Network Analysis:

The Hidden Structures behind the Webs We Weave

17-338 / 17-668

Connectedness and Random Networks

Tuesday, September 3, 2024

Patrick Park & Bogdan Vasilescu

Carnegie Mellon University
School of Computer Science

S3D
Software and Societal Systems Department

2-min Quiz, on Canvas

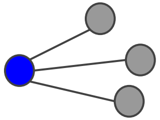
Quick Recap – Last Thursday's Lecture

Graph theory as our basic formalism for modeling networks

Basic building blocks: nodes and links

Most basic structure: dyads

Degree and degree distribution



Paths (shortest paths)

- The Breadth-first search algorithm to compute distances

Adjacency matrices as an algebraic representation of networks

- Network properties as matrix operations!

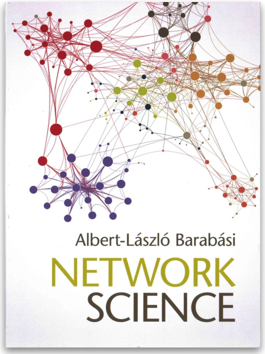
Plan for Today

More on connectedness and connected components

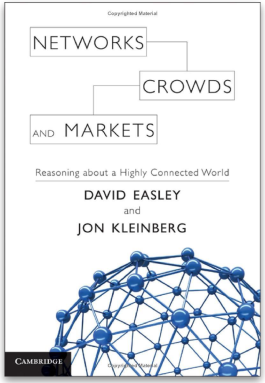
Random graphs, revisiting Six Degrees of Kevin Bacon

Larger building blocks: from dyads to triads

(B Ch. 2.9–2.10, Ch. 3 except 3.9)



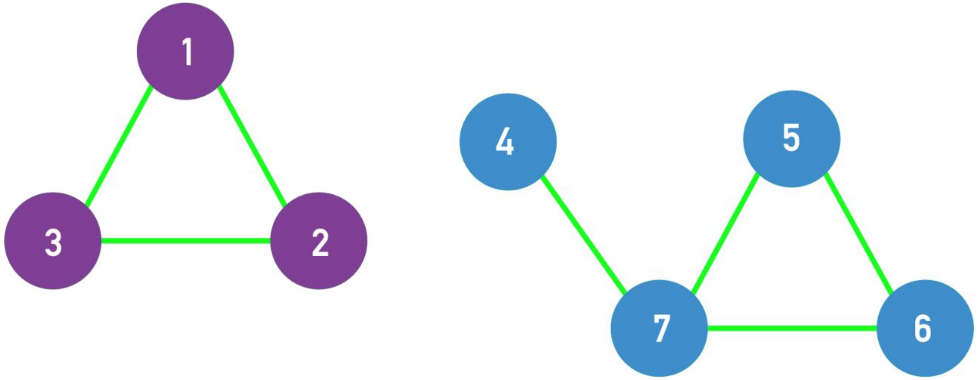
(E&K Ch. 4)



Connectedness

In a “Connected” Graph, There Is a Path Between Every Pair of Nodes

This example shows two disconnected components. If a network has disconnected components, the adjacency matrix (right) can be rearranged into a block diagonal form.

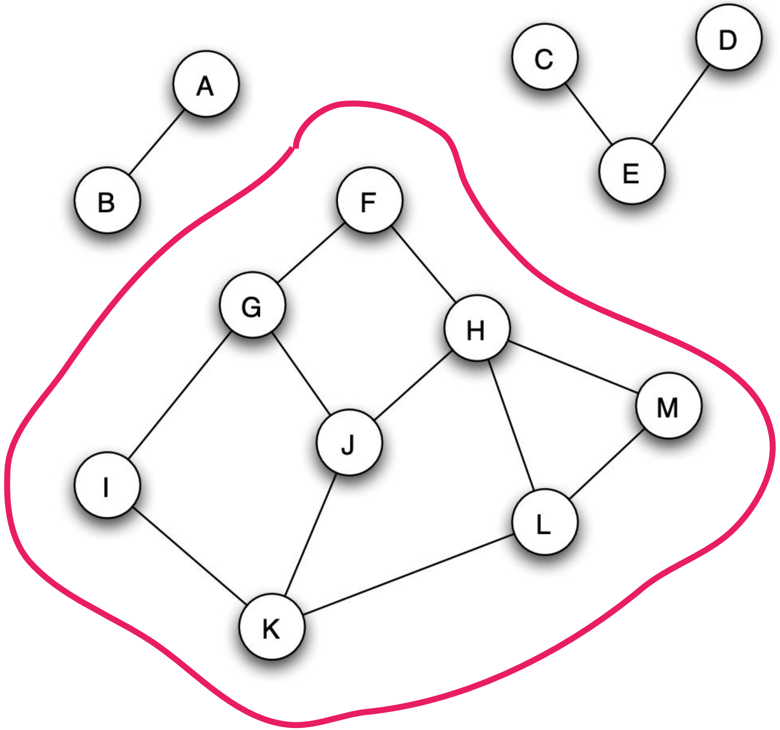


$$
\begin{pmatrix}
0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 & 0
\end{pmatrix}
$$

(Barabasi, 2016)

When a Network Contains a Giant Component, It Almost Always Contains Only One

Why?



When a Network Contains a Giant Component, It Almost Always Contains Only One

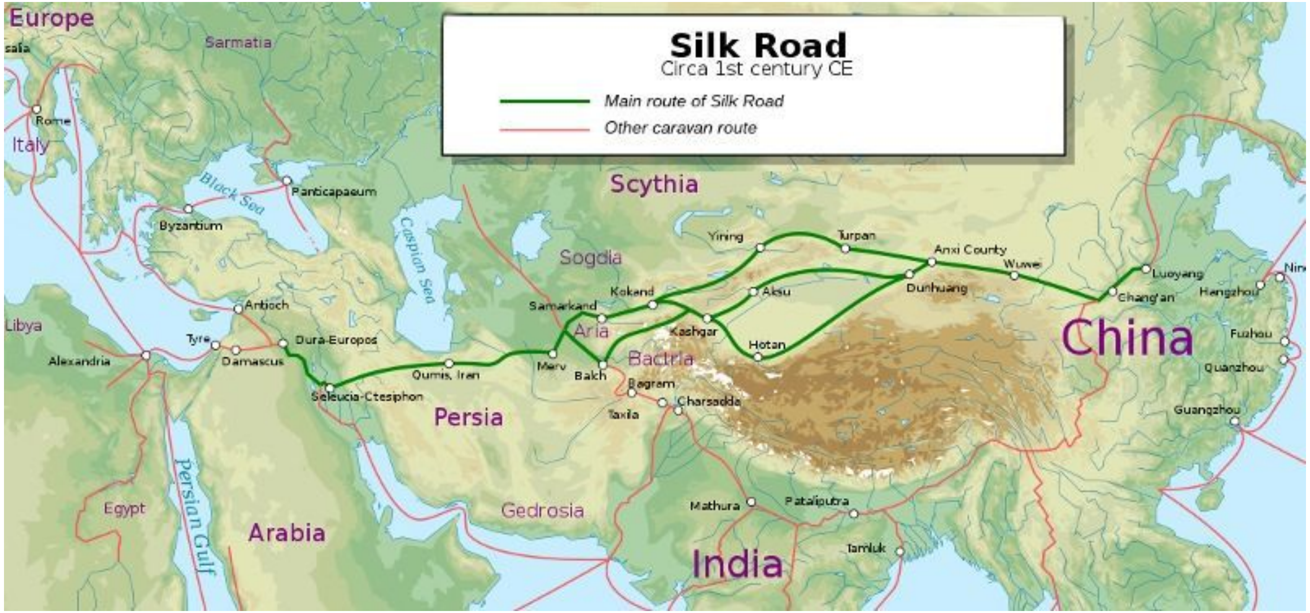
Imagine there were two giant components in the global friendship network example, each with hundreds of millions of people.

All it would take is a single edge from someone in the first of these components to someone in the second, and the two giant components would merge into a single component!

It’s essentially inconceivable that some such edge wouldn’t form, and hence two co-existing giant components are almost never seen in real networks.

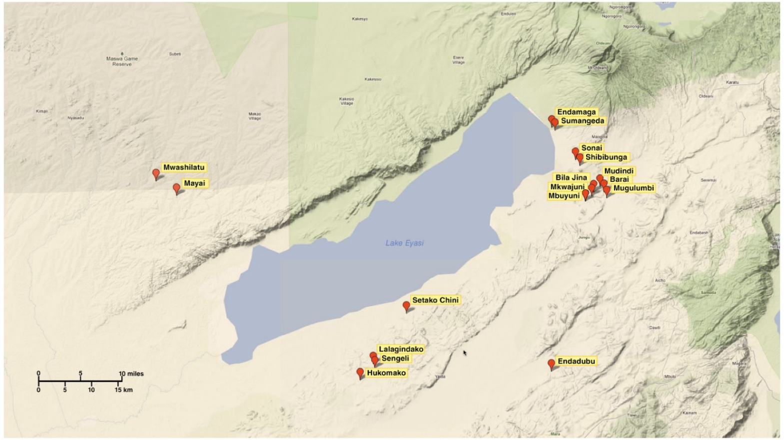
When a Network Contains a Giant Component, It Almost Always Contains Only One

Example: Silk Road



When a Network Contains a Giant Component, It Almost Always Contains Only One

Example: Hunter-gatherer society (Apicella et al. 2012)



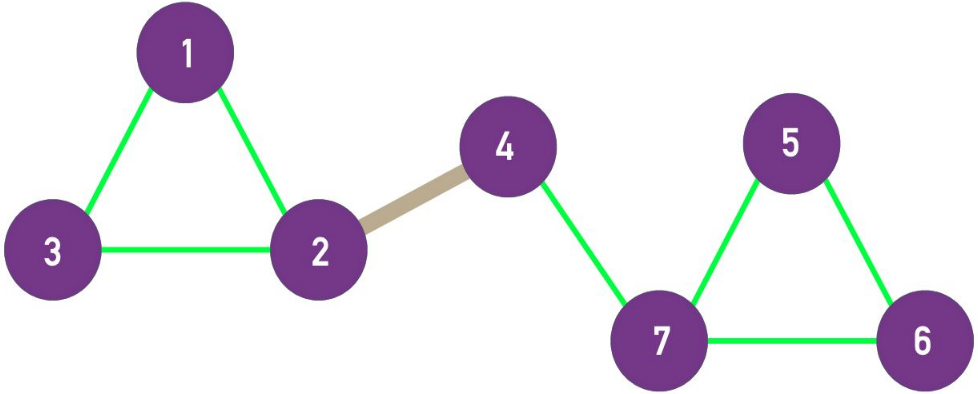
Supplementary Figure S1: Map showing the location of 17 different Hadza camps visited around Lake Eyasi in Tanzania.

Nominations Between Camps

	Barai	Bila Jina	Endadubu	Endamaga	Hukomako	Lalagindako	Mayai	Mbuyuni	Mizeu	Mkwajuni	Mudindi	Mugulumbi	Mwashilatu	Sengeli	Setako Chini	Shibibunga	Sonai	Sumangeda
Barai	23			2		1				4	6	7					1	
Bila Jina	2	18	2	1	3			1	1	2	3	1		1		3	2	
Endadubu			39		6	1						1					1	
Endamaga	1			25					1	3	2	4		1		3		
Hukomako			8		36	4	1			5				4	2		1	
Lalagindako				1	5	4			1	3	1	1		4	2		1	
Mayai							5	2					2			1		
Mbuyuni		1				1	1	11		2	1	5		2	1	2	1	1
Mizeu						1			3		1			4	1			1
Mkwajuni	1		5		4	2		3	1	44	1	1		2	3	2		2
Mudindi	3		2			2		2	1	3	25	4		1		5	2	
Mugulumbi	4	1	2					2		5	2	27				3	2	
Mwashilatu			1				3	3		1		1	50	1		1		1
Sengeli			1	1	5	2		1	1	1				13	1	1		
Setako Chini	1				1	3		1	4	4	1	2		4	22		1	
Shibibunga	3	3		1			1	3	1	1	5	3	2	1		30	1	2
Sonai	1	5		2				3		3	2	7			1	3	12	1
Sumangeda	1			1				2		1	1	3		2		4		6

A “Bridge” (2–4) Can Turn a Disconnected Network Into a Single Connected Component.

Note: The adjacency matrix cannot be written in a block diagonal form.



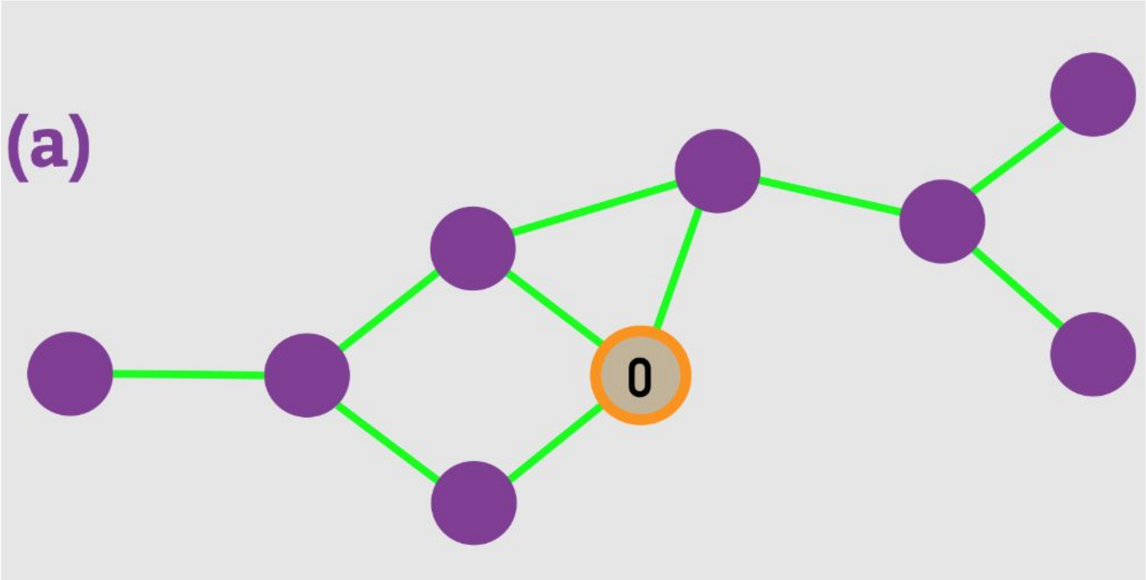
$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

(Barabasi, 2016)

Recall the BFS Algorithm

Assume we're starting from the orange node, labeled "0."

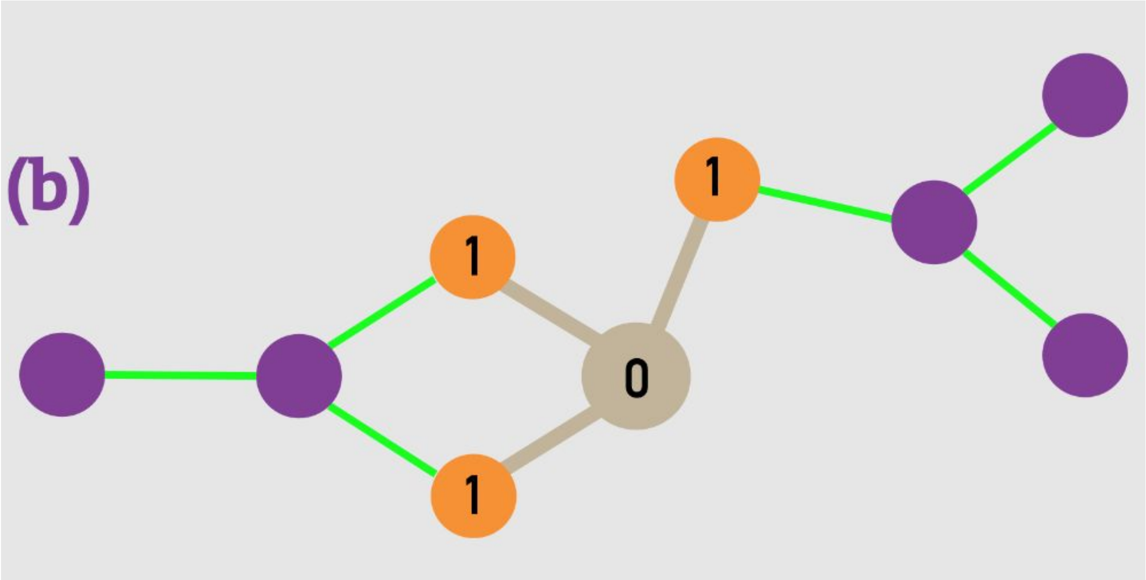
First, we identify all its neighbors, labeling them "1".



(Barabasi, 2016)

Recall the BFS Algorithm

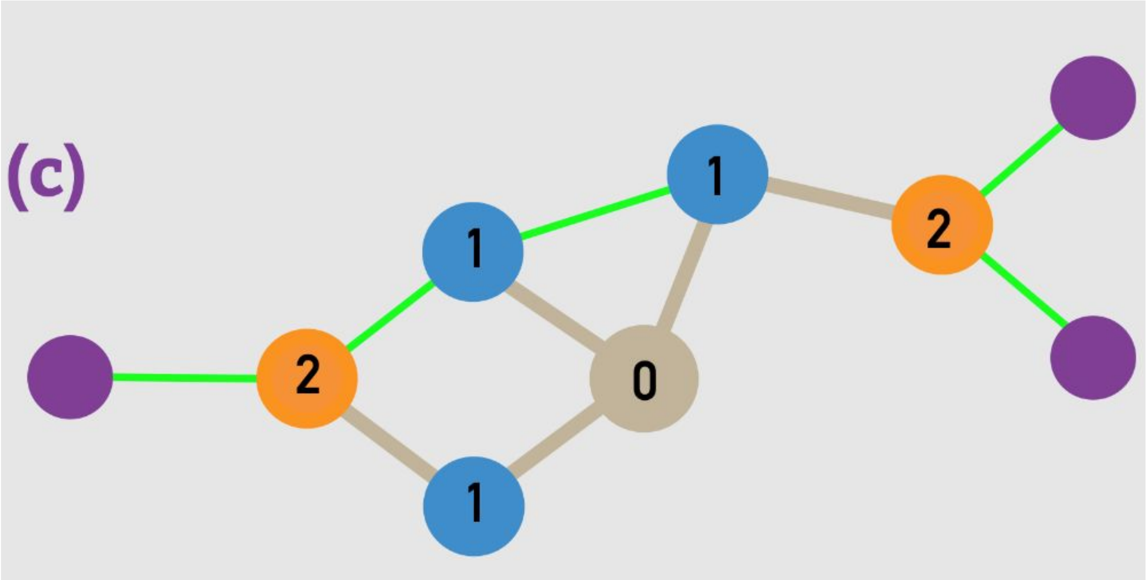
Next we label “2” the unlabeled neighbors of all nodes labeled “1”, and so on, in each iteration increasing the label number, until no node is left unlabeled.



(Barabasi, 2016)

Recall the BFS Algorithm

Next we label "2" the unlabeled neighbors of all nodes labeled "1", and so on, in each iteration increasing the label number, until no node is left unlabeled.

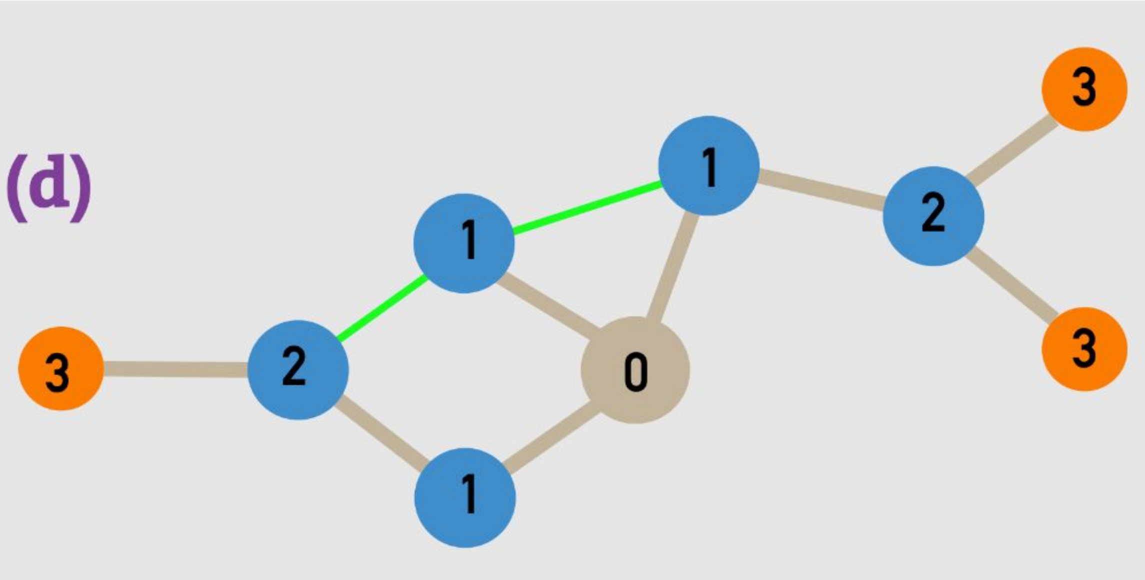


(Barabasi, 2016)

Recall the BFS Algorithm

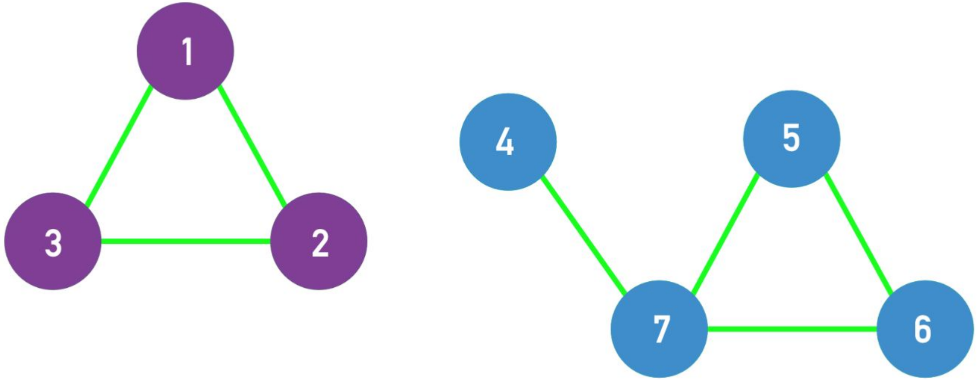
Ultimately, the length of the shortest path (or the distance d_{0i} between node 0 and any other node i in the network is given by the label of node i.

For example, the distance between node 0 and the leftmost node is d = 3.



(Barabasi, 2016)

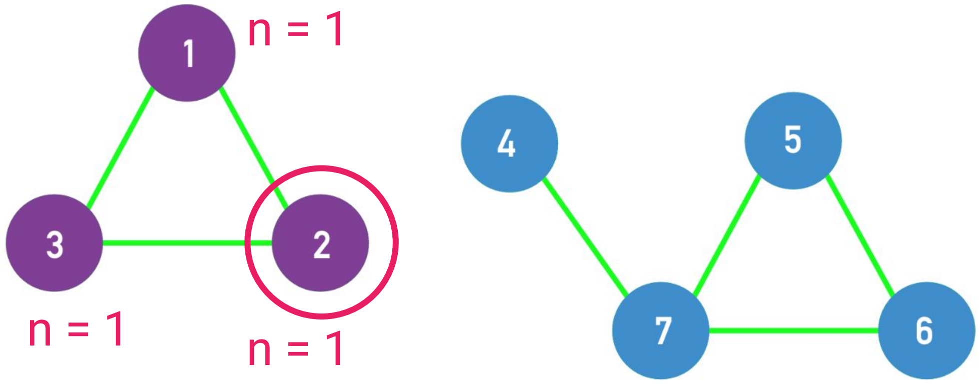
Can We Identify Connected Components Using BFS?



(Barabasi, 2016)

We Can Identify Connected Components Using BFS!

(1) Start from a randomly chosen node i and perform a BFS. Label all nodes reached this way with n = 1.

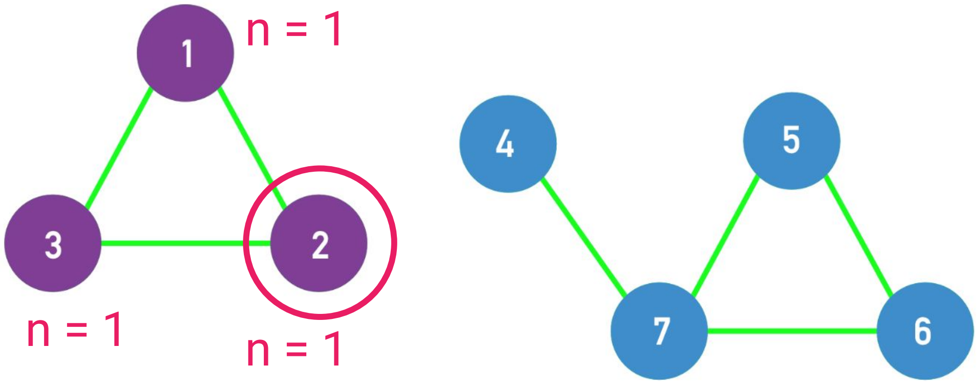


(Barabasi, 2016)

We Can Identify Connected Components Using BFS!

(2) If the total number of labeled nodes equals N, then the network is connected.

If the number of labeled nodes is smaller than N, the network consists of several components.



(Barabasi, 2016)

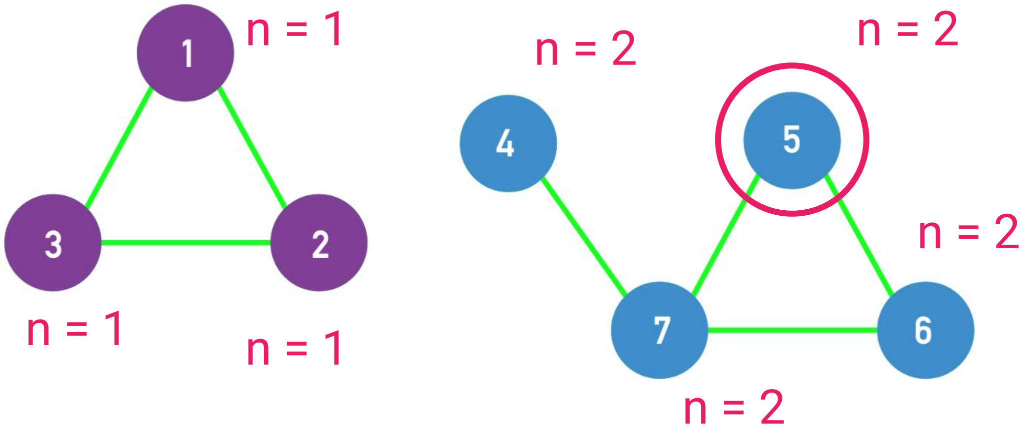
We Can Identify Connected Components Using BFS!

(3) Increase the label n → n + 1.

Choose an unmarked node j, label it with n.

Use BFS to find all nodes reachable from j, label them all with n.

Return to step 2.

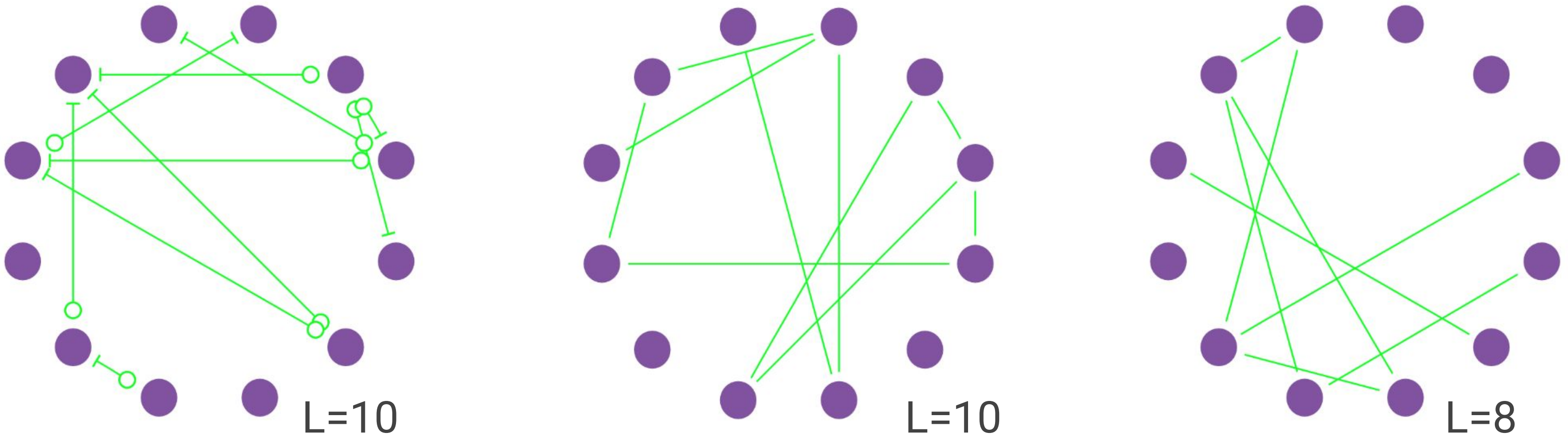


(Barabasi, 2016)

The Random Network Model

A random network consists of N nodes where each node pair is connected with probability p

Aka “Erdős-Rényi network” – from random graph theory (1959–1968)



Three realizations of a random network generated with the same parameters p=1/6 and N=12.

A random network consists of N nodes where each node pair is connected with probability p

Three realizations of a random network with p=0.03 and N=100. Several nodes have degree k=0, shown as isolated nodes at the bottom.

Why Random Network Models?

Q: Are the edges in social networks random?

Q: How are ties in social networks created?

Q: If a social tie is not formed by a coin toss (i.e., random), why should we study random networks?

(Barabasi Ch. 3.3)

Common question: How many links can we expect for a particular realization of a random network with fixed *N* and *p*?

The probability that a random network has exactly *L* links is:

$$\langle L \rangle = p \frac{N(N-1)}{2}$$

(note, the second term is the max possible number of pairs)

The average degree of a random network is:

$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1)$$

(note, the second term is the max possible node degree)

(Barabasi Ch. 3.3)

Common question: How many links can we expect for a particular realization of a random network with fixed N and p?

The number of links in a random network varies between realizations.

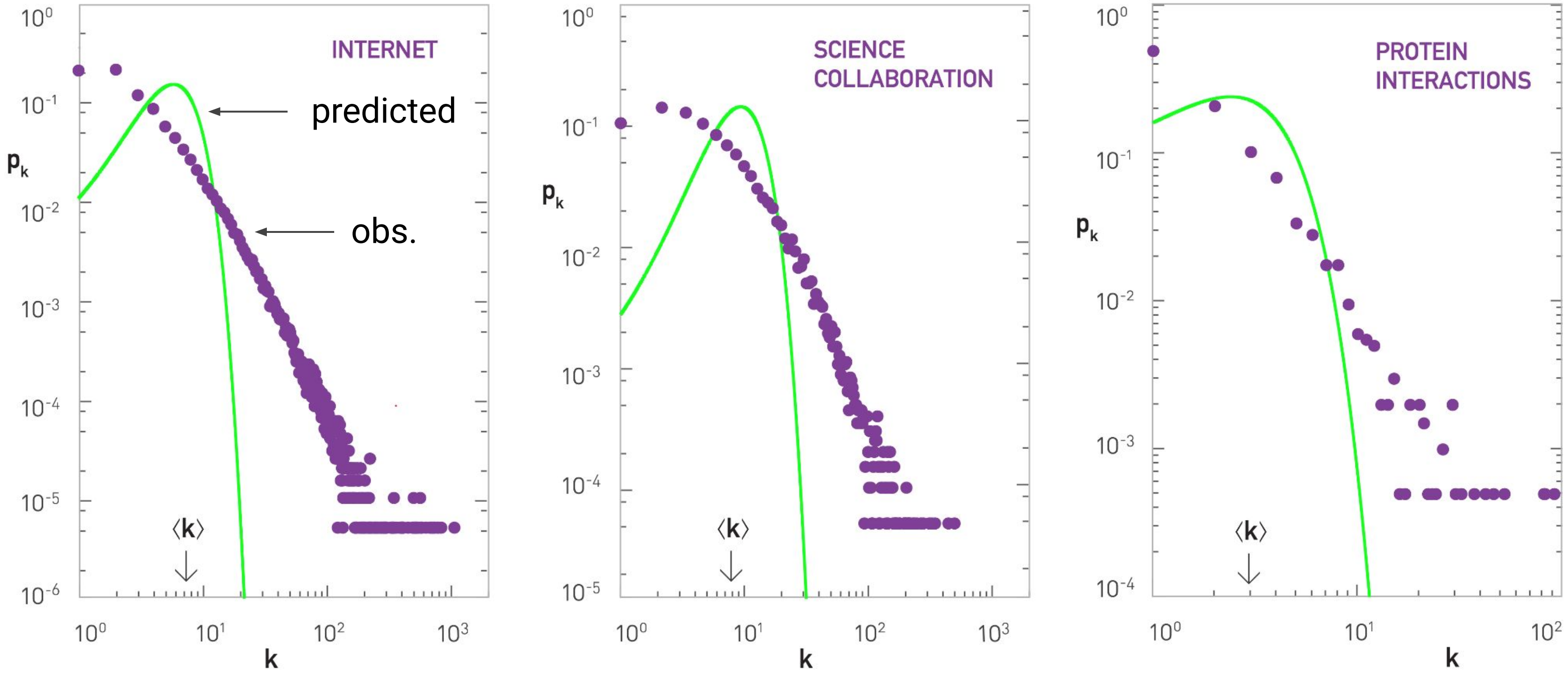
Its expected value is determined by *N* and *p*.

With larger *p*, a random network becomes denser:

- The average number of links increases linearly from $<L> = 0$ to L_{max}
- The average degree of a node increases from $<k> = 0$ to $<k> = N-1$.

(Barabasi Ch. 3.3)

The random network model underestimates the size and frequency of the high degree nodes, and the number of low degree nodes.



(Barabasi Ch. 3.5)

Connected Components in Random Networks

Let's inspect how the size of the largest connected component within the network, N_G, varies with <*k*>

For $p = 0$ we have $\langle k \rangle = 0$, hence all nodes are isolated. Therefore the largest component has size $N_G = 1$ and $N_G/N \rightarrow 0$ for large N.

For $p = 1$ we have $\langle k \rangle = N-1$, hence the network is a complete graph and all nodes belong to a single component. Therefore $N_G = N$ and $N_G/N = 1$.

The average degree of a random network is:

$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1)$$

Let's inspect how the size of the largest connected component within the network, N_G, varies with <*k*>

For $p = 0$ we have $\langle k \rangle = 0$, hence all nodes are isolated. Therefore the largest component has size $N_G = 1$ and $N_G/N \rightarrow 0$ for large N.

For $p = 1$ we have $\langle k \rangle = N-1$, hence the network is a complete graph and all nodes belong to a single component. Therefore $N_G = N$ and $N_G/N = 1$.

The average degree of a random network is:

$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1)$$

One would expect that the largest component grows gradually from $N_G = 1$ to $N_G = N$ if $\langle k \rangle$ increases from 0 to $N-1$. Right?

Let’s inspect how the size of the largest connected component within the network, N_G, varies with $\langle k \rangle$

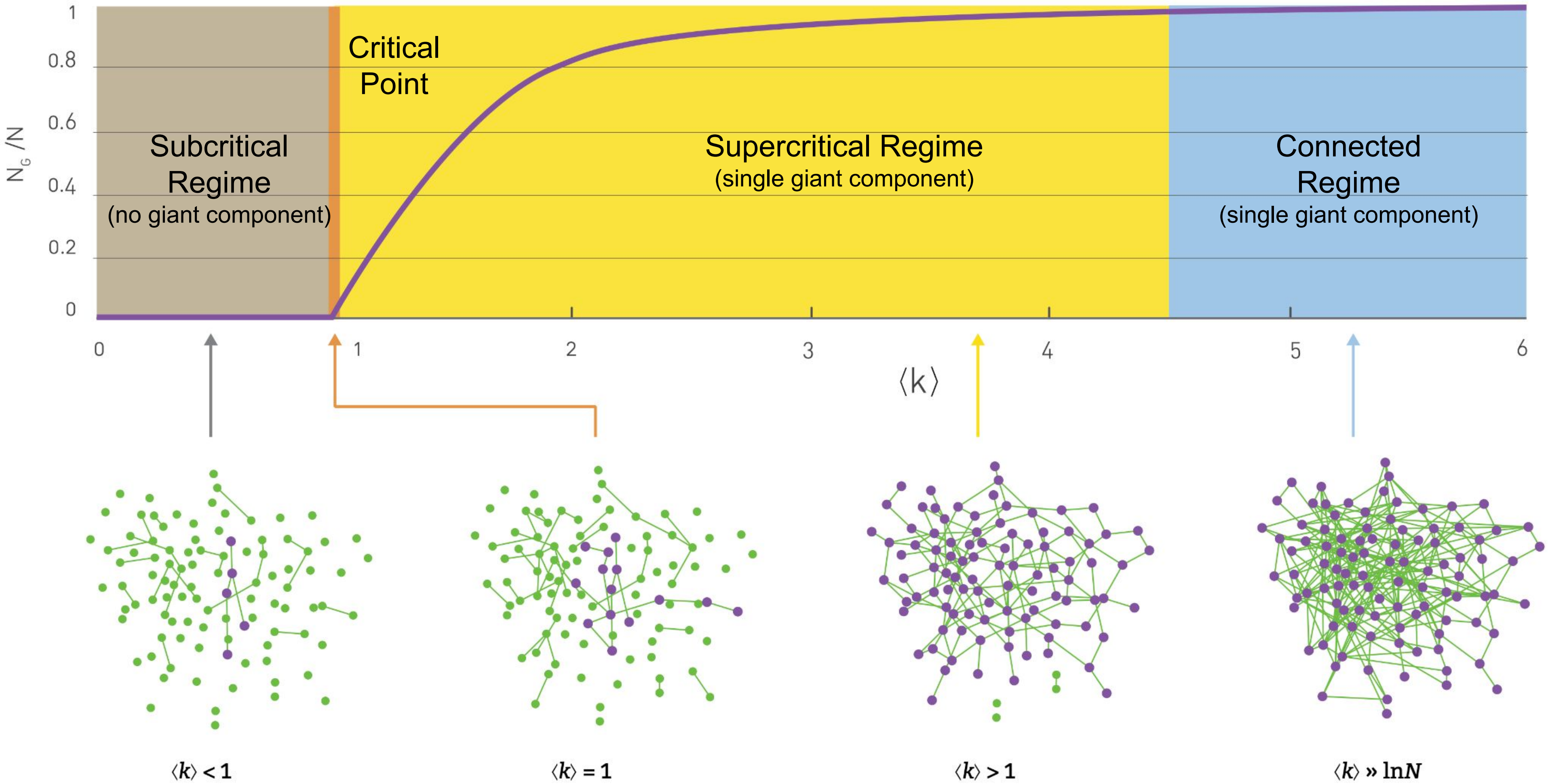
For $p = 0$ we have $\langle k \rangle = 0$, hence all nodes are isolated. Therefore the largest component has size $N_G = 1$ and $N_G/N \rightarrow 0$ for large N.

For $p = 1$ we have $\langle k \rangle = N-1$, hence the network is a complete graph and all nodes belong to a single component. Therefore $N_G = N$ and $N_G/N = 1$.

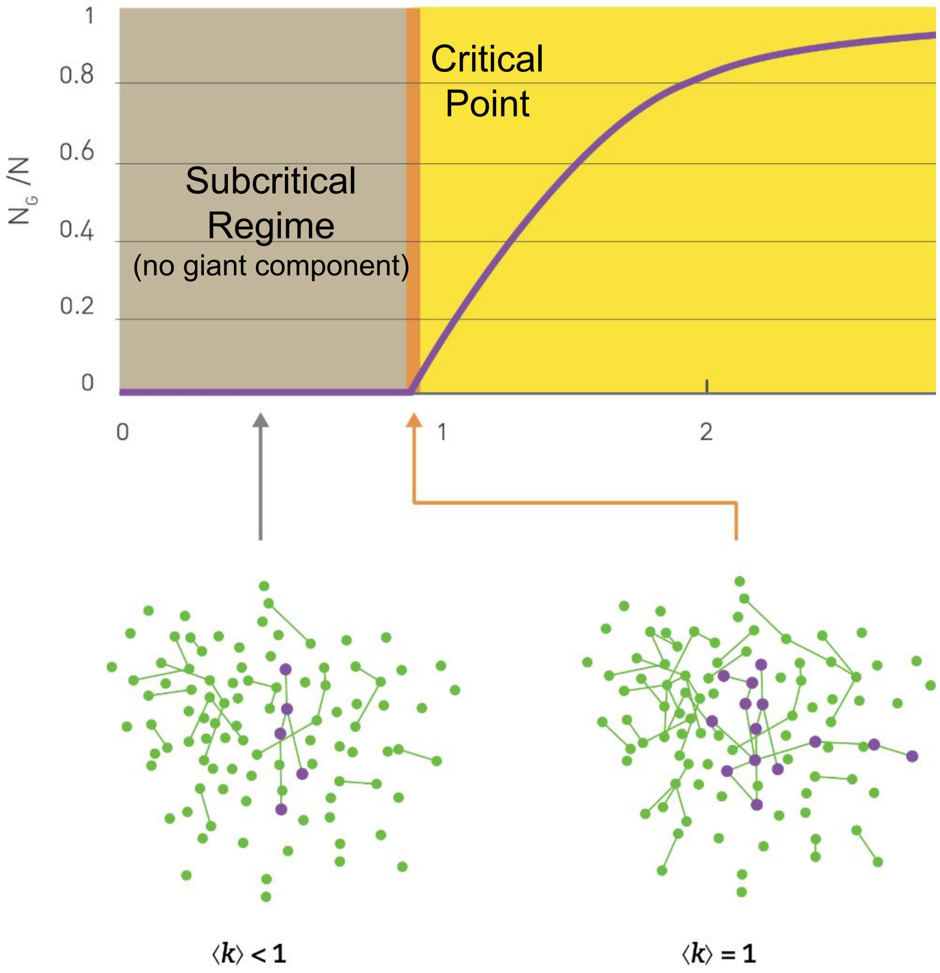
The average degree of a random network is:

$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1)$$

One would expect that the largest component grows gradually from $N_G = 1$ to $N_G = N$ if $\langle k \rangle$ increases from 0 to N-1. ~~Right~~? **Wrong**



(Barabasi Ch. 3.6; Erdős & Rényi, 1959)

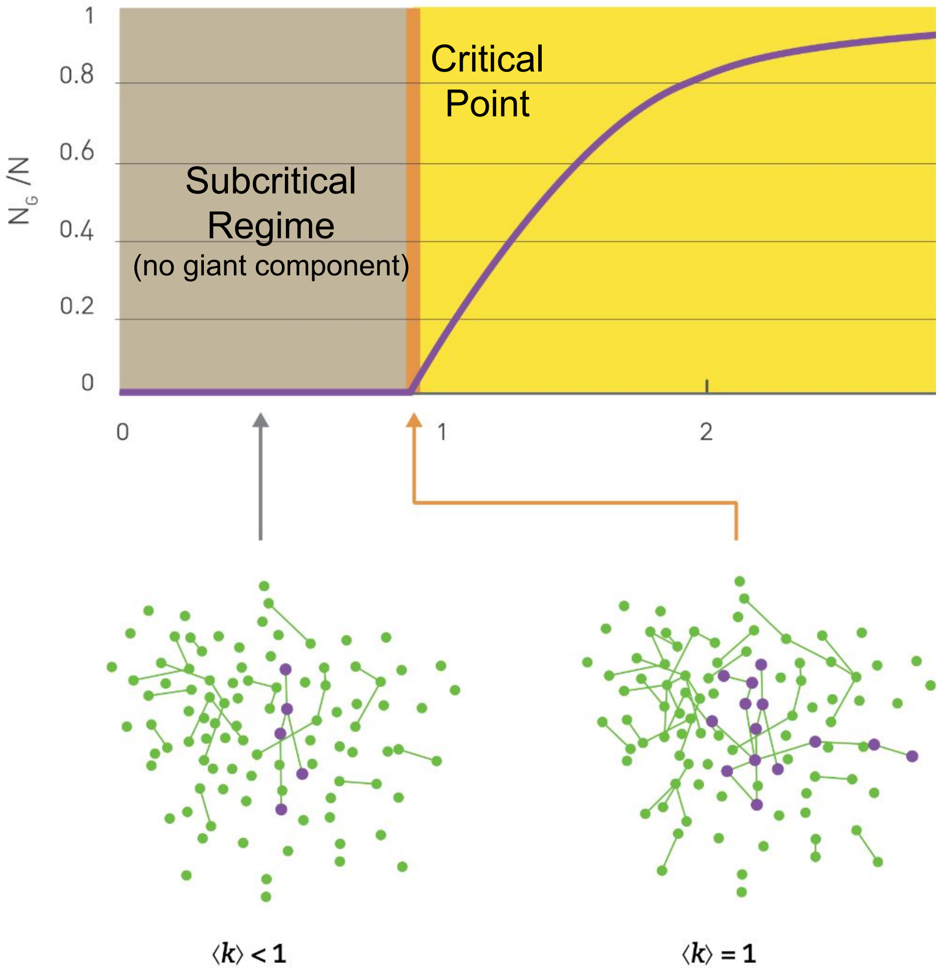


We have one giant component iff each node has on average more than one link.

That we need *at least* one link per node to observe a giant component is not unexpected.

But it is arguably counter-intuitive that one link is *sufficient* for its emergence.

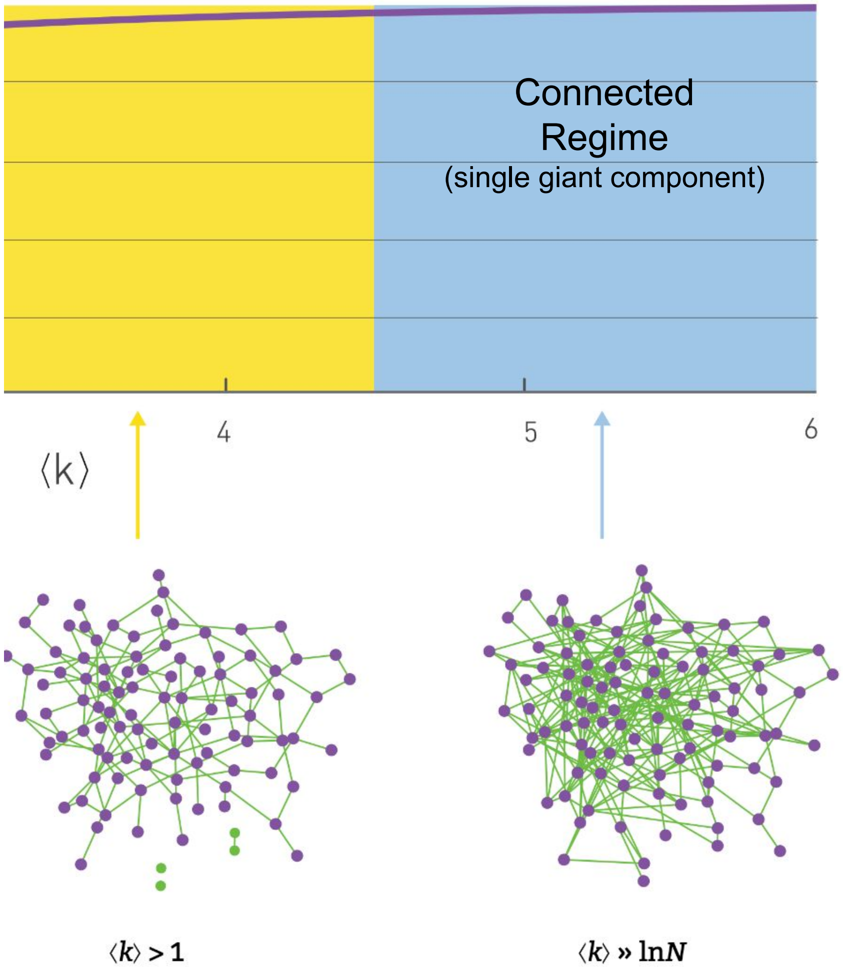
(Barabasi Ch. 3.6; Erdős & Rényi, 1959)



(Barabasi Ch. 3.6; Erdős & Rényi, 1959)

What's the average degree <k> in the HW1 networks?

- Is <k> > 1? Implying that they have a giant component.



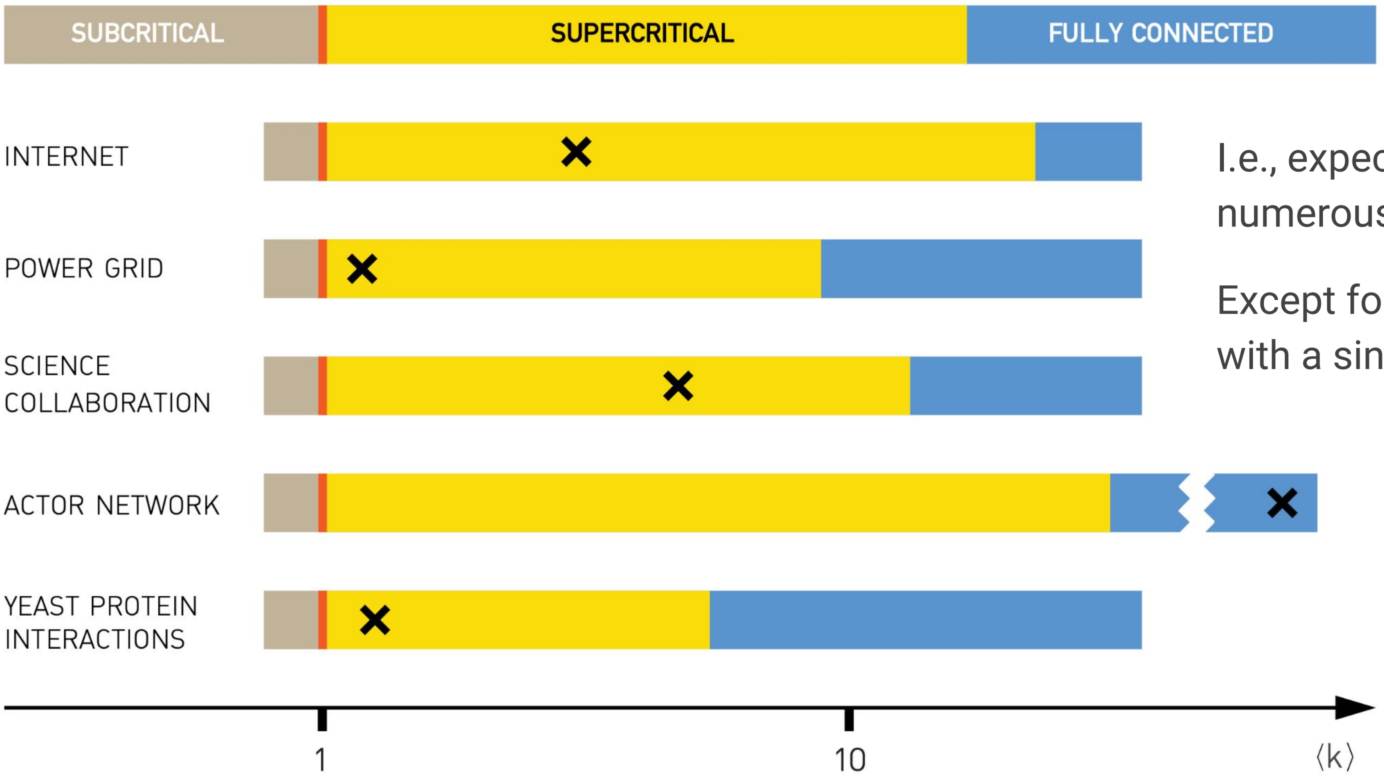
(Barabasi Ch. 3.6; Erdős & Rényi, 1959)

What's the average degree <k> in the HW1 networks?

- Is <k> > 1? Implying that they have a giant component.
- Is <k> > lnN? Implying that they have a *single* giant component.

(For the world population, if the average individual has more than $\ln(7 \times 10^9) \approx 22.7$ acquaintances, then the global network must have a single component)

Most real networks are supercritical



I.e., expected to be broken into numerous isolated components.

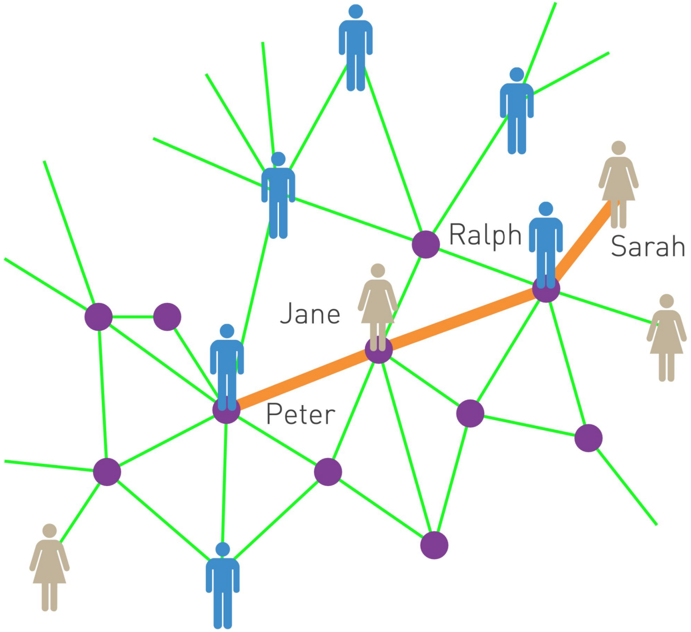
Except for the actor network, with a single giant component.

Back to Six Degrees of Kevin Bacon
(Aka the “Small world” phenomenon)

Small world property: The distance between any two nodes in a network is small.

Consider a random network with average degree <k>. A node in this network has on average:

- How many nodes at distance one (d=1)?

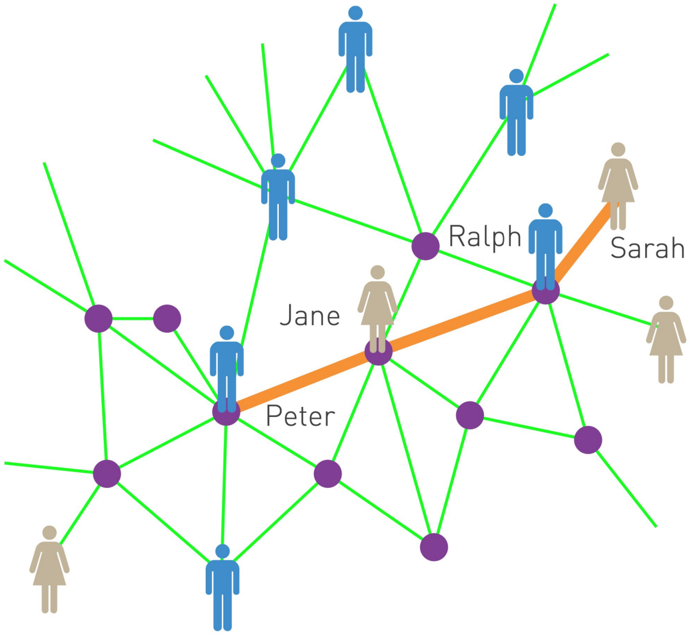


(Barabasi Ch. 3.8)

Small world property: The distance between any two nodes in a network is small.

Consider a random network with average degree <k>. A node in this network has on average:

- <k> nodes at distance one (d=1)
- How many nodes at distance two (d=2)?

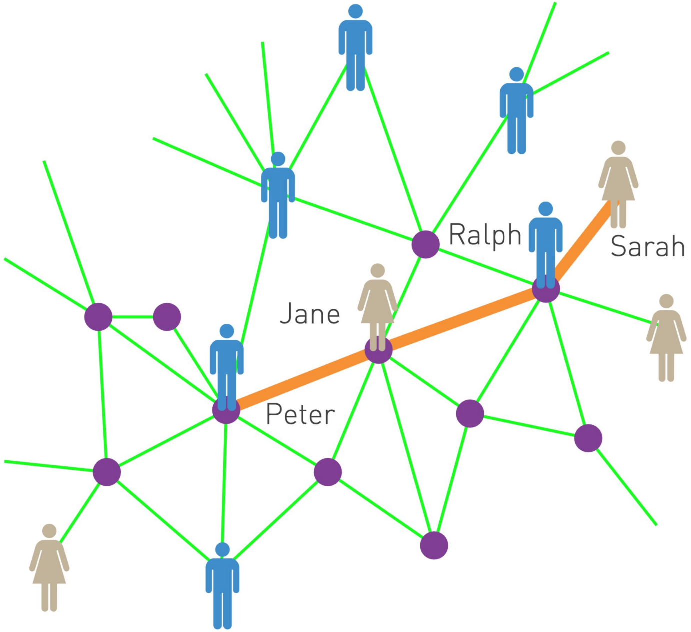


(Barabasi Ch. 3.8)

Small world property: The distance between any two nodes in a network is small.

Consider a random network with average degree <k>. A node in this network has on average:

- <k> nodes at distance one (d=1)
- $<k>^2$ nodes at distance two (d=2)



(Barabasi Ch. 3.8)

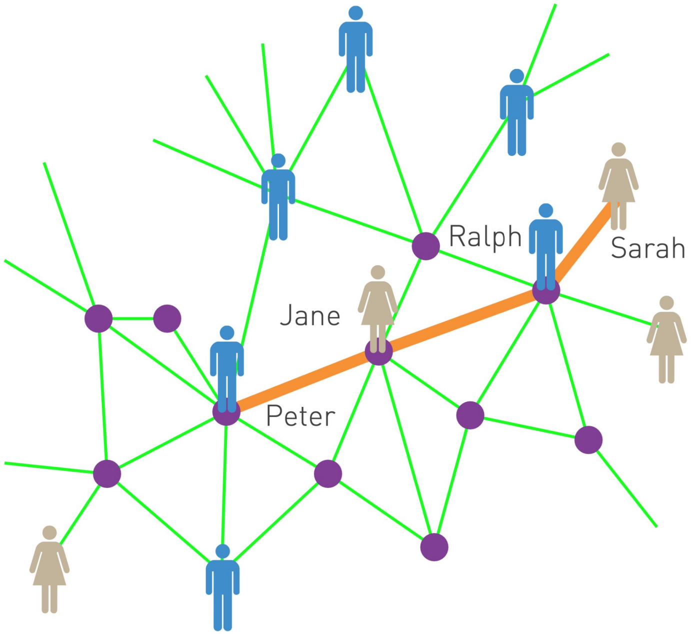
Small world property: The distance between any two nodes in a network is small.

Consider a random network with average degree <k>. A node in this network has on average:

- <k> nodes at distance one (d=1)
- $<k>^2$ nodes at distance two (d=2)
- $<k>^3$ nodes at distance three (d =3)

  ...
- $<k>^d$ nodes at distance d

E.g., if <k> ≈ 1,000 (the estimated number of acquaintances an individual has), we expect 10^6 individuals at d=2 and about a billion, i.e. almost the whole earth's population, at d=3 from us.



(Barabasi Ch. 3.8)

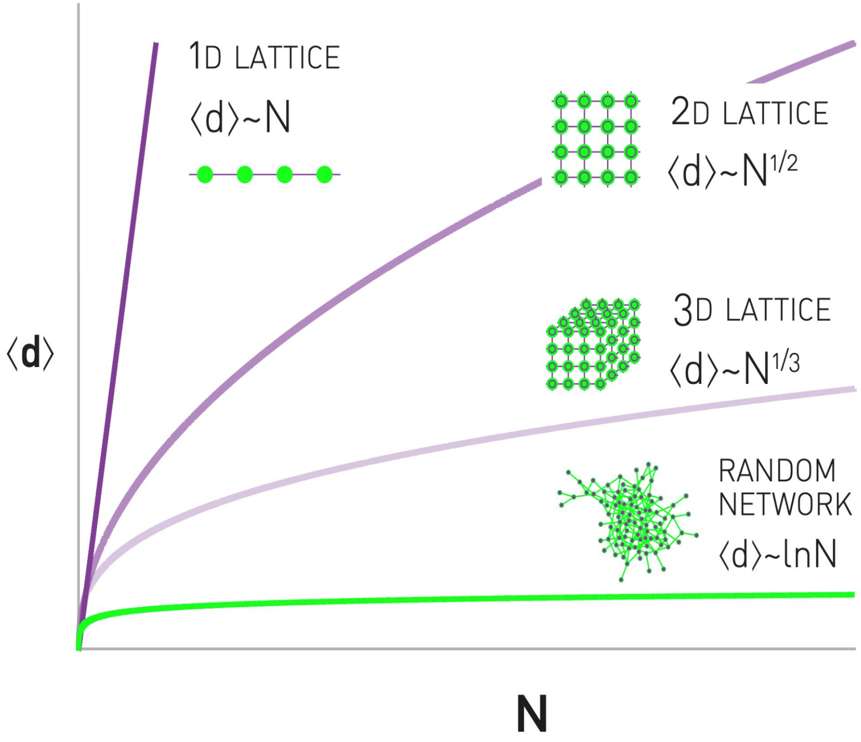
“Small” as in proportional to lnN, rather than N (or a power of N)

The dependence of the average distance in a random network on N and <k>:

$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$$

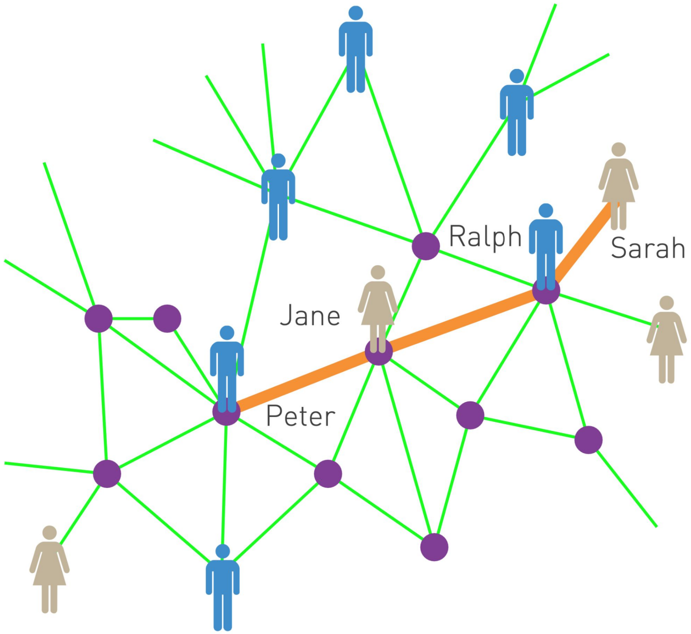
The distances in a random network are orders of magnitude smaller than the size of the network.

(For our world social network, if $N \approx 7 \times 10^9$ and $\langle k \rangle \approx 10^3$, we get $\langle d \rangle \approx 3.28$.)



“Small” as in proportional to lnN, rather than N (or a power of N)

Why $\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$?

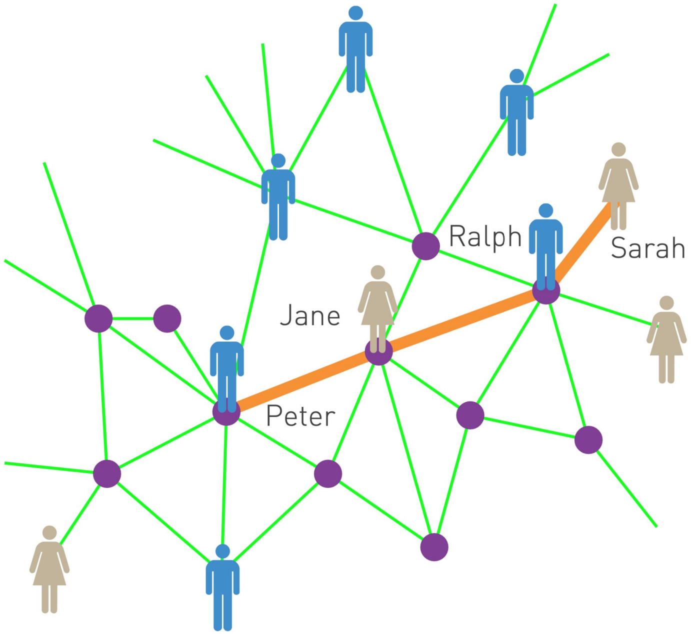


(Barabasi Ch. 3.8)

“Small” as in proportional to lnN, rather than N (or a power of N)

Why $\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$?

How many steps does it take from Jane to reach all *N*-1 people in the network?



(Barabasi Ch. 3.8)

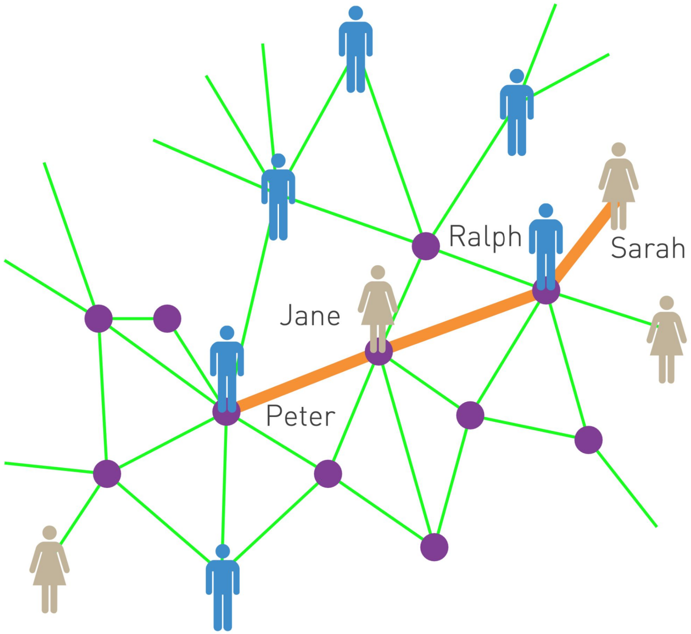
"Small" as in proportional to lnN, rather than N (or a power of N)

Why $\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$?

How many steps does it take from Jane to reach all N-1 people in the network?

$<k> + <k>^2 + <k>^3 + \ldots + <k>^d = N-1$

$\ln(<k> + <k>^2 + <k>^3 + \ldots + <k>^d) = \ln(N-1)$



(Barabasi Ch. 3.8)

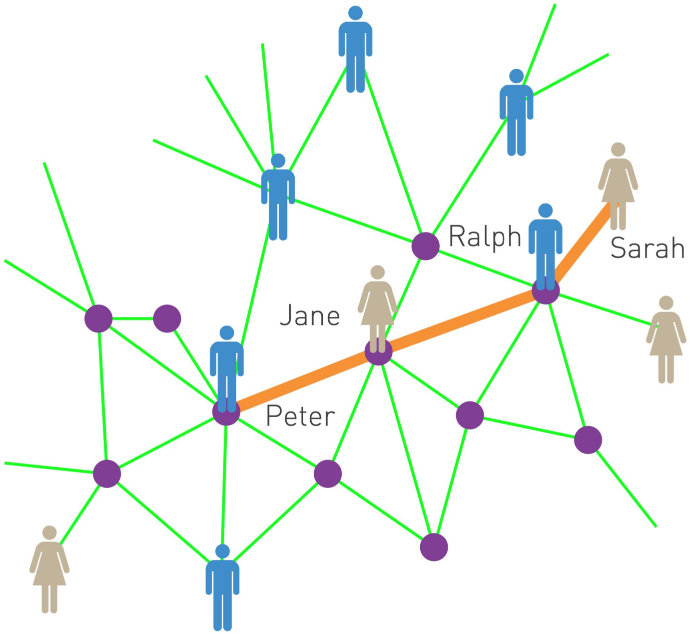
“Small” as in proportional to lnN, rather than N (or a power of N)

Why $\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$?

How many steps does it take from Jane to reach all N-1 people in the network?

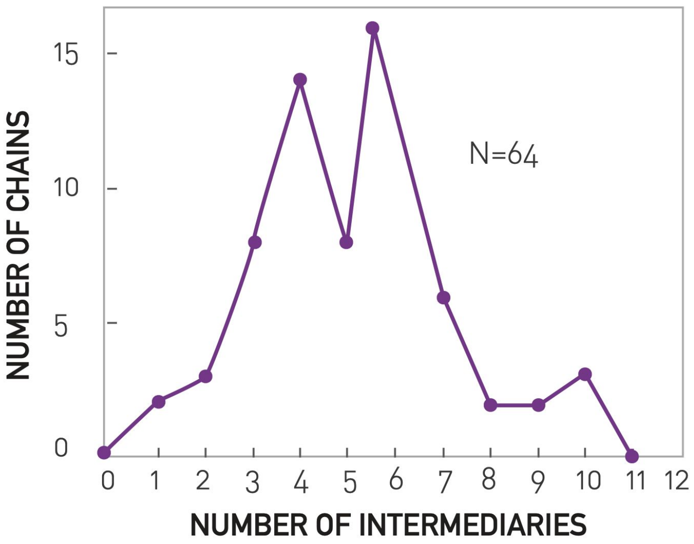
$<k> + <k>^2 + <k>^3 + \ldots + <k>^d = N-1$
$\ln(<k> + <k>^2 + <k>^3 + \ldots + <k>^d) = \ln(N-1)$

For large N,
$\ln<k>^d \sim \ln N$
$d * \ln<k> \sim \ln N$
$d \sim \ln N / \ln<k>$

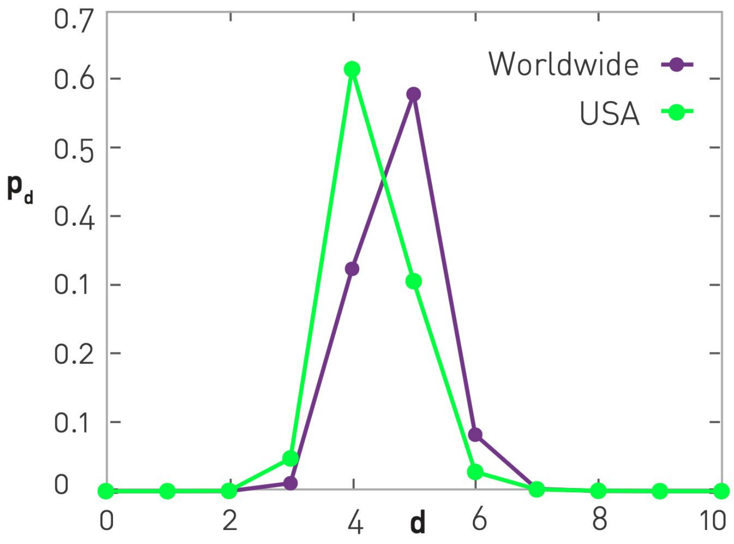


(Barabasi Ch. 3.8)

Six degrees: Experimental confirmation



Recall (Milgram, 1967) – the letter forwarding study: median 5.2 hops



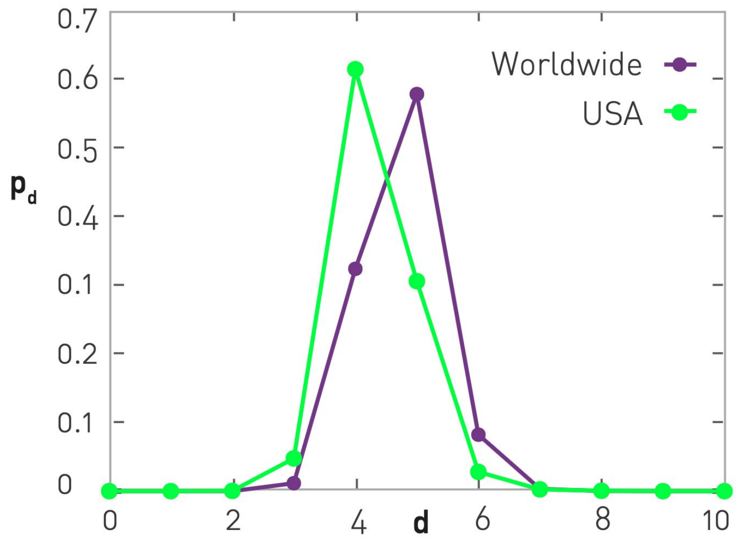
Facebook 2011 network (721M active users, 68B symmetric friendship links): average distance 4.74

(Backstrom et al, 2012)

Six degrees: Experimental confirmation

Q: If the Facebook friendship network were a random graph, what is the average shortest path length?

$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$$



Facebook 2011 network (721M active users, 68B symmetric friendship links): average distance 4.74

(Backstrom et al, 2012)

Six degrees: Experimental confirmation

Q: If the Facebook friendship network were a random graph, what is the average shortest path length?

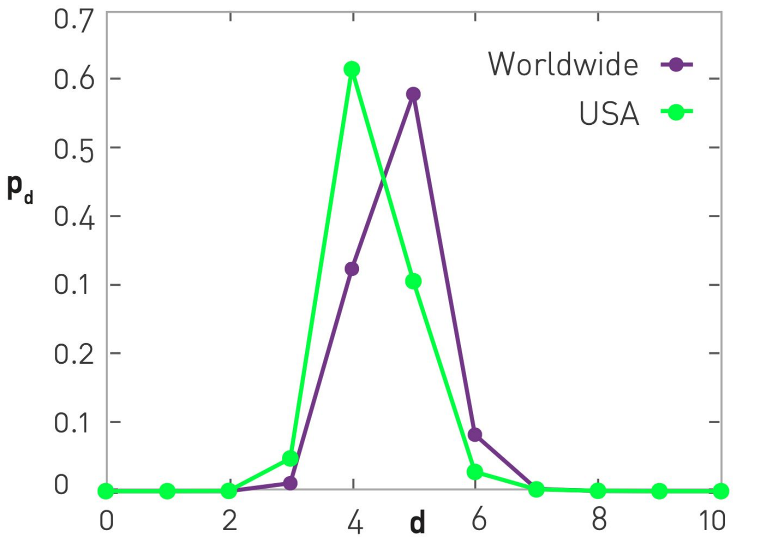
$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$$

N = 721,000,000

L = 68,000,000,000

$\langle k \rangle$=2L/N = 188.6

<d>~ln(721,000,000)/ln(188.6) = 3.892



Facebook 2011 network (721M active users, 68B symmetric friendship links): average distance 4.74

(Backstrom et al, 2012)

Six degrees: Experimental confirmation

Q: If the Facebook friendship network were a random graph, what is the average shortest path length?

$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$$

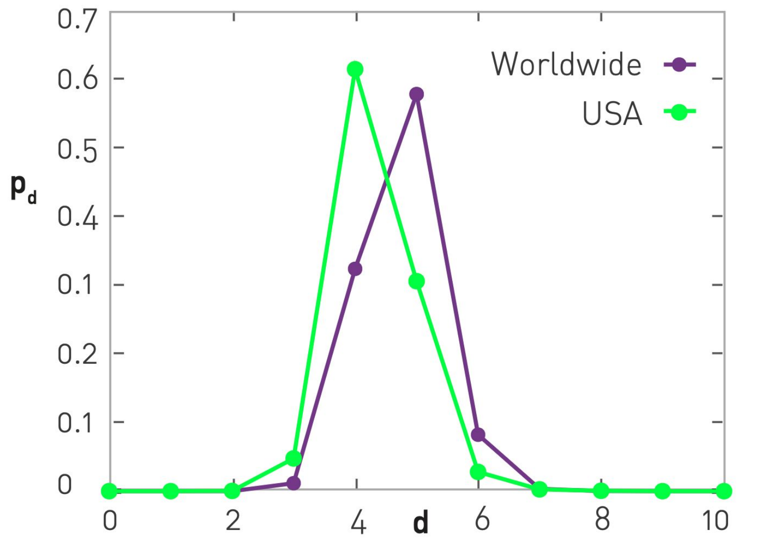
N = 721,000,000

L = 68,000,000,000

$<k>$=2L/N = 188.6

<d>~ln(721,000,000)/ln(188.6) = **3.892**

Q: Why is the actual distance longer?



Facebook 2011 network (721M active users, 68B symmetric friendship links): average distance **4.74**

(Backstrom et al, 2012)

Six degrees: Experimental confirmation

Random graph:

$\langle d \rangle \sim \ln(721{,}000{,}000)/\ln(188.6) =$ **3.892**

Facebook 2011 observed network:

$d \sim$ **4.74**

We can use the random graph as a baseline model to compare against actually observed networks.

Here, the observed network is not as small a world as the random graph!

Q: Why is the actual distance longer?

Today’s Summary

Giant components

The random graph model

An explanation for Six Degrees of Kevin Bacon

A teaser for the next smallest building block – edges vs. social ties