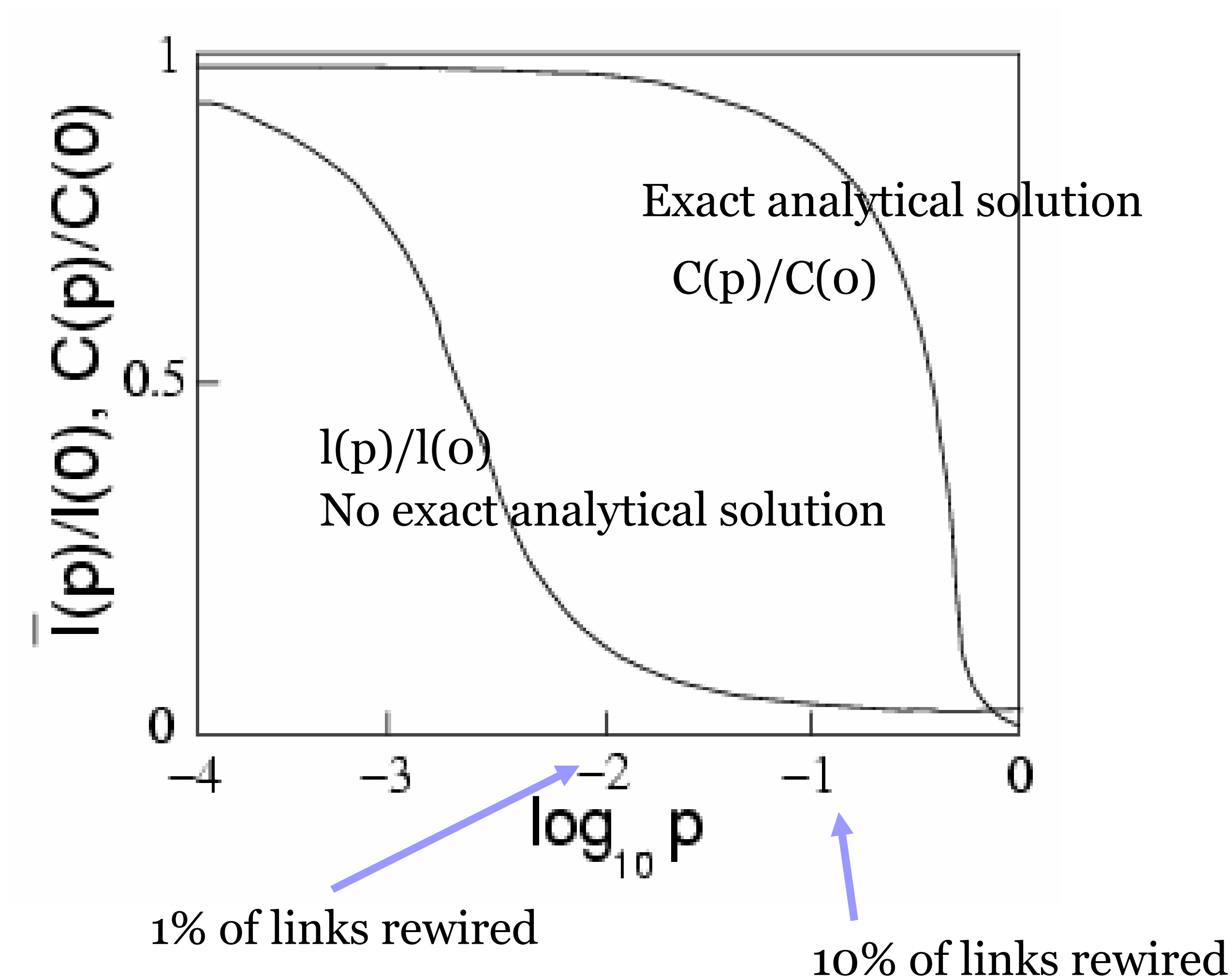
Almost no clustering

- $C \sim k/n$

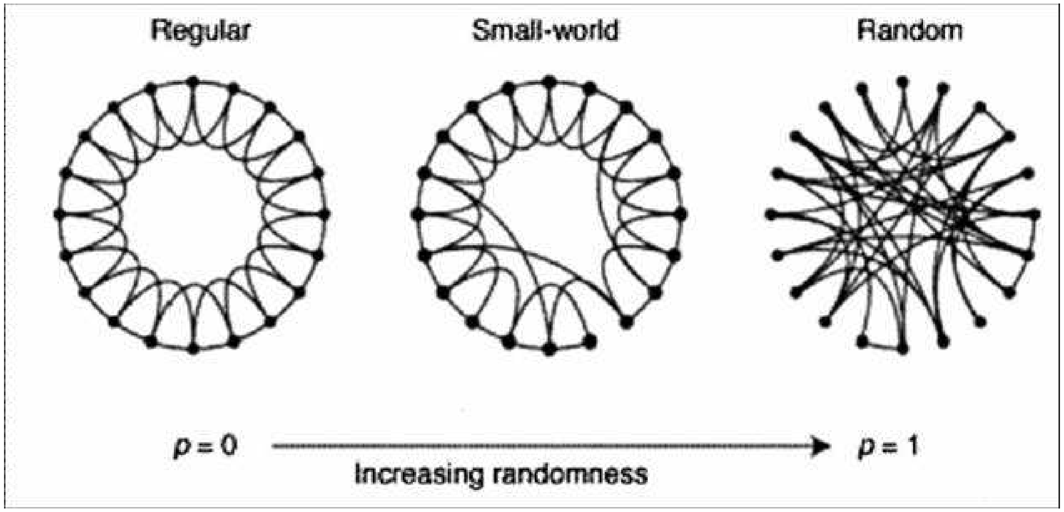


Tune *p*

- Change in clustering coefficient and average path length as a function of the proportion of rewired edges

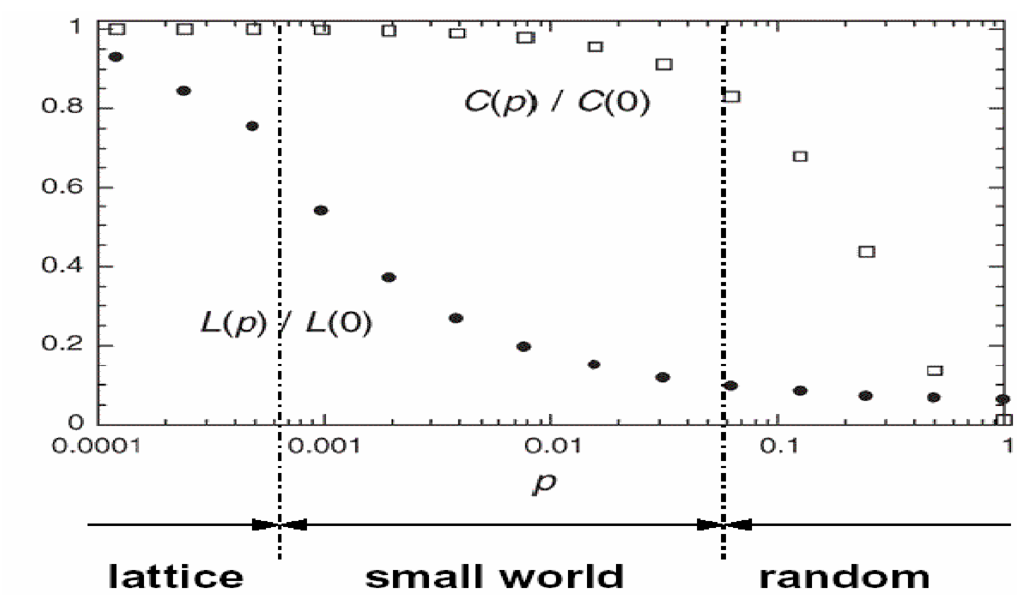


Small-World is “Sweet Spot”



If $n \gg k \gg \ln n \gg 1$ then

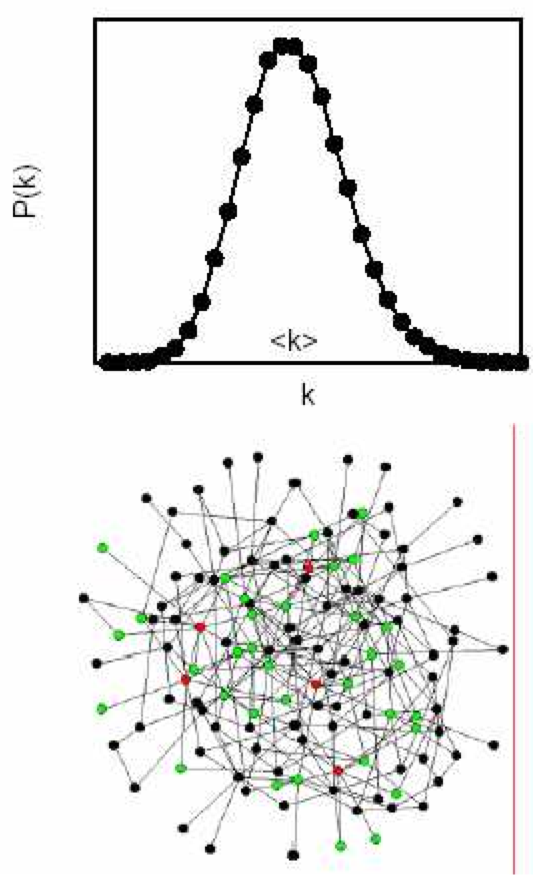
- for $p \approx 0$:
  $l \sim n/2k$ and $C \sim 3/4$
  Ordered lattice
- for $p \approx 1$:
  $l \sim \ln n / \ln k$ and $C \sim k/n$
  Random Network
- for $0.001 \lesssim p \lesssim 0.01$:
  $l \sim \ln n / \ln k$ and $C \sim 3/4$
  This is a **small world network**

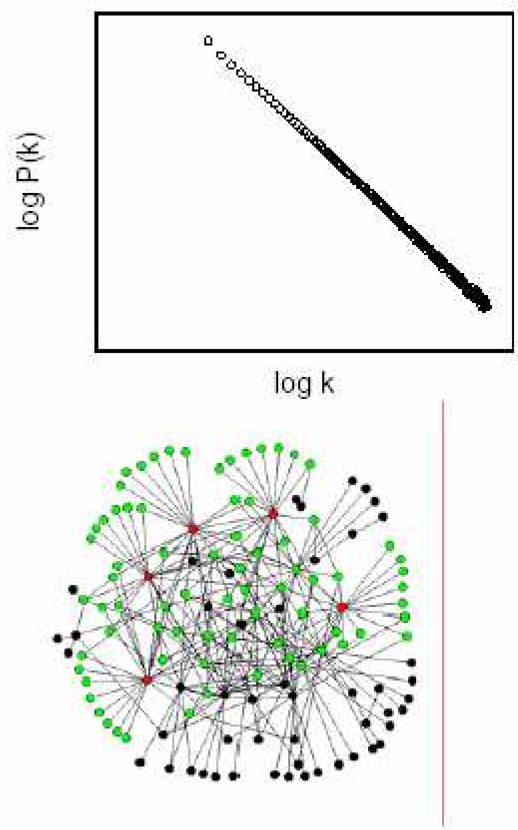


- ➔ Does this model real life networks?
- ➔ Is there still a fly (or two) in the ointment?

Fly #1: Degree distribution

- Watts-Strogatz
  - <k> approx. K
  - P(k) ~ Poisson(K)
- Real-Life
  - P(k) ~ $k^{-\alpha}$





Fly #2: Mechanism

- W-S assume
  - Fixed N
    - But Networks grow and shrink
  - Equal (rewiring/addition of link) probability p
    - Does not sound right either .. what about the “rich getting richer” ?

For next week

- Monday
  - Reread Watts (1998)
- Thursday
  - 60-90 min: Alberich (2002) "Marvel Universe looks almost like a real social network"
    - Write down terms, concepts that are new/unclear
    - Draw concept map to hand in to me