Network Analysis (INFOMNWA-2021)

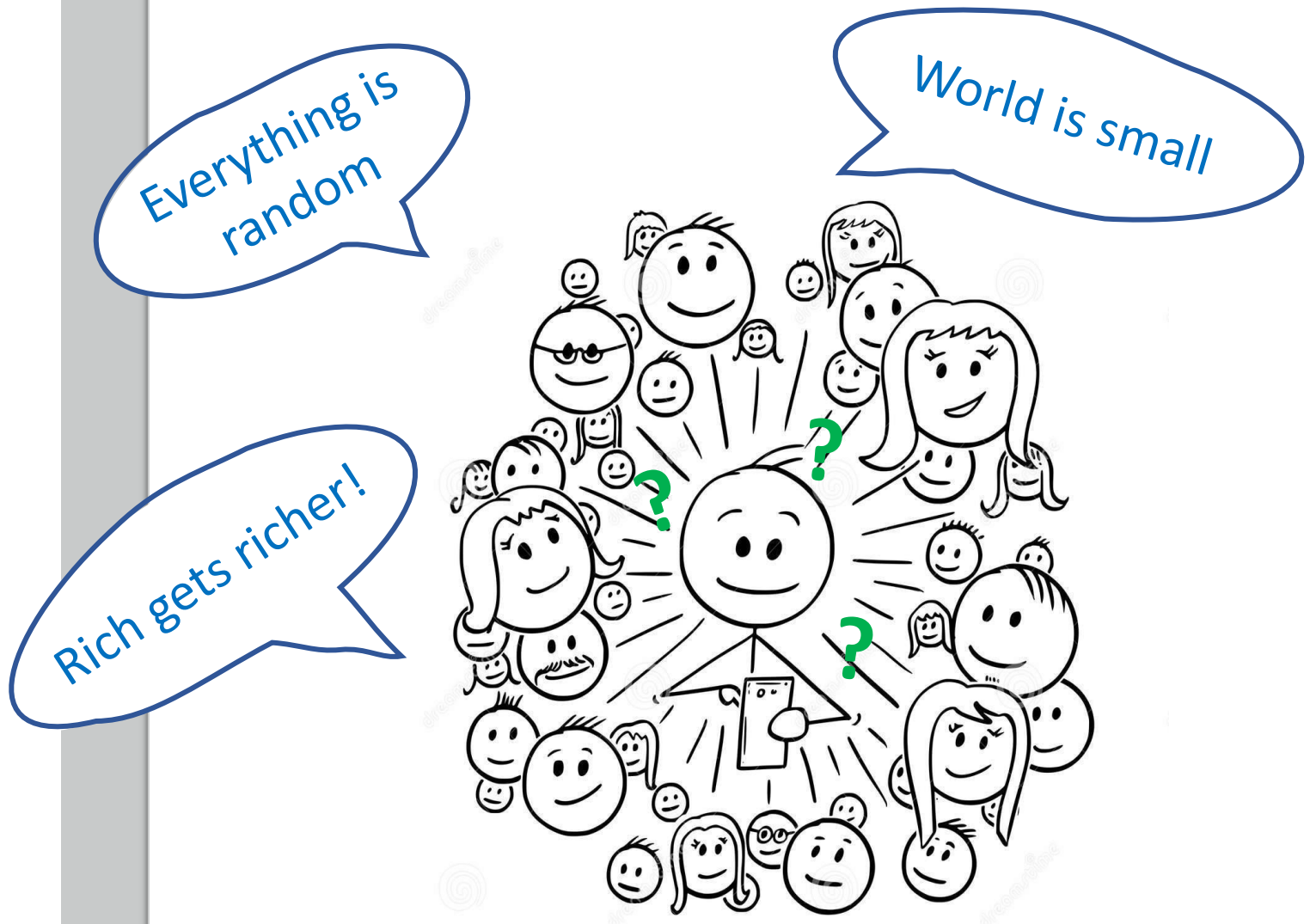
Lecture 6: Network formation

Jiamin Ou

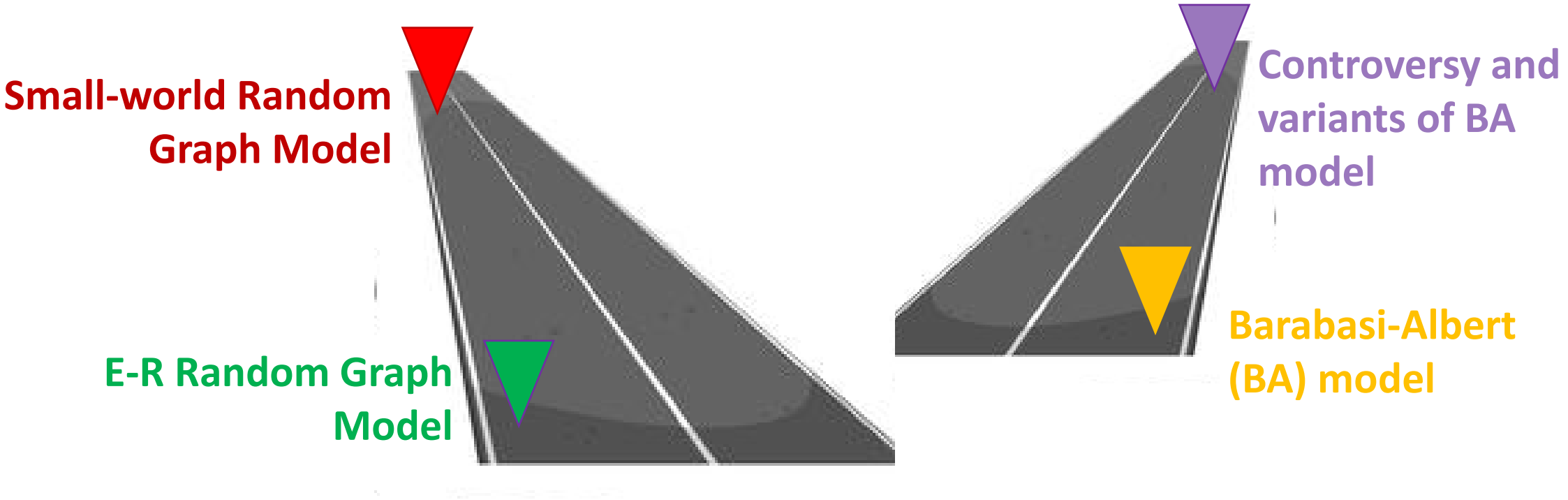
Today's programme

- **Erdös–Rényi Random Graph Model**
- **Small-world Random Graph Model**
- **Barabasi-Albert (BA) model**

How do we form friendships and other social connections?



Network formation models inspired by what we observe in real-world networks



A groundbreaking finding in 1970s: Human society is a small world

- **Sample**: People in the United States
- **Procedure**:

1) Information packets were initially sent to "randomly" selected individuals in Omaha, Nebraska, and Wichita, Kansas

2) The recipient was asked whether he or she personally knew a specific contact person in Boston or Massachusetts

3) If yes, the person was to forward the letter directly to that person.

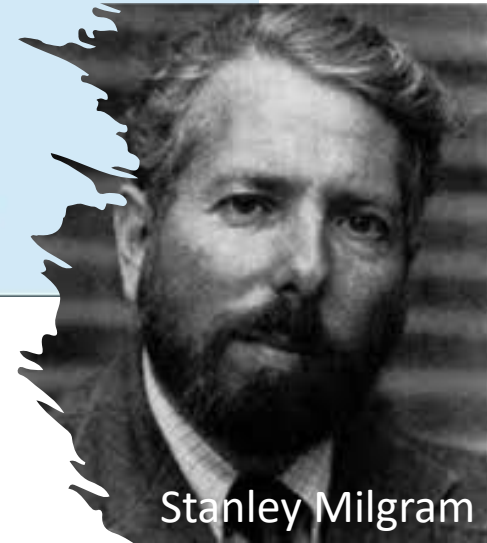
4) If no, the person was to think of a friend or relative who was more likely to know the target. They were then directed and forward the packet to that person.

5) When and if the package eventually reached the contact person in Boston or Massachusetts, count the number of times it had been forwarded from person to person.

- **Results:**

Among the successful chains, the average times it had been forwarded (path length) was five and a half or six. (six degrees of separation")

- **Limitations:** very limited samples and high drop-out rates


Stanley Milgram

Modern replication of the small world experiment: A global, Internet-based social search experiment

- **Procedure**:

1) Participants registered online (http://smallworld.sociology.columbia.edu) and were randomly allocated one of 18 target persons from 13 countries

2) Participants were informed that their task was to help relay a message to their allocated target by passing the message to a social acquaintance whom they considered "closer" than themselves to the target.

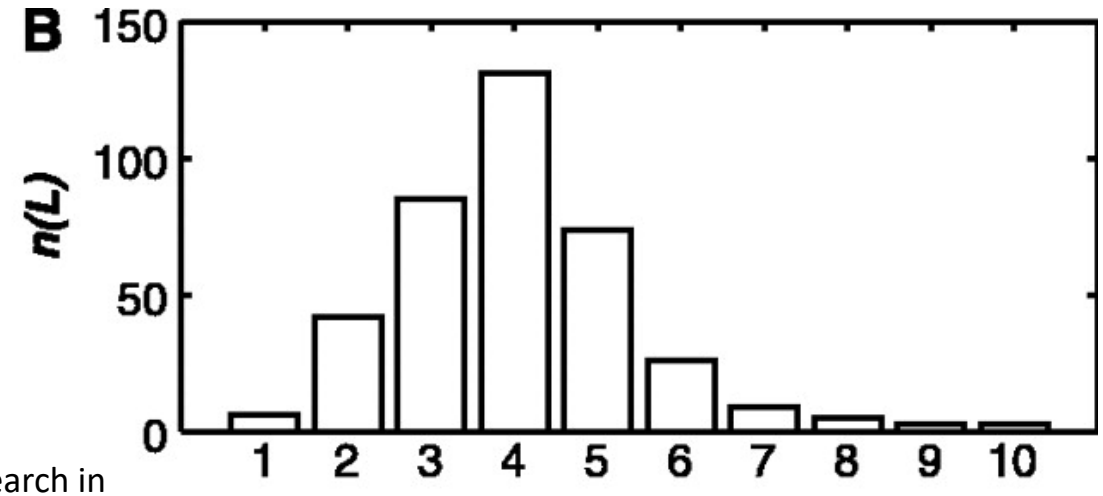
- **Results:**

1) Data were recorded on 61,168 individuals from 166 countries, constituting 24,163 distinct message chains

2) >50% of all participants resided in North America

3) All targets may in fact be reachable from random initial senders in only a few steps

4) But small differences in either participation rates or the underlying chain lengths can have a dramatic impact on the apparent reachability of different targets



Peter Sheridan Dodds, Roby Muhamad, Duncan J. Watts. An Experimental Study of Search in Global Social Networks. ***Science*** 08 Aug 2003.

Other attempts

Erdős number (1969 till now): **the "collaborative distance"** between mathematician Paul Erdős and another person

Erdős wrote 1,500 mathematical articles in his lifetime and had 512 direct collaborators: the people with Erdős number 1.

The people who have collaborated with them (but not with Erdős himself) have an Erdős number of 2 (12,600 people as of 7 August, 2020),

Those who have collaborated with people who have an Erdős number of 2 (but not with Erdős or anyone with an Erdős number of 1) have an Erdős number of 3...

– The median Erdős number across all mathematicians is **5**, with an extreme value of **13**.

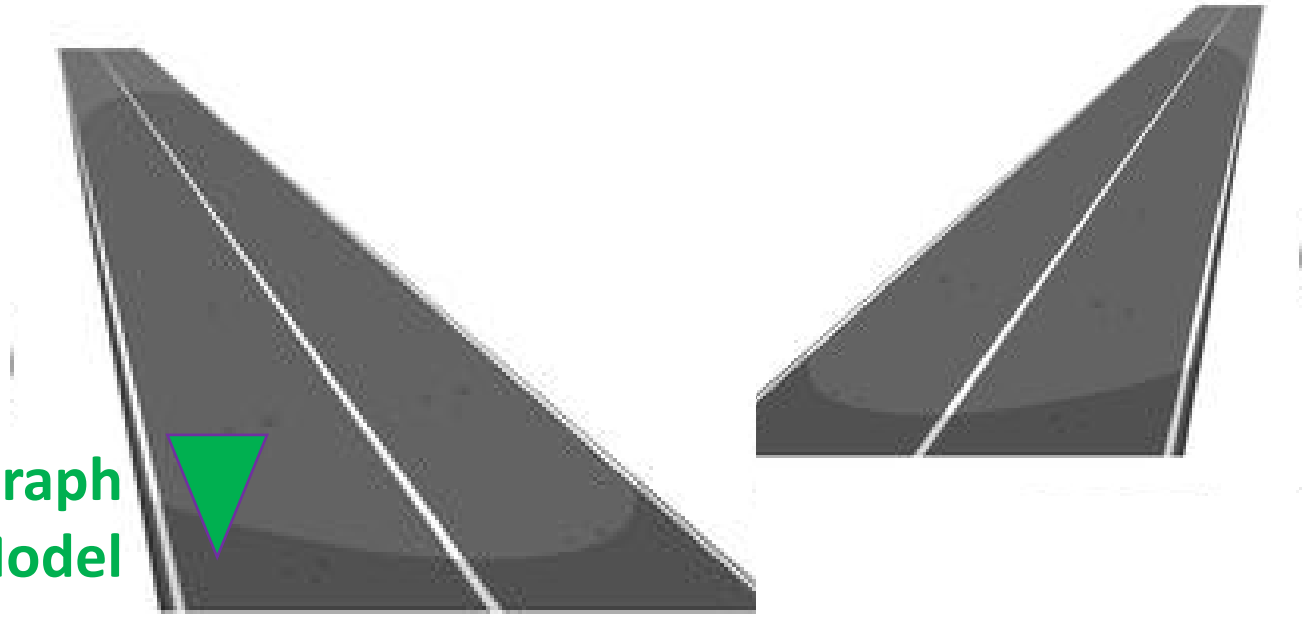
Erdős spent a large portion of his later life living out of a suitcase, visiting his over 500 collaborators around the world

The international film actor network:

Bacon number: Number of steps from **Hollywood movie stars** to **Kevin Bacon**

– As of Dec 2007, the highest Bacon number reported is **8**

Network formation models inspired by what we observe in real-world networks



E-R Random Graph Model

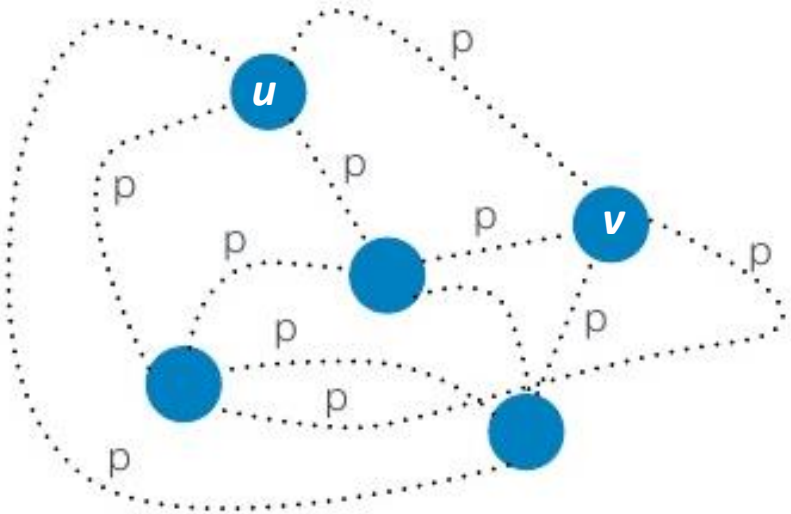
Small path length

“Everything is random”: Pál Erdös and Alfréd Rényi, 1959 (ER **Random Graph Model)**

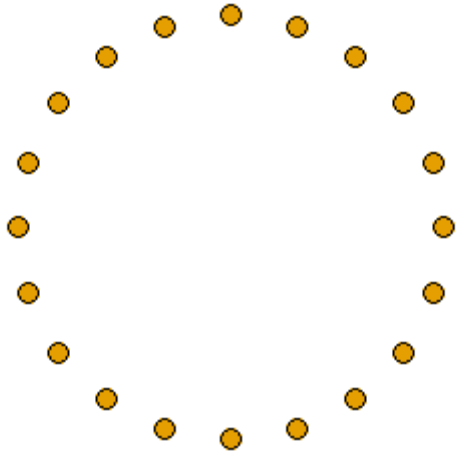
Edges are independent & each edge is equally likely

$G_{n,p}$: A graph of ***n*** nodes and each edge *(u,v)* appears *i.i.d* with probability ***p***

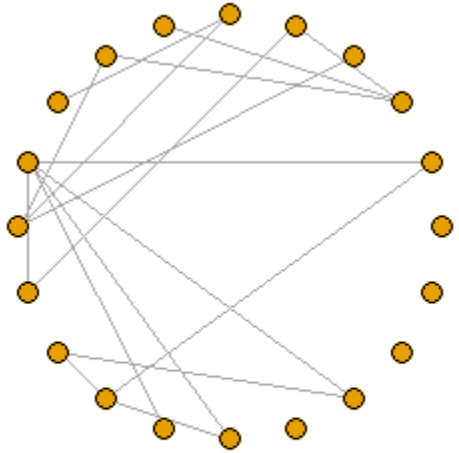
$G_{n,M}$: A graph is chosen uniformly randomly from the collection of all graphs which have ***n*** nodes and ***M*** edges.



As *p* increases….



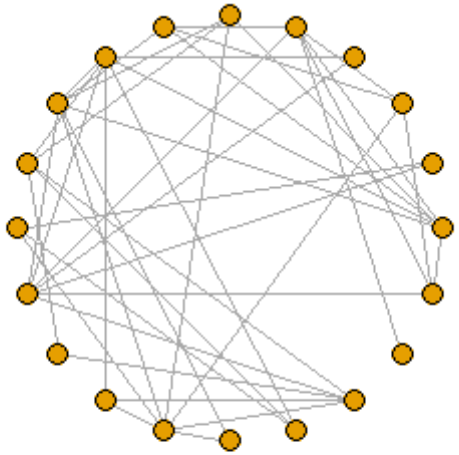
ER Random Network: G(N=20,p=0) model



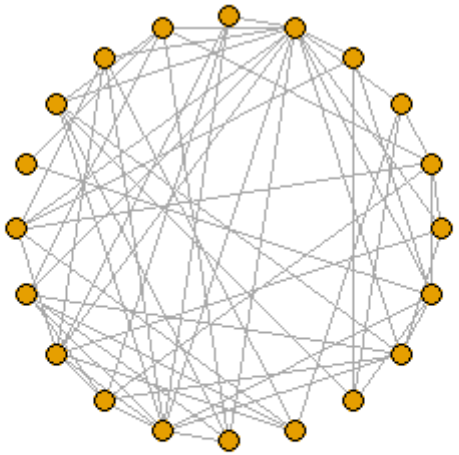
ER Random Network: G(N=20,p=0.1) model

$M \sim (n*(n-1)/2)*p$

Number of all the possible edges you can have * p



ER Random Network: G(N=20,p=0.2) model

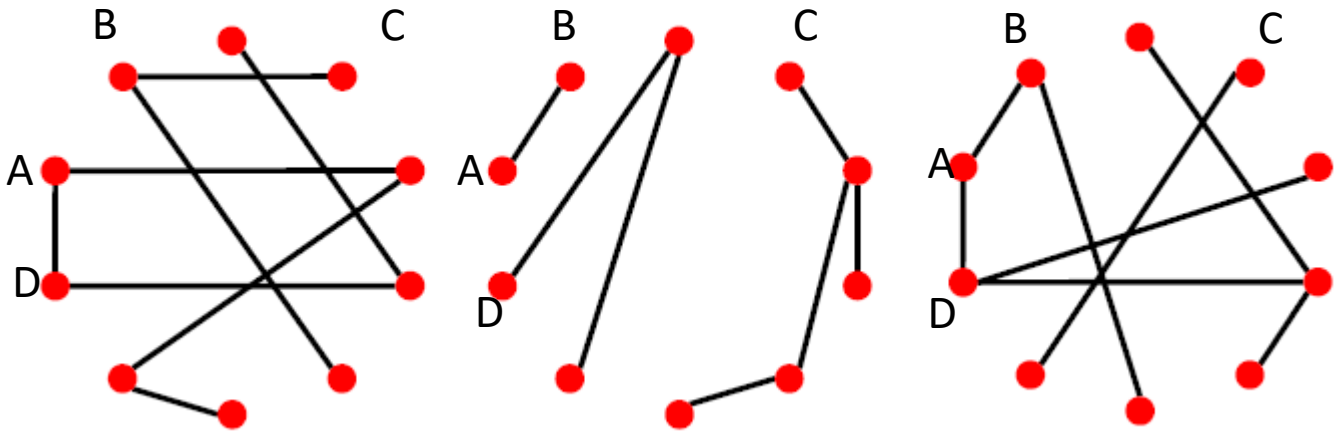


ER Random Network: G(N=20,p=0.3) model

For a set of n and p, do we have a unique graph?

- *n* and *p* do not uniquely determine the graph!

We can have many different realizations given the same *n* and *p*

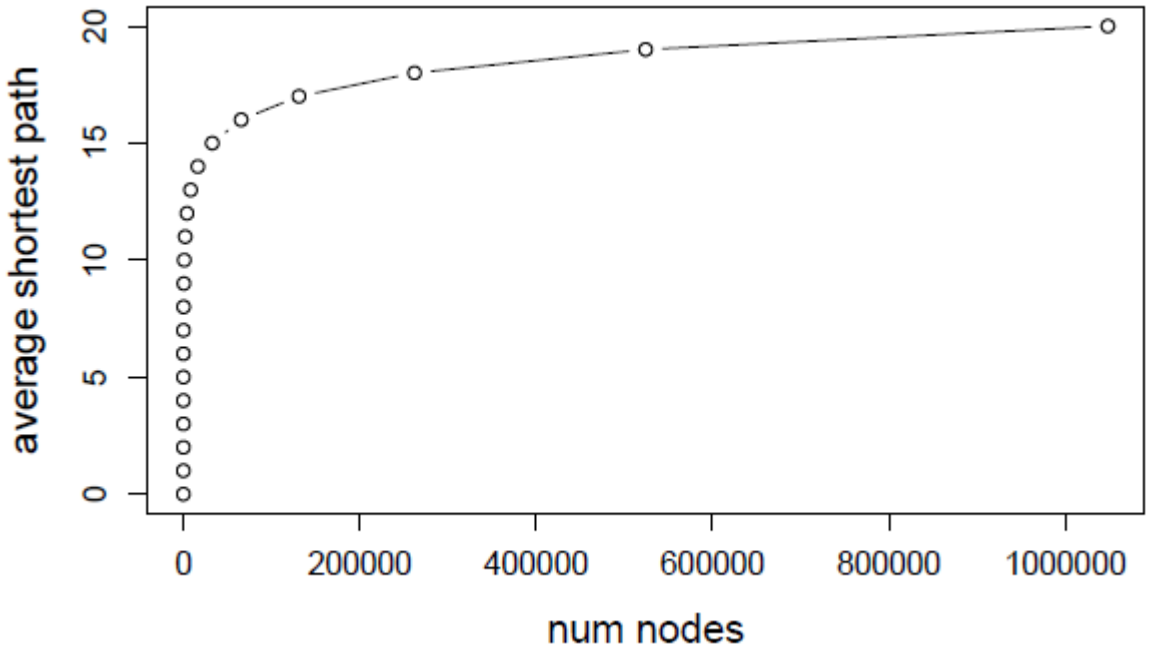


n=10 p=1/6

Key properties of random graph

1. Average shortest path

Average path length = *O* (log N)



N*p is constant
Or average degree k is constant

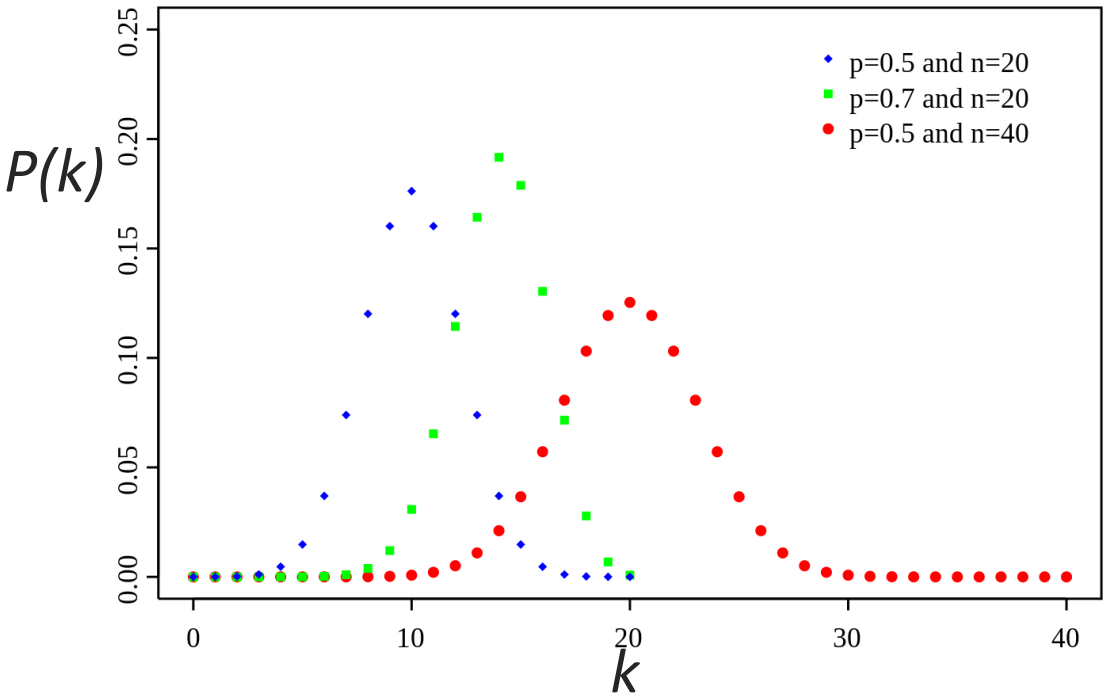
- Networks can grow to be very large but nodes will be just a few hops apart
- ER-graphs have a small average shortest path length

Key properties of random graph

2. Degree distribution: binomial

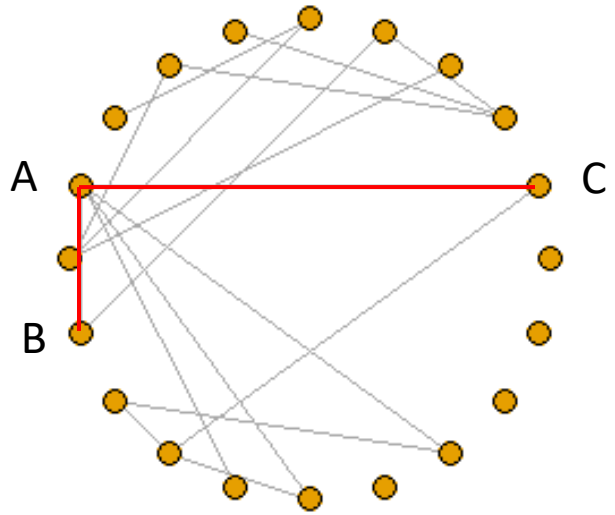
The degree by nodes (d_i) follow a distribution of $\{(p_{k(i)}, d_i), i \in N\}$

Average degree: p(n−1)

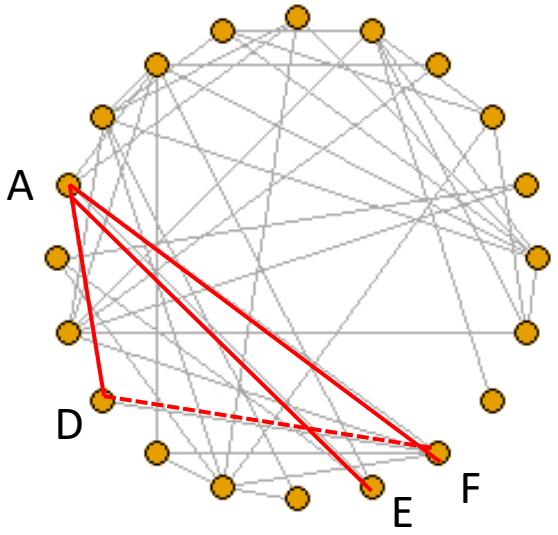


Key properties of random graph

3. Clustering coefficient: small or large?



ER Random Network: G(N=20,p=0.1) model



ER Random Network: G(N=20,p=0.2) model

Lack of local structure
(e.g., no triadic closure:
My friend don't know each other)

- The expected mean local clustering coefficient is *p*

Some networks: Short path length and high clustering

Network	Average shortest path length		Average clustering coefficient	
	Actual	Random	Actual	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*

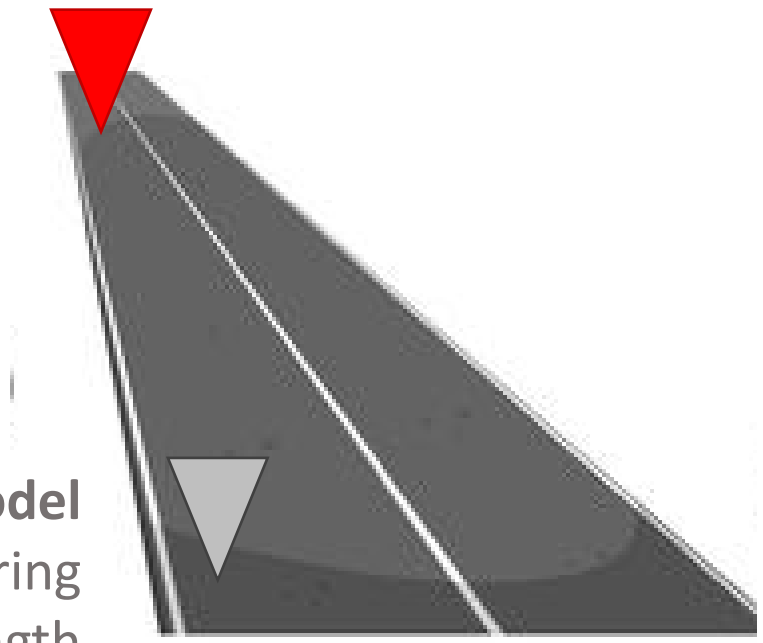
C. elegans: a neural network of neurons and synapses in *C. elegans*, a type of worm

Duncan J. Watts & Steven H. Strogatz, 1998. Collective dynamics of 'small-world' networks, Nature, 393, 440–442.

Network formation models inspired by what we observe in real-world networks

Small-world Random Graph Model
High clustering
Small average shortest path length

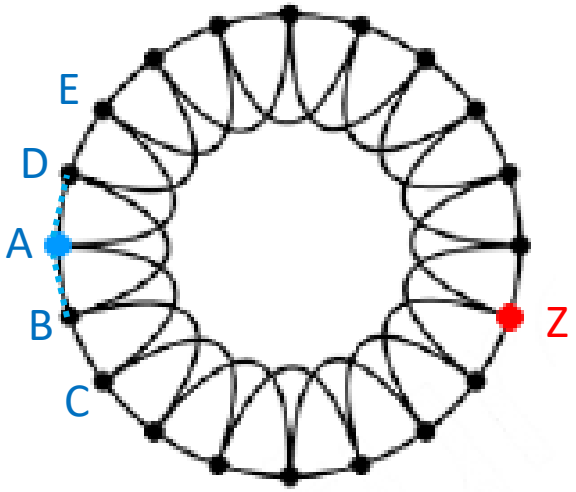
E-R Random Graph Model
Low clustering
Small average shortest path length



Small average shortest path length
High clustering

Regular Network

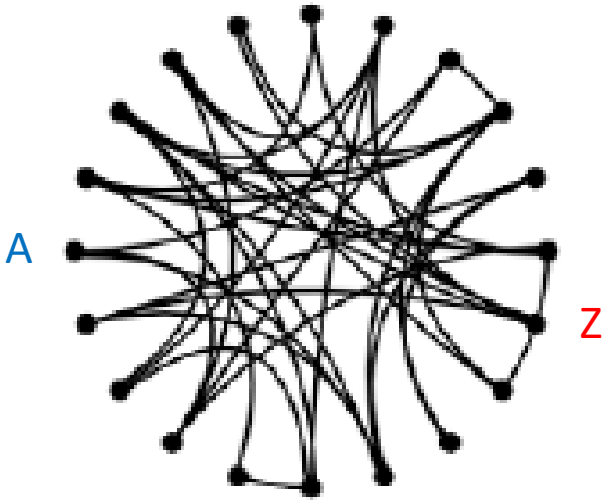
Completely regular



High clustering
High diameter

ER random Network

Completely random



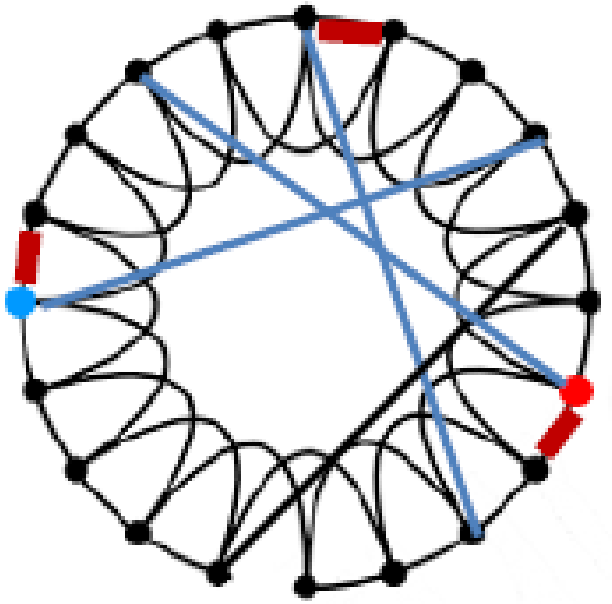
Low clustering
Low diameter

Can a network with high clustering be at the same time with small diameter?

The small-world model -Watts-Strogatz graphs

(1) Start with a low-dimensional regular lattice

- Use a ring as a lattice
- Has both high clustering coefficient and high diameter



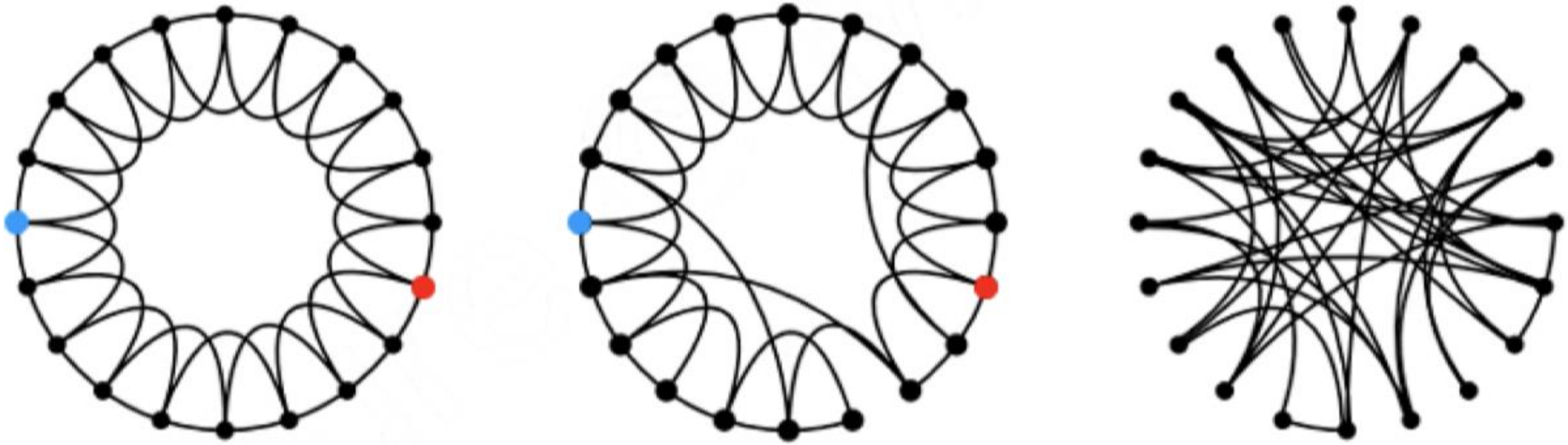
(2) Rewire: Introduce shortcuts between clusters to reduce diameter

- For each edge with a rewire probability p_r move the end of the edge to a node on the other side
- Add/remove edges to creat shortcuts to join remote parts

Regular Network

Small World Network

ER random Network



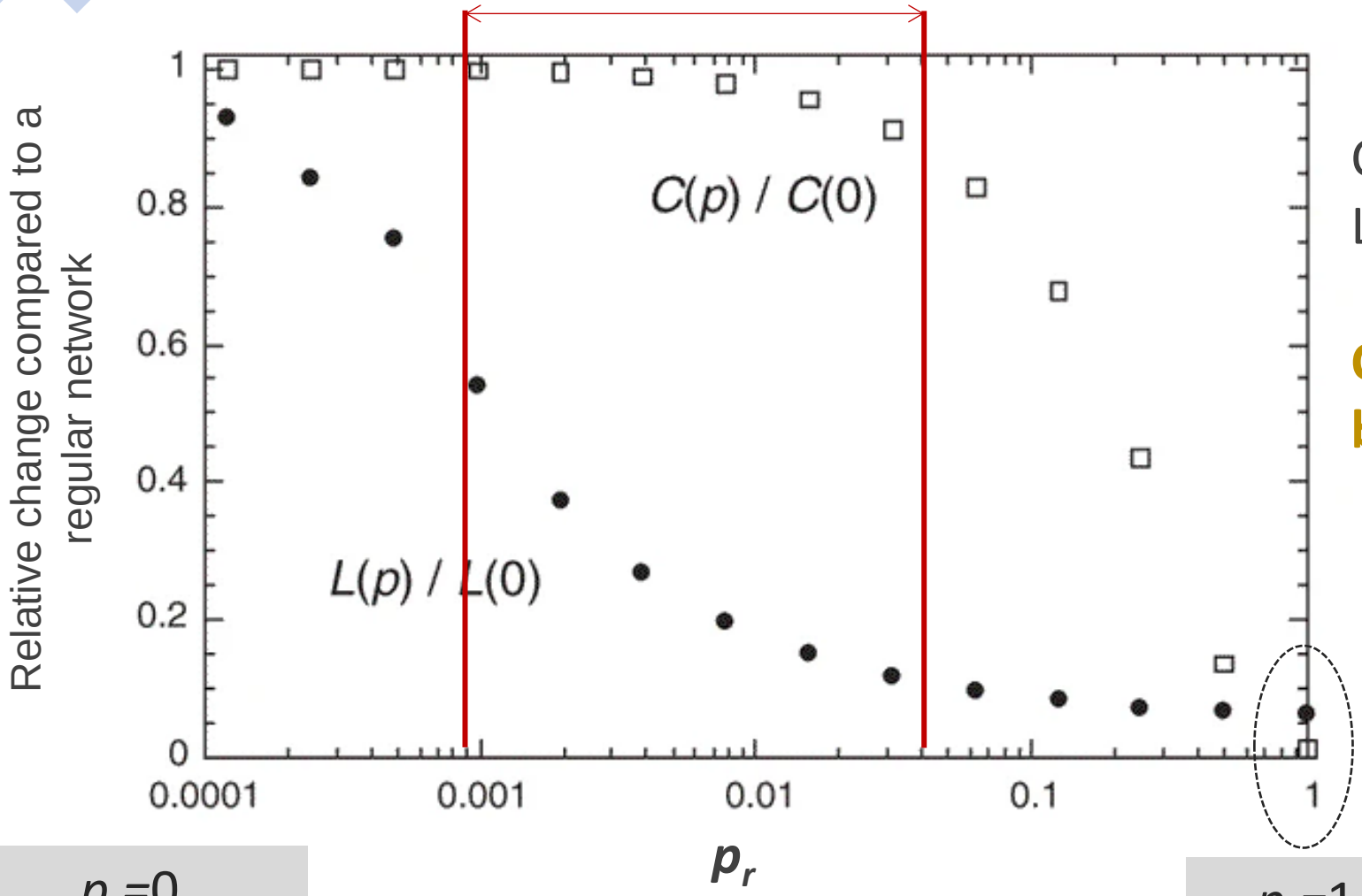
P=0 → INCREASING RANDOMNESS → P=1

High clustering High diameter	High clustering Low diameter	Low clustering Low diameter

Having both locality and shortcuts

Duncan J. Watts & Steven H. Strogatz, 1998. Collective dynamics of 'small-world' networks, ***Nature***, 393, 440–442.

How to get the rewire probability right



C: global clustering coefficient
L: average shortest path length

Clustering is hard to reduce, but not path length

p_r=0
Regular graph

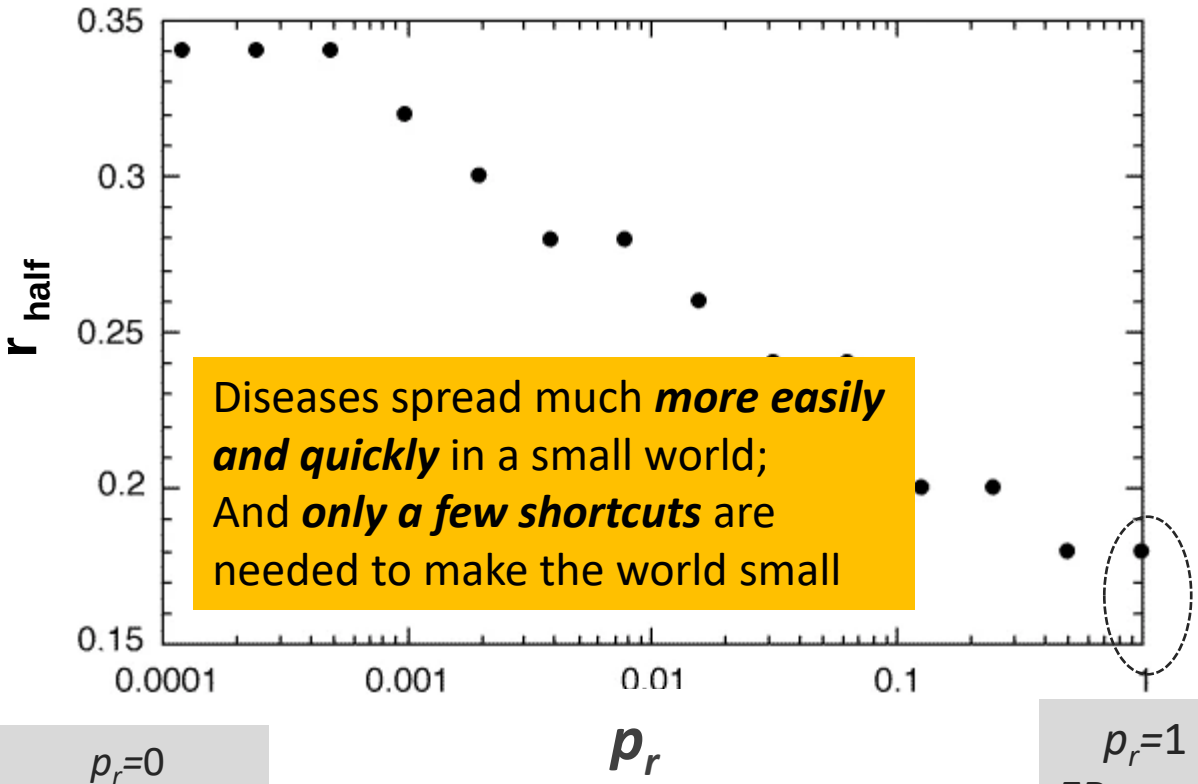
P_r <<1

p_r=1
ER graph

Infectious disease in a small-world network

At time t = 0, a single infective individual is introduced into an otherwise healthy population.
Infective individuals are removed permanently (by immunity or death) after one unit of time.
During this time, each infective individual can infect each of its healthy neighbours with probability ***r***.
On subsequent time steps, the disease spreads along the edges of the graph until it either infects the entire population, or it dies out.

r_{half} at which the disease infects 50% the population

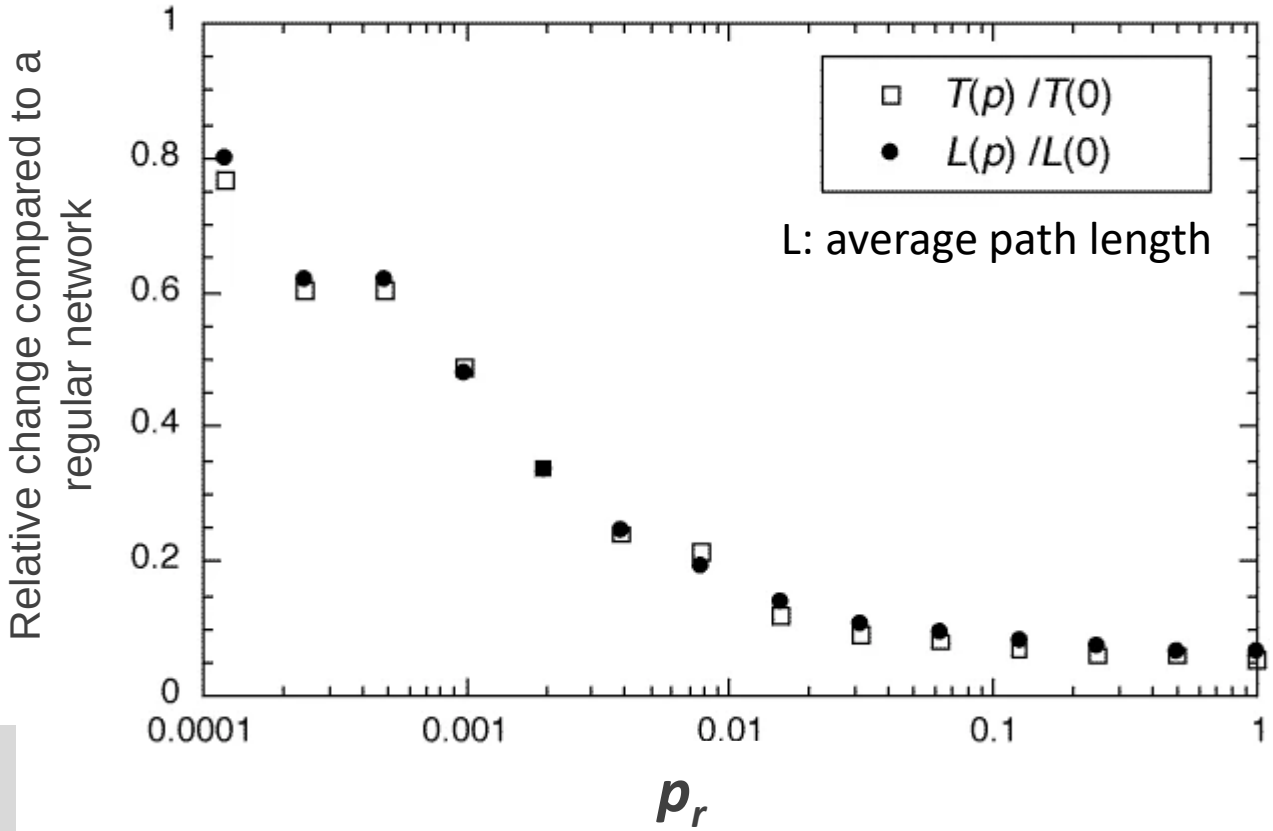


Diseases spread much ***more easily and quickly*** in a small world;
And ***only a few shortcuts*** are needed to make the world small

p_r=0
Regular graph

p_r=1
ER graph

The time $T(p)$ required for a maximally infectious disease (r = 1) for global population

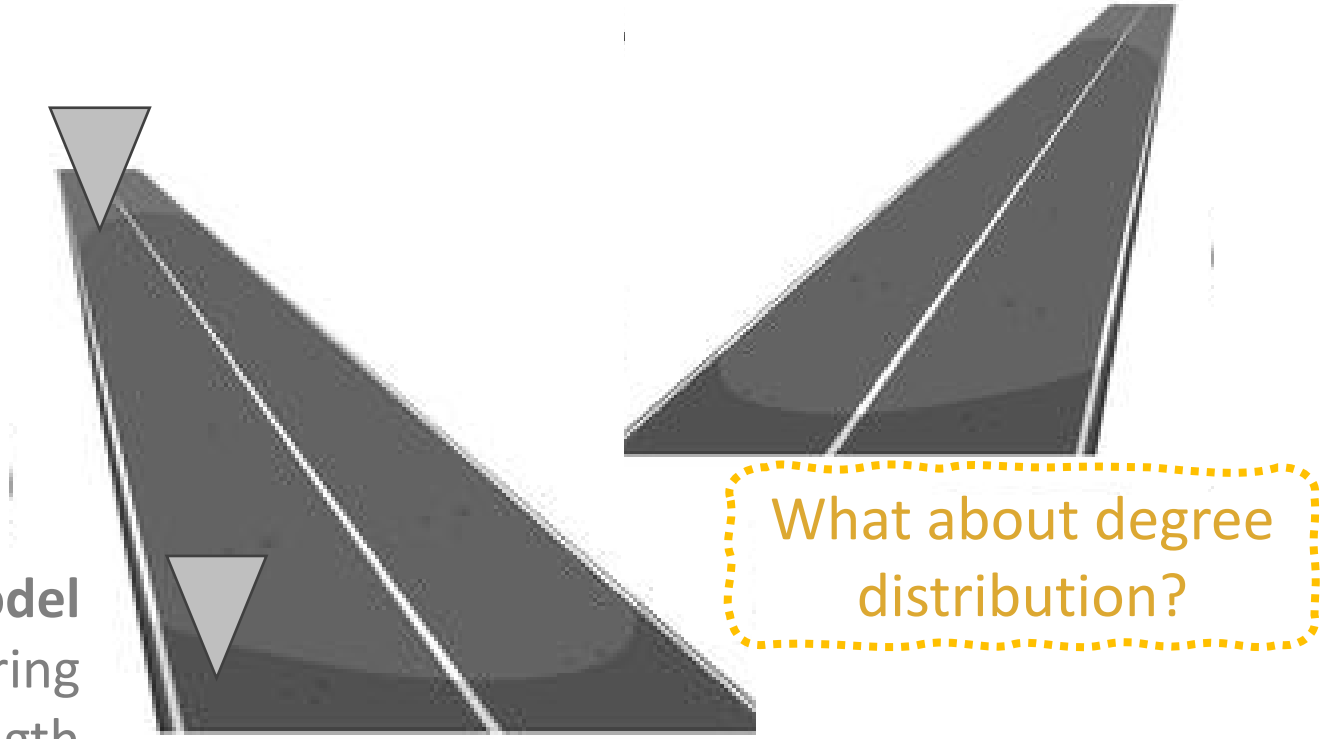


Duncan J. Watts & Steven H. Strogatz, 1998. Collective dynamics of 'small-world' networks, Nature, 393, 440–442.

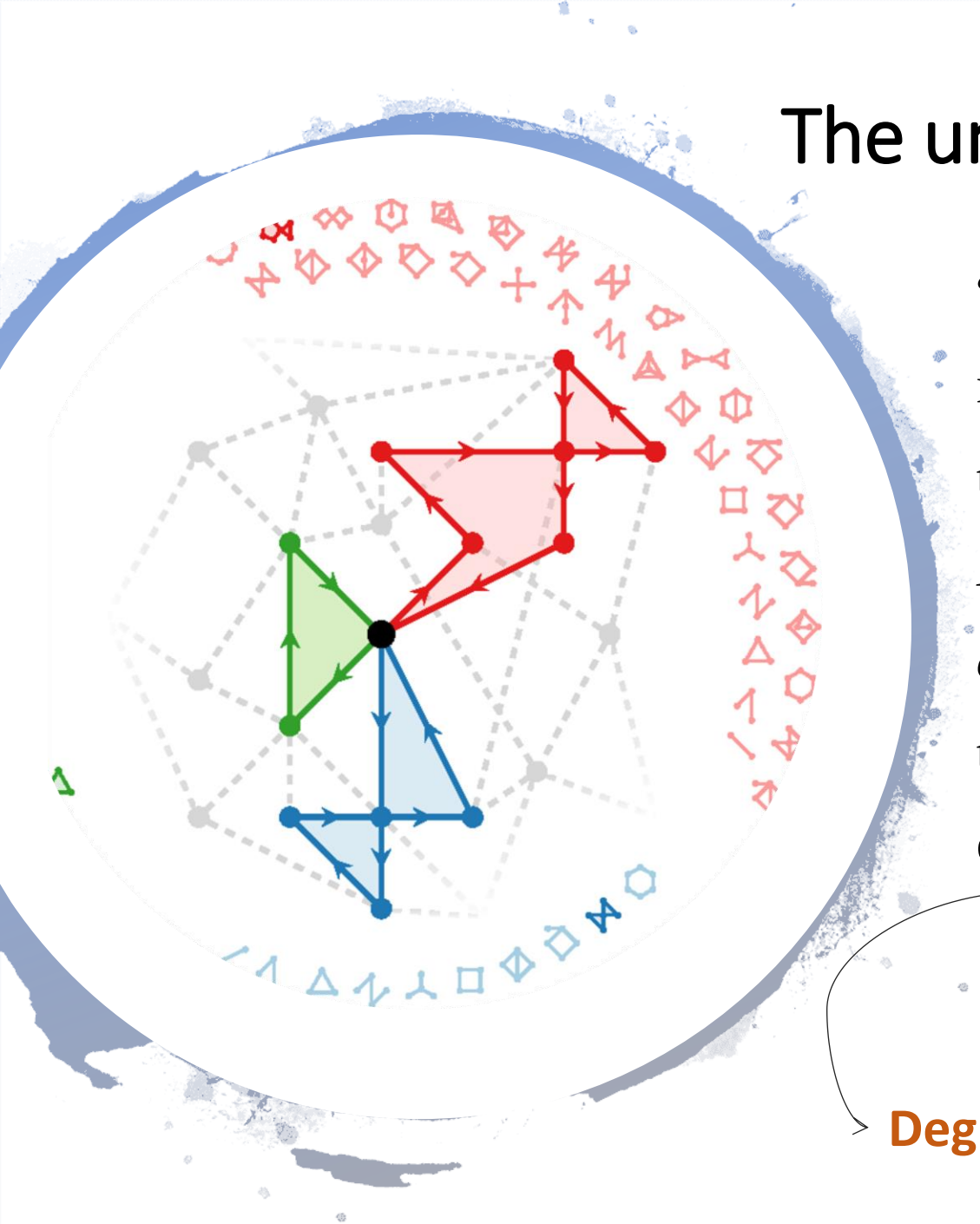
Network formation models inspired by what we observe in real-world networks

Small-world Random Graph Model
High clustering
Small average shortest path length

E-R Random Graph Model
Low clustering
Small average shortest path length



What about degree distribution?



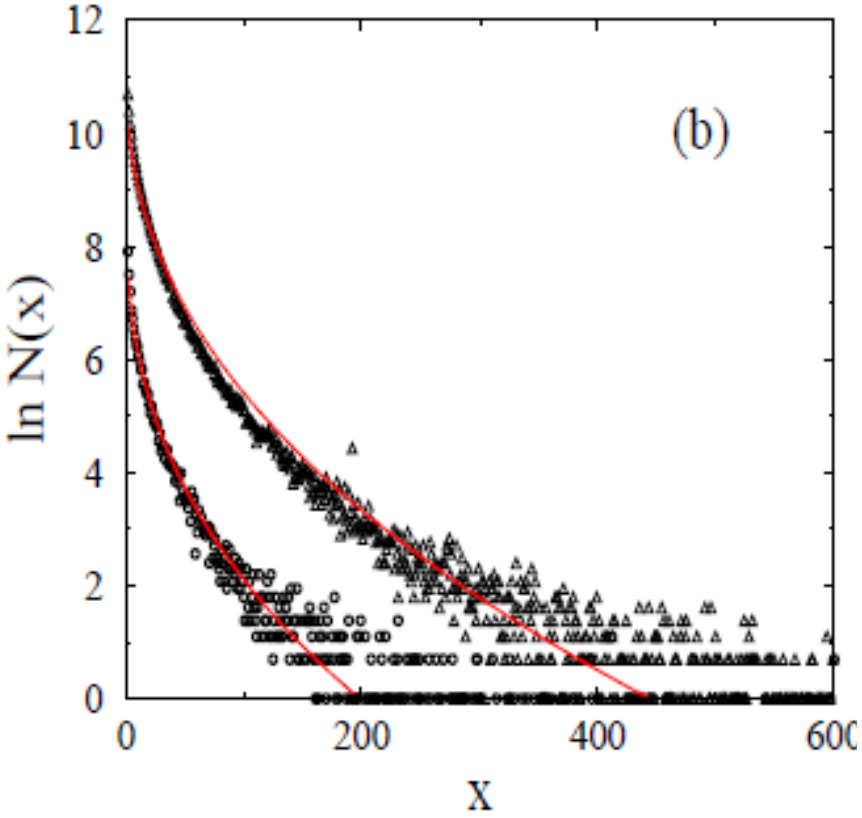
The universality of the network topology

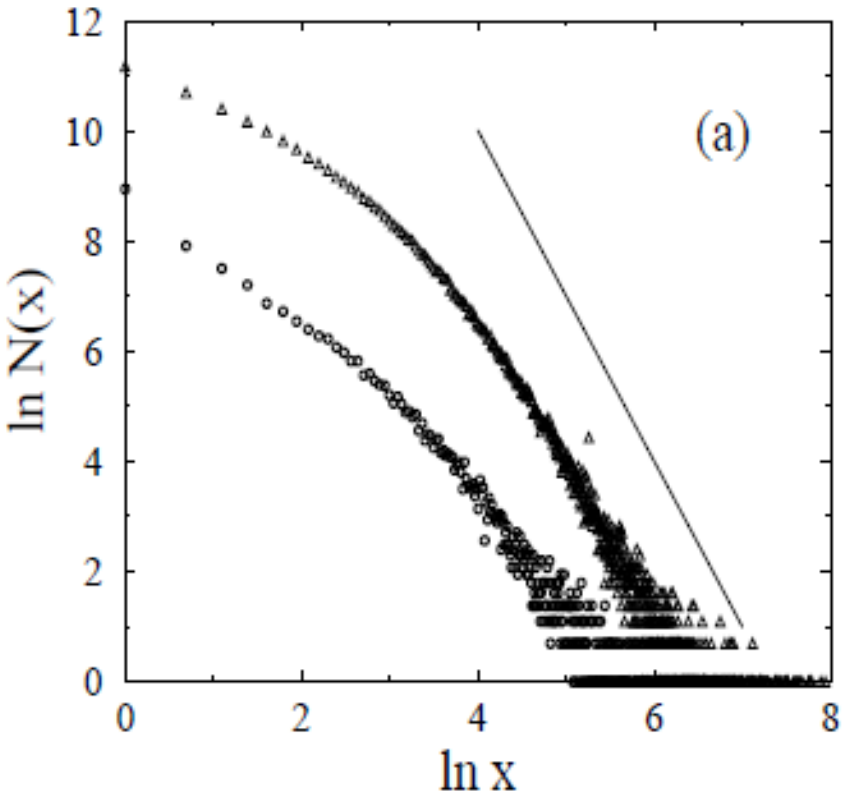
"Probably the most surprising discovery of modern network theory is the universality of the network topology: *Many real networks, from the cell to the Internet, independent of their age, function, and scope, converge to similar architectures*. It is this universality that allowed researchers from different disciplines to embrace network theory as a common paradigm."

------ Albert-László Barabási, 2009, ***Science***, 325: 412-4.

Degree distribution

Tail of degree distribution follows a power law: Citation networks





N(k): the number of papers with *k* citations

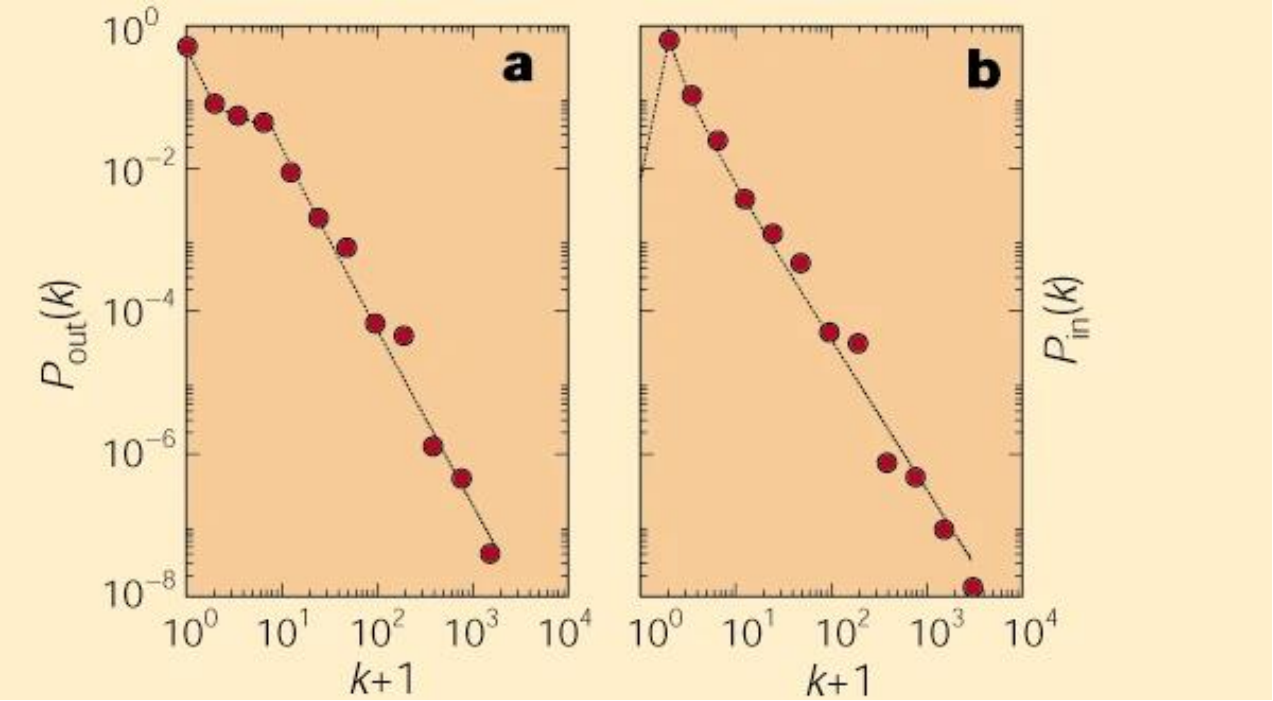
For large k, $N(k) \sim k^{-\gamma}$, with $\gamma \approx 3$.

Δ: Citation distribution of the 783,339 papers in journals catalogued by the Institute for Scientific Information

o: Citation distribution of the 24,296 papers in Physical Review D, vols. 11-50

S. Redner, Eur. How popular is your paper? An empirical study of the citation distribution. Phys. J. B 4, 131 (1998).

Tail of degree distribution follows a power law: World-Wide Web



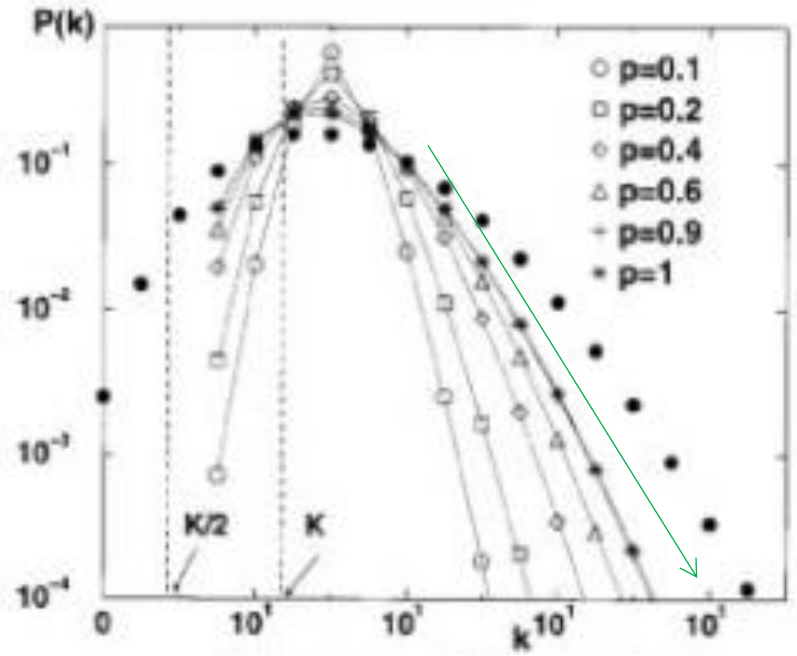
Outgoing links (URLs found on an HTML document)
Incoming links (URLs pointing to a certain HTML document)

The tail of the degree distribution follows $p(k) \approx k^{-\gamma}$, with $\gamma_{out}=2.45$ and $\gamma_{in}=2.1$.

R. Albert, H. Jeong, A.-L. Barabási, Nature 401, 130 (1999).

Degree distribution of ER and SW networks

ER and SW models: Binomial degree distribution

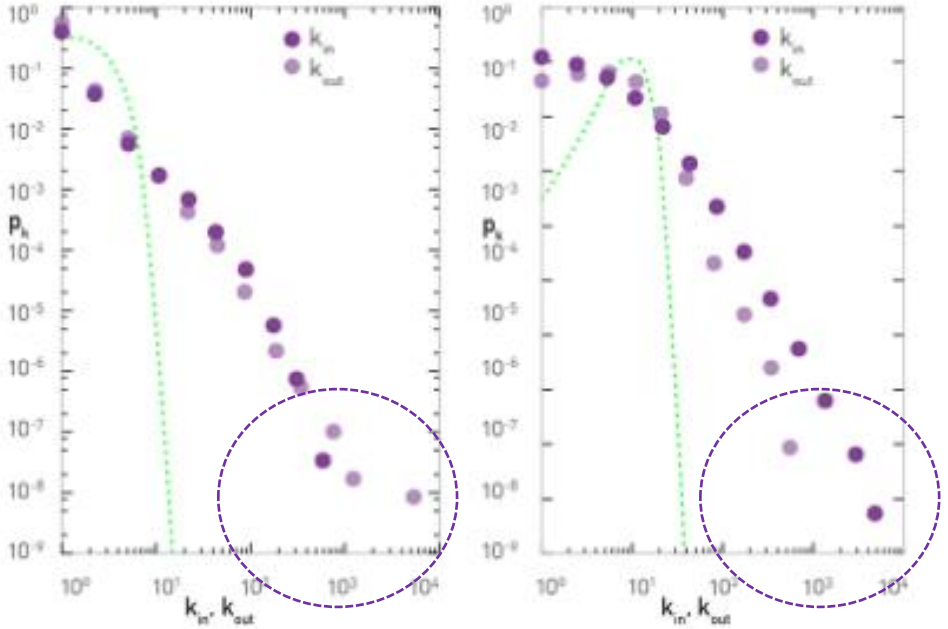


$p(k)$ that peaks at the average and **decays exponentially for large *k***

Many node have about the same number of links around the mean; No highly-connected or poorly-connected nodes.

"exponential networks" or **"homogeneous networks"**

Degree distribution follows **a power law distribution** and produced by ***ER models***



Inhomogeneous: a few nodes have many connections ("hubs")

From https://www.barabasilab.com/post/love-is-all-you-need

Barabasi-Albert (BA) model

Growth

- Networks are open and they form by the continuous addition of new vertices to the system
- Thus the number of vertices *N* increases throughout the lifetime of the network. In contrast the random network model assumes *N* is fixed.

Preferential attachment

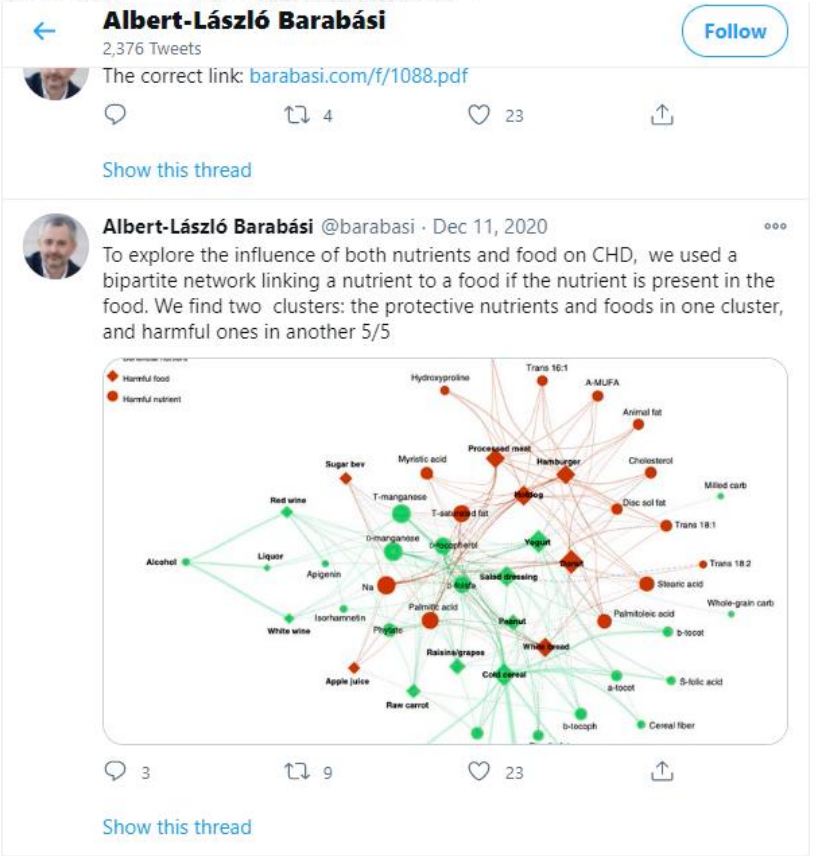
- Real-world network construction is a continues growth (path dependence)
- New nodes in most real networks prefer to link to the more connected nodes
- “Cumulative advantage", "the rich get richer" ("Matthew effect“)

Barabasi-Albert (BA) model

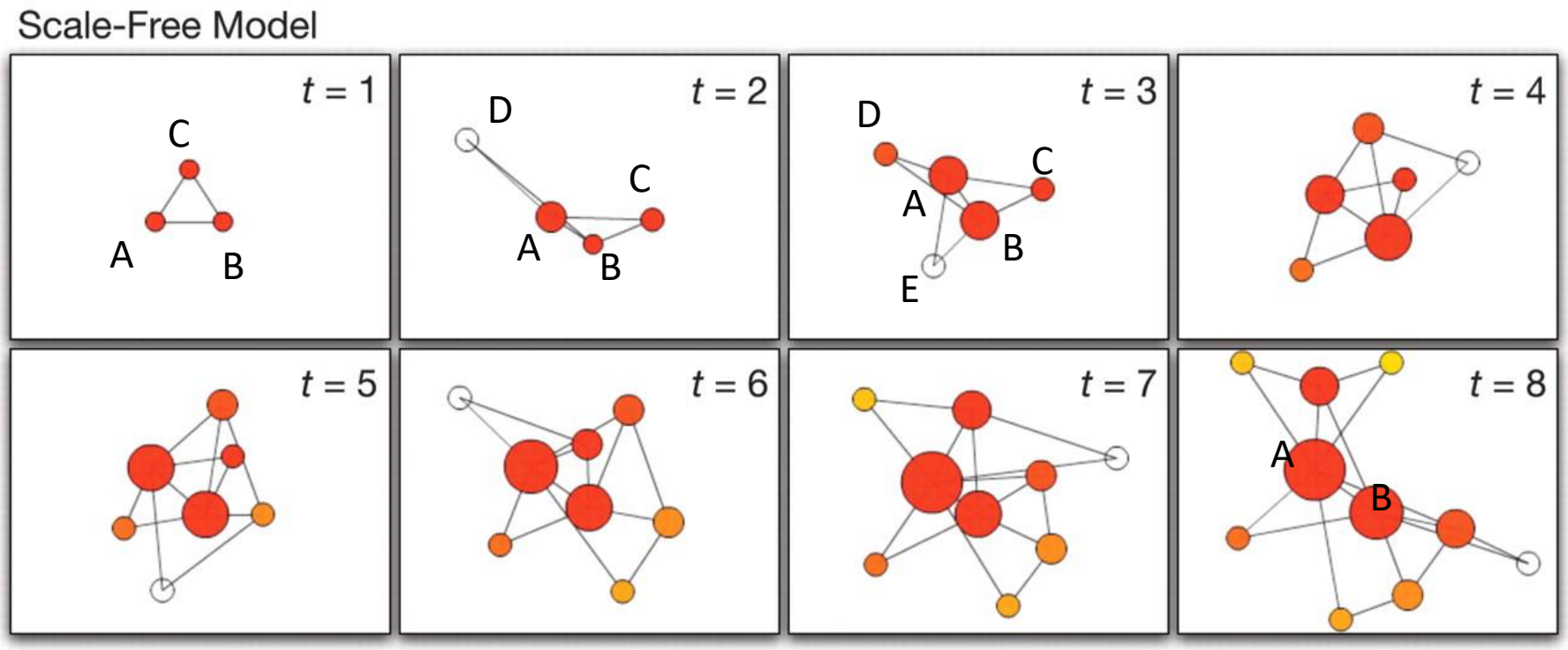
- **Growth:** At each timestep we add a new node with m ($\leq m_0$) links that connect the new node to m nodes already in the network.
- **Preferential attachment**: The probability Π (k_i) that a link of the new node connects to node i depends on the degree k_i,

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

A probabilistic mechanism: A new node is free to connect to *any* node in the network, whether it is a hub or has a single link. However, if a new node has a choice between a degree-two and a degree-four node, it is twice as likely that it connects to the degree-four node.



The birth of a scale-free network under the BA model

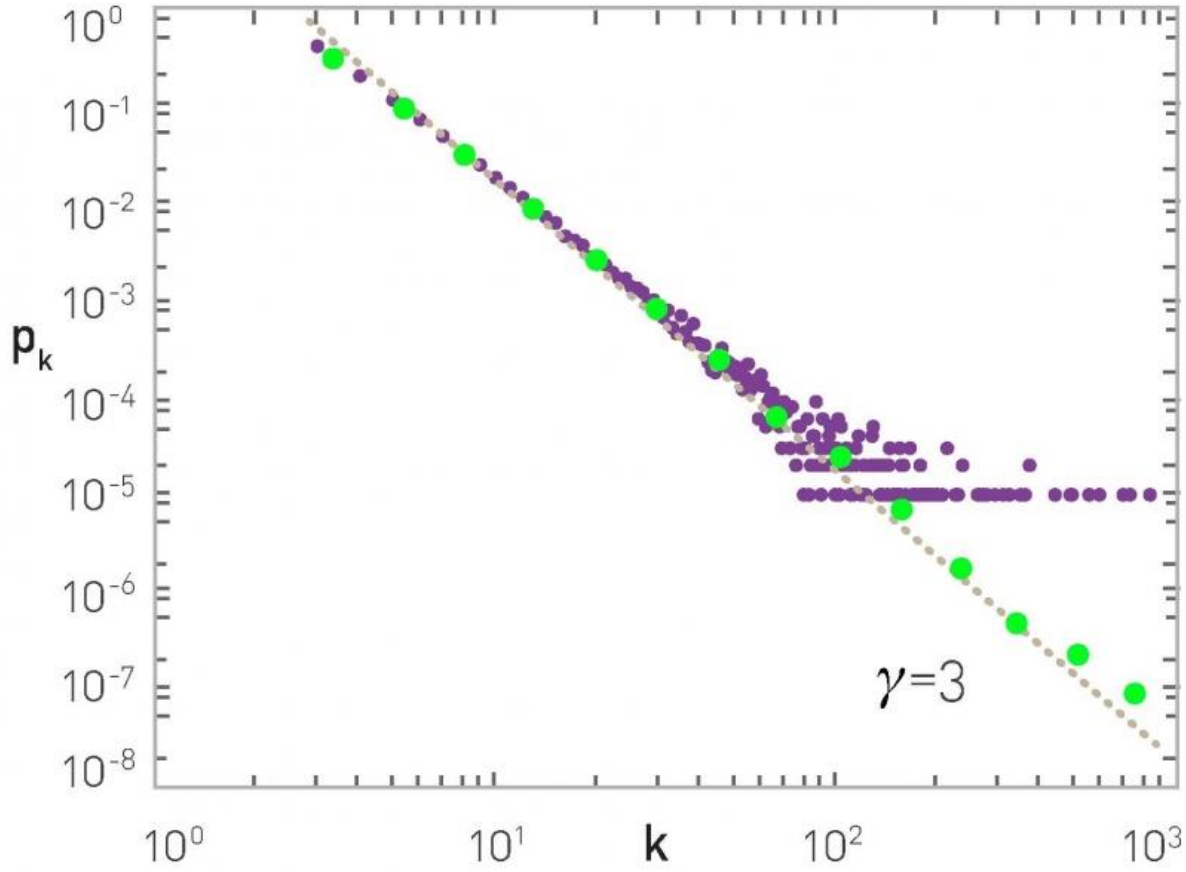


An **initial difference** in the connectivity between two vertices will **increase further as the network grows**;

The earlier node *i* is added, the higher will be its degree $k_i(t)$. Hence, hubs are large because they arrived earlier, a phenomenon called ***first-mover advantage*** in marketing and business.

Properties for a standard BA model ("Scale-free model")

Degree distribution $p(k) \sim 2m^2k^{-3}$



linearly-binned (purple); log-binned (green) of pk

Intersect is controlled by m

Average path length ~ *ln(N)/ln(ln(N))*

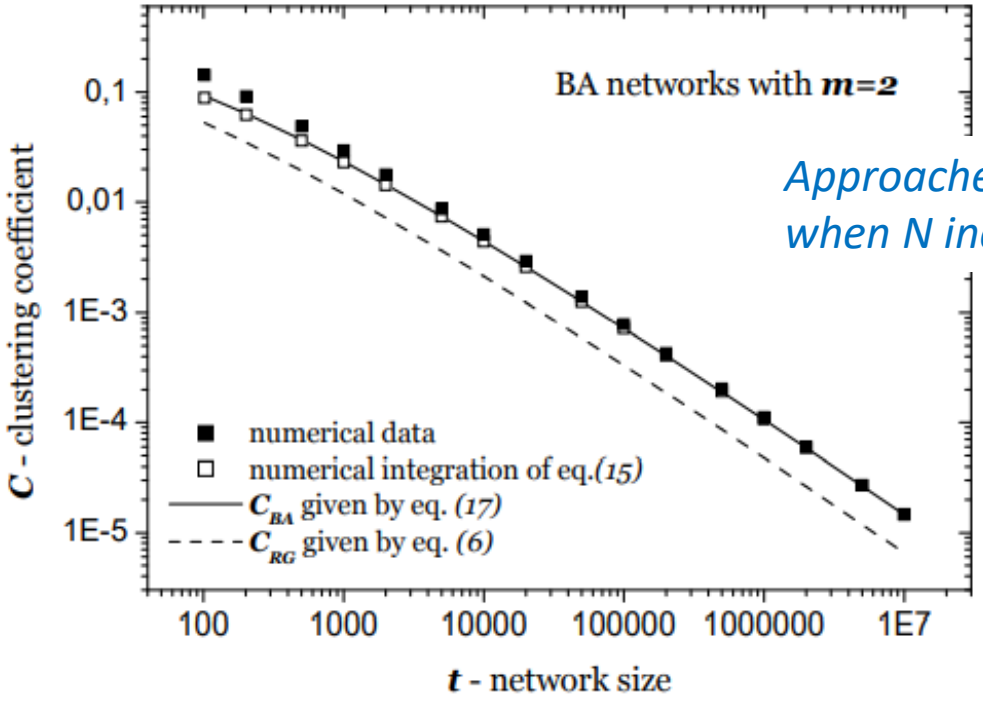
A systematically shorter average path length than a random graph

"Ultra small world"

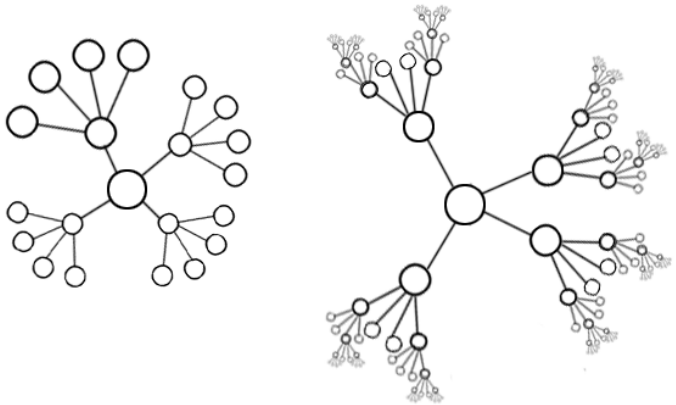
From http://networksciencebook.com/chapter/5#barabasi-model;
Reuven Cohen, and Shlomo Havlin, 2003. Scale-Free Networks are Ultrasmall.

Clustering coefficient:

Global clustering coefficient rapidly decreases with the network size



Approaches to 0 when N increases.



Typically several orders of magnitude lower than found empirically:

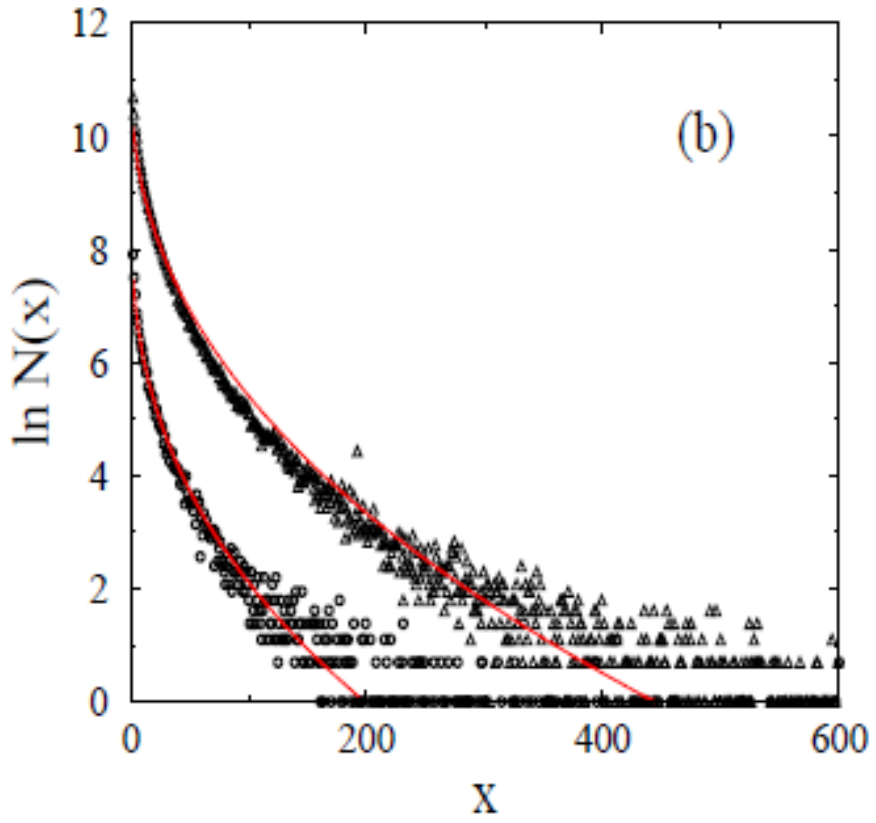
Gnutella peer-to-peer network: **0.0062** (size=6301)
EU email communication network (All): **0.0671** (size=265214)
Youtube: **0.0808** (size=1134890)
Higgs Twitter Dataset: **0.1887** (size=456626)
Amazon *Customers Who Bought This Item Also Bought*: **0.4177** (size=403394)

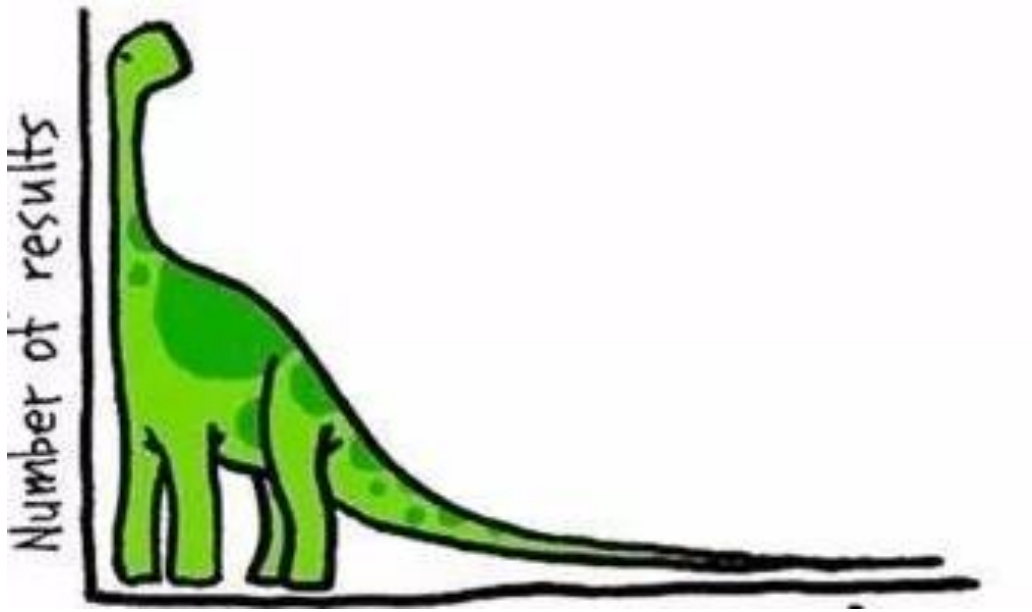
With some corrections clustering coefficient can increase significantly!

Agata Fronczak, Piotr Fronczak and Janusz A. Ho lyst. Mean-field theory for clustering coefficients in Barab´asi-Albert.

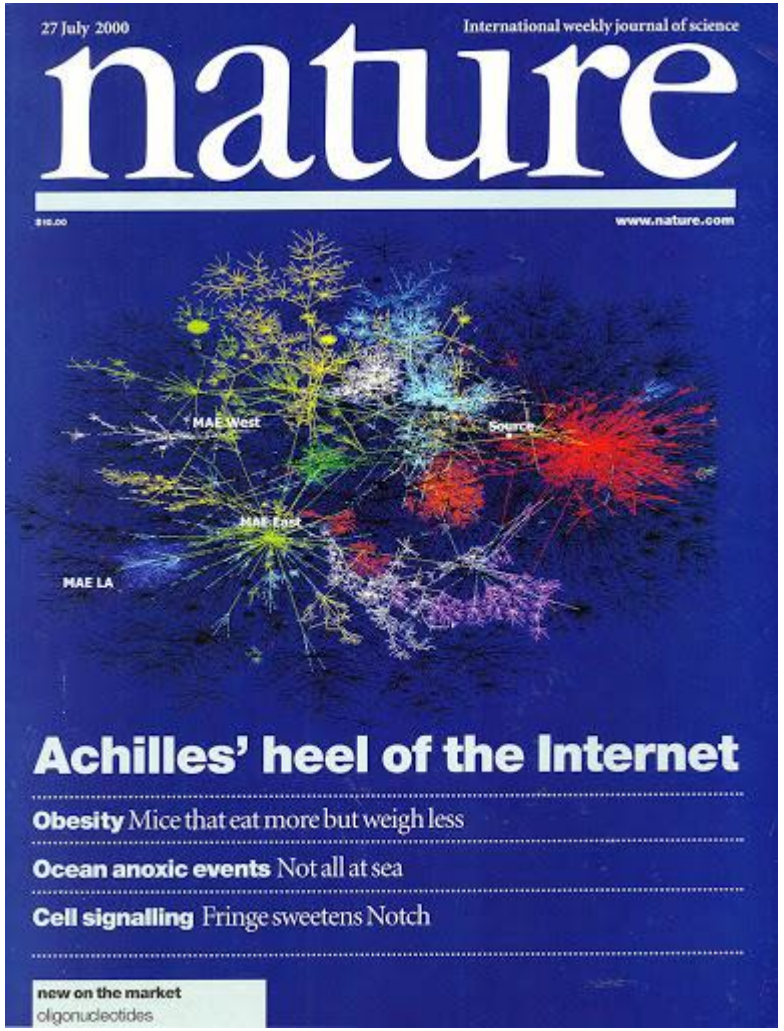
Important implications for a network produced by the BA model

(i.e., a scale-free network with high inhomogeneity in degree)





Implication 1: A system's ability to survive random failures or targetted attacks



Its overall strength:

High error tolerance:

The communication ability of the nodes are unaffected even under very high failure rates

Its only weakness:

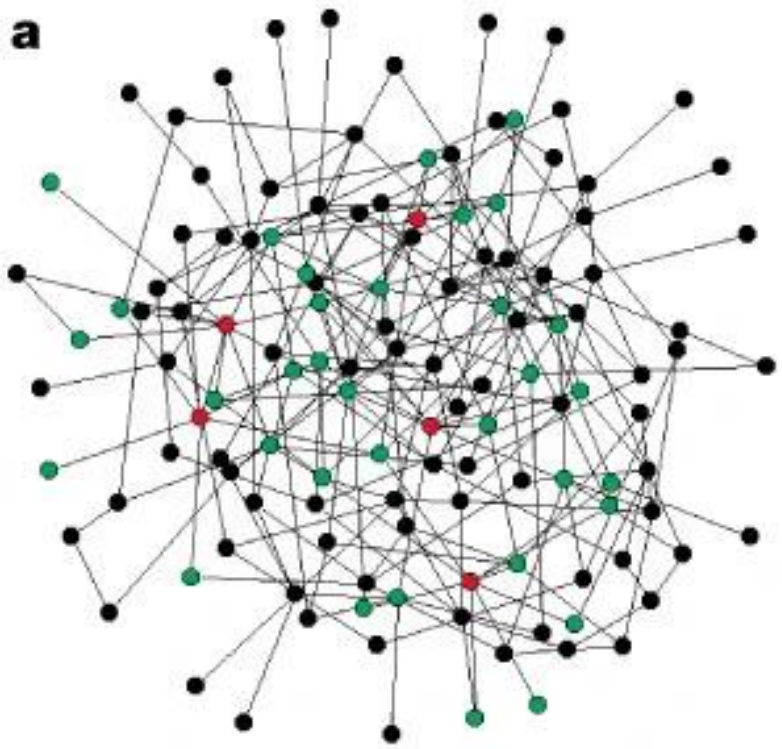
Extremely vulnerable to attacks (that is, to the selection and removal of a few nodes that play a vital role in maintaining the network's connectivity)

"Overall it is very strong but it has one deathly weakness"

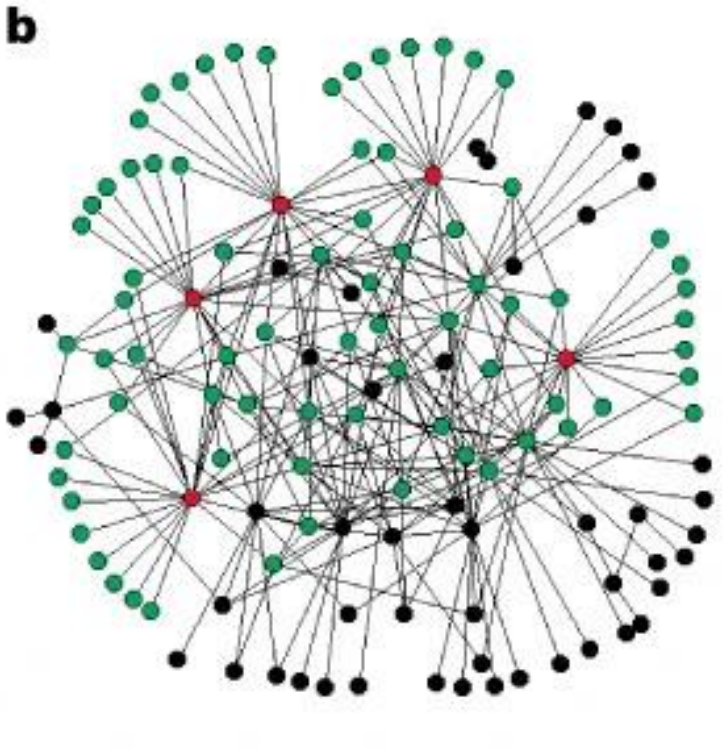
Réka Albert, Hawoong Jeong & Albert-László Barabási., 2000. Error and attack tolerance of complex networks. Nature, 406, pages378–382

Red, the five nodes with the highest number of links
Green, their first neighbours

27% of the nodes are reached by the five most connected nodes



Random network by ER model



Scale-free network by BA model

> 60% are reached by the five most connected nodes

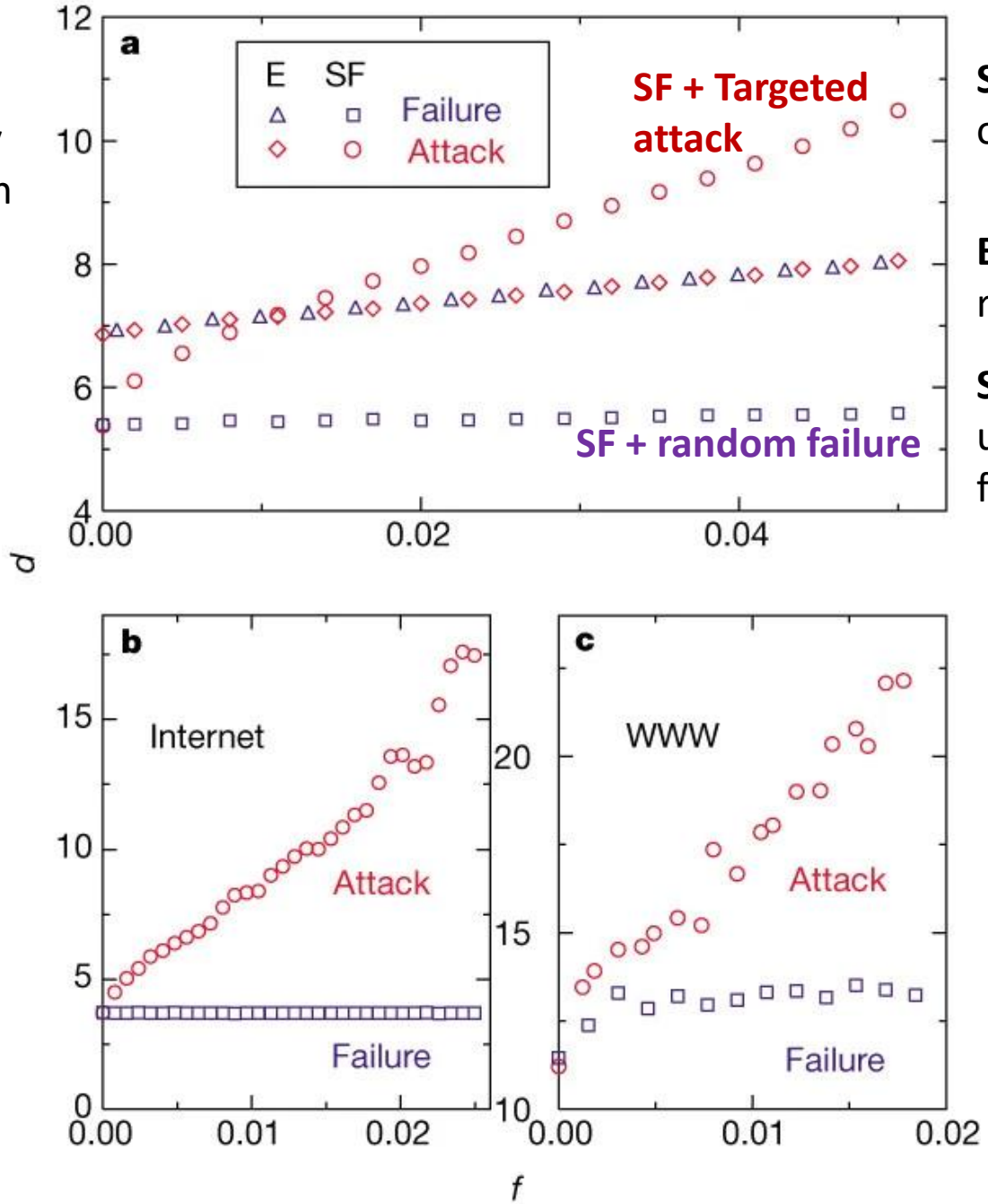
Both networks contain 130 nodes and 215 links

Réka Albert, Hawoong Jeong & Albert-László Barabási., 2000. Error and attack tolerance of complex networks. Nature, 406, pages378–382

Diameter *d :* connectivity of a network; characterizes the ability of two nodes to communicate with each other: the smaller *d* is, the shorter is the expected path between them.

Changes in diameter *d* when a small fraction *f* of the nodes is removed.

Random Failure: nodes are removed randomly
Targeted Attack: the most connected nodes are removed (attacked by an informed agent)



SF: Diameter doubles by removing 5% of most connected nodes

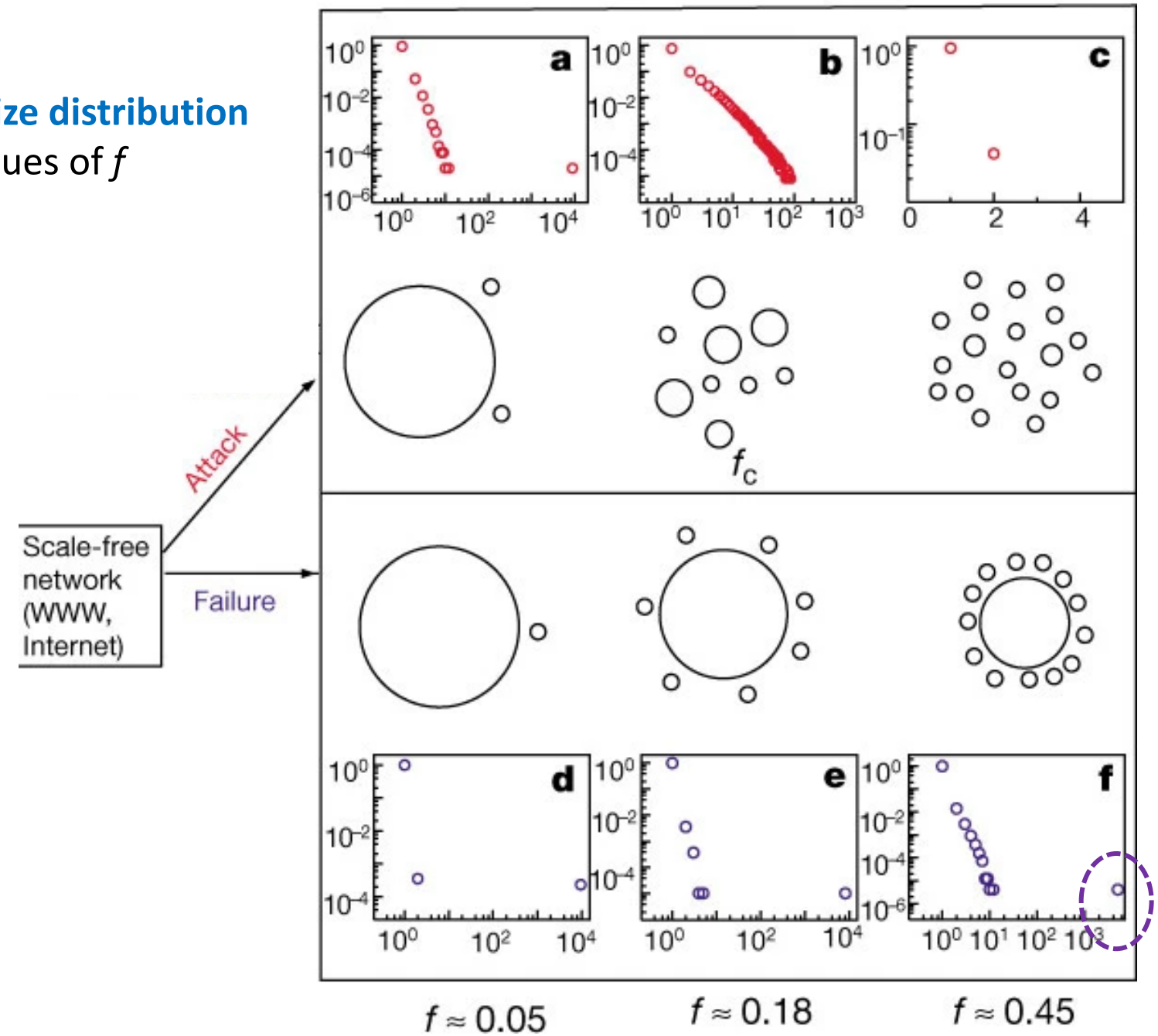
ER: Similar effects from removal of random or most connected nodes

SF: Diameter remains unchanged under an increasing level of random failure

Réka Albert, Hawoong Jeong & Albert-László Barabási., 2000. Error and attack tolerance of complex networks. ***Nature***, 406, pages378–382

Component size distribution

for various values of *f*



Even for an unrealistically high error rate of *f = 0.45* the large cluster persists

Réka Albert, Hawoong Jeong & Albert-László Barabási., 2000. Error and attack tolerance of complex networks. Nature, 406, pages378–382

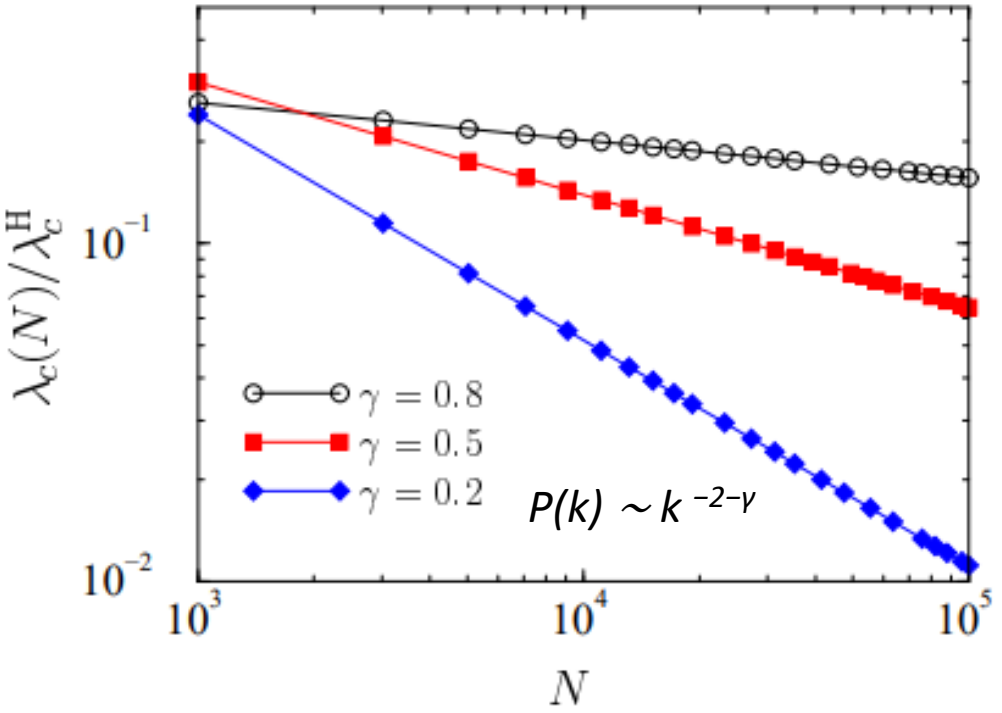
Implication 2: Prone to the spreading and persistence of virus

Each susceptible (healthy) node is infected with rate v;

Infected nodes are cured and become again susceptible with rate ß;

An effective spreading rate ƛ=v/ ß;

ƛc, at which infection spreads and becomes persistent.



Compared to homogeneous networks with the same average connectivity

Random immunization: Requires immunizing a very large faction of population (Measles: 95%; Computer virus: ~100%)

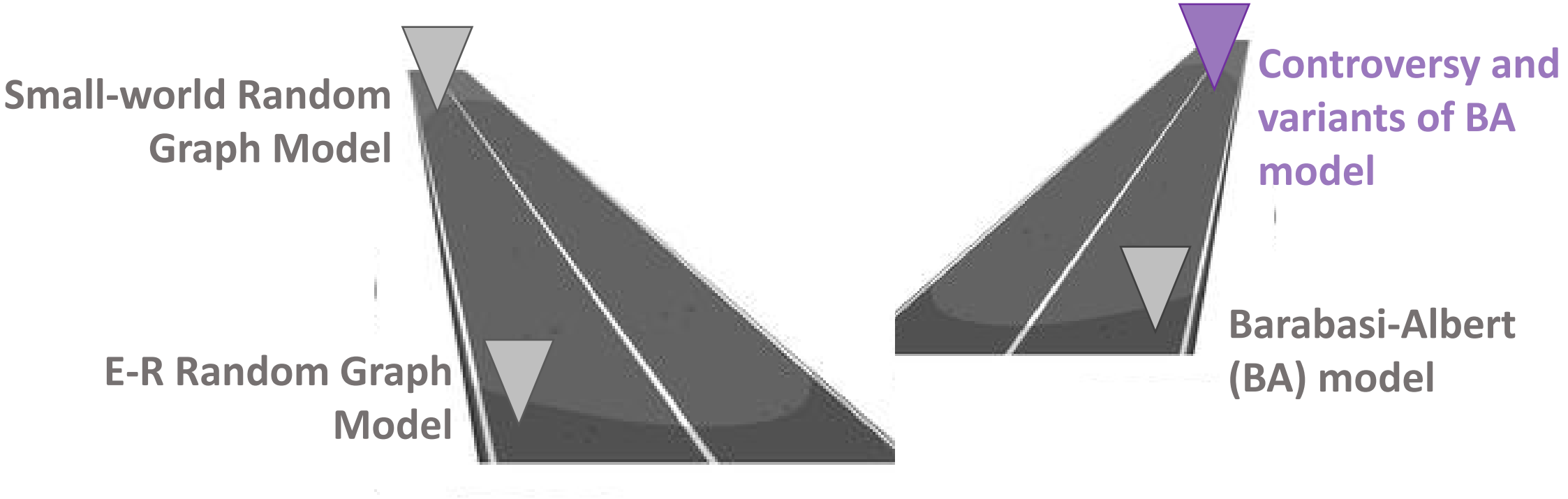
Targeted immunization: Effective but require information of network

“Acquaintance immunization”: The immunization of random acquaintances of random nodes (individuals).

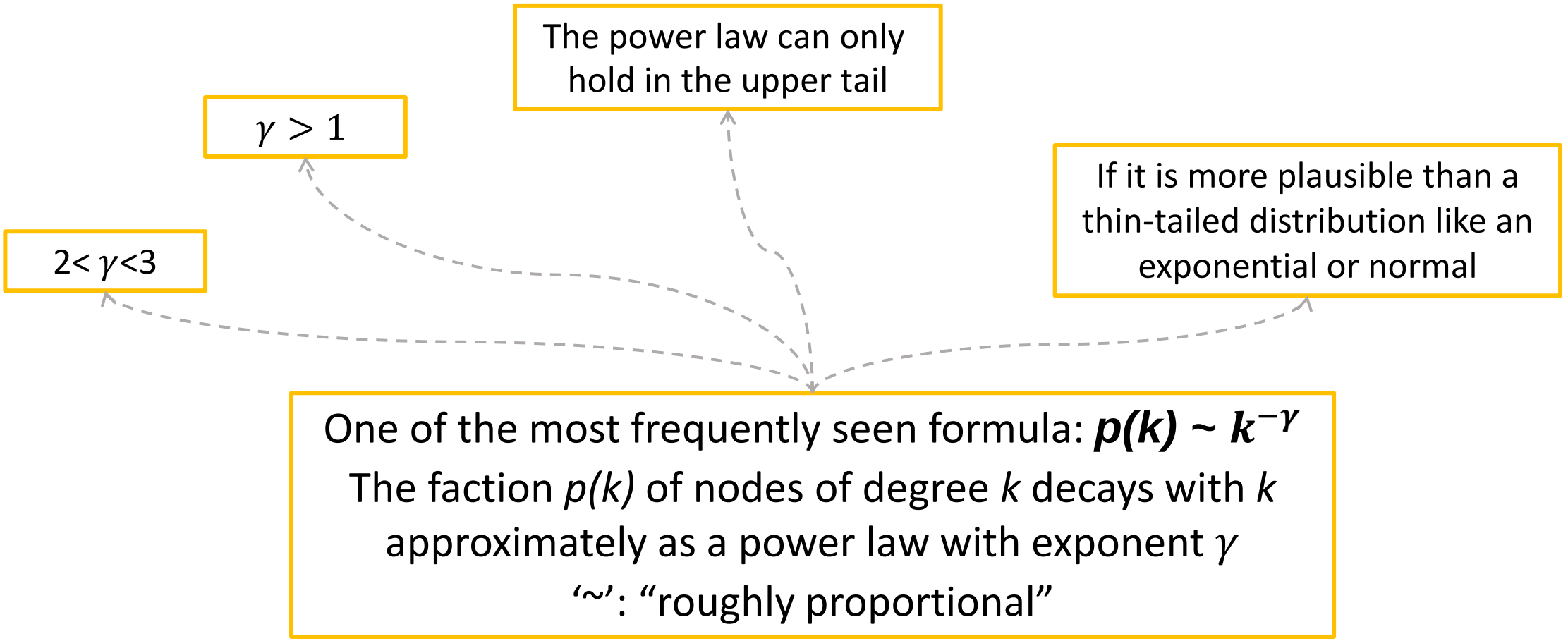
Romualdo Pastor-Satorras, Alessandro Vespignani, 2002. Epidemic dynamics in finite size scale-free networks.

Reuven Cohen, Shlomo Havlin, and Daniel ben-Avraham, 2003. Efficient Immunization Strategies for Computer Networks and Populations

Network formation models inspired by what we observe in real-world networks



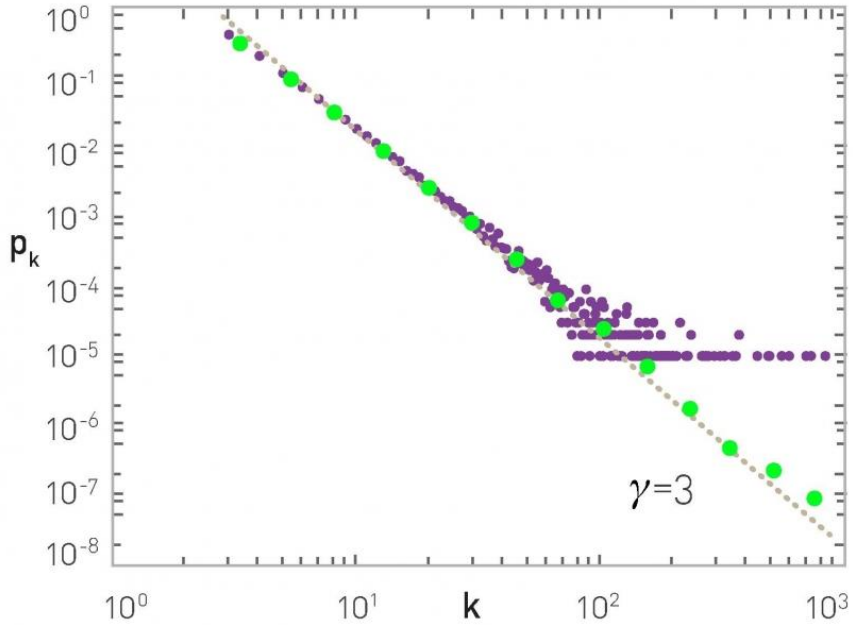
No widely agreed-upon rigorous definition of SF networks



Different researchers can use the same term to refer to slightly different concepts

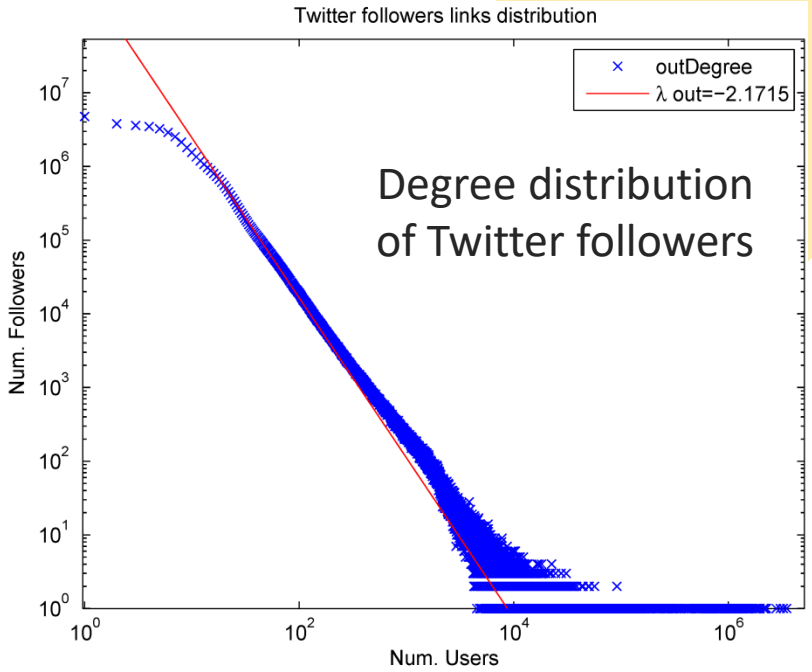
γ in many real-world networks: 2-8

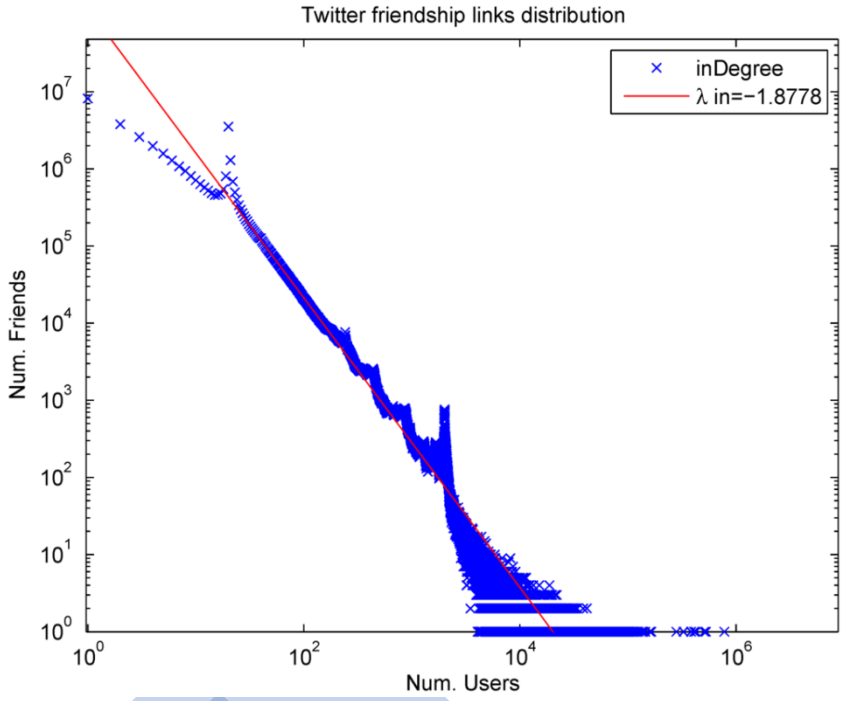
A standard degree distribution for networks created by BA model



linearly-binned (purple); log-binned (green) of pk

From http://networksciencebook.com/chapter/5#barabasi-model





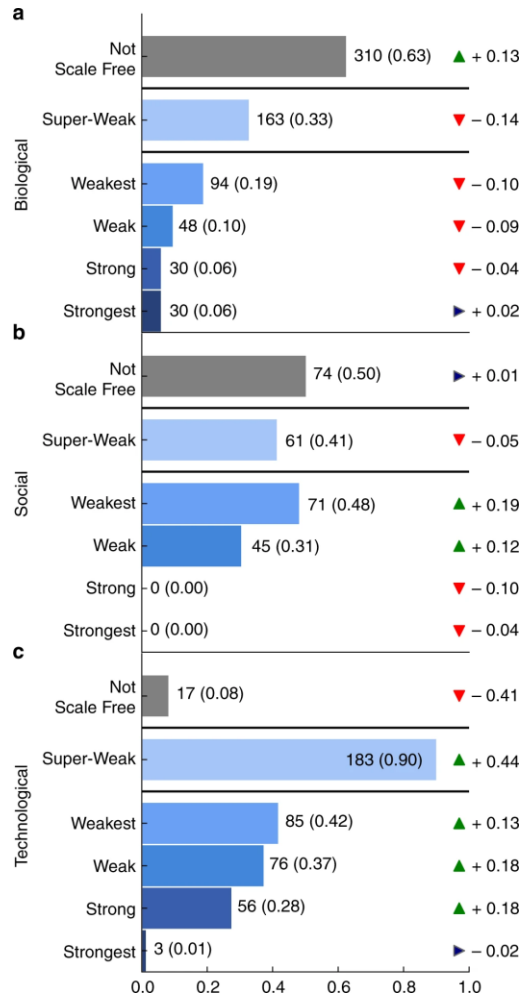
“Strongly scale-free structure is empirically rare”

Broido & Clauset, 2019. ‘Scale-free networks are rare’, Nature Communication:

- For the degree distribution of real networks, estimated the best-fitting power-law distribution, test its statistical plausibility, and compare it to non-scale-free distribution
- Strongly scale-free structure is empirically rare, while for most networks, **log-normal distributions fit the data as well or better than power laws**

Barabasi fight back, https://www.barabasilab.com/post/love-is-all-you-need

- Even a synthetic SF network will fail for the strict procedure produced by Broido & Clauset
- The scale-free model is a mechanistic model. It is not a model of the Internet, or the WWW or the cell—it only aims to explain the mechanism behind the scale-free nature of a network



“BA/SF model is not our final destination as our understanding of the network formation”

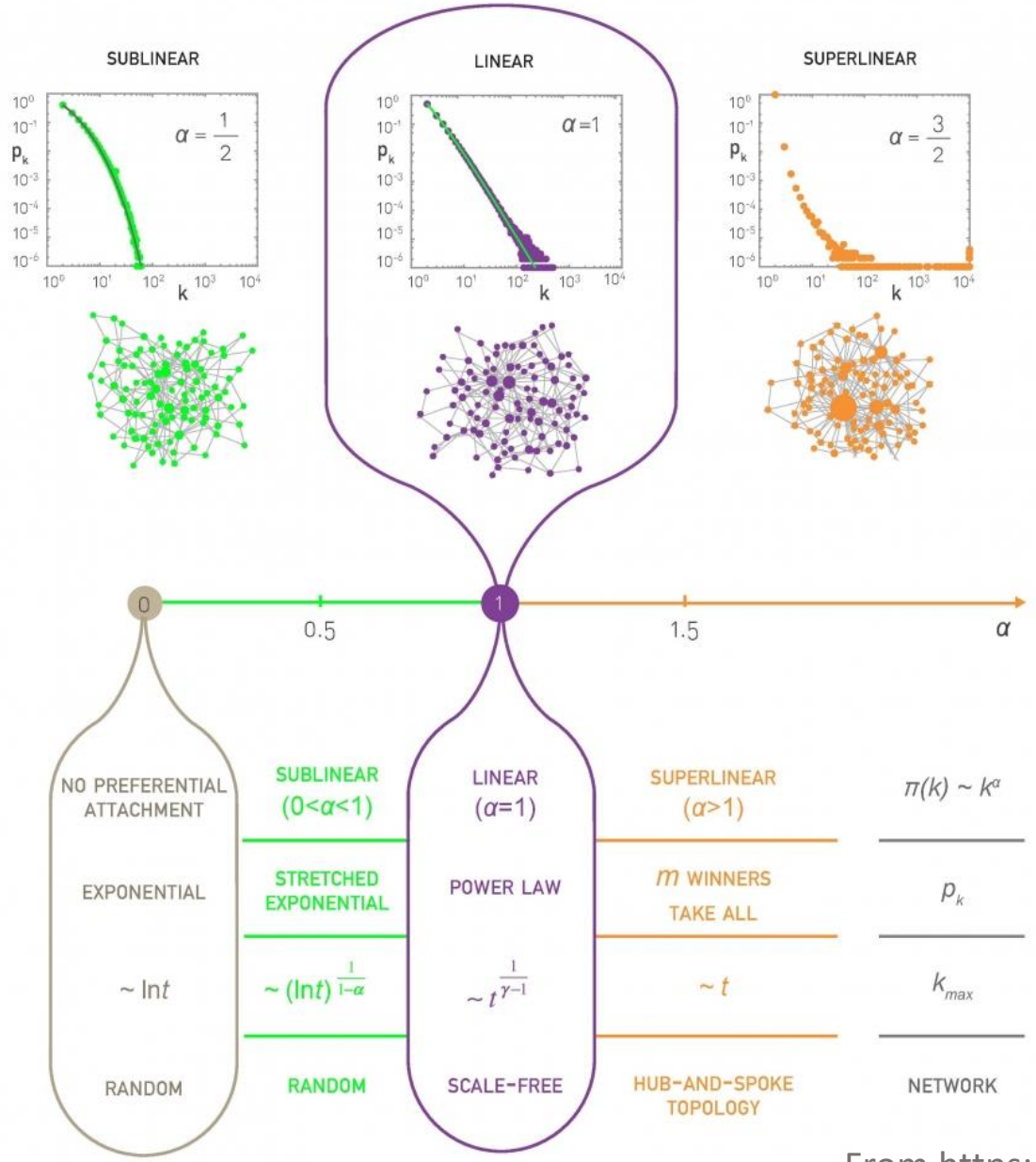
The need for a formulized definition of SF network and test of its universality

More ‘corrections’/variants of BA model to capture network varieties

Variant 1:Non-linear preferential attachment (NLPA) model

$$\Pi(k_i) = \frac{k_i^{\alpha}}{\sum_j k_j^{\alpha}}$$

Sublinear: Fewer and smaller hubs than in a scale-free network



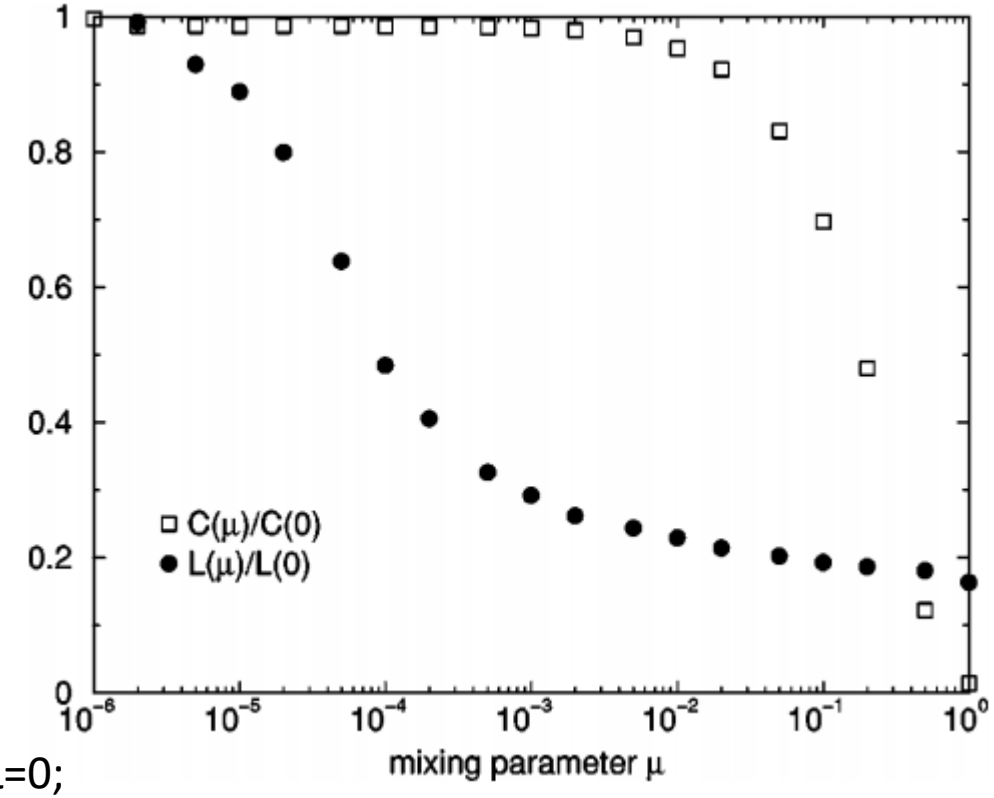
Superlinear:
High degree nodes are even more attractive;
A *winner-takes-all* dynamics leads to a hub-and-spoke topology

From https://barabasi.com/f/622.pdf

Variant 2: Scale-free networks with small-world behavior

A network with **scale-free distribution of degree and high clustering coefficient**

μ: For each of the *m* links of the new node it is decided with probability *μ*, whether the link connects to the active node (controlled by deactivation process) (*1- μ)* or connects to a random node according to linear preferential attachment (*μ*)



μ=0;
high clustering model

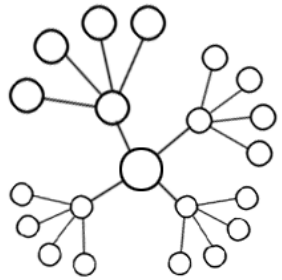
μ=1;
BA model

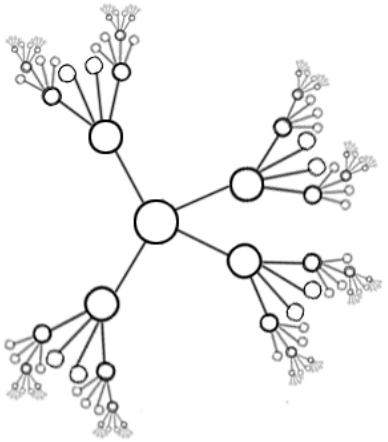
Konstantin Klemm and Vı´ctor M. Eguı´luz, 2001. Growing scale-free networks with small-world behavior. PHYSICAL REVIEW E, VOLUME 65, 057102

More variants

Internal links: In any system where two nodes that are already in the network can choose to connect to each other (partly solve the low clustering problem)

High degree cutoffs: A cut-off in the degree distribution of a finite size network due to structural limitations

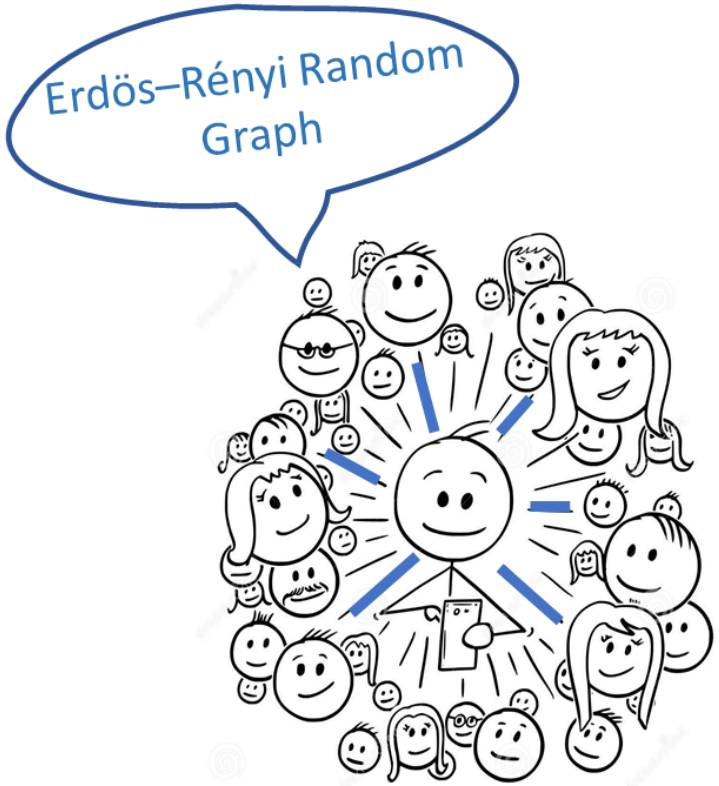


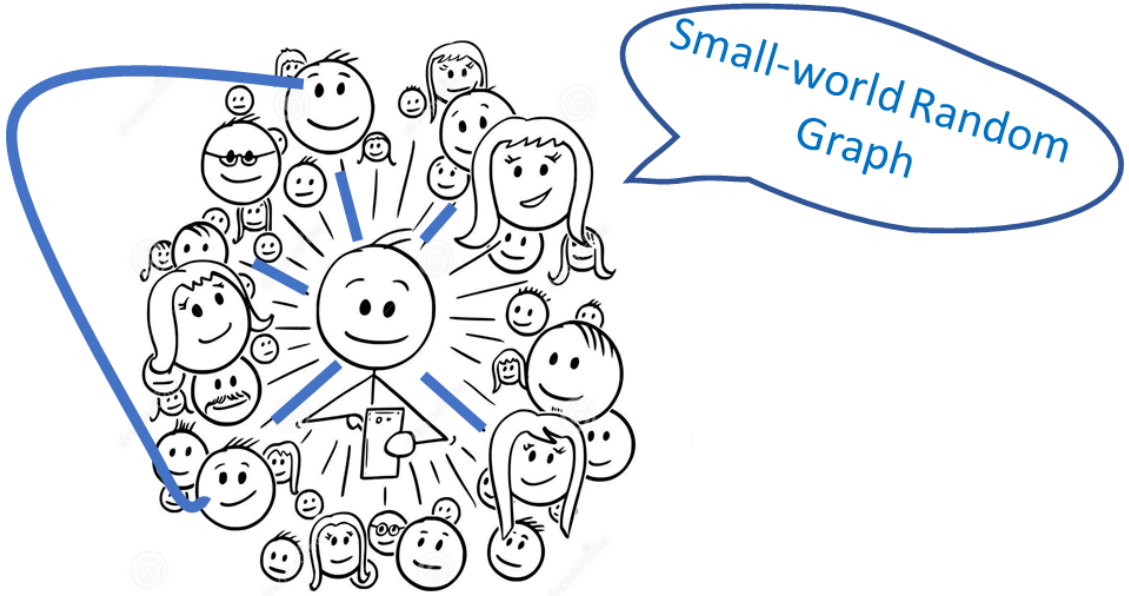


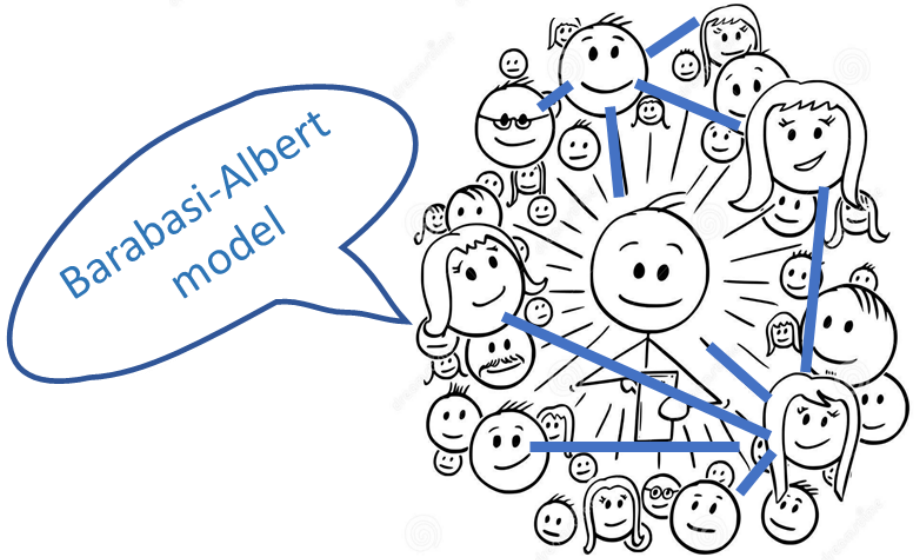
Age correction: ‘First mover advantage’ is not always true. For citation network, researchers not only like to cite the most cited ones but also have a preference towards the newly published ones

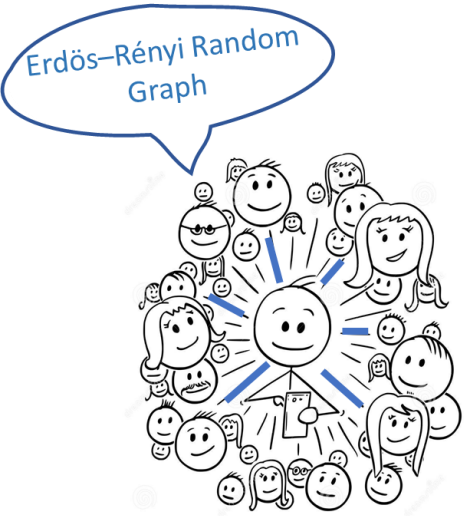
Local-world evolving network model: Considers preferential attachment , but only for existing nodes in the same area

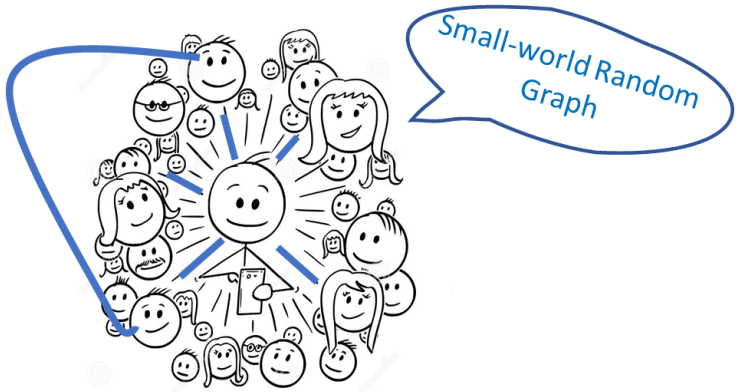
Recap: How do we form friendships and other social connections?

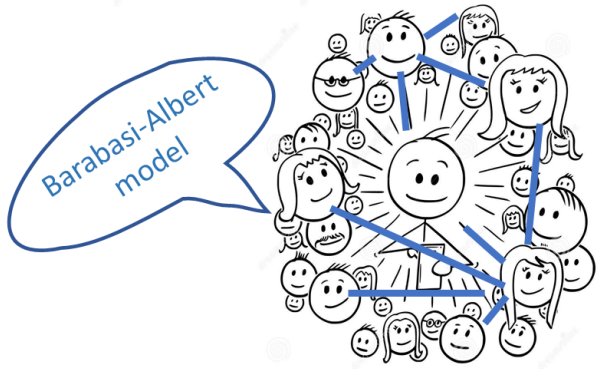


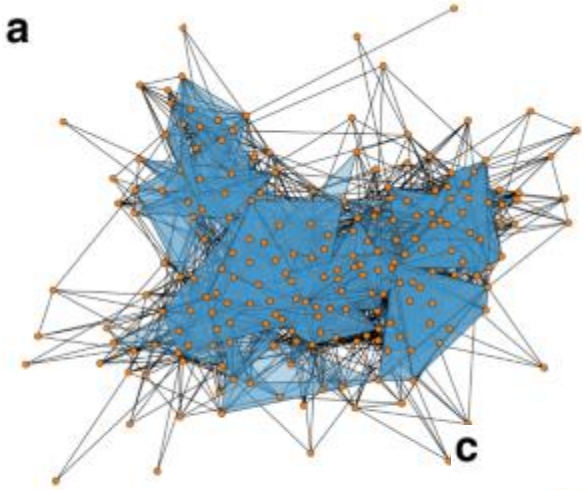


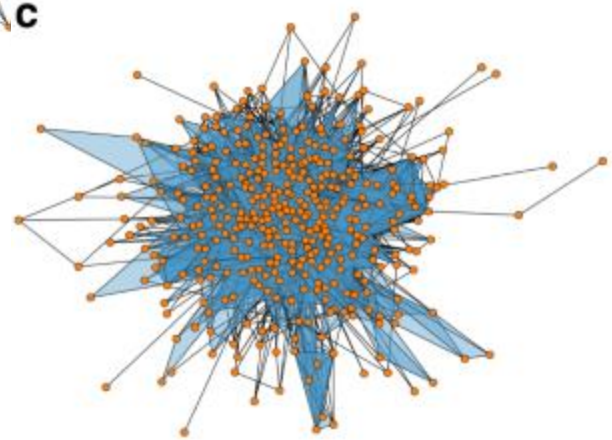












Theoretical networks

Calibration

Real-world networks

Toy (synthetic) networks to test new algorithms or solutions;
to study the effect of network typology (controlled experiment)

Questions?