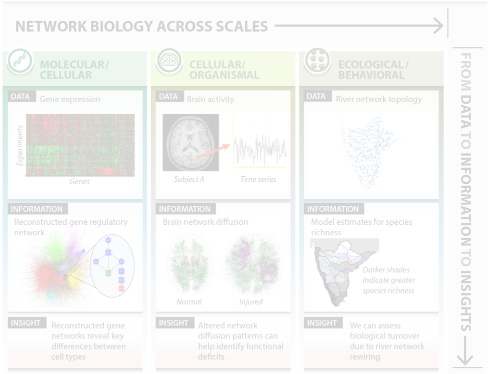

NETWORK BIOLOGY ACROSS SCALES
MOLECULAR/ CELLULAR
CELLULAR/ ORGANISMAL
ECOLOGICAL/ BEHAVIORAL
FROM DATA TO INFORMATION TO INSIGHTS
DATA
Gene expression
Experiments
Genes
Brain activity
Subject A
Time series
River network topology
INFORMATION
Reconstructed gene regulatory network
Brain network diffusion
Normal
Injured
Model estimates for species richness
Darker shades indicate greater species richness
INSIGHT
Reconstructed gene networks reveal key differences between cell types
Altered network diffusion patterns can help identify functional deficits
We can assess biological turnover due to river network rewiring

Random Networks

Héctor Corrada Bravo

University of Maryland, College Park, USA

CMSC828O 2019-09-23

The core of data analysis

Given observations (entities), and their characteristics (attributes)

Understand how those characteristics are *distributed* in the population

- means, variances, empirical distributions
- data generative models with parameters

Hope we can infer something useful about that population from the understanding of the distribution of characteristics

Thinking of networks in terms of data analysis

One way of thinking about networks:

- Entities and their characteristics (nodes and their attributes)
- Interactions and their characteristics (edges and their attributes)

How is the *presence or absence of interactions* distributed in the population?

Presence/absence models

Consider the analysis of a binary attribute in the population.

E.g., is gene PTEN expressed in normal breast cells?

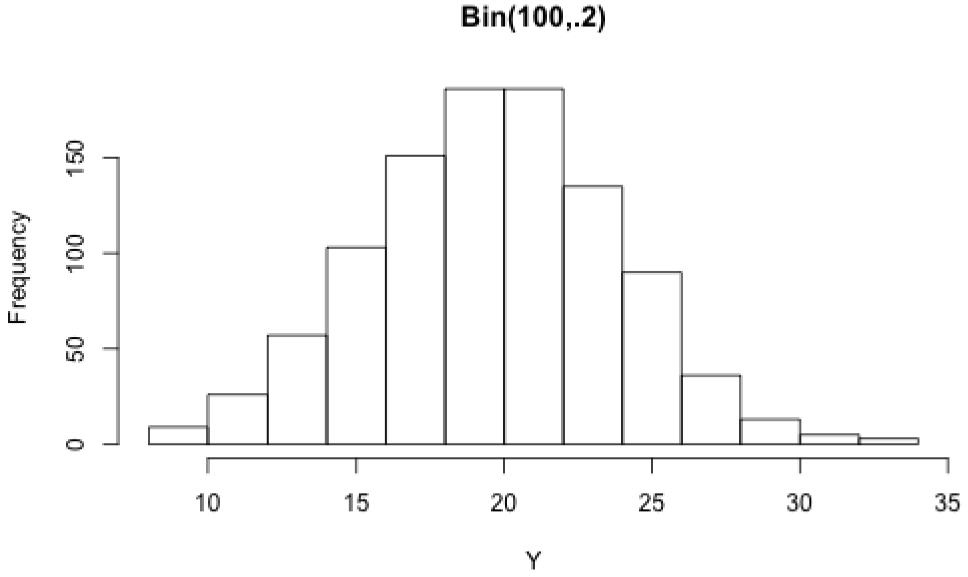
What is a model we tend to use? Suppose the gene is expressed in p cells in the population.

I take a sample of N cells, measure how many express PTEN.

Presence/absence models

If we do this sampling experiment many times, we expect to see Binomial distribution

$$Y \sim Bin(N, p)$$



Presence/absence models

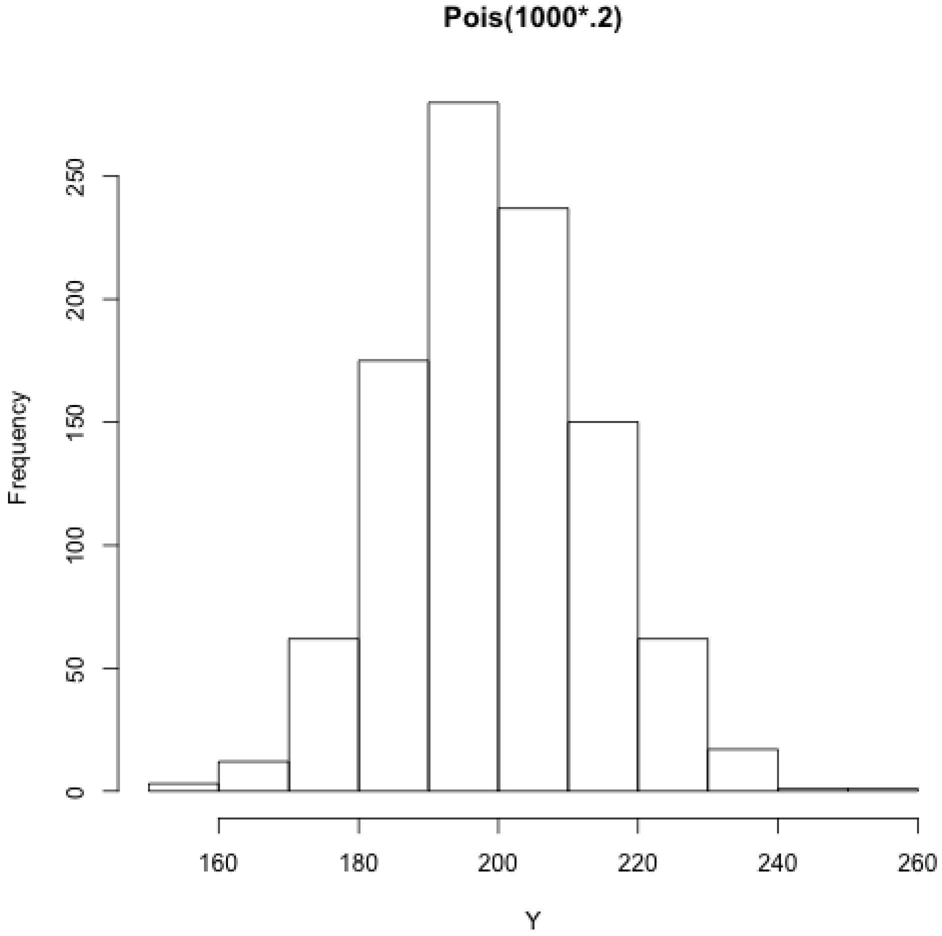
What is our expectation for Y? Variance?

Presence/absence models

What is our expectation for Y? Variance?

For very large N, same distribution is well approximated by $Y \sim \text{Poisson}(\lambda = pN)$

Presence/absence models



Presence/absence models

We can follow a similar idea to model presence/absence of an *interaction* in the network

Given a network over N entities (vertices)

What is the population of interest? All possible pairwise interactions

Presence/absence models

Assumptions:

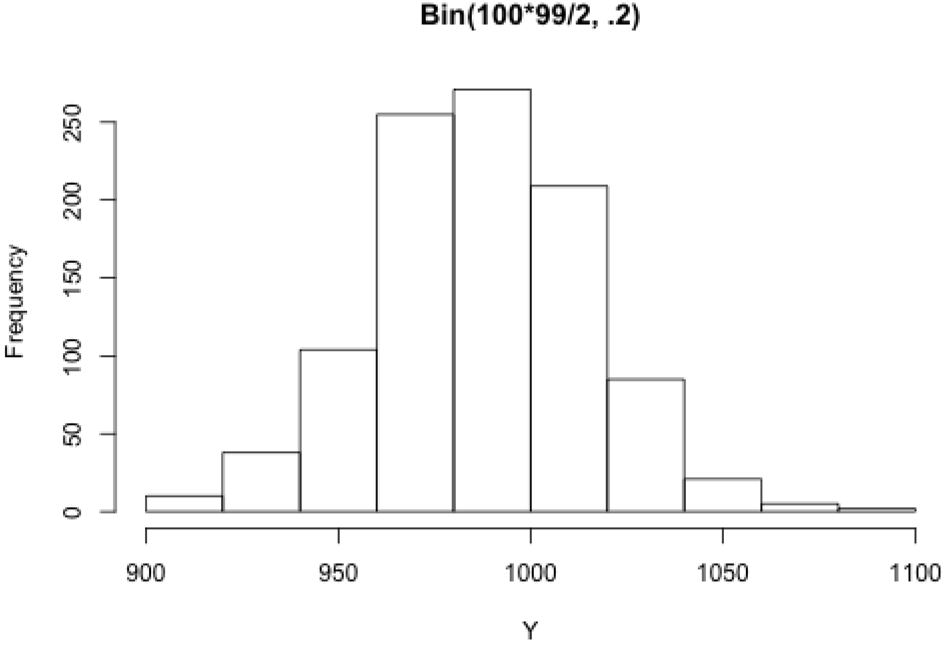
Fraction p of interactions actually occur in the population

The observed edges is a *realization* of random occurrences in measurement of those interactions

How can we model the number of edges in a network?

Presence/absence models

$Y \sim Bin(N(N-1/2),p)$



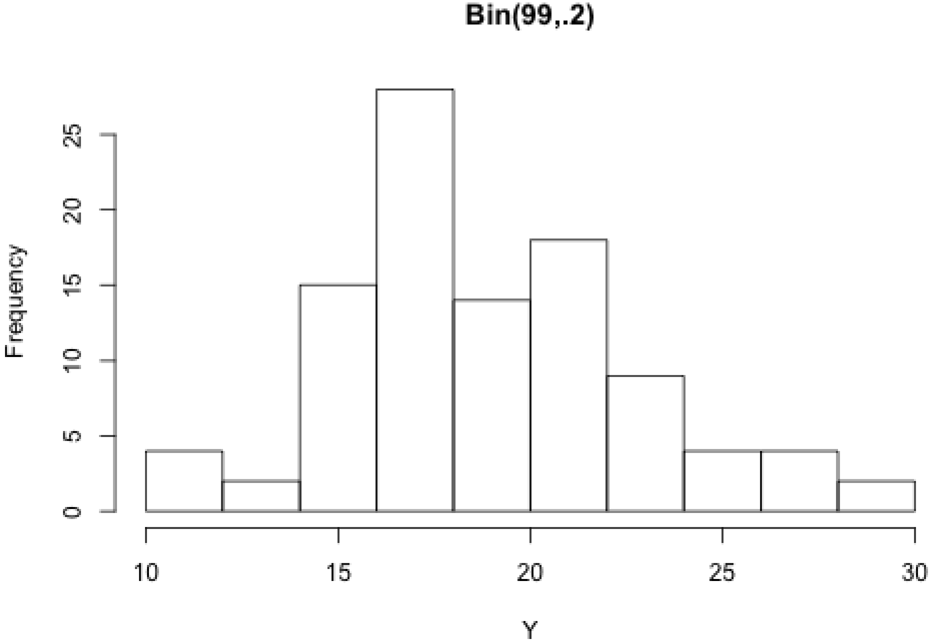
Presence/absence models

What if we ask this question *one node at a time*?

If proportion p of interactions actually occur in the population, then what is the distribution of the number of interactions (edges) for a single node?

Presence/absence models

$k_i \sim Bin(N-1, p)$



Important observations

Expected degree is $p(N-1)$.

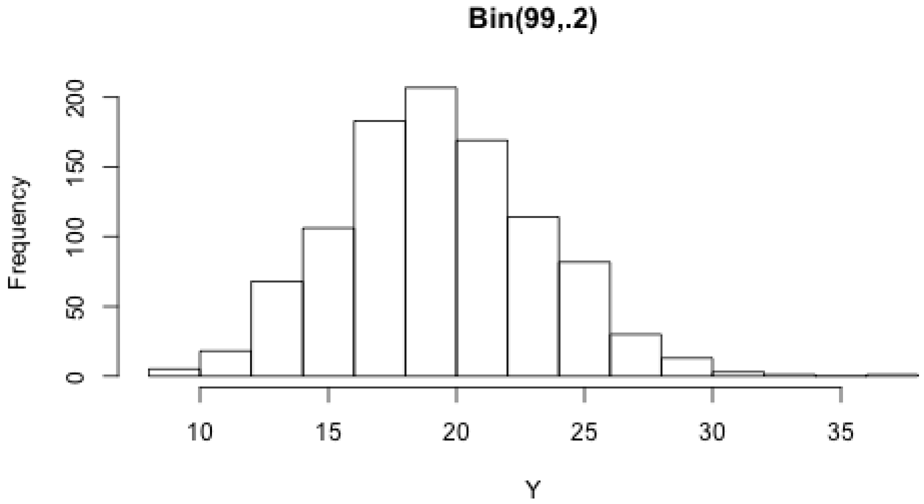
Important observations

Expected degree is $p(N-1)$.

One reasonable estimate from data would be $< k >$ (average degree)

Important observations

- Degrees concentrated around expectation
- Low degree and high degree nodes rare



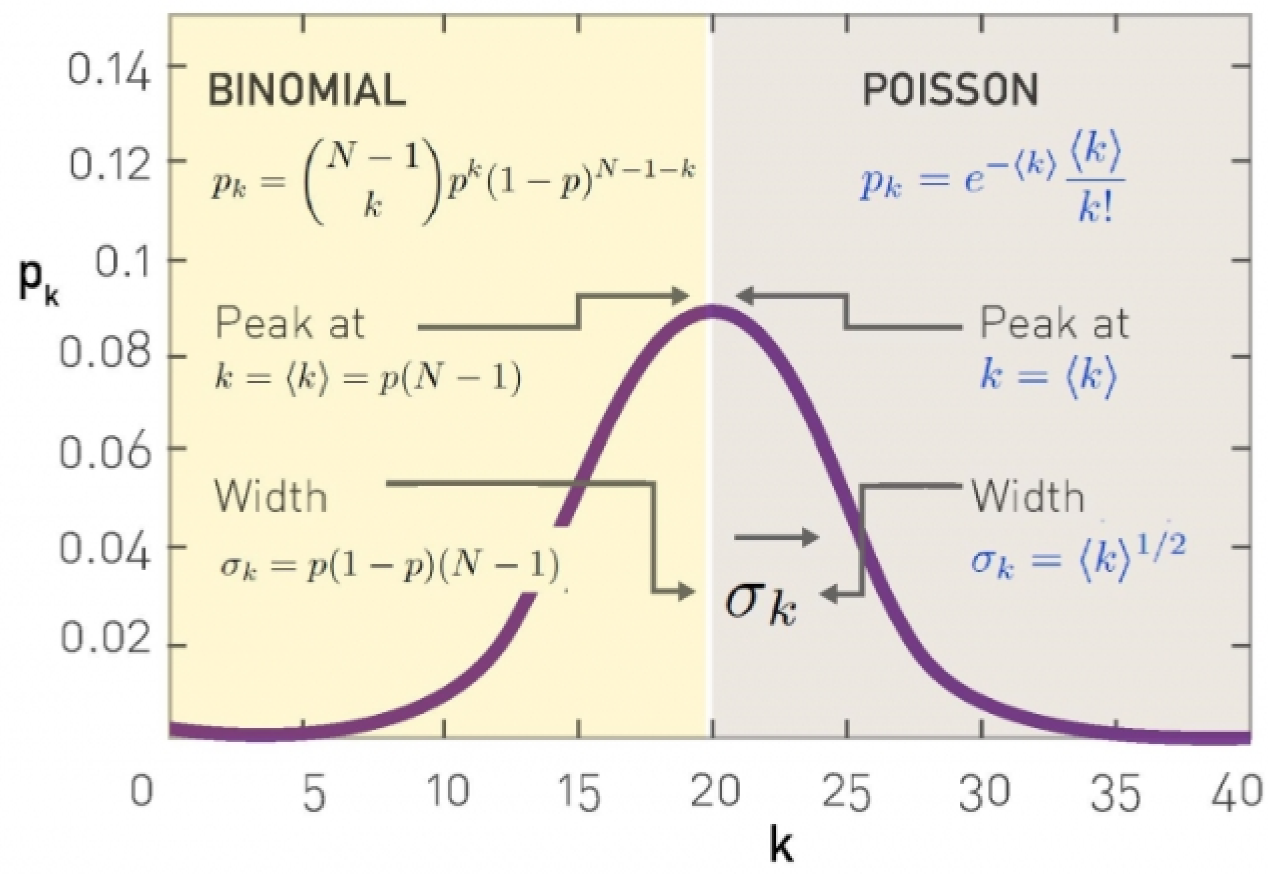
Important observations

Back to Poisson

$$k_i \sim \mathrm{Pois}(<\mathrm{k}>)$$

$$p_k = P(k_i = k) = e^{-<k>} \frac{<k>^k}{k!}$$

Important observations



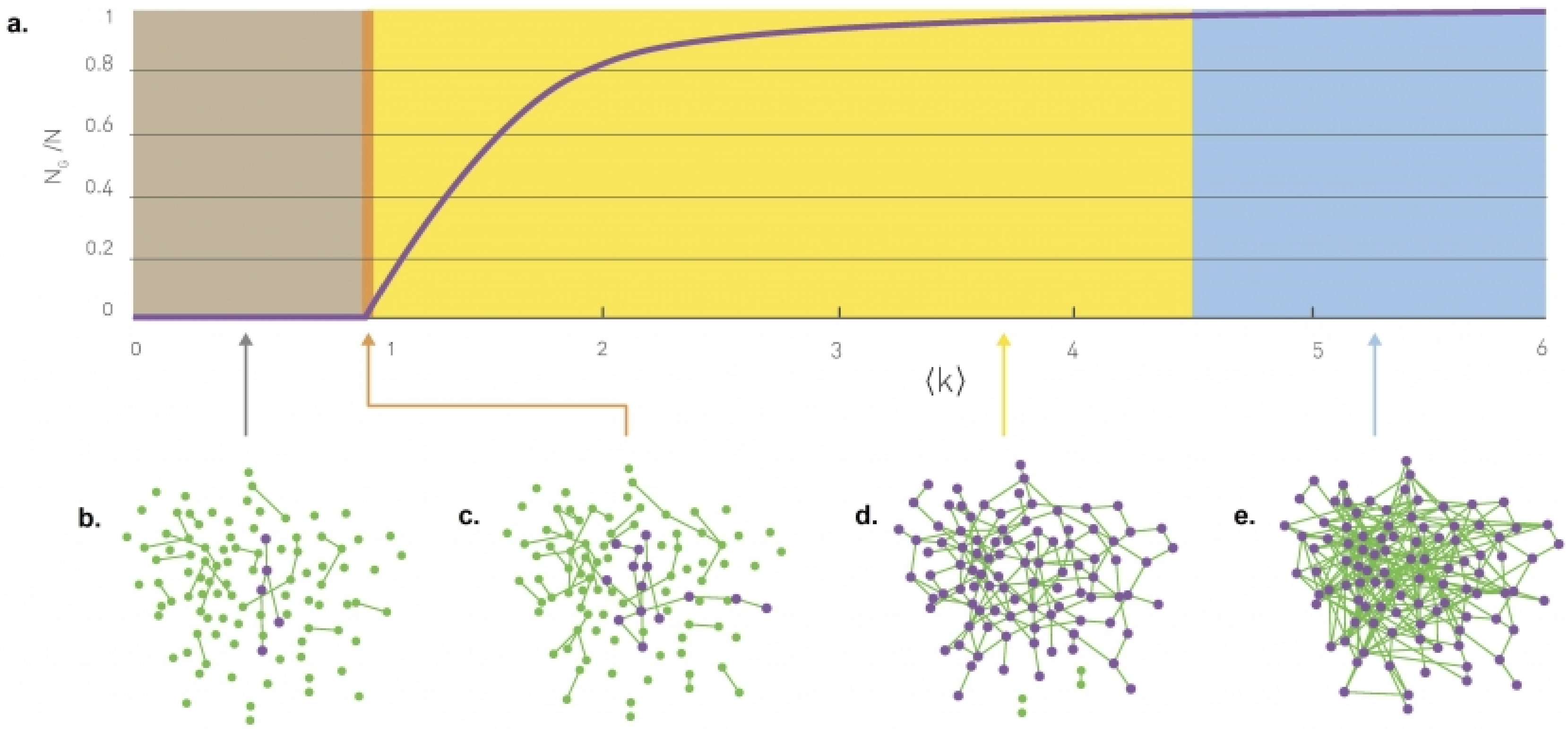
Note: equation for p_k is incorrect (numerator is $< k >^k$)

ER network model

Erdos-Renyi Construct a network with N vertices given parameter p by connecting vertices with probability p

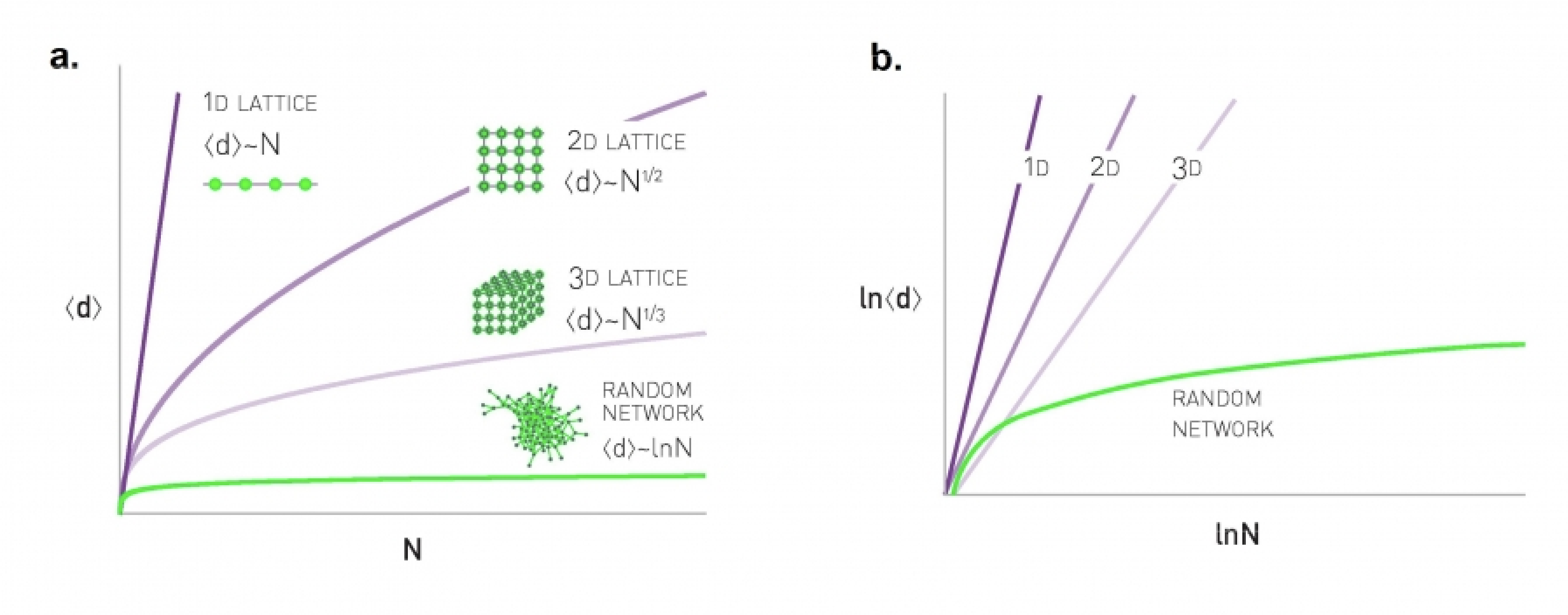
ER model

Emmergence of connected component



ER model

Small world



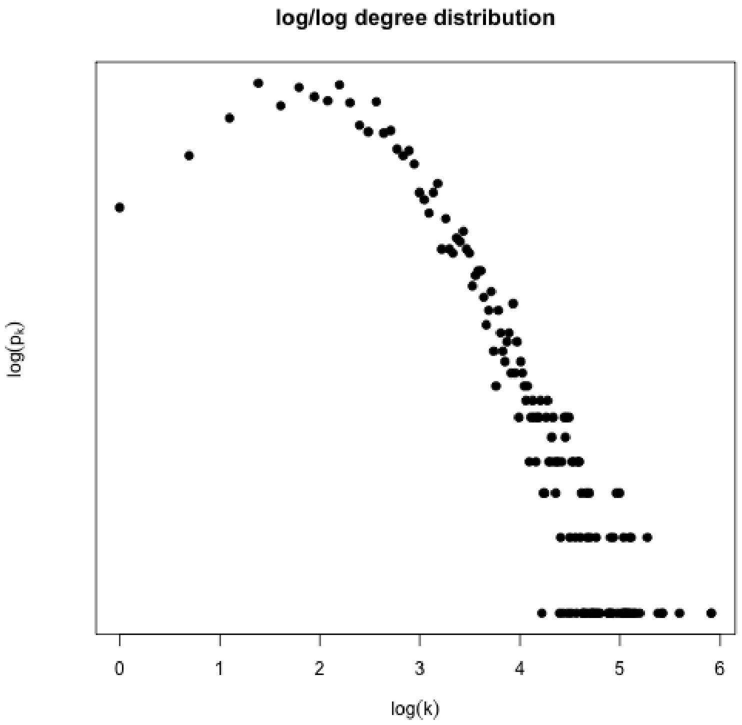
ER model

Clustering coefficient?

On the board

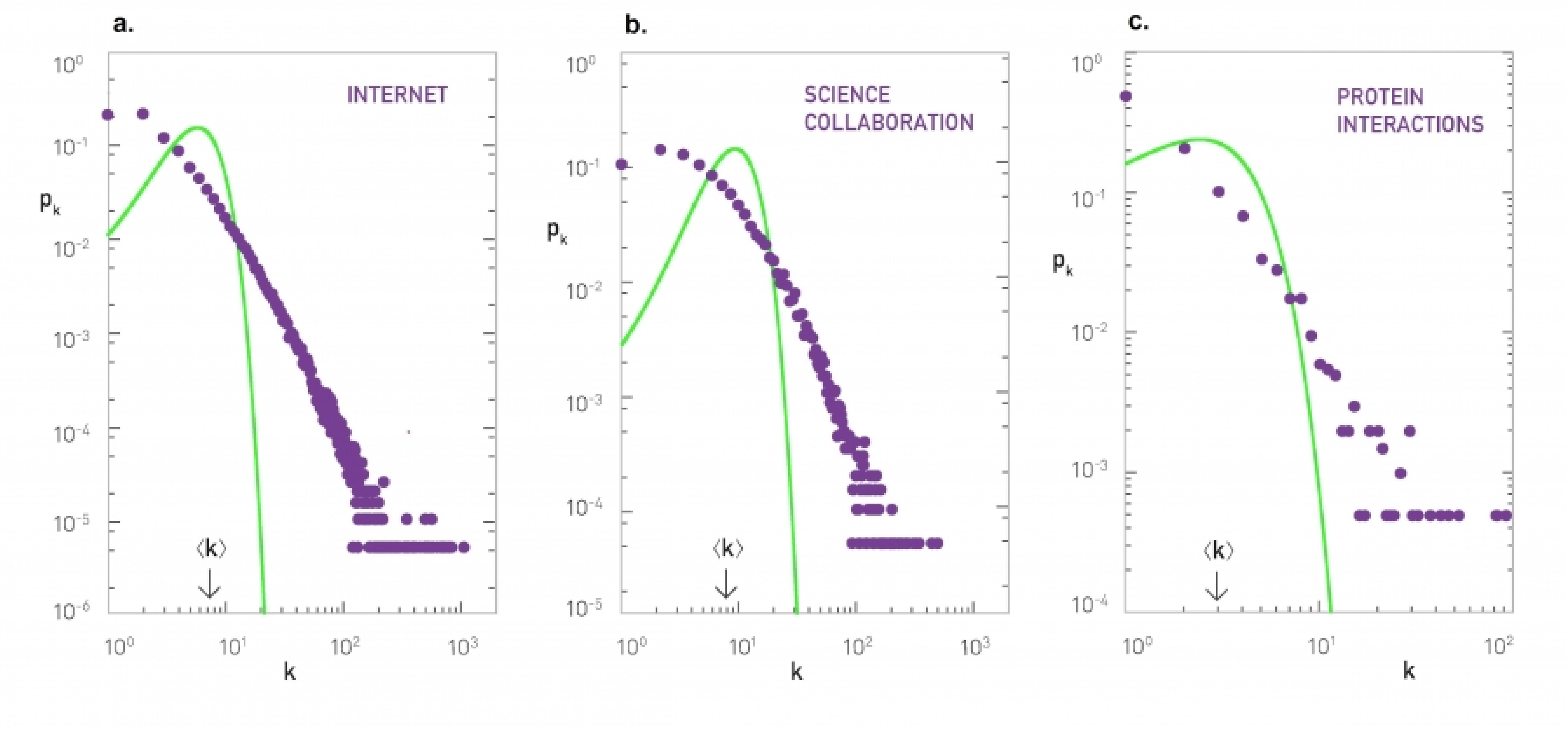
ER model

How well does this model fit data?



ER model

How well does this model fit data?



Power law

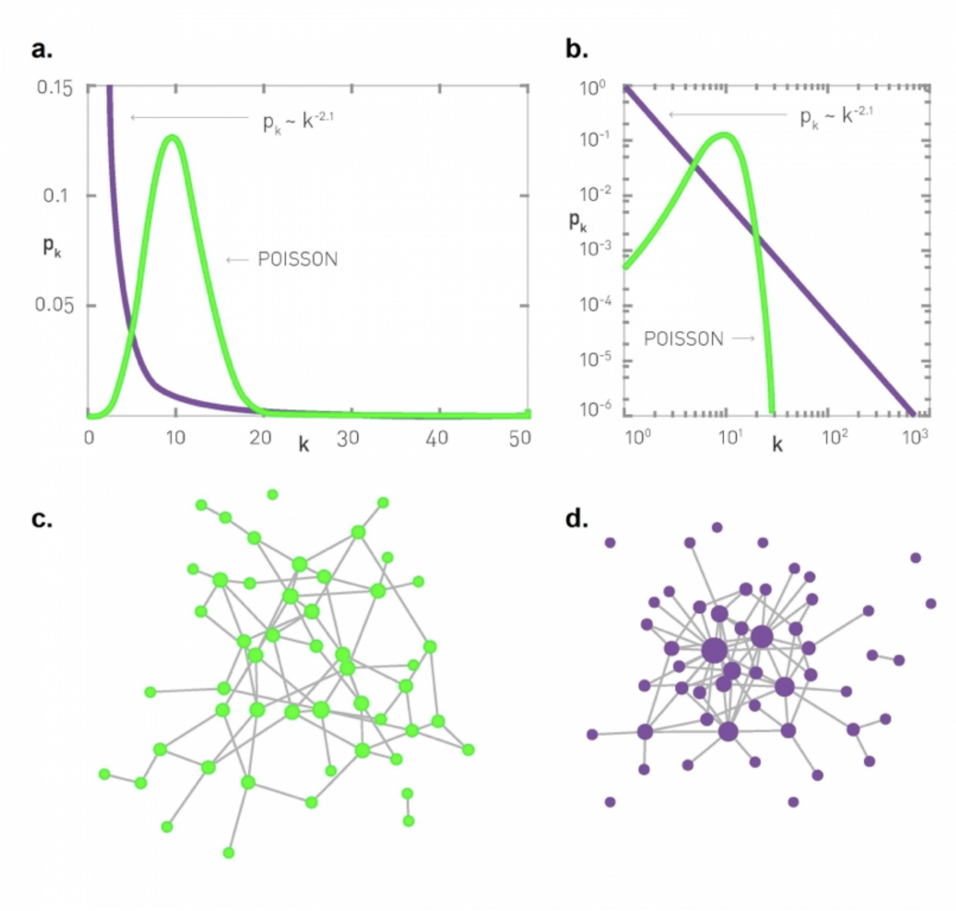
We see that exponential decay of high-degree probability does not fit data well.

Log-log plot suggests a linear relationship

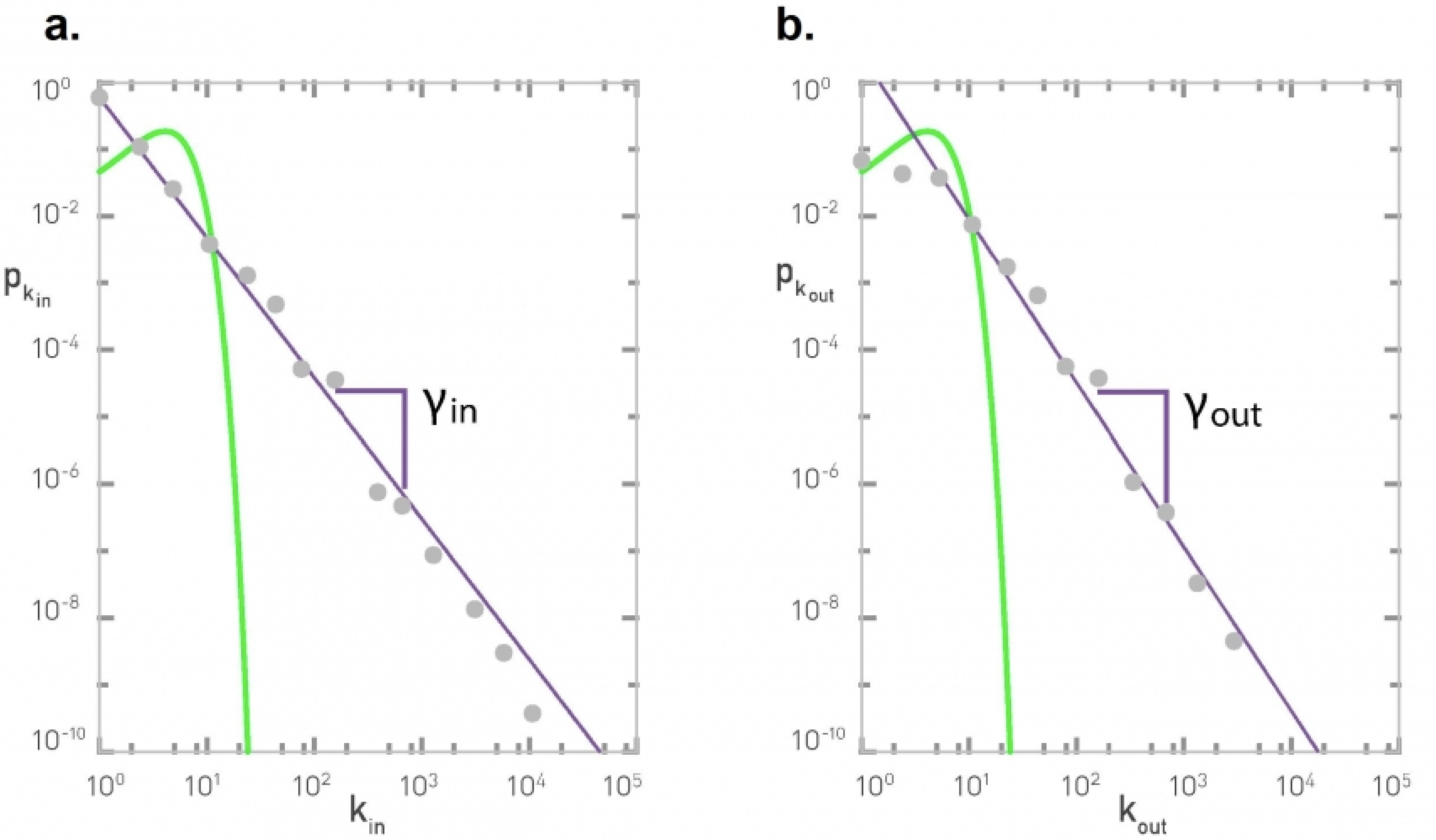
$$\log p_k \approx -\gamma \times \log k$$

$$p_k \approx k^{-\gamma}$$

Power law

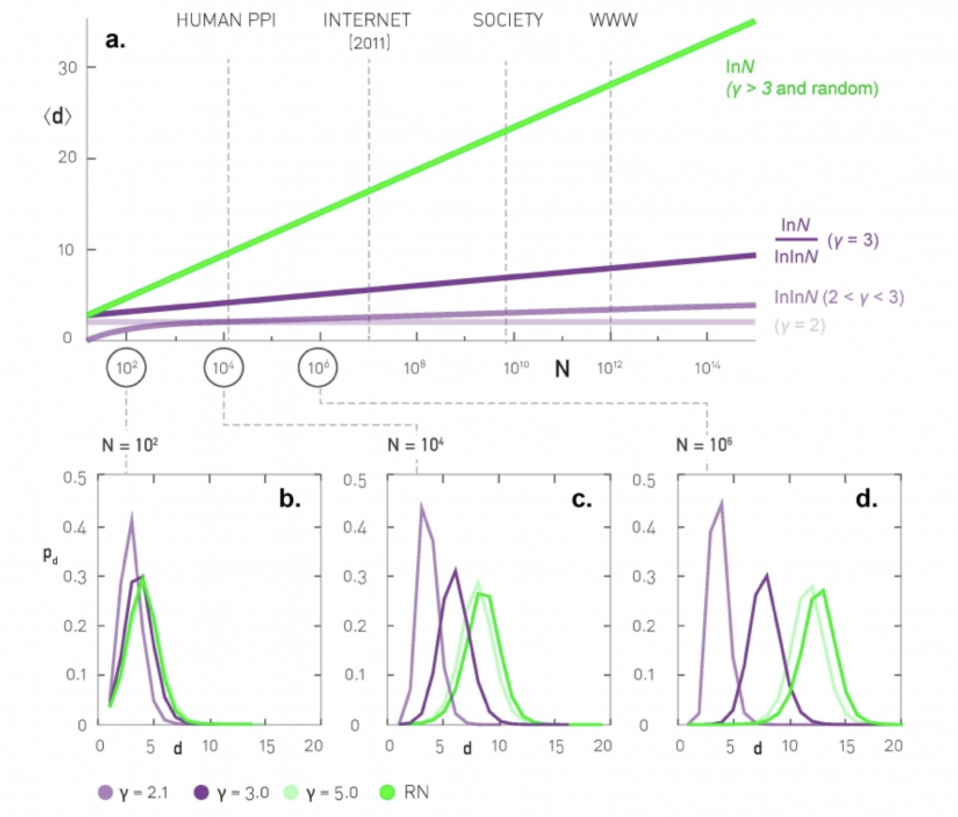

a.
0.15
0.1
p_k
0.05
$p_k \sim k^{-2.1}$
POISSON
0
10
20
k
30
40
50
b.
10^0
10^{-1}
10^{-2}
p_k
10^{-3}
10^{-4}
10^{-5}
10^{-6}
$p_k \sim k^{-2.1}$
POISSON
10^0
10^1
k
10^2
10^3
c.
d.

Power law

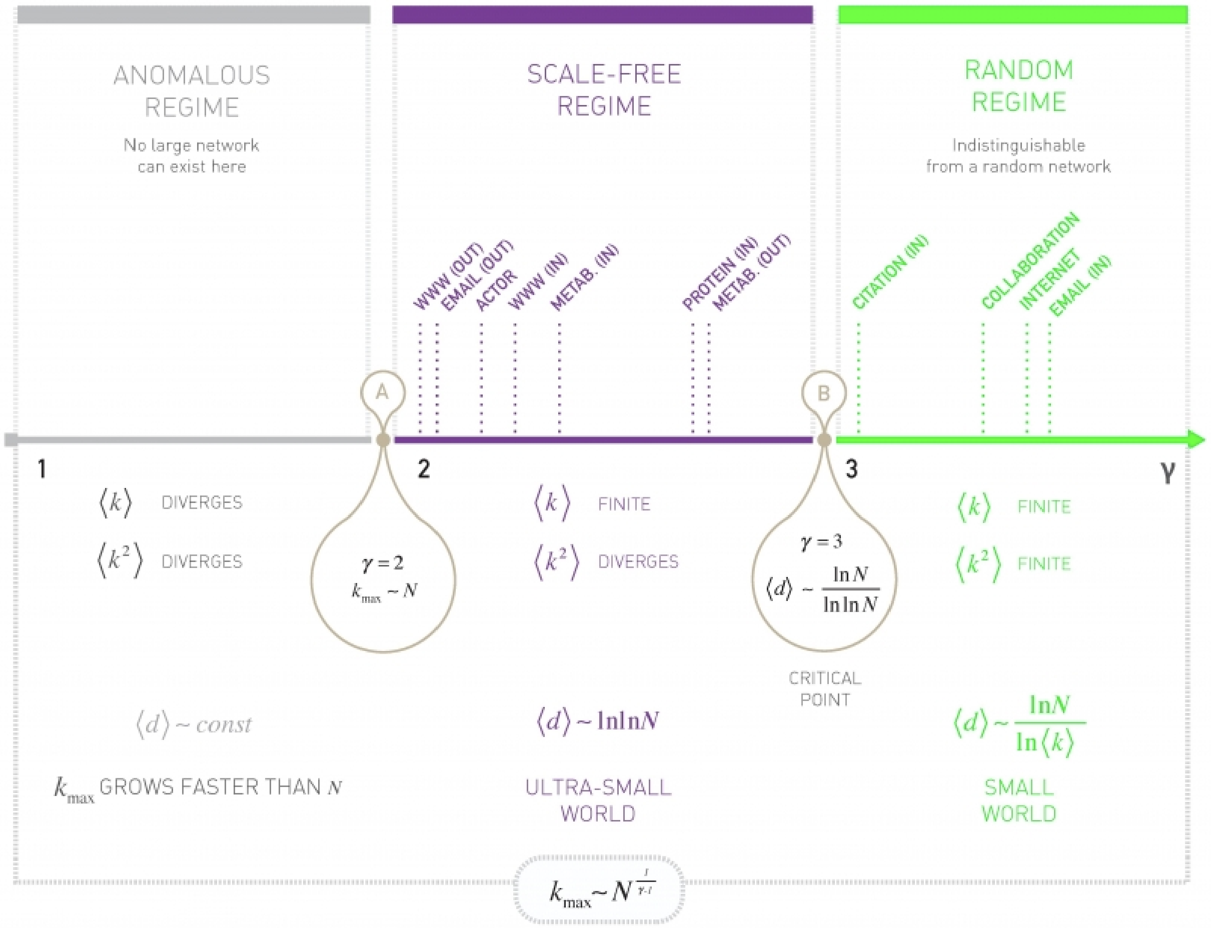


a.
b.
10^0
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
$p_{k_{in}}$
$p_{k_{out}}$
γ_{in}
γ_{out}
10^0
10^1
10^2
10^3
10^4
10^5
k_{in}
k_{out}

Power law (ultra-small world)



Power law (scale-free)



ANOMALOUS REGIME

No large network can exist here

SCALE-FREE REGIME

RANDOM REGIME

Indistinguishable from a random network

$\langle k \rangle$ DIVERGES

$\langle k^2 \rangle$ DIVERGES

$\gamma = 2$

$k_{max} \sim N$

$\langle k \rangle$ FINITE

$\langle k^2 \rangle$ DIVERGES

$\gamma = 3$

$\langle d \rangle \sim \frac{\ln N}{\ln \ln N}$

CRITICAL POINT

$\langle k \rangle$ FINITE

$\langle k^2 \rangle$ FINITE

$\langle d \rangle \sim const$

k_{max} GROWS FASTER THAN N

$\langle d \rangle \sim \ln \ln N$

ULTRA-SMALL WORLD

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

SMALL WORLD

$k_{max} \sim N^{\frac{1}{\gamma - 1}}$

Power Law

Generative random model for ER (poisson, random) network was straightforward

Next topics:

- Why scale-free?
- Fitting power law networks to data
- Generative random model for power law networks
- More sophisticated models to fit empirical data closer
- Biologically-plausible random models (evolution/fitness etc.)

Power Law

Why scale-free? A couple of ways to think about it:

- The variance of the degree distribution is infinite
- The right-hand tail of the degree distribution does not go to 0

Power Law

Why scale-free? A couple of ways to think about it:

- The variance of the degree distribution is infinite
- The right-hand tail of the degree distribution does not go to 0

Moments of an RV, the expected value of powers of an RV

$E[k]$ - center of the distribution

$E[k^2]$ - spread of the distribution (variance) **this is the important one**

$E[k^3]$ - skew of the distribution

and more...

Power Law

Why scale-free?

$$E[k^2] = \sum_{k_{min}}^{\infty} k^2 p_k = \sum_{k_{min}}^{\infty} k^{2-\gamma}$$

Power Law

Why scale-free?

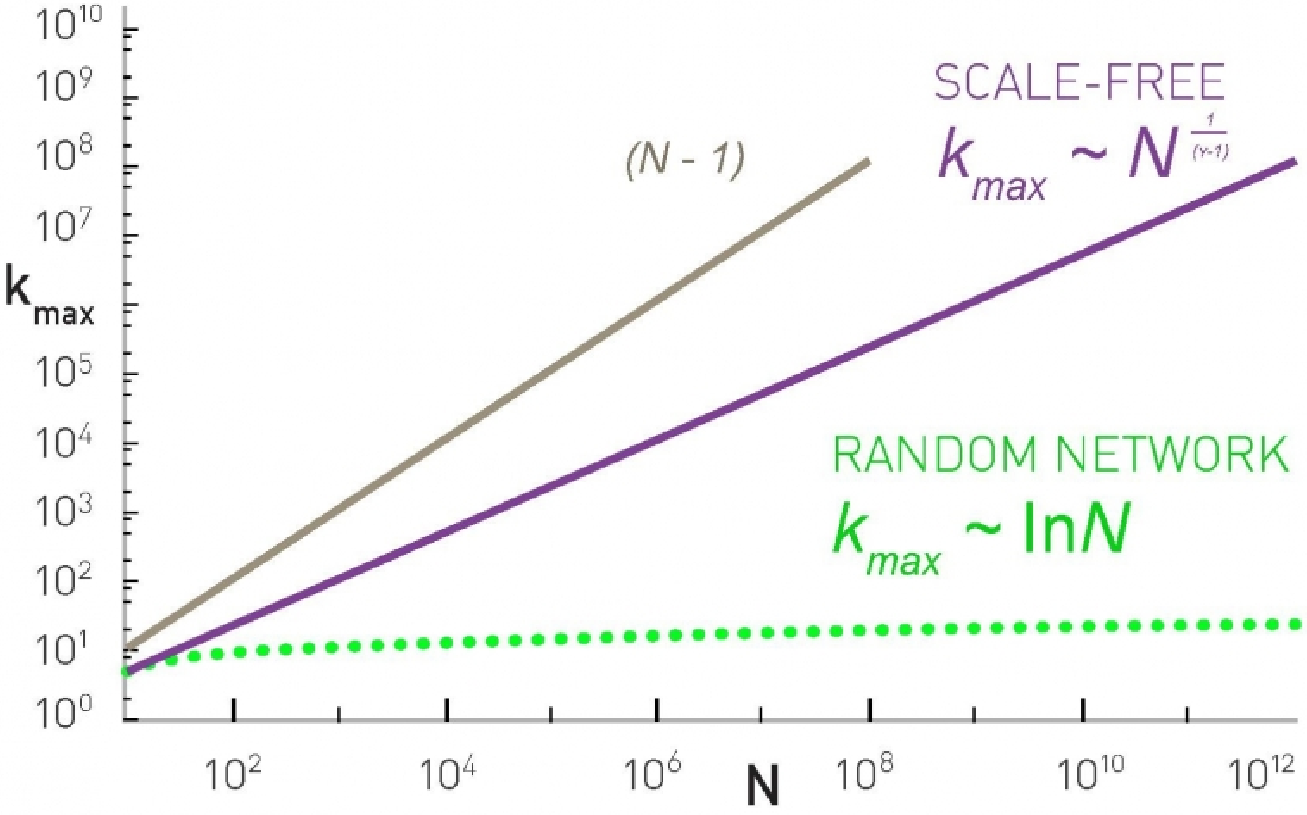
$$E[k^2] = \sum_{k_{min}}^{\infty} k^2 p_k = \sum_{k_{min}}^{\infty} k^{2-\gamma}$$

Let's take some liberties to make this easier to think about

$$E[k^2] = \int_{k_{min}}^{k_{max}} k^{2-\gamma} dk$$

and see what happens as $k_{max} \to \infty$

Power law

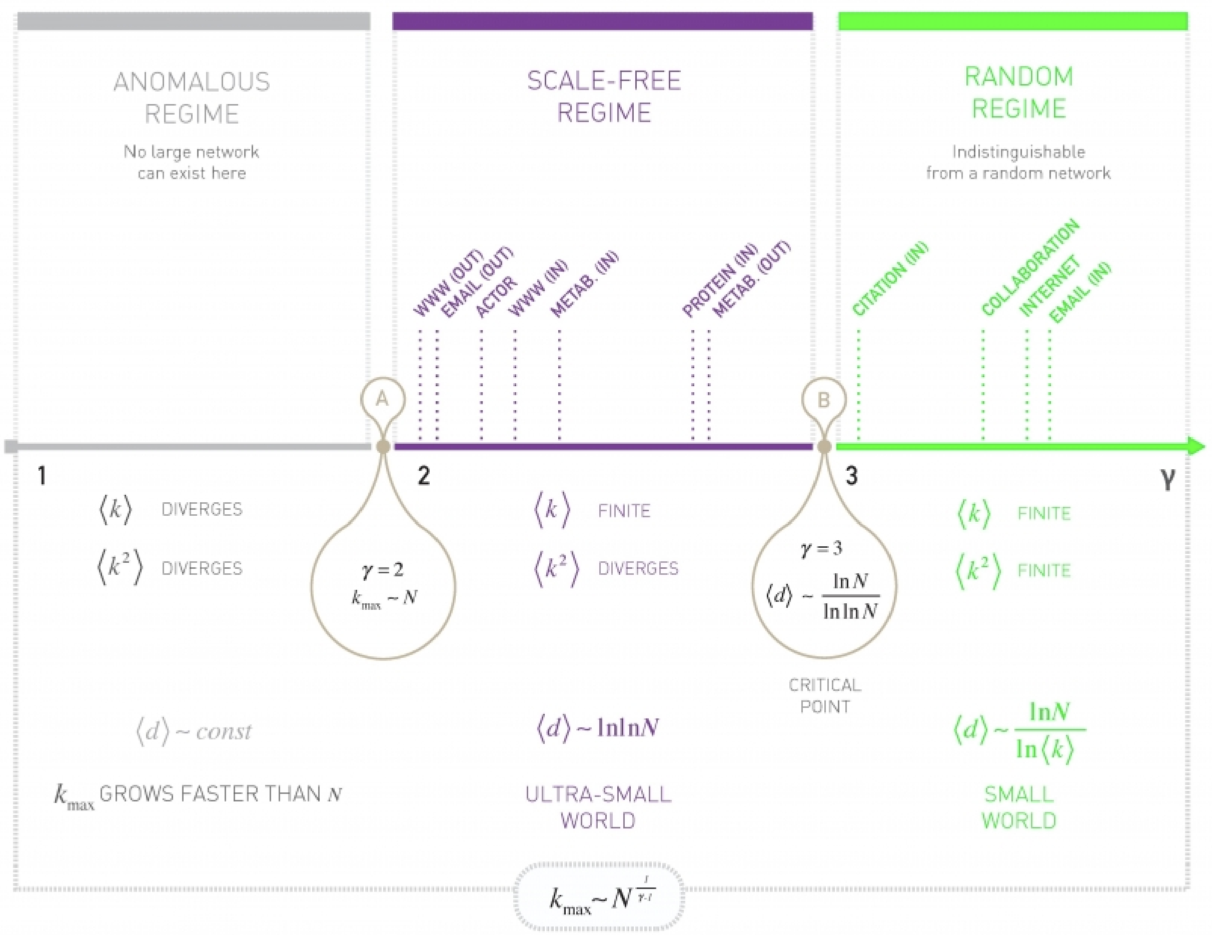

10^{10}
10^9
10^8
10^7
k_{max}
10^5
10^4
10^3
10^2
10^1
10^0
10^2
10^4
10^6
N
10^8
10^{10}
10^{12}
$(N - 1)$
SCALE-FREE
$k_{max} \sim N^{\frac{1}{(\gamma-1)}}$
RANDOM NETWORK
$k_{max} \sim \ln N$

Power law

$$\begin{aligned} E[k^2] &= \quad C\frac{k_{max}^{3-\gamma} - k_{min}^{3-\gamma}}{3-\gamma} \\ &= \quad a k_{max}^{3-\gamma} + b \end{aligned}$$

- For $\gamma > 3$: scale is finite value as k_{max} grows
- For $2 \leq \gamma \leq 3$: scale is infinite value as $k_{\max}$ grows (scale-free)

Power law

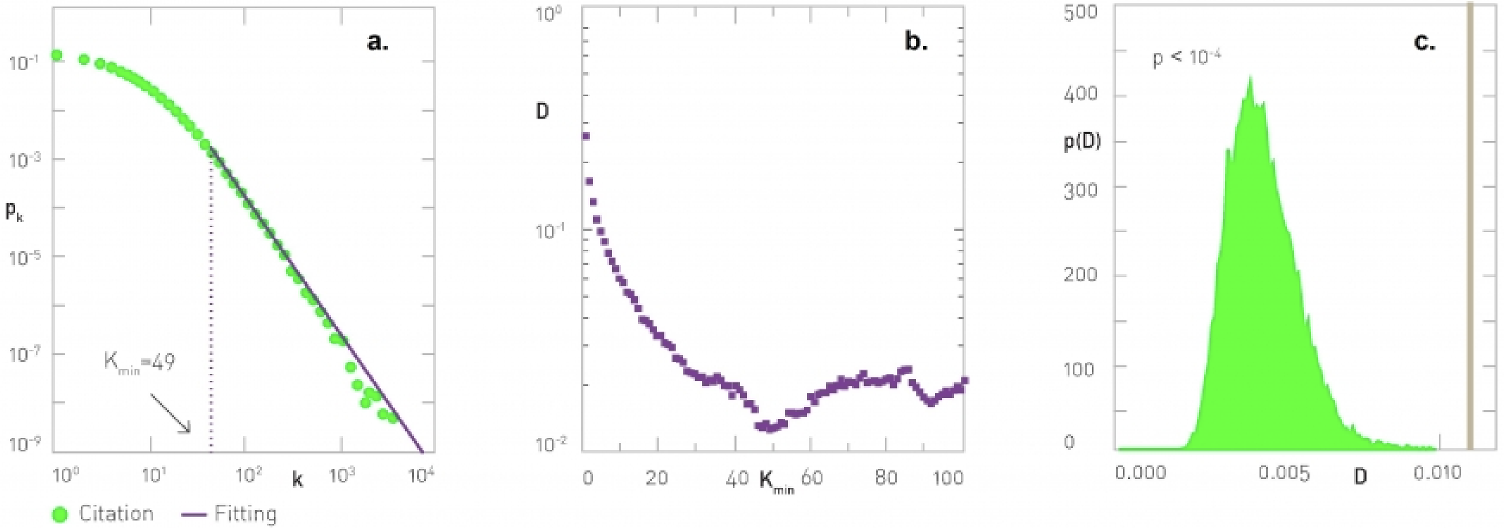

ANOMALOUS REGIME
No large network can exist here
SCALE-FREE REGIME
RANDOM REGIME
Indistinguishable from a random network
WWW (OUT)
EMAIL (OUT)
ACTOR
WWW (IN)
METAB. (IN)
PROTEIN (IN)
METAB. (OUT)
CITATION (IN)
COLLABORATION
INTERNET
EMAIL (IN)
A
B
1
2
3
γ
$\langle k \rangle$ DIVERGES
$\langle k^2 \rangle$ DIVERGES
$\gamma = 2$
$k_{max} \sim N$
$\langle k \rangle$ FINITE
$\langle k^2 \rangle$ DIVERGES
$\gamma = 3$
$\langle d \rangle \sim \frac{\ln N}{\ln \ln N}$
$\langle k \rangle$ FINITE
$\langle k^2 \rangle$ FINITE
CRITICAL POINT
$\langle d \rangle \sim const$
k_{max} GROWS FASTER THAN N
$\langle d \rangle \sim \ln \ln N$
ULTRA-SMALL WORLD
$\langle d \rangle \sim \frac{\ln N}{\ln \langle k \rangle}$
SMALL WORLD
$k_{max} \sim N^{\frac{1}{\gamma - 1}}$

Estimating the degree coefficient

Since so many structural properties of network depend on this value, estimating it properly matters a lot.

Here is a procedure based on Maximum Likelihood Estimation

Estimating the degree coefficient



Estimating the degree coefficient

Step 1 - The estimate

Given $\boldsymbol{k}_{min}$, estimate γ as

$$\gamma_{\boldsymbol{k}_{min}} = 1 + N\left[\sum_{i=1}^{N} \log \frac{\boldsymbol{k}_i}{\boldsymbol{k}_{min} - \frac{1}{2}}\right]^{-1}$$

Estimating the degree coefficient

Step 2 - How good is it?

$$D_{k_{min}} = \max_{k \geq k_{min}} |S(k) - p_k|$$

with p_k given by $\gamma_{k_{min}}$

Estimating the degree coefficient

Step 3 - Best model fit

Estimate γ as $\gamma_{k_{min}}$ that *minimizes* $D_{k_{min}}$

Estimating the degree coefficient

Step 4 - Model check

Is the optimal observed deviance $D_{k_{min}}$ consistent with the estimated γ?

Generate a distribution of deviance statistics from power law model, check where optimal deviance falls within generated distribution.

Network Generation

The ER model gave use a generative algorithm to create a network that satisifies "Poisson" degree distribution?

Let's see a generative model to obtain a power law network

The Barabasi-Albert Model

Procedure:

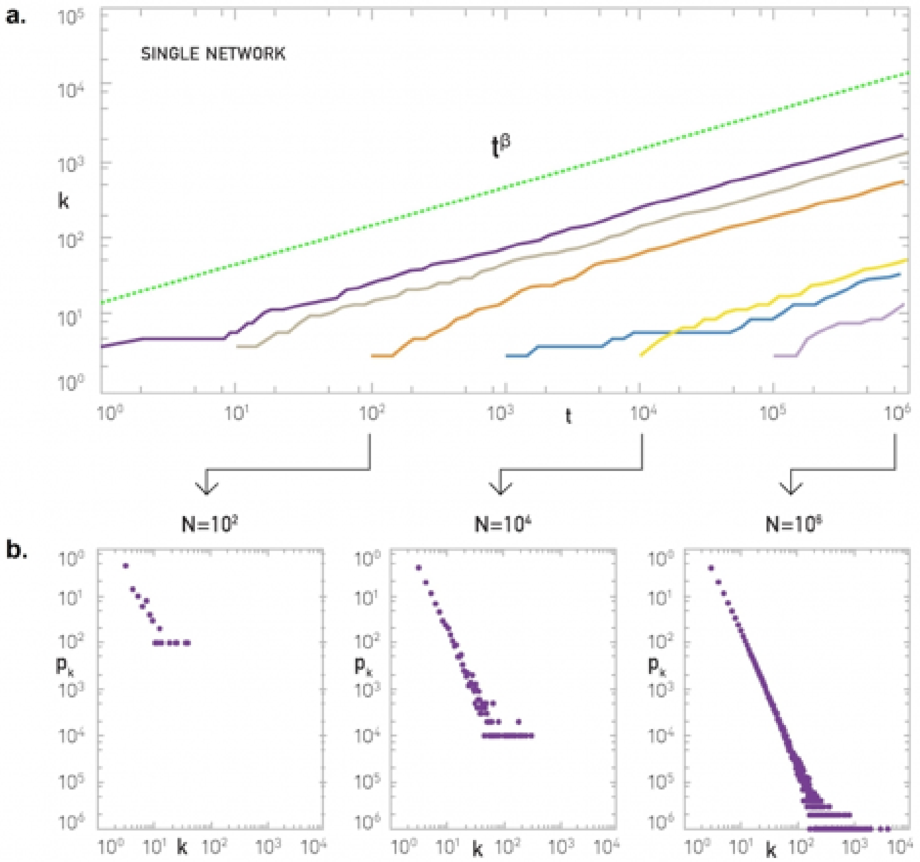
- Start with m_0 nodes (with arbitrary links, or connected)
- At each time step t
  - Growth: Add a new node with m edges
  - Preferential attachment: probability new node links to node i $\Pi(k_i) = \frac{k_i}{\sum_i k_i}$

Some simple questions to warm up: (a) how many nodes after t steps? (b) how many edges after t steps?

The Barabasi-Albert Model

Degree dynamics

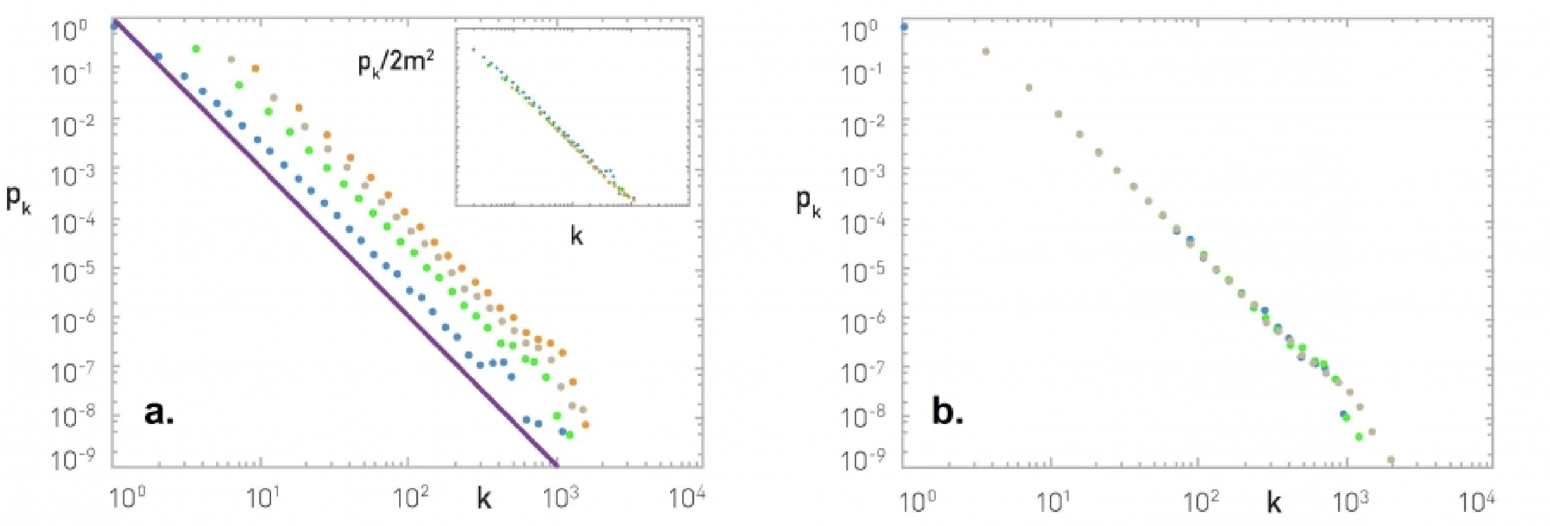
$$k_i(t) = m\left(\frac{t}{t_i}\right)^{1/2}$$



The Barabasi-Albert Model

Degree distribution

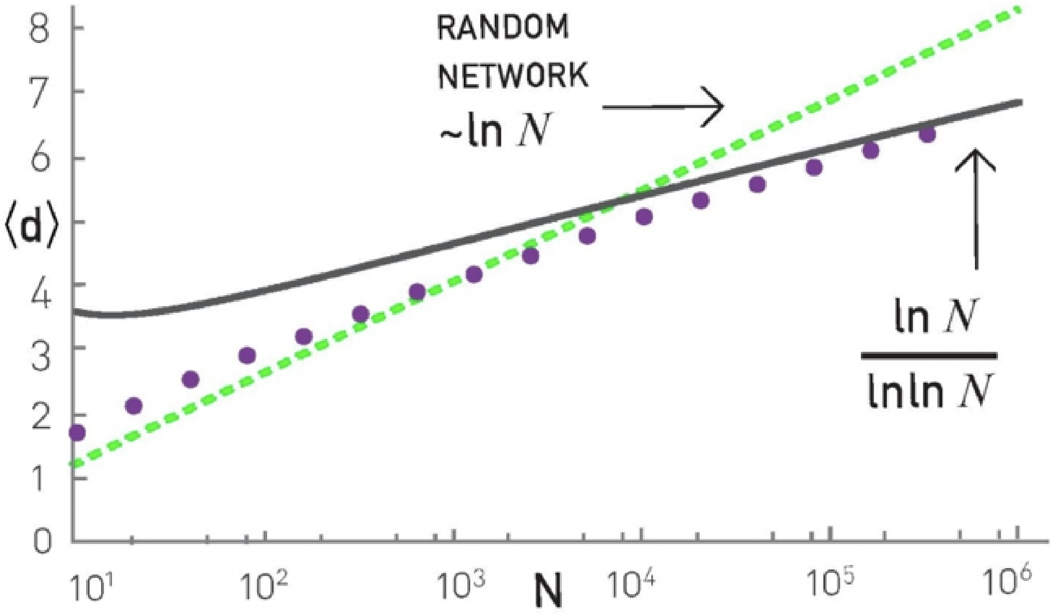
$$p_k \sim k^{-3}$$



The Barabasi-Albert Model

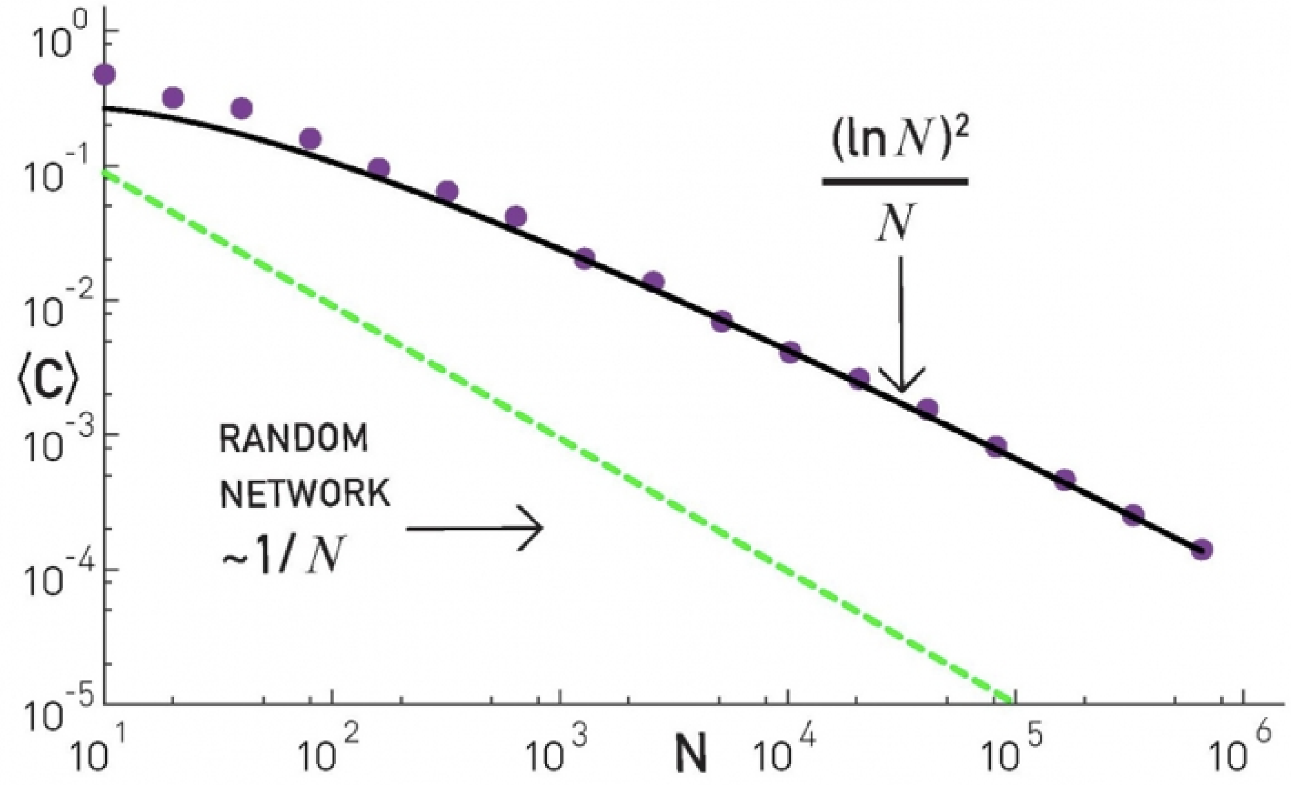
Distance

$$< d >= \frac{\log N}{\log \log N}$$



The Barabasi-Albert Model

Cluster coefficient

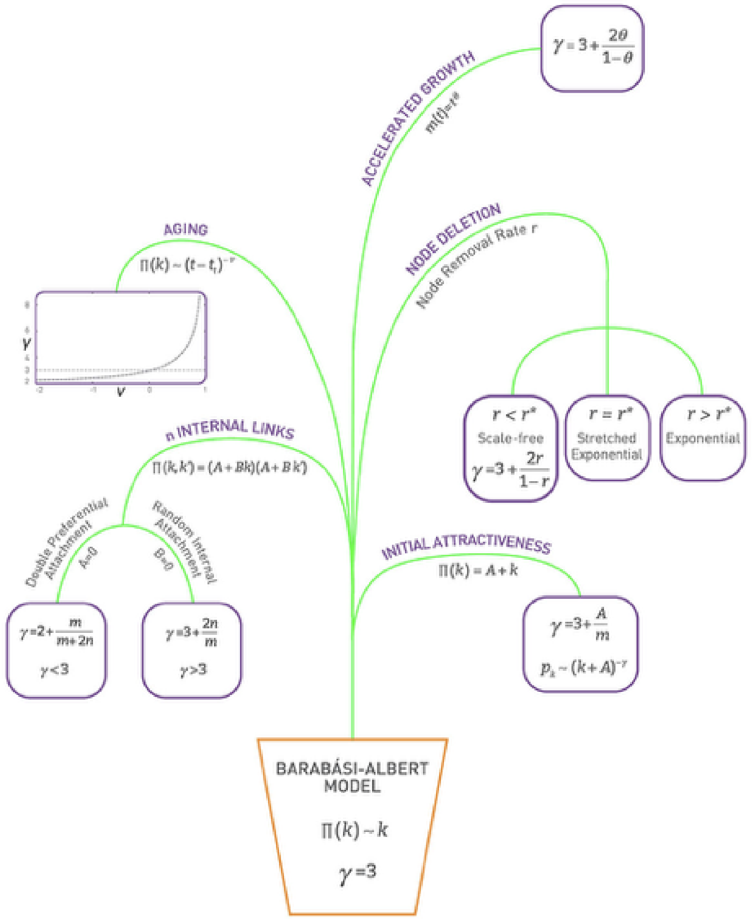


Network growth models

Barabasi-Albert, some missing ingredients

- Fitness, properties beyond degree that control preferential attachment
- Age, preferential attachment depending on time
- Death, how to model nodes that go away?

Network growth models



Network growth models

More importantly: how well does this describe biologically plausible emergence.

Is this consistent with evolution?

Next time...

Fitness: Bianconi-Barabasi

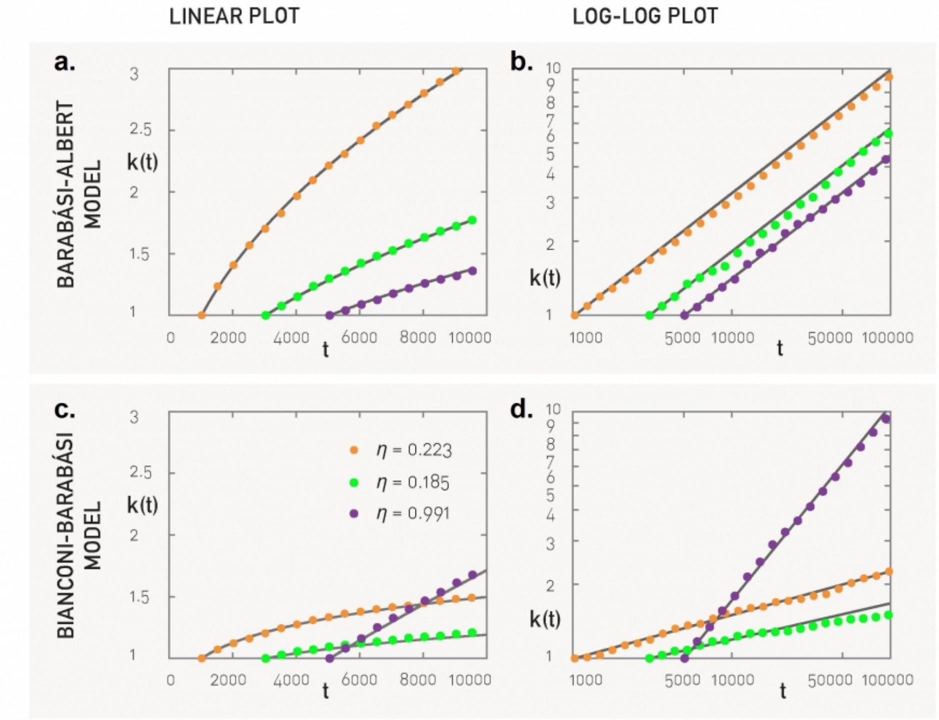
Goal: not all nodes of the same degree are equally attractive

Each node has fitness η_i

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

Dynamics:

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\eta_i/C}$$



Fitness: Bianconi-Barabasi

Note: if you have data of network growth, you can estimate fitness coefficient under BB model assumption

Node deletion

In biological networks we often encounter systems in which nodes are removed from networks:

- deleterious mutations in molecular networks
- neuron loss (e.g. AD)
- extinction in ecological networks

Model: at each time step, remove nodes with rate r

Node deletion

Some observations:

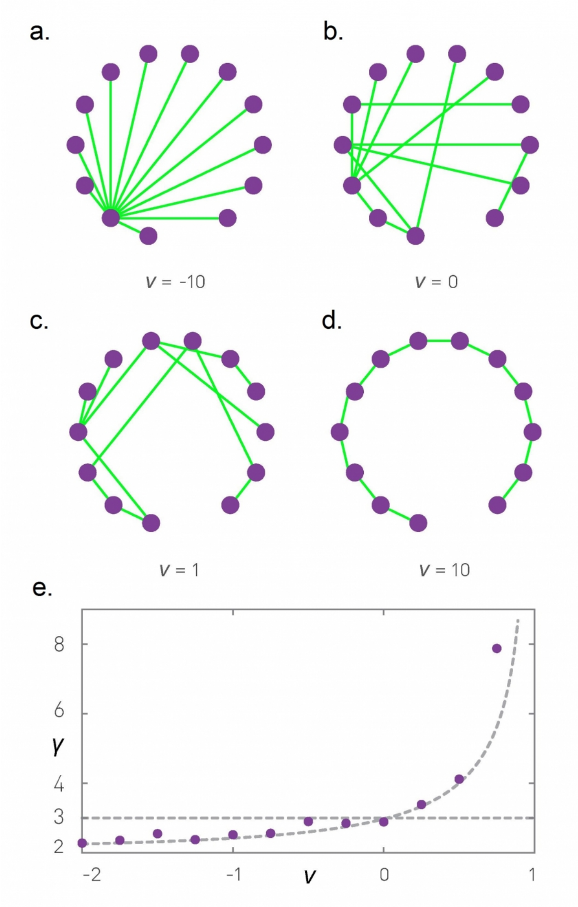
- For $r < 1$ network grows slighly slower to BA model, so power-law with $\gamma = 3 + \frac{2}{1-r}$
- $r = 1$ exactly BA model
- For $r > 1$ network decays as it grows, analytical model not so clean...

Regimes for deletion important to think about: e.g., removal probability depends on degree (weak-get-weaker)

Aging

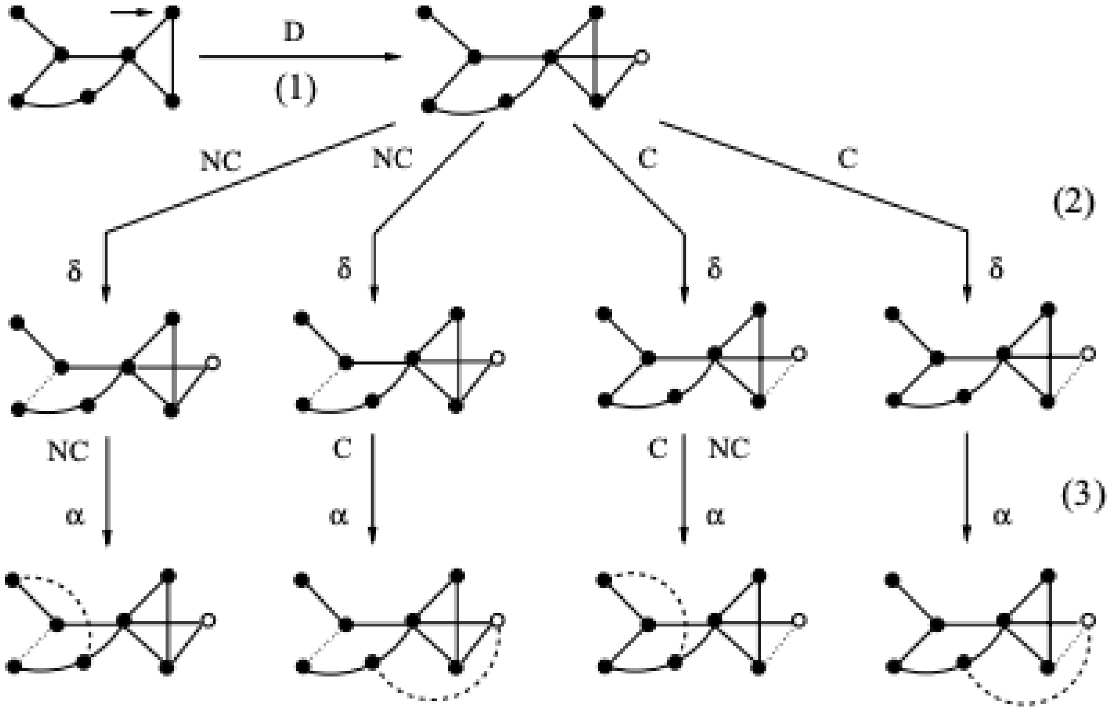
Probability of attachment depends on node age

$$\Pi(k_i, t_i) \sim k_i(t - t_i)^{\nu}$$



Node duplication

An alternative, more biologically plausible network growth model. *New nodes evolve from existing nodes through duplication*



Node duplication

Procedure:

- Start with seed graph G_0 with n_0 nodes
- At each time step t
  - Growth: Create graph G_t by selecting a vertex v_t uniformly at random from graph G_{t-1} and duplicate it by adding new node v with same links as v_t
  - Deletion: Delete each edge connecting v with probability $(1 - p)$
  - Addition: For each vertex u not connected to v, attach with probability $\frac{r}{t+n_0}$

Node duplication

For large k can show power law

$$p_k \approx k^{-\gamma}$$

$$\gamma = 1 - \frac{2r}{1-p}$$

Node duplication

Arguments of fitness, and to a greater degree, node deletion still hold in this case.

Analytical question

Given observed network structure, can we posit a growth model (with associated parameters) that recapitulate observed network structure.

For many biological applications, node duplication with a node deletion regime is a good first place to start

Next time

We briefly saw analysis of some network properties in terms of dynamics

Next we use similar ideas to model processes over data:

- network diffusion
- network epidemics
- network robustness (percolation)