Network Science

#3 Scale-free networks

MIME.

Why a network model?

Want to devise a model for generating a network that looks like a real one

- The model should capture meaningful network parameters
- The model should explain how networks emerge
- The model can be used to test our algorithms: many vs. few instances

MIME.

Random networks

Random networks

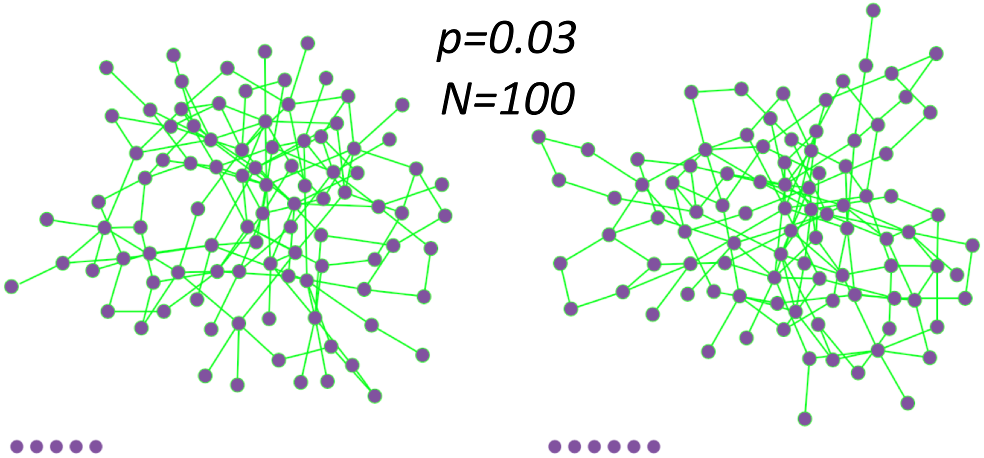
It seems that many connections arising in real networks are unpredictable

Idea:

- links are randomly generated, i.e., each link is active with probability p
- random = i.i.d. distributed

This seems sensible as we often observe unexpected links

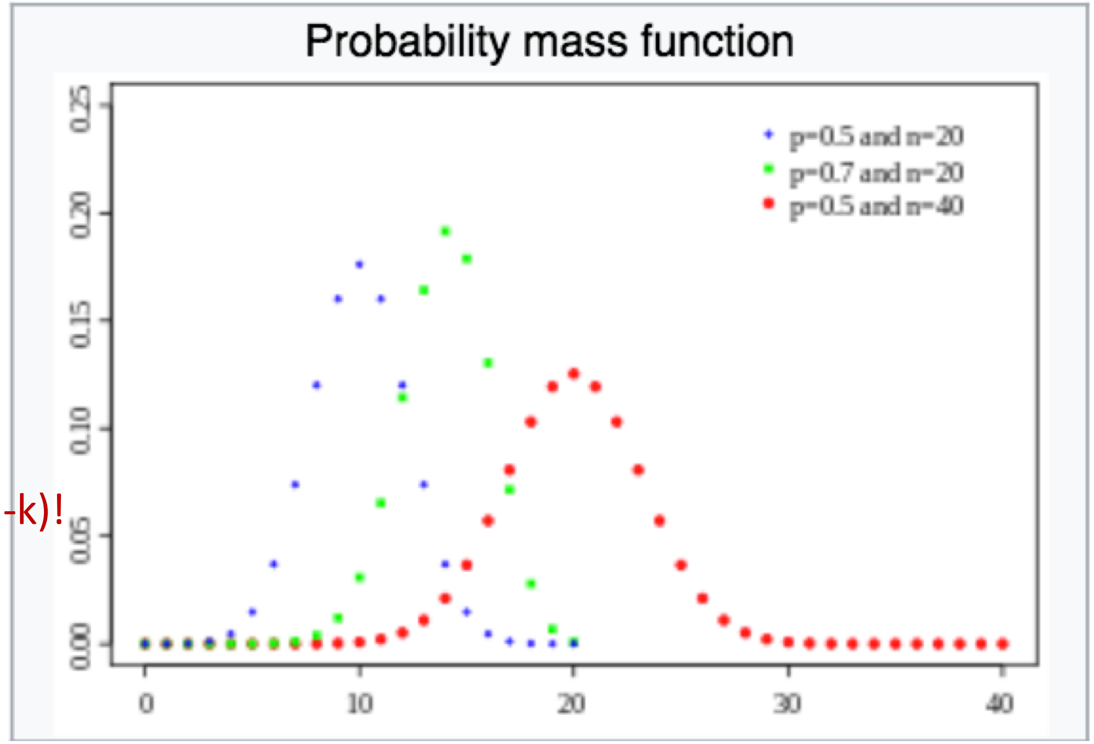
Erdös-Rényi model [1959/60]



- The random network is the simplest model:
  - pick a probability *p*, with *0<p<1*
  - activate each link (*i,j*) with probability p
- The number of links is variable
- There might be isolates
- Easy to calculate fundamental parameters

Binomial distribution

Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial $q = 1 - p$
Support	$k \in \{0,1,\ldots,n\}$ – number of successes
PMF	$\binom{n}{k} p^k q^{n-k}$ ← binomial coefficient n!/k!(n-k)!
CDF	$I_q(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	npq
Skewness	$\frac{q-p}{\sqrt{npq}}$
Ex. kurtosis	$\frac{1-6pq}{npq}$



P(k;n,p) = probability that *k* out of *n* trials are positive, where each is positive with probability *p*

Degree distribution

- The number of neighbours is binomially distributed

  $P(k;n,p)$ = probability that a node has exactly k neighbours, with number of possible neighbours $n = N-1$

- Average # of neighbours

  this defines p

  $$\langle k \rangle = (N-1)p \quad \rightarrow \quad p = \langle k \rangle / (N-1)$$

  p is usually very small (since $\langle k \rangle \ll N$)

- Variance

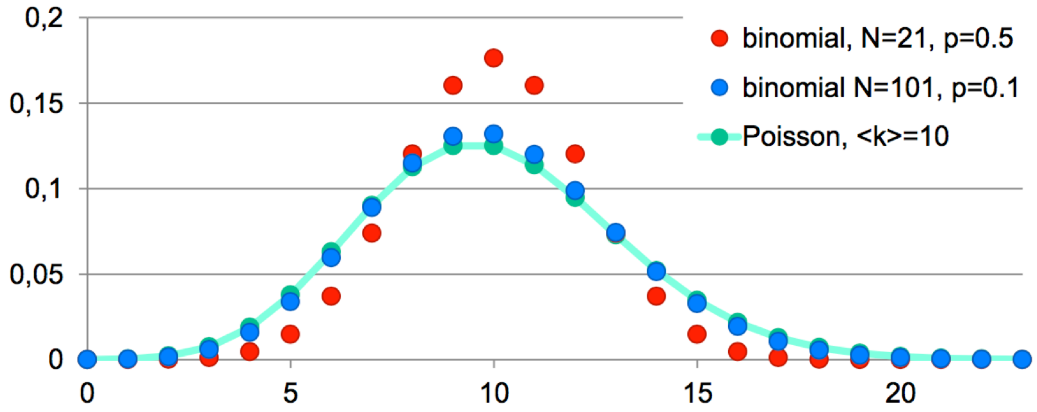
  $$\sigma_x^2 = (N-1)p(1-p) \simeq \langle k \rangle$$

  tight around the mean

Poisson approximation

- Poisson distribution (easier to use)

$$P[x = k] = \frac{m_x^k}{k!} \cdot e^{-m_x}$$

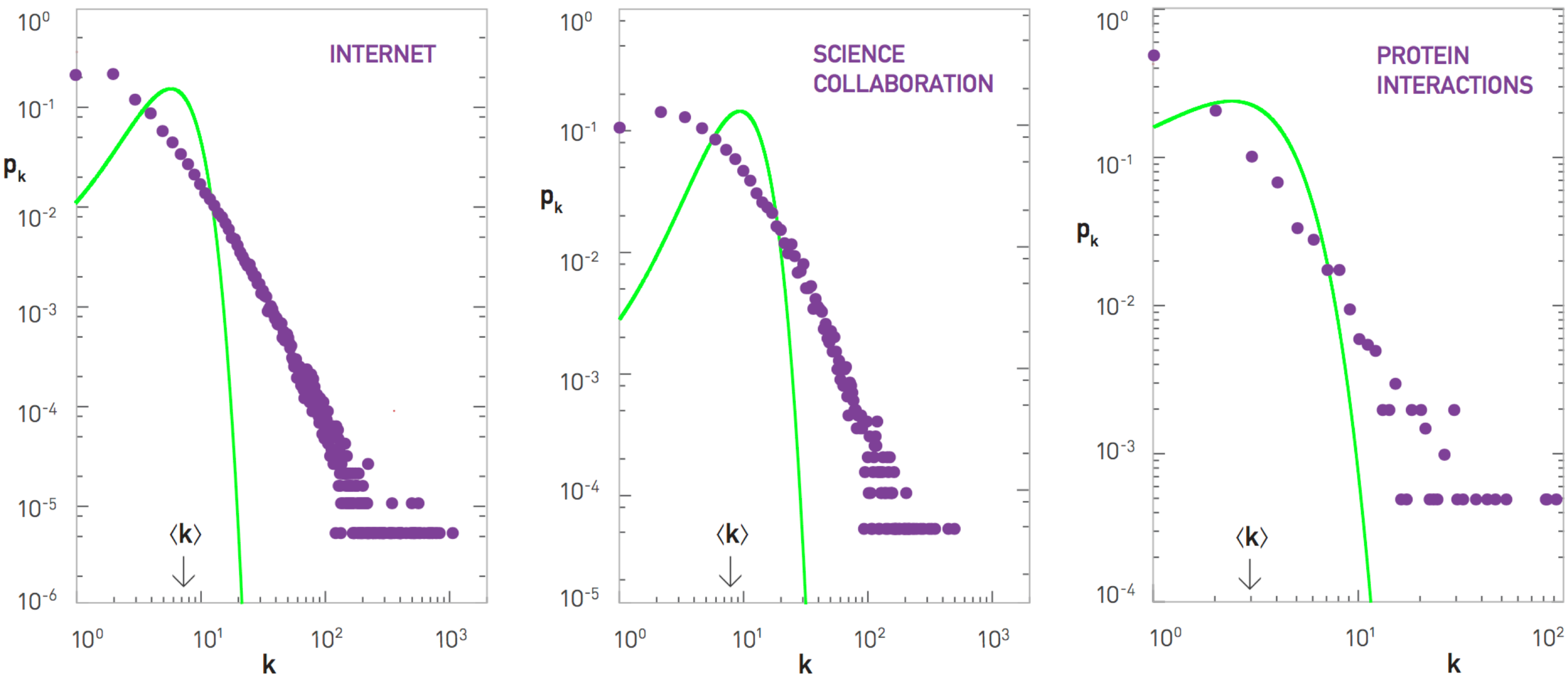


- Very good approximation of binomial for small *p* (and at small *k*)

active part   *p*

$$P[x = k] = \underbrace{\frac{(n-k+1)\dots(n-1)n}{n^k}}_{\simeq 1} \cdot \frac{m_x^k}{k!} \cdot \underbrace{\left(1 - \frac{m_x}{n}\right)^{n-k}}_{\simeq \text{const}}$$

Are real networks Poisson?

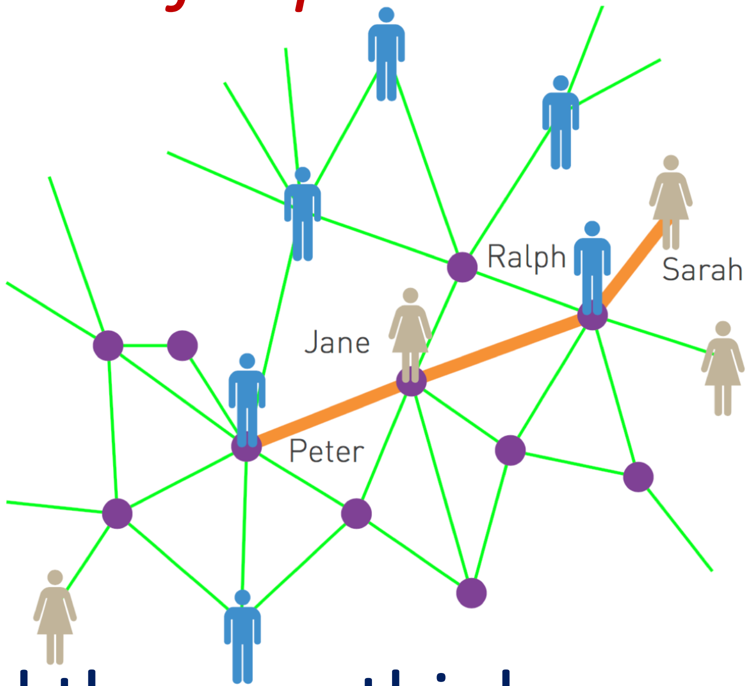


No! Poisson networks are deprived of hubs

… but, nevertheless, Poisson networks capture some aspects

Small world

- In real networks distance between two randomly chosen nodes is generally short
- Milgram [1967]: *6 degrees of separation*



- What does this mean?
  We are more connected than we think

Distances in random graphs

- we reach $\langle k \rangle$ nodes in one hop, $\langle k \rangle^2$ in two, $\langle k \rangle^3$ in three, etc.
- an estimate of the average distance $\langle d \rangle$ is found by solving for $N = \langle k \rangle^{\langle d \rangle}$ to have

$$\langle d \rangle = \ln(N)/\ln(\langle k \rangle)$$

- $\langle d \rangle$ is often taken as an estimate of the network diameter d_{max}

e.g.: on earth we are $N=7 \cdot 10^9$ individuals,
with $\langle k \rangle$=1000 acquaintances each → $\langle d \rangle = 3.28$

MIME.

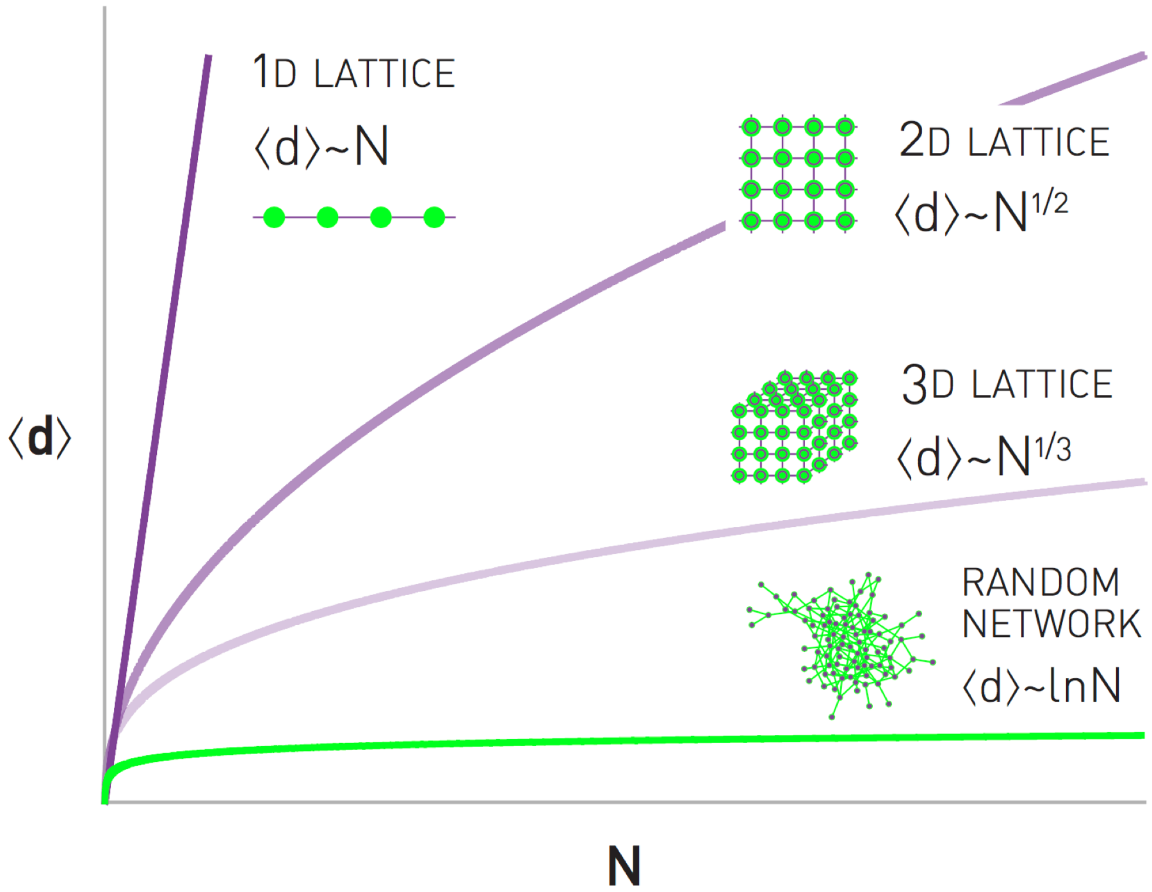
Fit in real networks

NETWORK	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\ln N}{\ln \langle k \rangle}$
Internet	192,244	609,066	6.34	6.98	26	6.58 ✓
WWW	325,729	1,497,134	4.60	11.27	93	8.31 ✓
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.72	39	11.42 ✓
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	93,439	8.08	5.35	15	4.81 ✓
Actor Network	702,388	29,397,908	83,71	3,91	14	3,04 ✓
Citation Network	449,673	4,707,958	10.43	11,21	42	5.55
E. Coli Metabolism	1,039	5,802	5.58	2.98	8	4.04
Protein Interactions	2,018	2,930	2.90	5.61	14	7.14 ✓

Very good fit ! Correct at least as order of magnitude

MIME.

What about structured networks?



No fit !

Are real networks random?

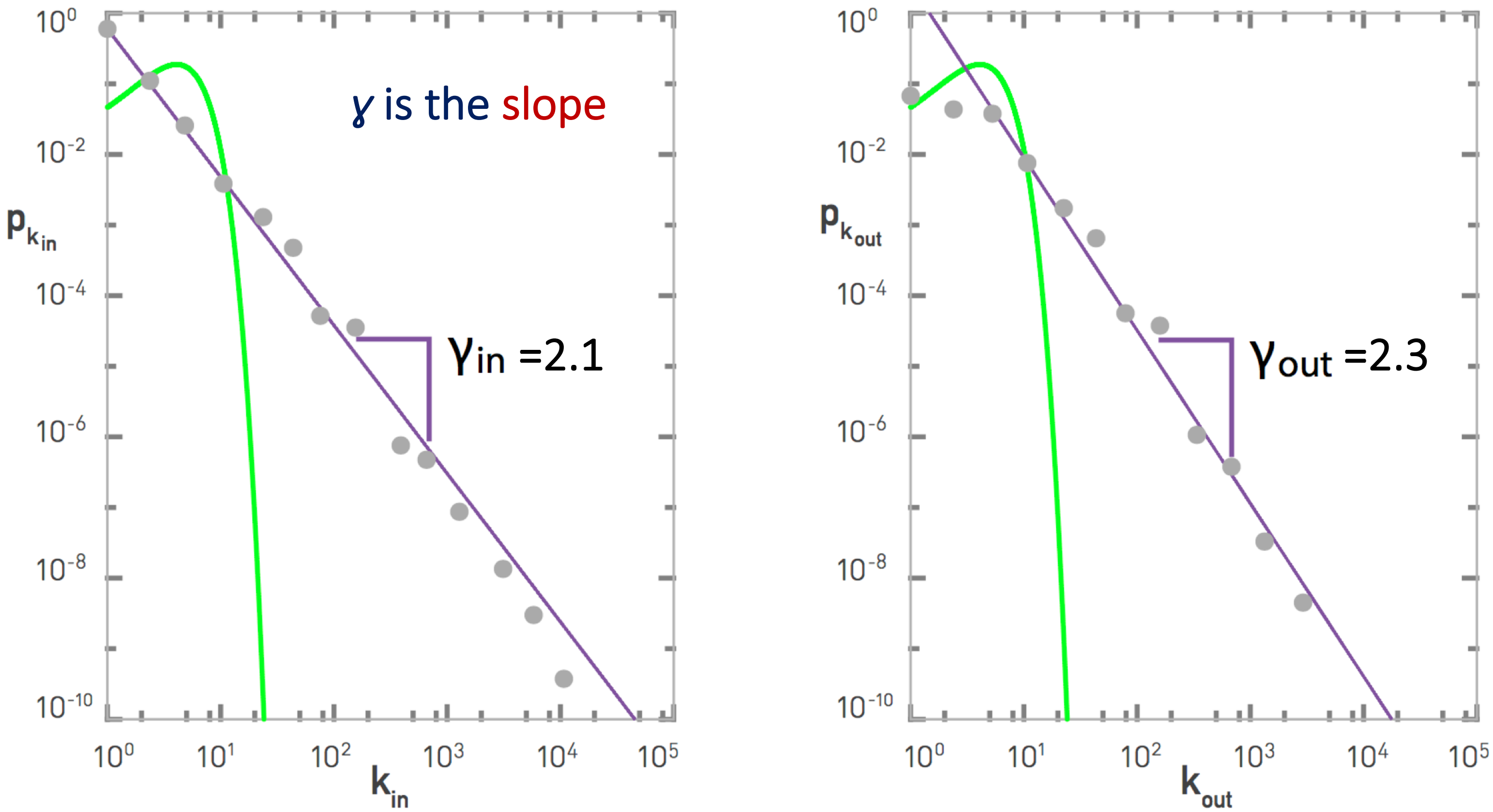
- Random networks are generally NOT a good model for real world scenarios, but
  - ✓ Are easy to describe and generate
  - ✓ Can make some correct predictions
  - ✓ Can serve as a general reference model (in the sense that a model is good if it deviates from the random model as much as real networks do)

MIME.

Scale-free networks

Scale free networks

The degree distribution of the www (Albert *et al.,* 1999)

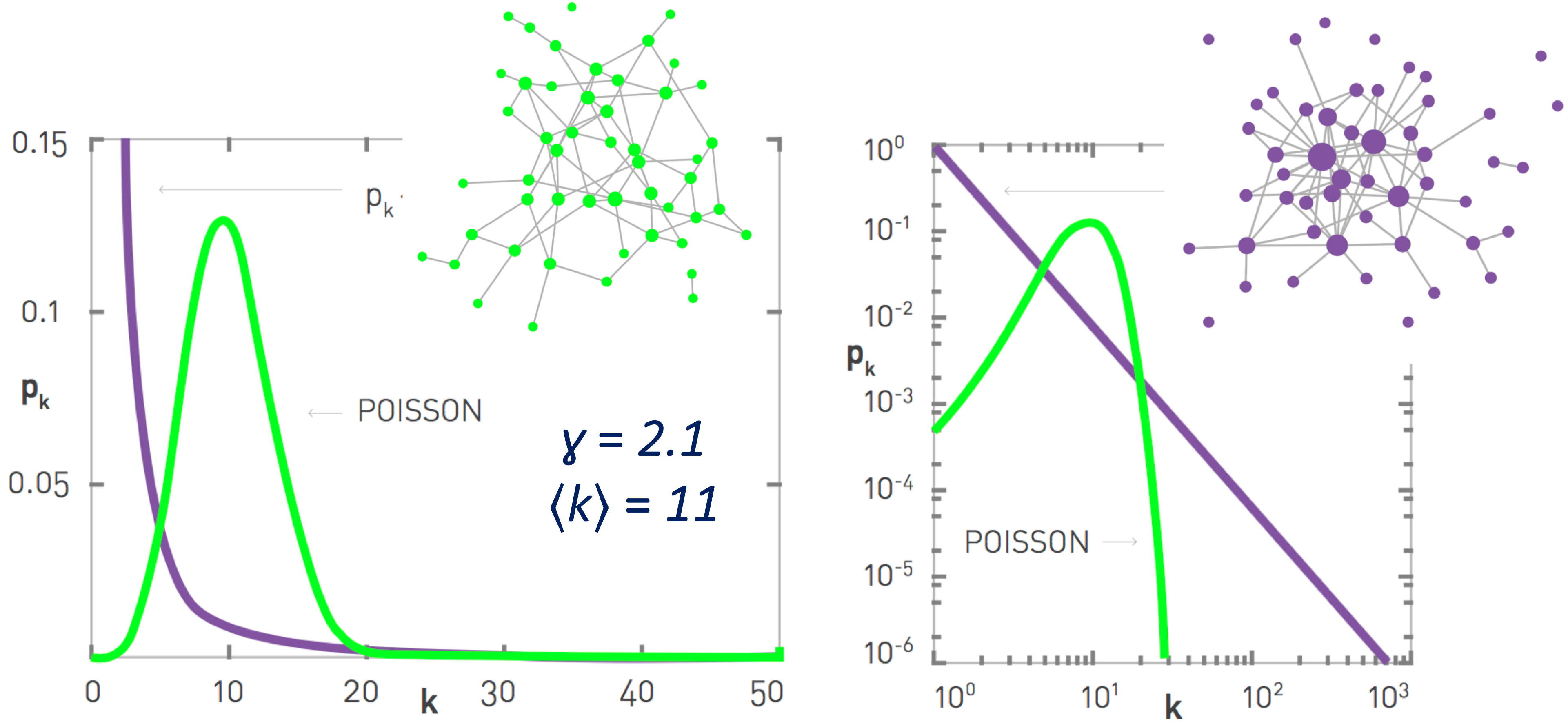


they follow a power-law

$\ln(p_k) = c - \gamma \cdot \ln(k)$ → $p_k = C \cdot k^{-\gamma}$

MIME.

Poisson versus power-law



Power-law is heavy tailed (presence of hubs) - like Weibull, lognormal, Lévy

MIME

The value of γ in real networks

NETWORK	N	L	$\langle k \rangle$	γ_{in}	γ_{out}	γ
Internet	192,244	609,066	6.34	-	-	3.42*
WWW	325,729	1,497,134	4.60	2.00	2.31	-
Power Grid	4,941	6,594	2.67	-	-	Exp.
Mobile Phone Calls	36,595	91,826	2.51	4.69*	5.01*	-
Email	57,194	103,731	1.81	3.43*	2.03*	-
Science Collaboration	23,133	93,439	8.08	-	-	3.35*
Actor Network	702,388	29,397,908	83.71	-	-	2.12*
Citation Network	449,673	4,689,479	10.43	3.03**	4.00*	-
E. Coli Metabolism	1,039	5,802	5.58	2.43*	2.9 0*	-
Protein Interactions	2,018	2,930	2.9 0	-	-	2.89*

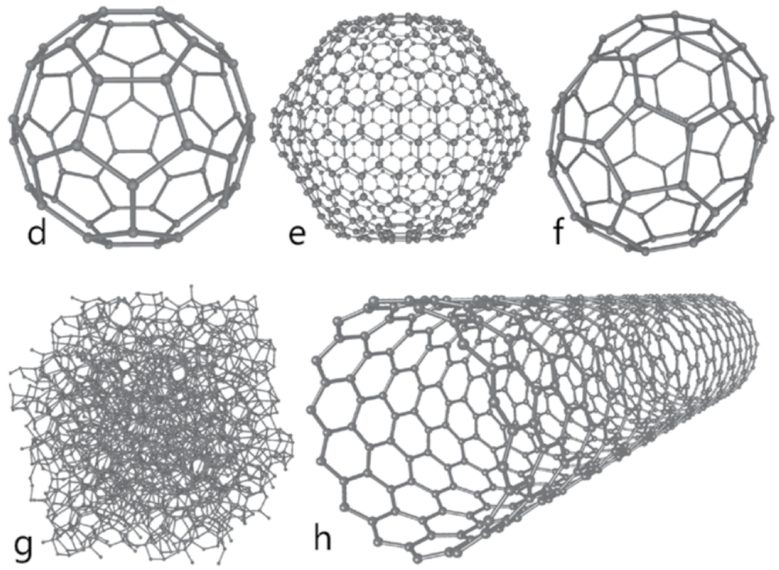
$\gamma \in [2,4]$, but the statistical fit is not always good (** = good fit with an exponential)

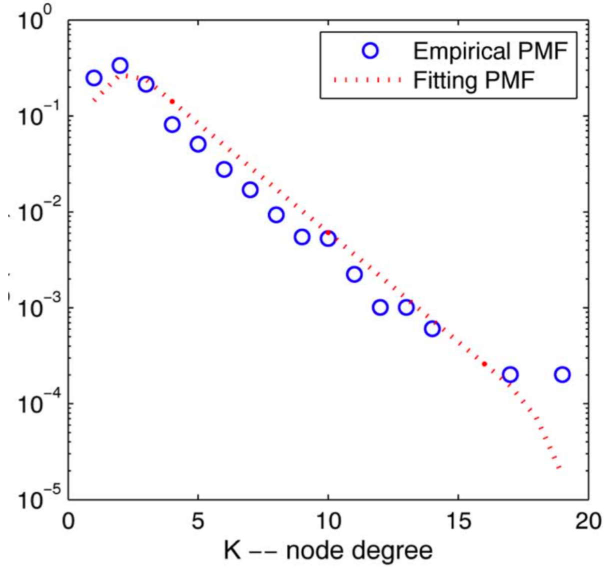
MIME.

Is everything scale-free?

Not all networks are scale-free, e.g.:

- Networks appearing in material science, where all nodes have the same degree
- The neural network of the *C.elegans* worm
- The power grid, consisting of generators and switches connected by power lines





Scale-free networks

- A scale free network is a network whose degree distribution follows a power law $p_k = C \cdot k^{-\gamma}$

- It is meaningful in an interval $[k_{min}, k_{max}]$

- The parameter γ (slope) is called the exponent

- C is determined by the total normalization condition $\sum_{k=k_{min}}^{k_{max}} p_k = 1$ which we approximate to

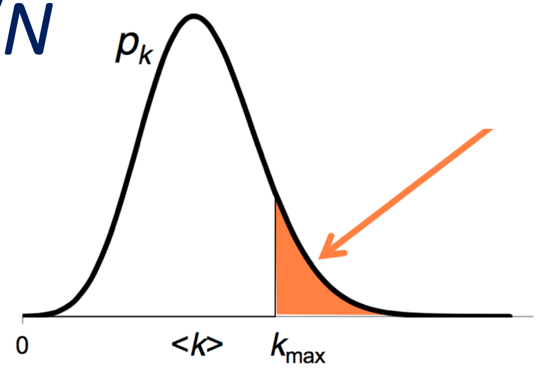
$$\int_{k_{min}}^{\infty} p_k \, dk = C \cdot k_{min}^{-(\gamma-1)} / (\gamma - 1) = 1$$

MIME.

The largest hub

The size of the largest hub is captured by

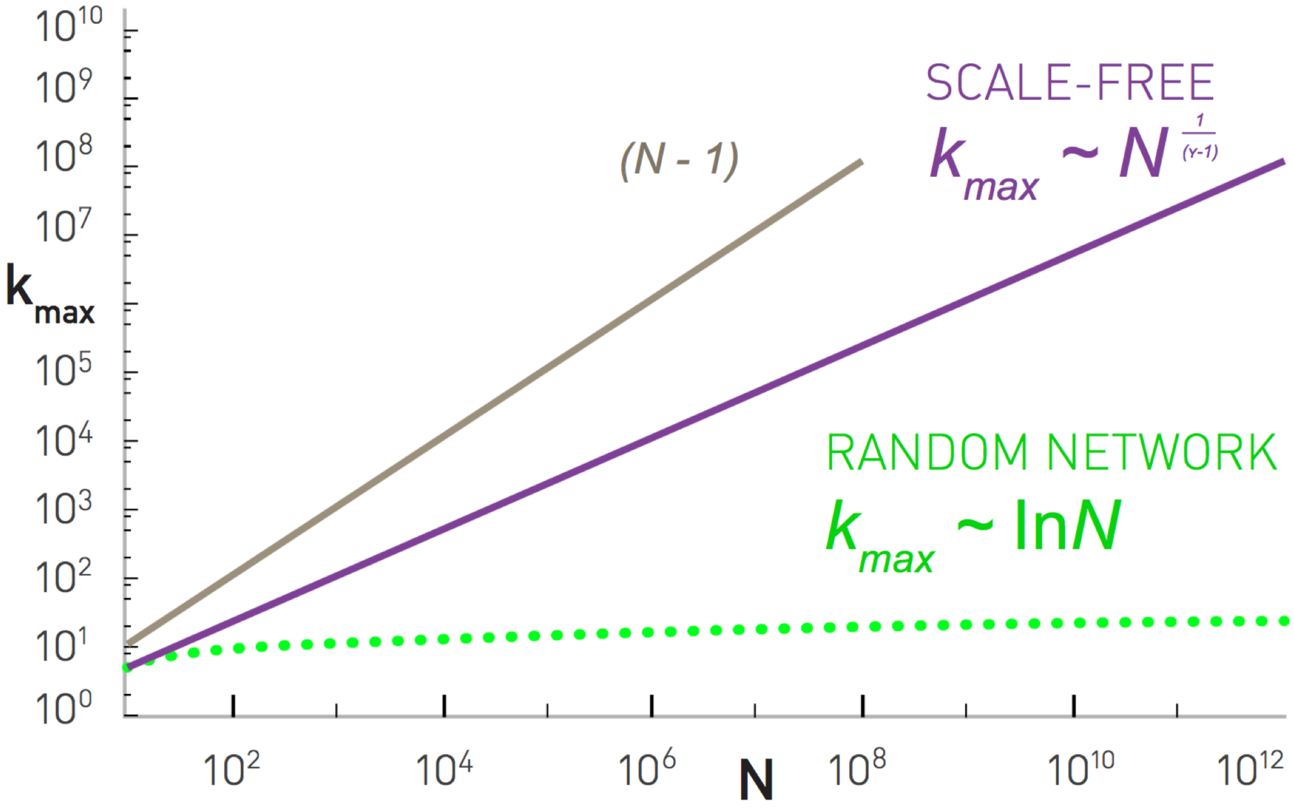
$$\int_{k_{max}}^{\infty} p_k \, dk = C \cdot k_{max}^{-(\gamma-1)} / (\gamma - 1) = 1/N$$



$$k_{max} = k_{min} N^{1/(\gamma-1)}$$

is the natural cutoff

Hubs are large in scale-free networks



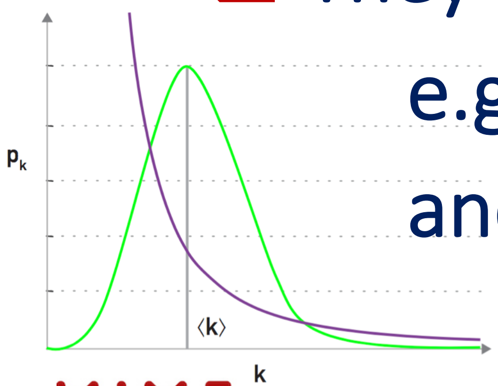
The highest degree increases in *N* polynomially (sub-linearly) fast = big hubs

Moments of scale-free networks

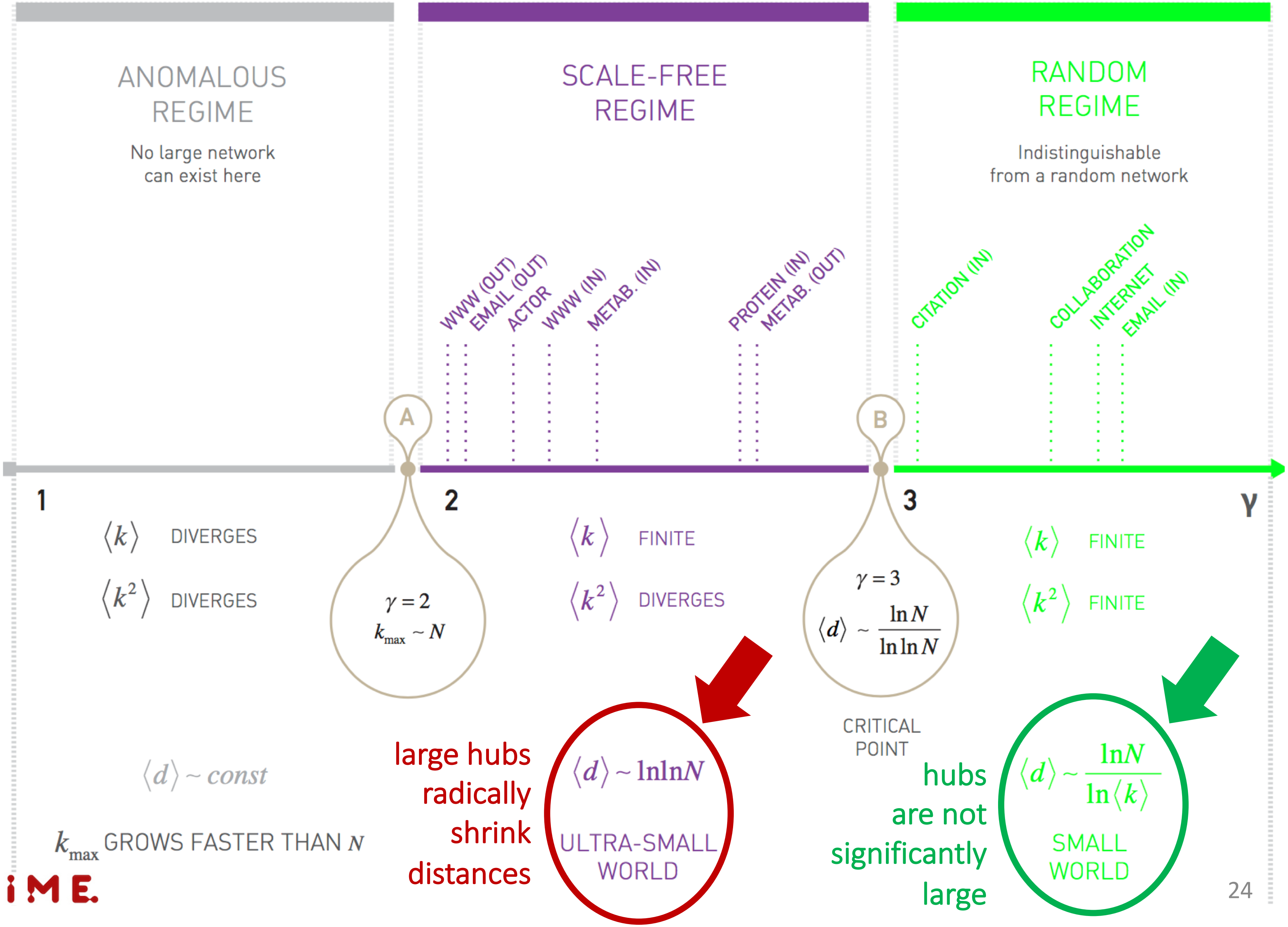
Moments of the power law $p_k = C \cdot k^{-\gamma}$

- $\langle k^n \rangle = \int_{k_{min}}^{k_{max}} k^n \, p_k \, dk$
  $= C \cdot (k_{max}^{n-\gamma+1} - k_{min}^{n-\gamma+1})/(n-\gamma+1)$
  $= C \, k_{min}^{n-\gamma+1} \cdot (N^{n/(\gamma-1)-1} - 1) / (n-\gamma+1)$

- They diverge with N if $\gamma < n+1$
  e.g. variance ($n=2$) diverges for $\gamma \in [2,3]$
  and the network does not have a scale



The scale-free regime

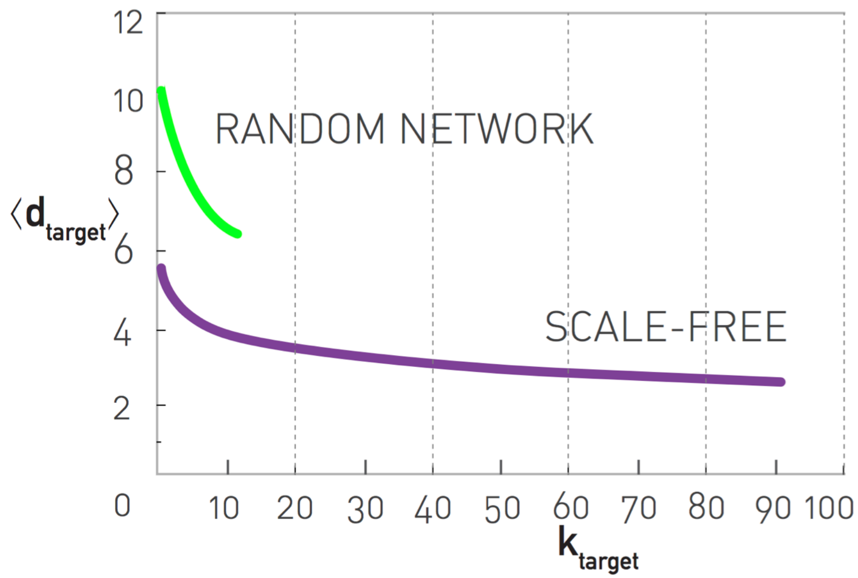


Ultra-small-world, 2<ɣ<3

- The average distance increases as ln(ln(*N*)), much slower than *N* or ln(*N*)

e.g. in www $N=7\cdot10^9$, ln(*N*)=*22.7*, ln(ln(*N*))=*3.12* (very small)

distance of a node of degree ⟨*k*⟩ from a node of degree k_{target}



- The large hubs radically shrink the distance between nodes → ultra small world

MIME.

Curiosity

In many social experiments people avoided hubs for entirely perceptual reasons (e.g., they assumed they are busy, better use them only if really needed)

We live in a ultra-small-world, but we perceive that we are more distant from others than we really are!

Friendship paradox

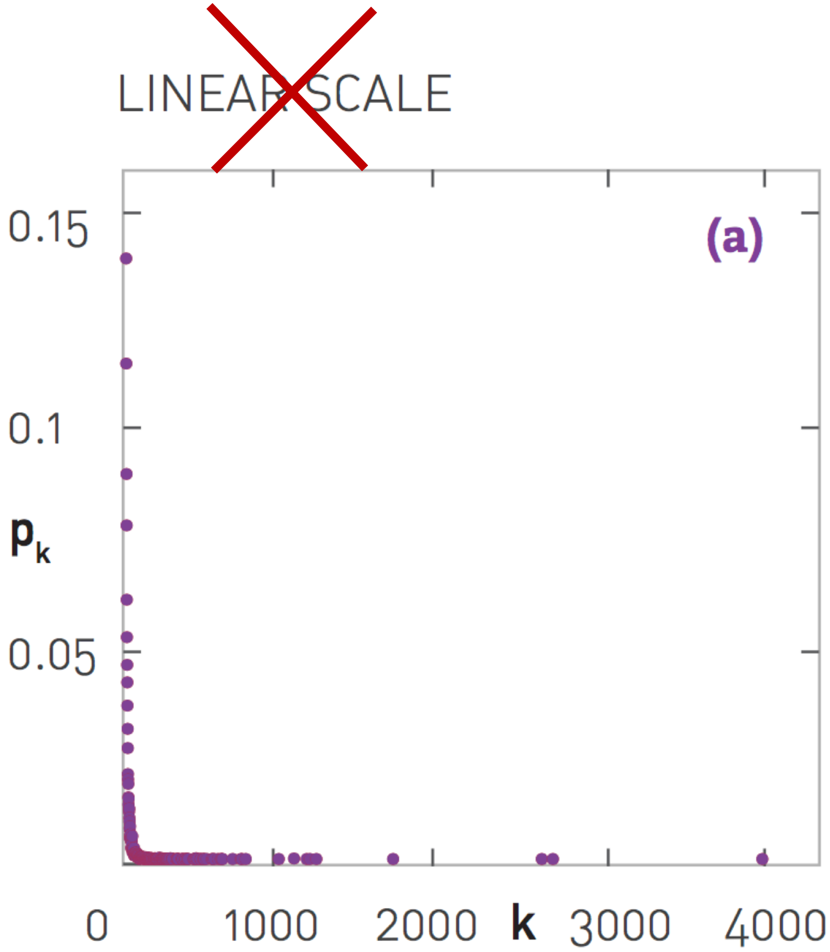
My friends are more popular than me! ☹

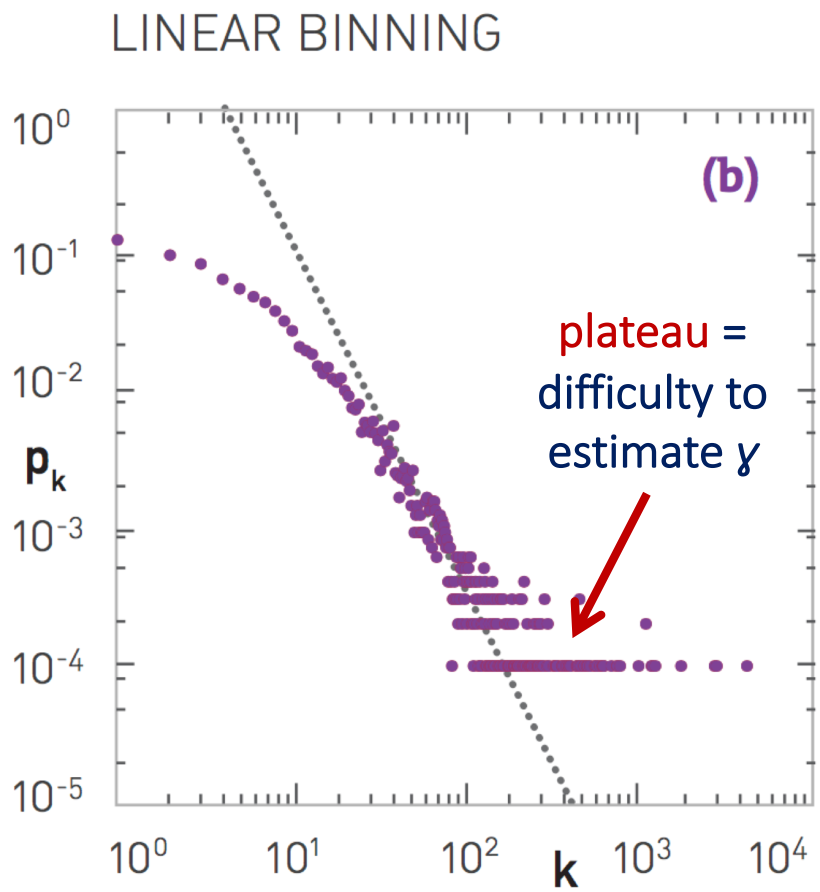
(Feld, 1991)

- Can be observed in the ultra-small-world under the presence of big hubs
- Rationale: a node is very likely to be connected to a big hub, having a very large number of connections
- # of friends (in the average) = $\langle k \rangle$
- # of friends of friends $\simeq N$

Estimating the exponent

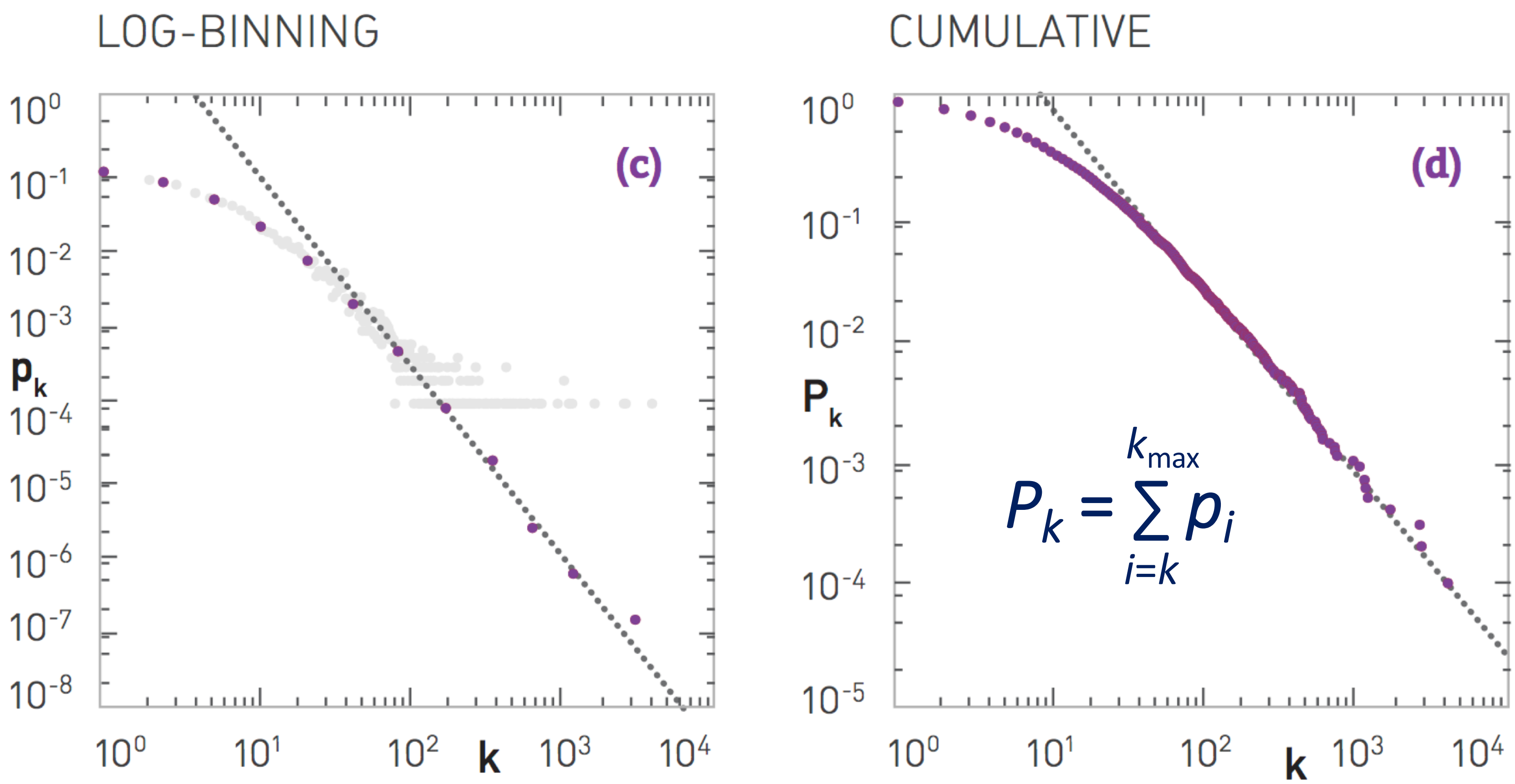
Plotting power laws





Better use a log-log scale

Plotting power laws (cont'd)



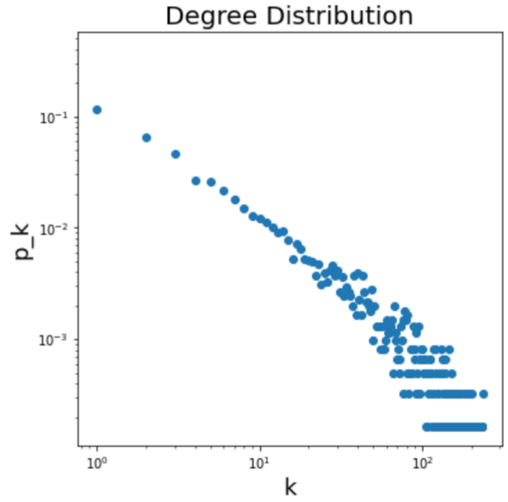
Even better: log-binning or complementary cumulative distr. function (CCDF) $P_k \sim k^{-(\gamma-1)}$

Pseudocode example

```
G = np.loadtxt('Wiki-Vote.txt').astype(int)

# adjacency matrix
N = np.max(G)
A = csr_matrix((np.ones(len(G)), (G[:, 1], G[:, 0])))


#distribution
which_deg = 0 # 0=out degree, 1=in degree
d = np.sum(A, which_deg) # out degree for each node
d = np.squeeze(np.asarray(d)) # from matrix to array
d = d[d>0] # avoid zero degree
k = np.unique(d) # degree samples
pk = np.histogram(d, k)[0] # occurrence of each degree
pk = pk/np.sum(pk) # normalize to 1
Pk = 1 - np.cumsum(pk) # complementary cumulative
```




```
fig = plt.figure()
plt.loglog(pk, 'o')
plt.title("Degree Distribution", size = 20)
plt.xlabel("k", size = 18)
plt.ylabel("p_k", size = 18)
plt.show()
```

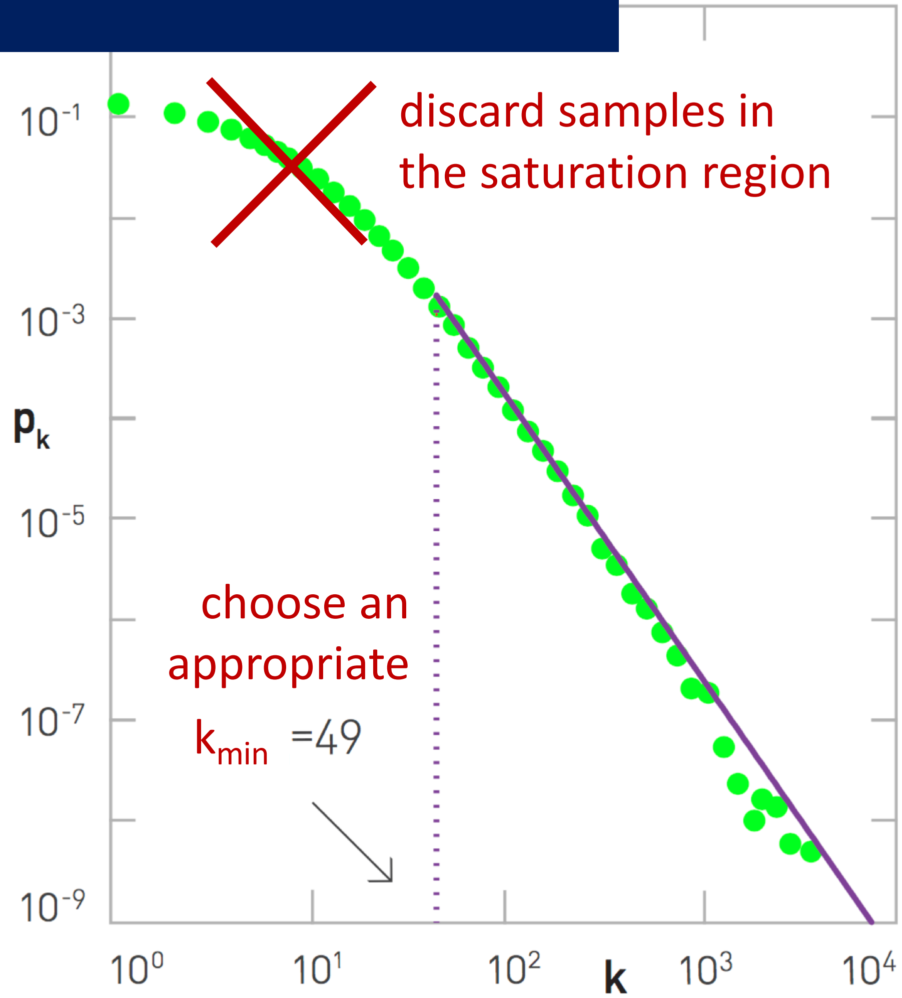
Estimating γ

Fit γ only in a wisely chosen interval

- $p_k = C\, k^{-\gamma}$
- $C = (\gamma-1)\, k_{min}^{\gamma-1}$

is determined by the normalization condition

$$\int_{k_{min}}^{\infty} p_k\, dk = C \cdot k_{min}^{-(\gamma-1)} / (\gamma-1) = 1$$



ML estimate for γ

- Target PDF $p(k|\gamma) = (\gamma\text{-}1)/k_{min} \cdot (k/k_{min})^{-\gamma}$
- ML criterion (k_i is the measured degree of node i)

$$\max_\gamma f(\gamma) = \sum_i \ln p(k_i|\gamma)$$

- $f(\gamma) = \sum_i \ln((\gamma\text{-}1)/k_{min}) - \gamma \ln(k_i/k_{min})$
- Solve $f'(\gamma) = \sum_i 1/(\gamma\text{-}1) - \ln(k_i/k_{min}) = 0$
- The result is

$$\gamma = 1 + \sum_i 1 / \sum_i \ln(k_i/k_{min})$$

Pseudocode example

```
which_deg = 1; % 1 = out degree, 2 = in degree
d = full(sum(A,which_deg));
d2 = d(d>=kmin); % restrict range
ga = 1+1/mean(log(d2/kmin)); % estimate the exponent
```

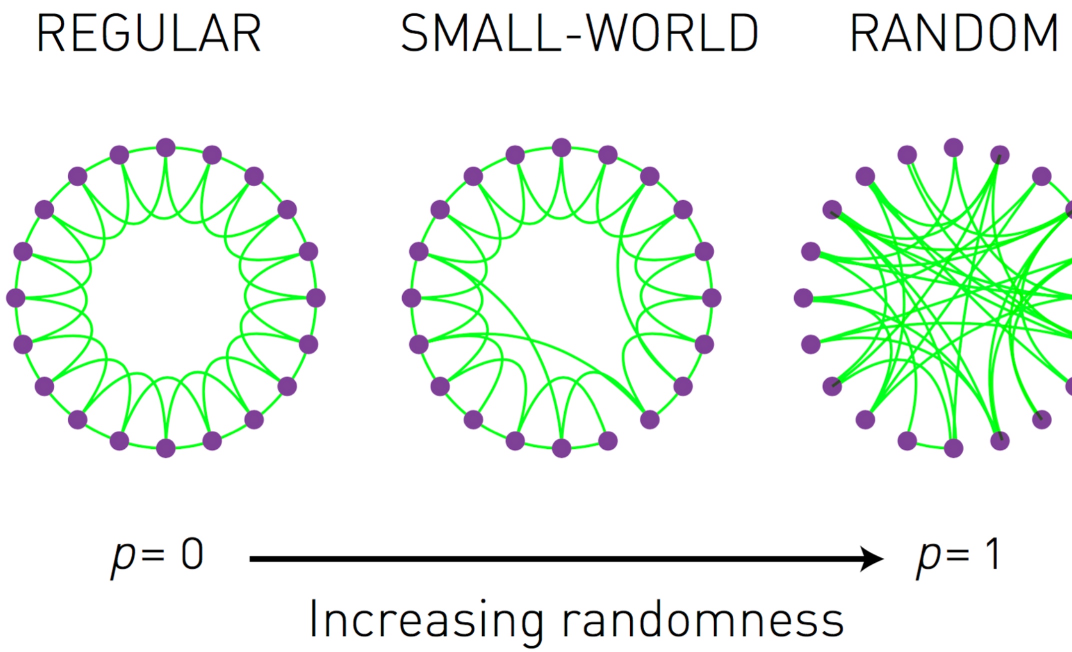
Other network models

Watts-Strogatz model [1998]

- It is the small-world model
  - ✓ Generalises the random network
  - ✓ It stresses the small-world property ☺
  - ✓ But predicts a Poisson like degree distribution ☹

MIME.

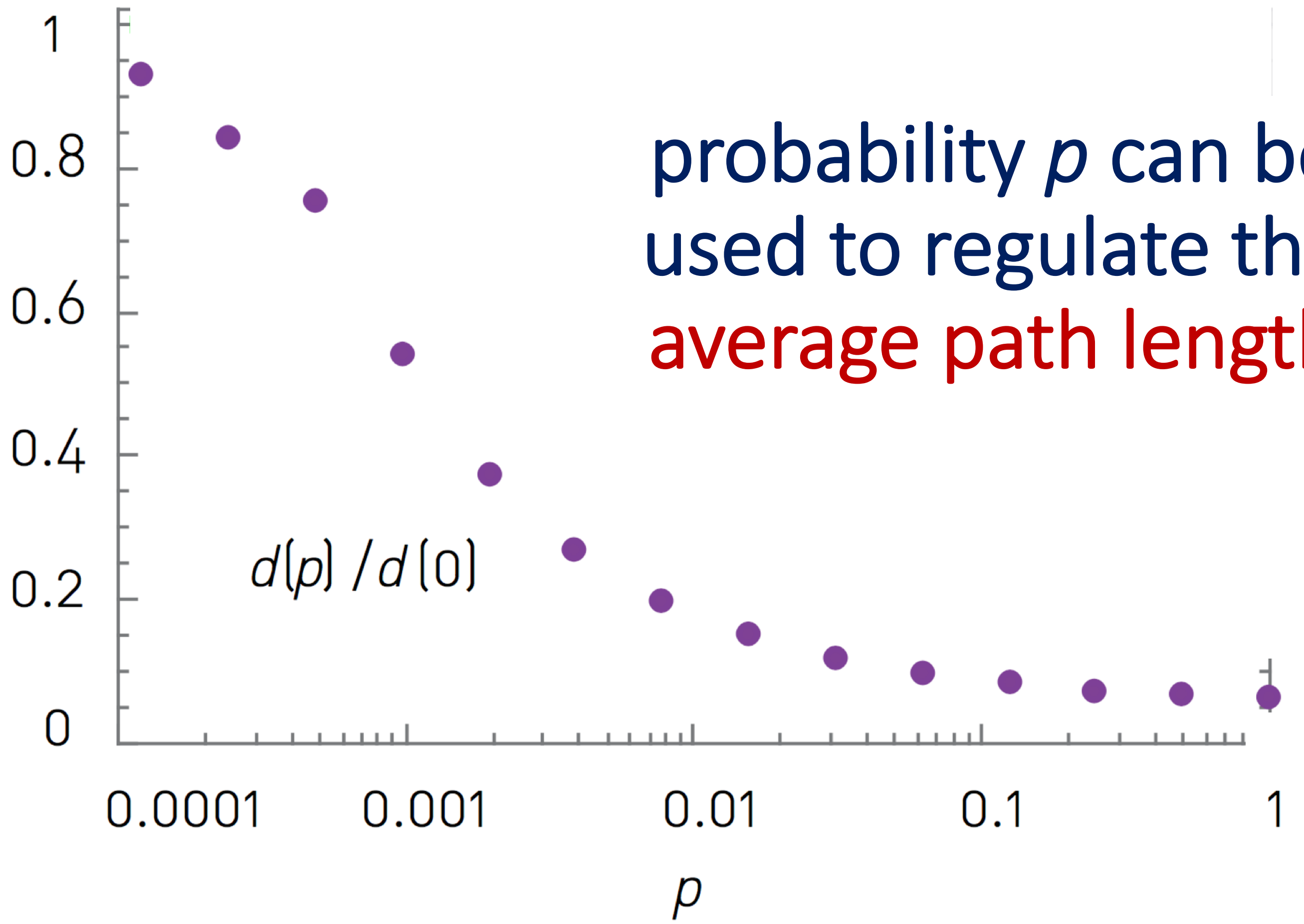
Watts-Strogatz model (cont'd)



- Start from a regular lattice
- Each node in the circle is connected to 2 neighbours on the right, and 2 on the left
- Rewire at random with probability *p*

MIME

Watts-Strogatz model (cont'd)

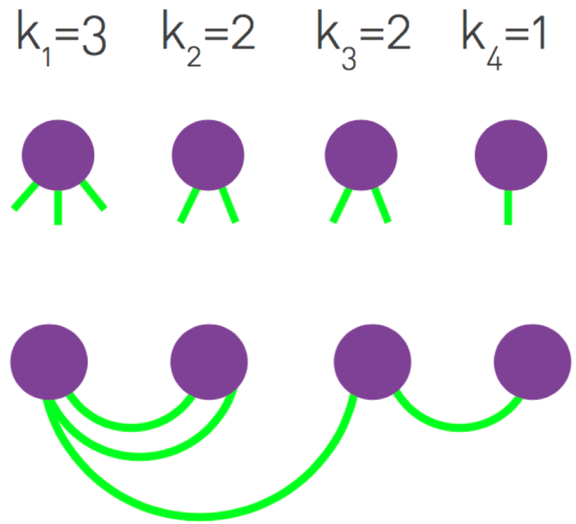


probability *p* can be used to regulate the **average path length**

Molloy-Reed model [1995]

1. assign a degree to each node
   make sure that the # of stubs is even, *2L*

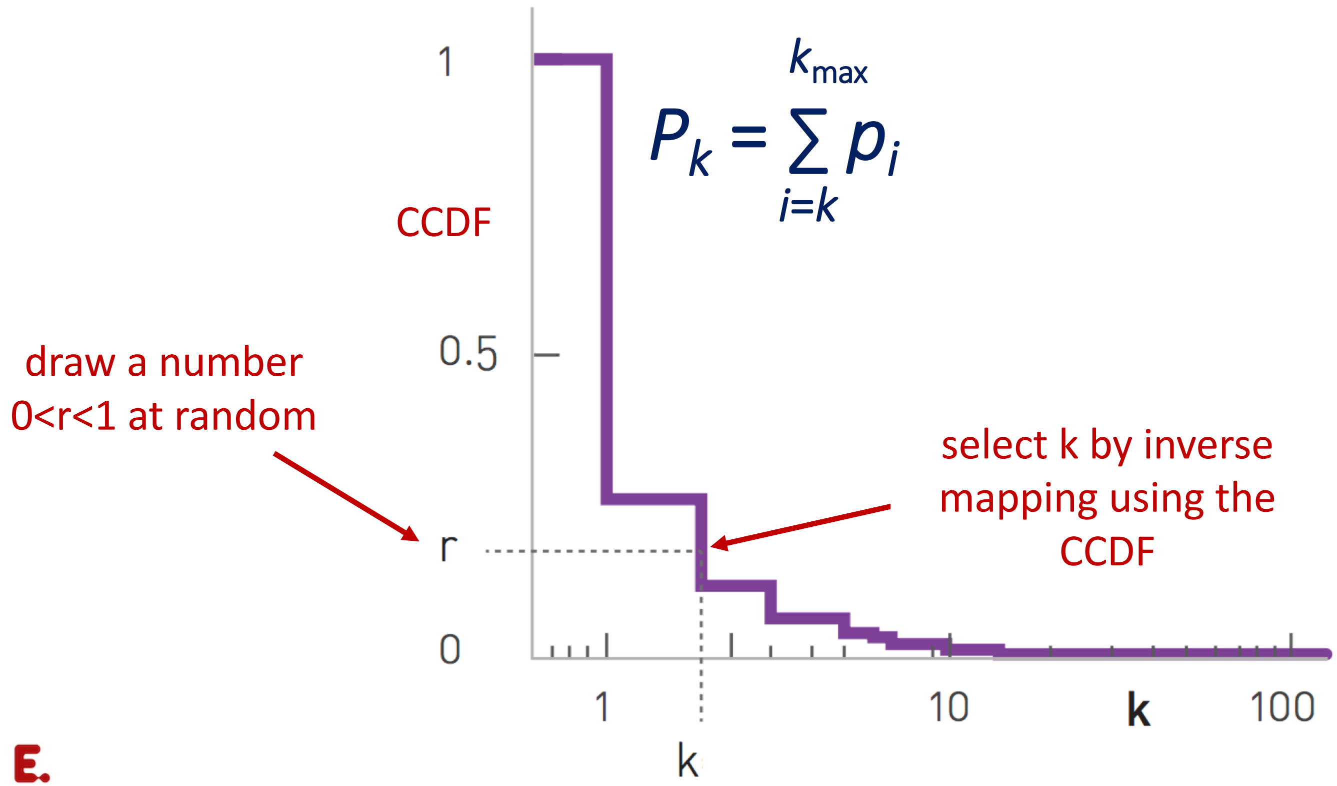
2. rewire stubs at random



- The resulting graph may contain cycles and multiple links (but are a few)
- ☹ cannot control features other than p_k

Molloy-Reed model [1995]

- How to generate degrees with given distribution p_k?
- Use the inverse CCDF method



Generalizations

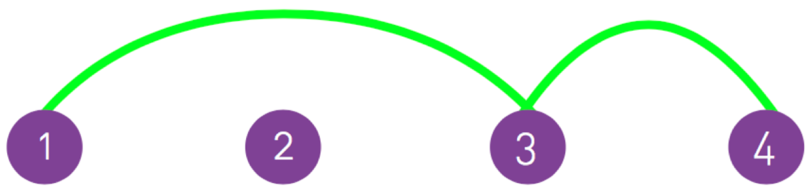
- Degree-preserving rewiring can help in reducing self/multiple-loops:
  1. select two links (T_1,S_1) and (T_2,S_2)
  2. rewire them as (T_1,S_2) and (T_2,S_1) = flip

- Can also be applied to directed networks by separately generating k_{in} and k_{out} stubs

Chung-Lu model

The hidden parameter model:

- activates a link with probability $p_{ij} = k_i k_j / \langle k \rangle N$



- prevents from having multi/self-loops
- ☹ has very high complexity $\propto N^2$
- ☹ cannot control features other than p_k
- can be extended to directed networks by considering $p_{i \to j} = k_i^{out} k_j^{in} / L$

Readings

- A.L. Barabási, Network science

  http://barabasi.com/networksciencebook

  Ch.3 “Random networks”

  Ch.4 “The scale-free property”