RANDOM GRAPH MODELS

WHY USING RANDOM GRAPH MODELS

- Several good reasons:
  - Study some properties in a "controlled environment"
    - How does property X behaves when increasing property Y ?
  - Compare an observed network with a randomized version
    - Is observed property X "exceptional", or any similar network with same property Y and Z ?
  - Explain a given phenomenon
    - Such simple mechanism can reproduce property X and Y
  - Generate synthetic datasets
    - Testing an algorithm on 100 variations of the same network

NULL MODELS

- Using Random Graphs as Null models
  - Assume some properties (X1, X2, …) of your data are given
  - And that everything else is random
    - =>Is what you are observing on property Y unexpected/random/exceptional ?
  - Principle of a reference point

NULL MODELS(NON-GRAPH)

- Total CO2 emissions 2017:
  - China: **37 000** Mt, Germany: 796 Mt, France: 338 Mt
    - So China emit "more" than Germany and France

- Considering variable *population*
  - China: 7.7t — Germany: **9.6t** — France: 4.8t
    - So Germany emit more (per person) than China, and then France

- Considering variable *Trade*.(consumption-based index)
  - China: 6.27t — Germany: **10.84t** — France: 6.93t
    - So China is the lowest of the three

- What about countries weather? Cumulated historical emissions? Land area? Geopolitical reasons (nuclear…)?

CLASSES OF SYNTHETIC NETWORKS

Synthetic networks types

There are three main types of synthetic networks:

- **Deterministic models** are instances of famous graphs or, more commonly, repeated regular patters. e.g.,*Caveman graph, grids, lattices*.
- **Generative models** assign to each pair of nodes a probability of having an edge according to their properties (degree, label, etc.). e.g., *Erdős Rényi, Configuration model, etc.*
- **Mechanistic models** create networks by following a set of rules, a process defined by an algorithm. e.g., *Preferential attachment, Forest fire, etc.*

Fundamental network models

Central quantities in network analysis

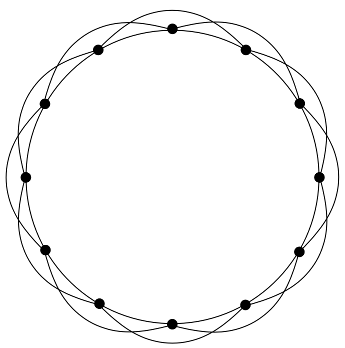
- Degree distribution: $P(k)$
- Clustering coefficient: C
- Average path length: $<d>$

Network	Degree distribution	Path length	Clustering coefficient
Real world networks	broad	short	large

Regular lattices

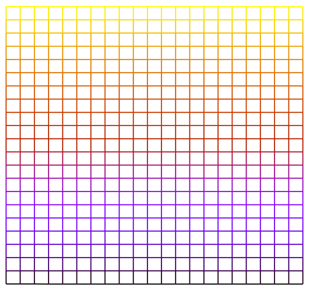
- Graphs where each node has the **same degree** k
- Translational symmetry in n directions

1D

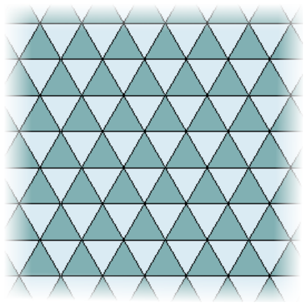


k=4

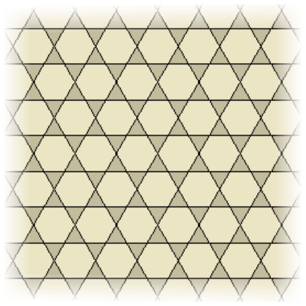
2D lattices



k=4

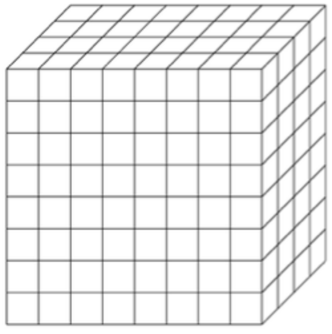


k=6

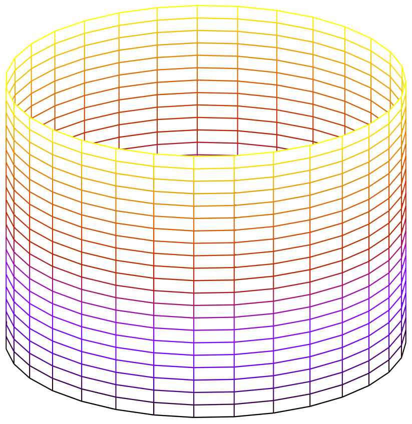


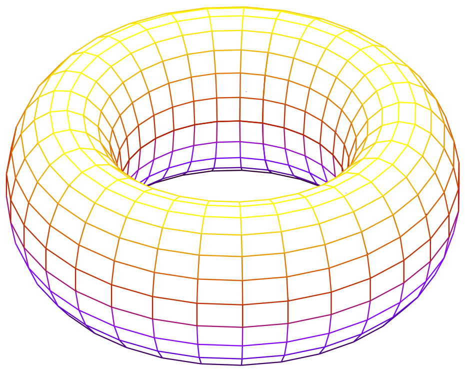
k=4

3D lattices



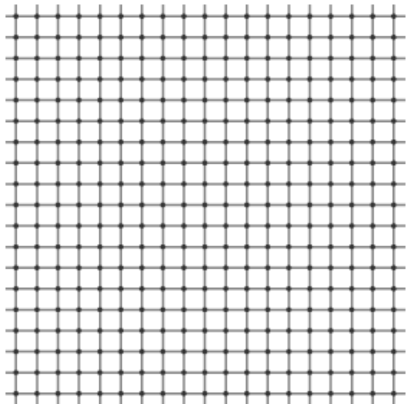
k=6



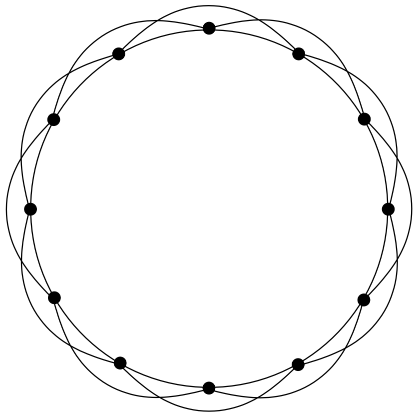


Regular lattices

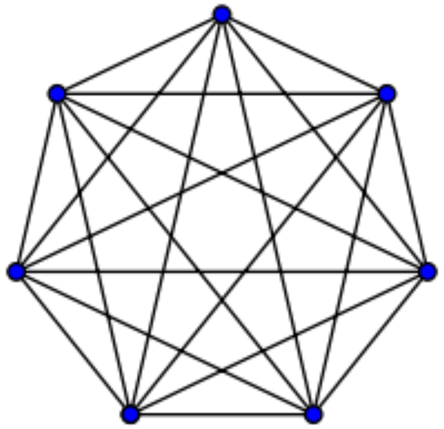
Clustering coefficient



C=0



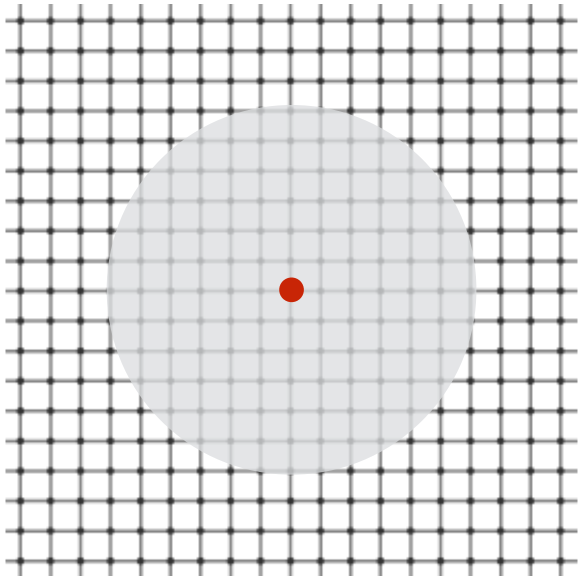
C=3/6



C=1

- Clustering coefficient depends on the structure (can be large or not)
- It is constant for each node

Path length



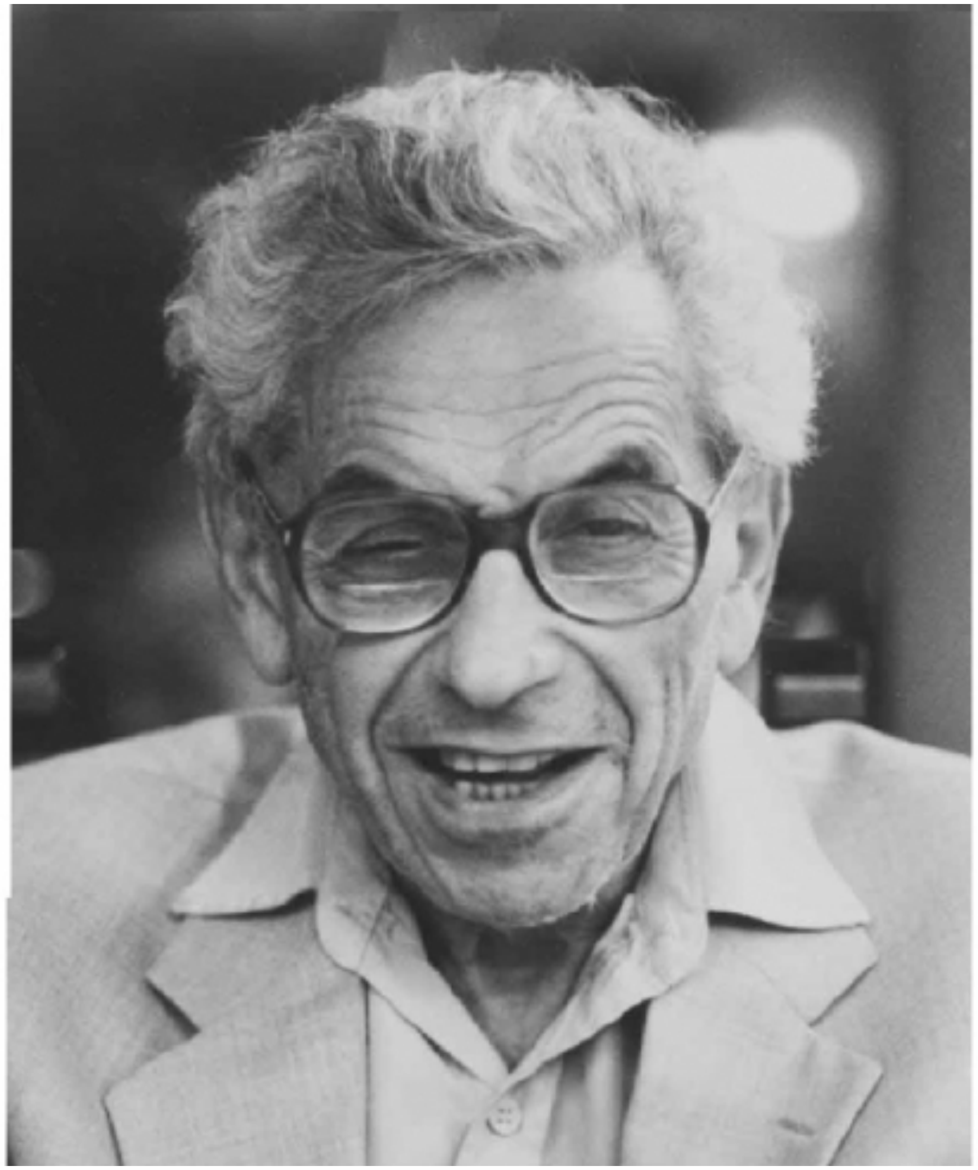
- Average path length grows quickly with n when k << n
- In a *large* graph with *realistic* average degrees, will be large

Regular lattices

Network	Degree distribution	Path length	Clustering coefficient
Real world networks	broad	short	large
Regular lattices	constant	long	can be large

The Erdős-Rényi Random Graph model (ER)

Random Graphs



Pál Erdős
(1913-1996)

Alfréd Rényi
(1921-1970)

"If we do not know anything else than the number *n* of nodes and the number *L* of links, the simplest thing to do is to put the links at random (no correlations)"

P. Erdős and A. Rényi. On random graphs, I. Publicationes Mathematicae (Debrecen), 6:290-297, 1959.
P. Erdős and A. Rényi. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci., 5:17-61, 1960.

ER Random Graphs

Erdős-Rényi model: simple way to generate random graphs

- The *G(n,L)* definition

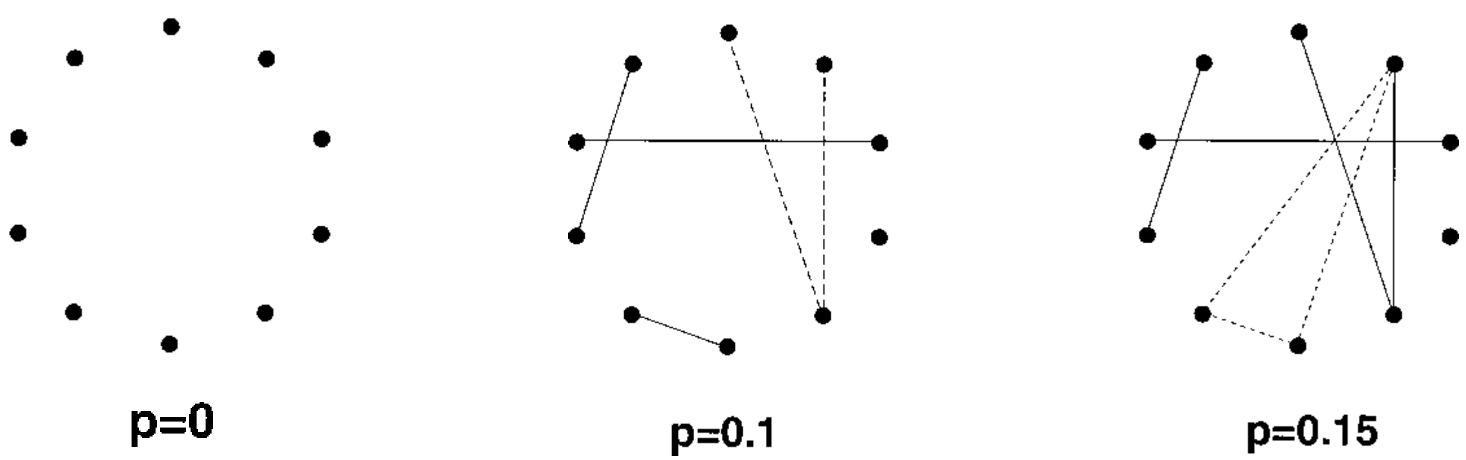
  1. Take n disconnected nodes
  2. Add L edges uniformly at random

  Alternatively:

  - pick uniformly randomly a graph from the set of all graphs with n nodes and L links

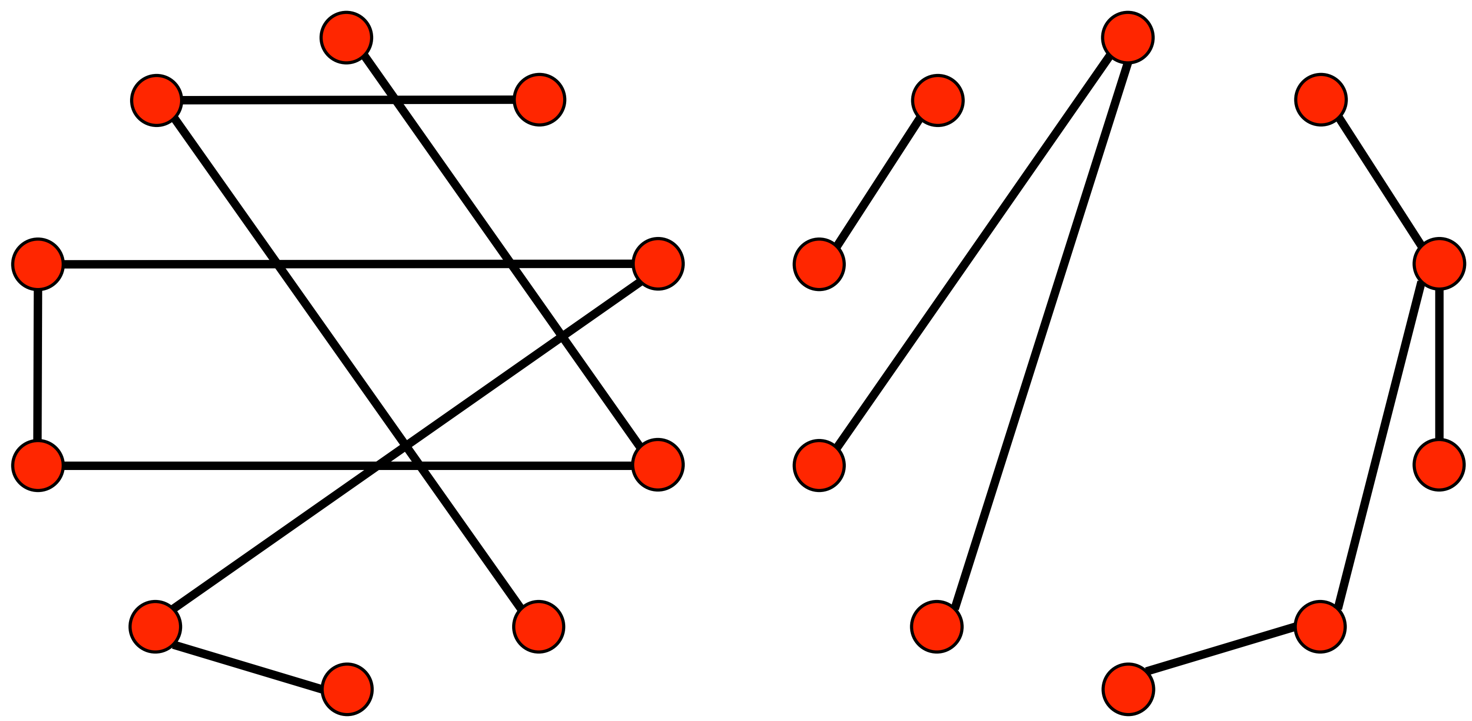
- The *G(n,p)* definition

  1. Take n disconnected nodes
  2. Add an edge between any of the nodes independently with probability p



Random Graphs

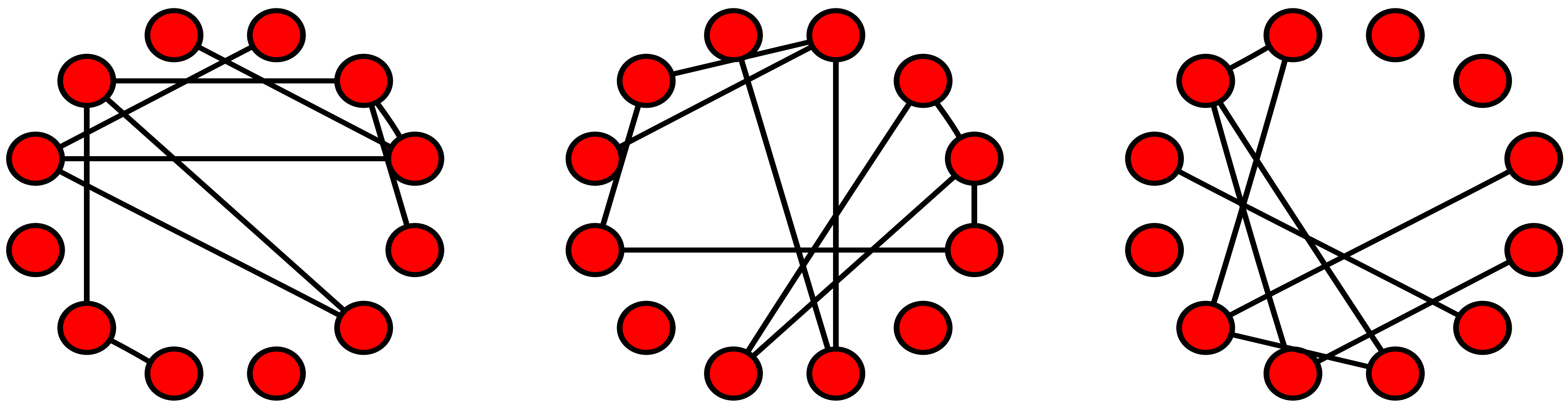
In the G(n,p) variant, the number of edges may vary



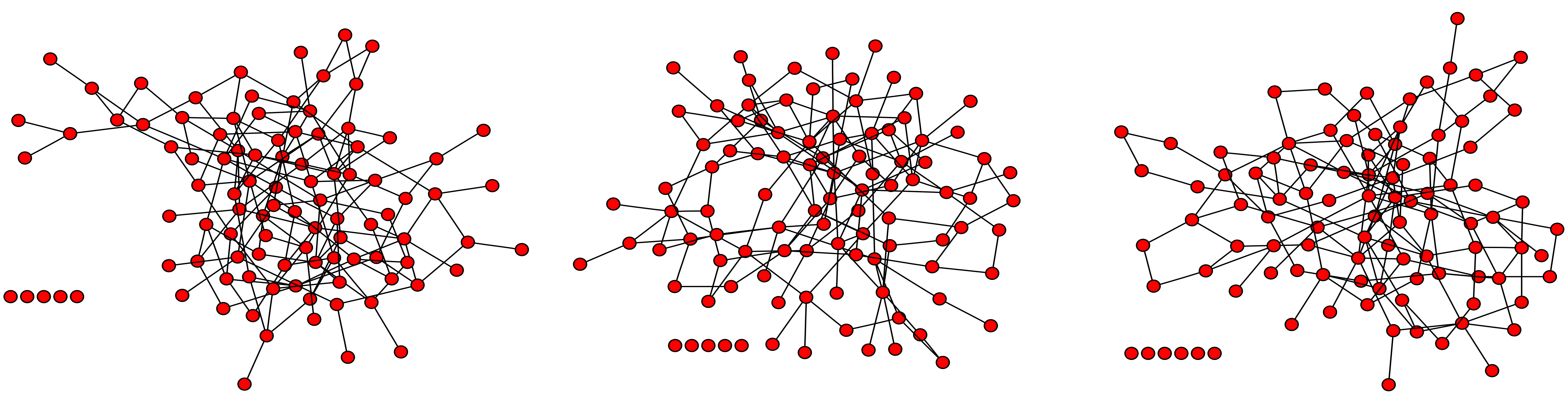
n=10
p=1/6

ER Random Graphs

p=1/6
N=12



p=0.03
N=100

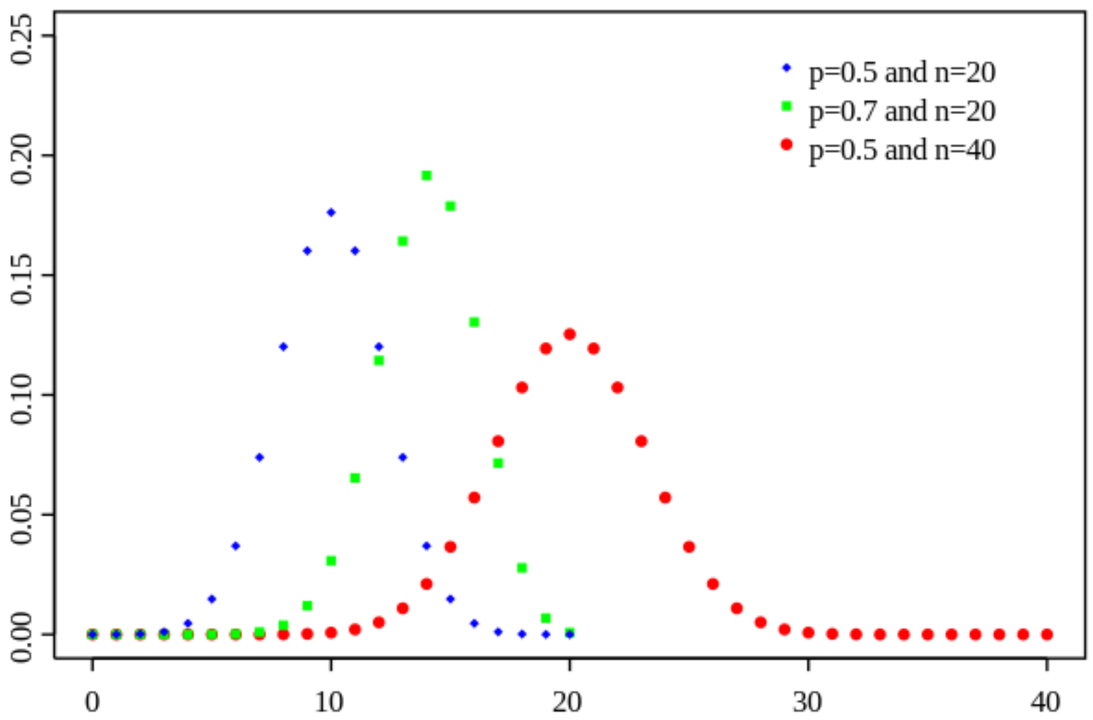


DESCRIBING ER RANDOM GRAPHS

Reminder

Binomial distribution:

Discrete probability distribution of the number of successes(**x**) in a sequence of **N** independent experiments, with success probability **p**



(PMF)

$$P(x) = \binom{N}{x} p^x (1-p)^{N-x}$$

Binomial coefficient:

Number of ways, disregarding order, that **k** objects can be chosen from among **n** objects

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Reminder

Binomial distribution:

Discrete probability distribution of the number of successes(**x**) in a sequence of **N** independent experiments, with success probability **p**

Properties of Binomial distribution

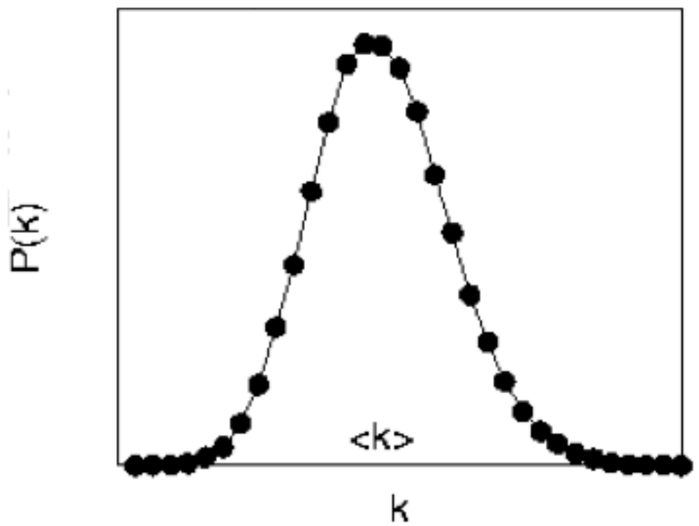
PMF $$P(x) = \binom{N}{x} p^x (1-p)^{N-x}$$

Mean $$< x > = pN$$

variance $$\sigma^2 = Np(1-p)$$

Degree distribution - Random Graphs

- *G(n,p)*



For each node,
independent probabilities to take each neighbor
=> Binomial distribution of degrees

P(k): probability to have exactly **k** links among **n-1** (total # of other nodes), with **p** the probability to have an edge

$$P(k) = \binom{n-1}{k} p^k (1-p)^{(n-1)-k}$$

Characteristics:

$$< k > = p(n-1)$$

$$\sigma_k^2 = p(n-1)(1-p)$$

Degree distribution - Random Graphs

For large **n** and small **k** (p,L), we can approximate the degree distribution using a poisson distribution of parameter (mean) $\lambda = <k>$

Poisson distribution

$$P(K) = \frac{\lambda^K e^{-\lambda}}{K!}$$

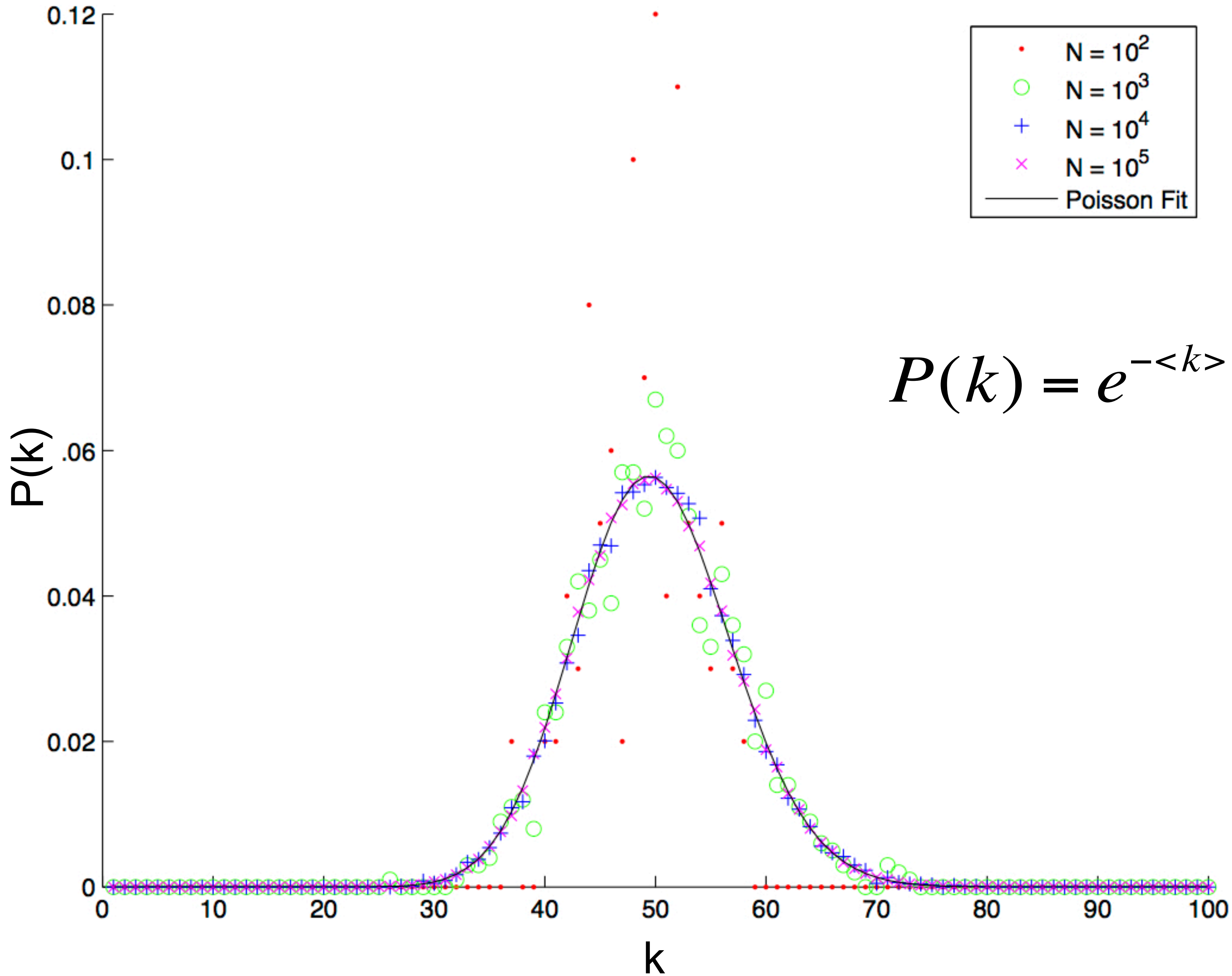
Distribution of degrees

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

standard deviation

$$\sigma = \sqrt{<k>}$$

Degree distribution - Random Graphs



$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

Degree distribution - Random Graphs

standard deviation $\sigma = \sqrt{<k>}$

$$\frac{\sigma}{<k>} = \frac{\sqrt{<k>}}{<k>}$$

High confidence to have degrees close to average degrees as degrees increase

Degree distribution - Random Graphs

Conclusion: degree distribution is **not**

-Heterogeneous

-Long tail

-Scale free

Clustering - Random Graphs

Local clustering of a node

Reminder, clustering coefficient

$$C_i \equiv \frac{2n_i}{k_i(k_i-1)}$$

where n_i is the number of links between the neighbours of node i

- Edges are independent and have the same probability p

$$n_i \cong p\frac{k_i(k_i-1)}{2}$$

possible links btw neighbors

$$p = \frac{<k>}{n-1}$$

$$C_i = \frac{2<k>}{n-1} \frac{k_i(k_i-1)}{2} \frac{1}{k_i(k_i-1)} = \frac{<k>}{n-1} = p$$

- For fixed average degree C is decreasing as N goes large

➡ Low clustering coefficient

➡ It is vanishing with the system size

Clustering - ER Random Networks

- **Small clustering coefficient** $C_i \equiv \frac{1}{N} < k >= p$

Real-world networks

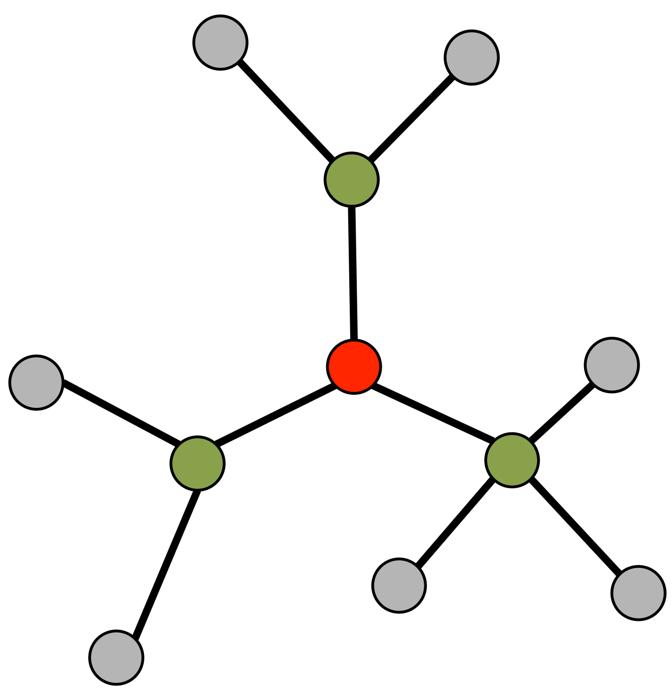
Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.*, 2001
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.*, 2001b
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998

Albert, R. et.al. Rev. Mod. Phy. (2002)

Distance - ER Random Graphs - Intuition

low clustering coefficient=>

Random graphs tend to have a tree-like topology with almost constant node degrees.



- nr. of first neighbors: $N(u)_1 = <k>$
- nr. of second neighbors: $N(u)_2 = <k>^2$
- nr. of neighbours at distance d: $N(u)_d = <k>^d$

Intuition: At which distance are all nodes reached?

$$n = <k>^d \Rightarrow \log_{<k>} n = d \Rightarrow d = \frac{\log n}{\log <k>}$$

Diameter, avg. distance in $\mathcal{O}(\log n)$

Distance - ER Random Graphs

- **Logarithmically short distance**

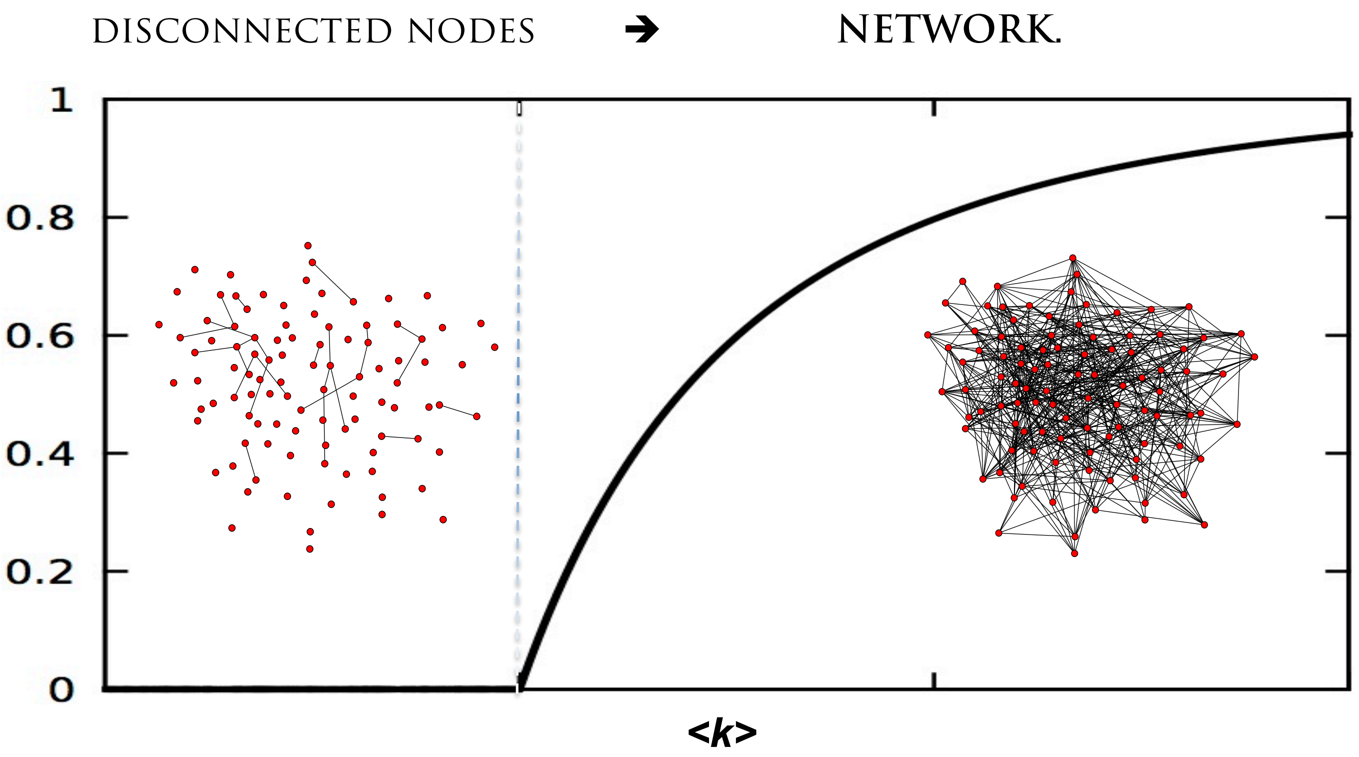
$$d = \frac{\log n}{\log\langle k\rangle}$$

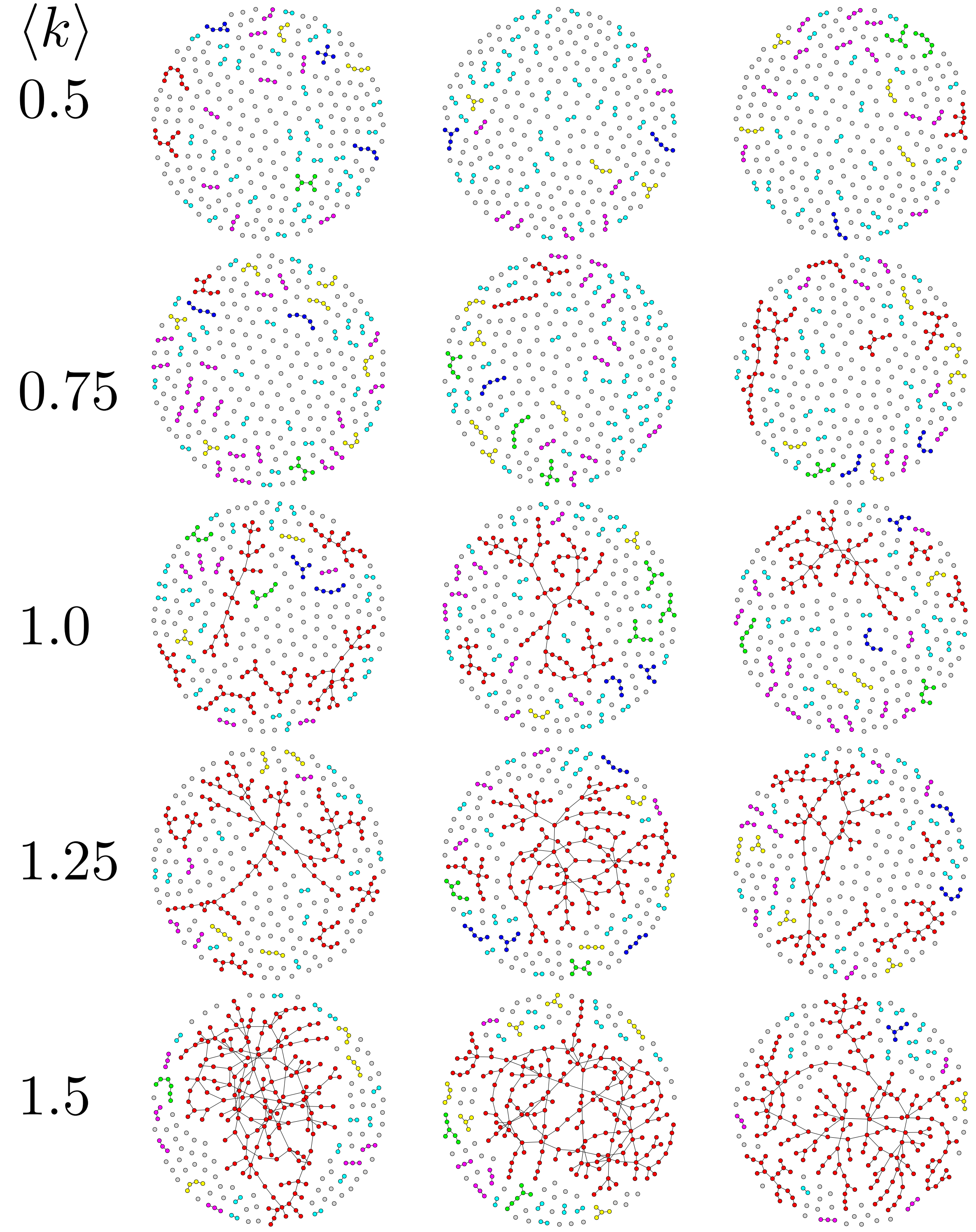
Real-world networks

Network	Size	$\langle k\rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.*, 2001
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.*, 2001b
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998

Albert, R. et.al. Rev. Mod. Phy. (2002)

Phase transition in connected components





- Network structure goes through a transition
- Question: How and when does this transition happen

Connected components of Random Graphs

https://www.complexity-explorables.org/explorables/the-blob/

ER Random Network - catch up

Network	Degree distribution	Path length	Clustering coefficient
Real world networks	broad	short	large
Regular lattices	constant	long	large
ER random networks	Poissonian	short	small

It is not capturing the properties of any real system

BUT

it serves as a reference system for any other network model

Configuration model

More details at [http://tuvalu.santafe.edu/~aaronc/courses/5352/fall2013/csci5352_2013_L11.pdf]

Random graphs with specified degrees

Problem

- The ER Random Graph model has a Poisson degree distribution
- Most real-world networks have heavy-tailed degree distributions
- **We need to generate networks having pre-determined degrees or degree distribution, but maximally random otherwise**
- The observed properties (clustering coefficient, etc.) might be due *only* to the difference in degree distribution

Random graphs with specified degrees

Configuration model

Based on an observed network

- Defined as $G(n, \vec{k})$ where $\vec{k} = \{k_i\}$ is a degree sequence on n nodes, with k_i being the degree of node i

Ad hoc degree distribution

- **The degree sequence $\vec{k} = \{k_i\}$ can be sampled from a probability distribution**
  - Delta/Dirac function => Random regular graph
  - Poisson => Similar to ER for proper parameters
  - Scale-free => Power-law random graph
- Only global condition to satisfy is: $\sum_i k_i$ **mod** $2 = 0$

(even degree sum) i.e. each edge has to have ending nodes

Random graphs with specified degrees

Configuration model

How much of some observed pattern is driven by the degrees alone?

Exact or approximate degree distribution

- The model can preserve the **expected** degree sequence, or the **exact** degree sequence
  - Chung-lu (appoximate)
  - Molloy-reed (Exact)

Random graphs with specified degrees

Chung-Lu model for configuration networks = Approximate degree distribution

- Probabilistic model which produce a network with degrees approximating (on average) the original degree
- It is a *"coin-flipping"* process as ER model but **the probability that two nodes i and j are connected depends on the degree k_i and k_j of the ending nodes**
- From the point of view of node i with degree k_i, the probability that one of its edges will connect to j with k_j:

$$k_j/2m$$

- This can happen via k_i links, thus the probability that they are connected:

$$p_{ij} = \frac{k_i k_j}{2m}$$

assuming that: $[\max(k_i)]^2 < 2m$

(/!\ inconsistent probability, it is rather expected number of edges)

- Chung-Lu model takes each pairs of nodes and connects them with this probability

$$\forall_{i>j} \qquad A_{ij} = A_{ji} = \begin{cases} 1 & \text{with probability } p_{ij} \\ 0 & \text{otherwise} \end{cases}$$

Random graphs with specified degrees

Chung-Lu model for configuration networks = Approximate degree distribution

$$\forall_{i>j} \qquad A_{ij} = A_{ji} = \begin{cases} 1 & \text{with probability } p_{ij} \\ 0 & \text{otherwise} \end{cases} \qquad \text{where} \qquad p_{ij} = \frac{k_i k_j}{2m}$$

- Each pairs of nodes are considered once, thus it produces a **simple graph** (without self-loops and multi edges)

- Degree of a node equals only in "expectation" to the originally assigned degree

Complexity:

- $O(n^2)$: We need $n(n-1)$ flips to test all node pairs

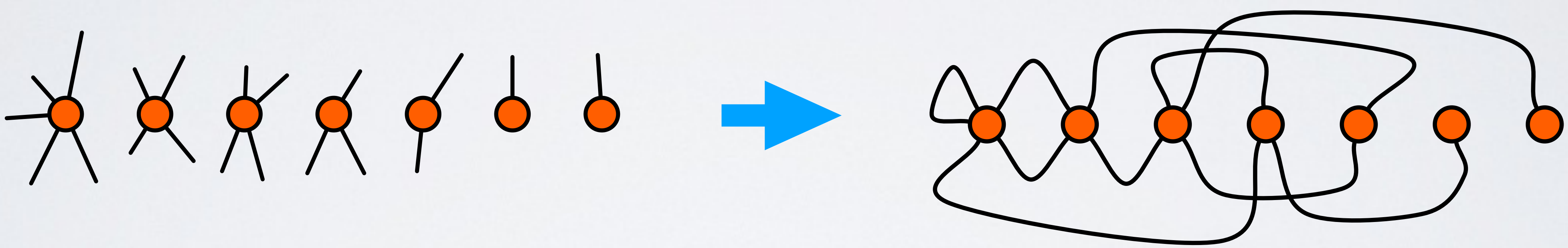
EXPENSIVE!

Random graphs with specified degrees

Molloy-Reed model for configuration networks = exact degree preservation

Original idea:

1. Given a degree sequence $\vec{k} = \{k_1, k_2, \ldots, k_n\}$
2. Assign each node $i \in V$ with k_i number of stubs
3. Select random pairs of unmatched stubs and connect them
4. Repeat 3 while there are unmatched stubs



- This process will produce a configuration model with exact degree sequence
- Possible to select multiple times stubs of the same pair of nodes ➜ **Multilinks**
- Possible to select the stubs of the same node to connect ➜ **Self-links**

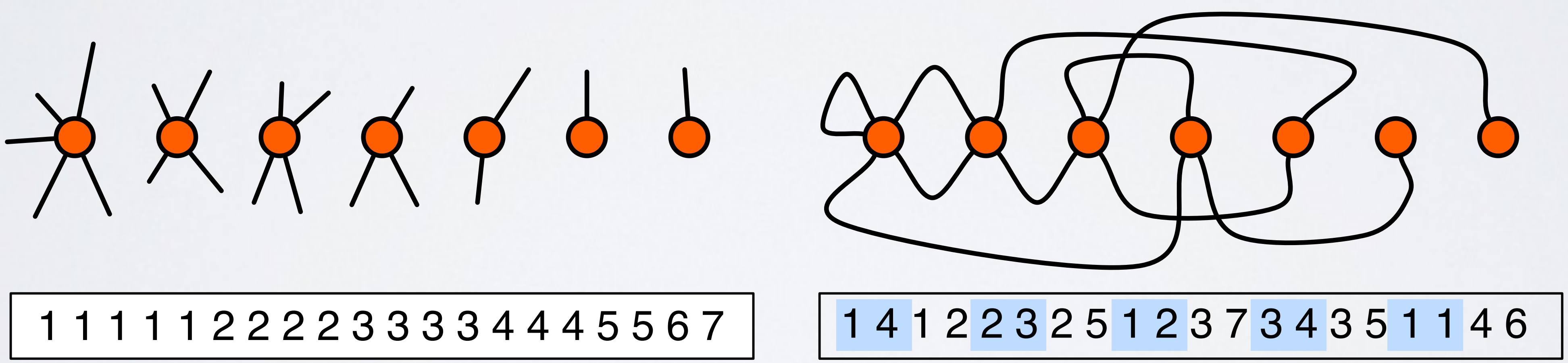
The obtained graph is not simple…but the density of multi and self-links ➜ 0 as N ➜ ∞

Random graphs with specified degrees

Molloy-Reed model for configuration networks = exact degree preservation

An effective algorithm:

1. Take an array $\vec{v}$ with length $2m$ and fill it with exactly k_i indices of each node $i \in V$
2. Make a random permutation of the array $\vec{v}$
3. Read the content of the array in an order and in pairs
4. Pairs of consecutive node indices will assign links in the configuration network



Complexity:

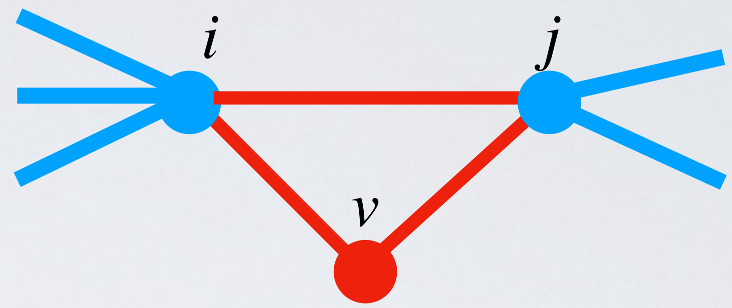
- *O(m)*: Random permutation of an array
- *O(m log m)*: assigning uniformly random variables to indices and quick-sort them

CHEAP!

More details at [http://tuvalu.santafe.edu/~aaronc/courses/5352/fall2013/csci5352_2013_L11.pdf]

Configuration model - mathematical properties

Expected clustering coefficient



It is the average probability that two neighbours of a vertex are connected

- Start at some vertex v (with degree $k \geq 2$)
- Choose a random pair of its neighbours i and j
- The probability that i and j are themselves connected is $k_i k_j / 2m$
- *But probabilities to encounter some degrees as neighbors depends on their degree: more complex than simply counting frequency of degrees (friendship paradox)*

Clustering coefficient

independent of network size

$$C = \ldots = \frac{1}{n} \frac{[\langle k^2 \rangle - \langle k \rangle]^2}{\langle k \rangle^3}$$

- It is a vanishing quantity $O(1/n)$ as long as the second moment is finite (not power law)

For details, see: http://tuvalu.santafe.edu/~aaronc/courses/5352/fall2013/csci5352_2013_L12.pdf

ER Random Network - catch up

Network	Degree distribution	Path length	Clustering coefficient
Real world networks	broad	short	large
Regular lattices	constant	long	large
ER random networks	Poissonian	short	small
Configuration Model	Custom, can be broad	short	small

Watts-Strogatz model of small-world networks

Small-world networks

- On of the founding papers of Network Science…

D.J. Watts and S. Strogatz,

"Collective dynamics of 'small-world' networks", Nature 393, 440–442, 1998

Contradiction: Real-world networks have

High clustering coefficient

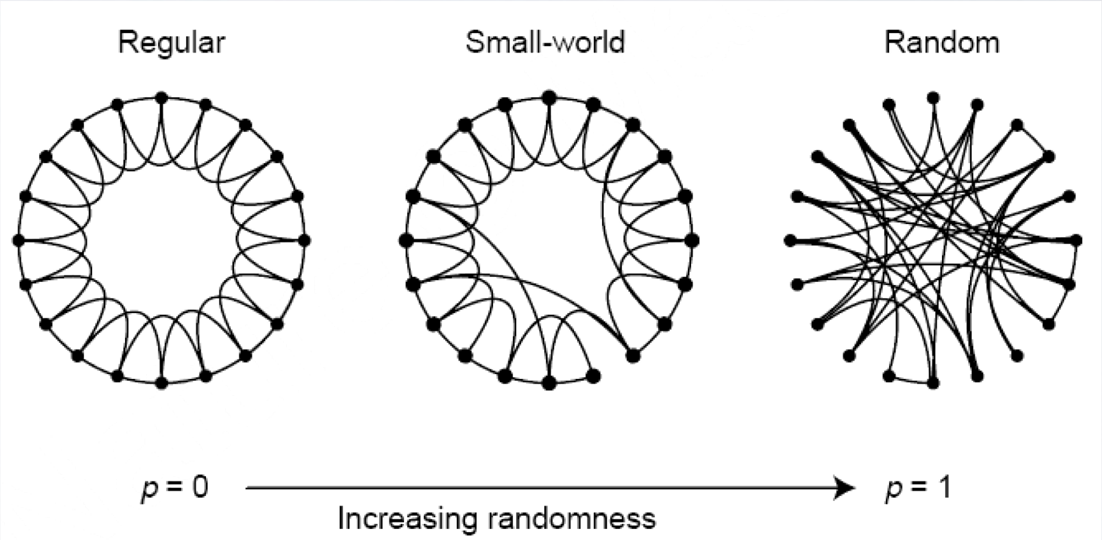
AND

Short distances

The Watts-Strogatz model

A model to capture large clustering coefficient and short distances observed in real networks

- It interpolates between an ordered finite lattice and a random graph
- Fixed parameters:
    - n - system size
    - K - initial coordination number
- Variable parameters:
    - p - rewiring probability



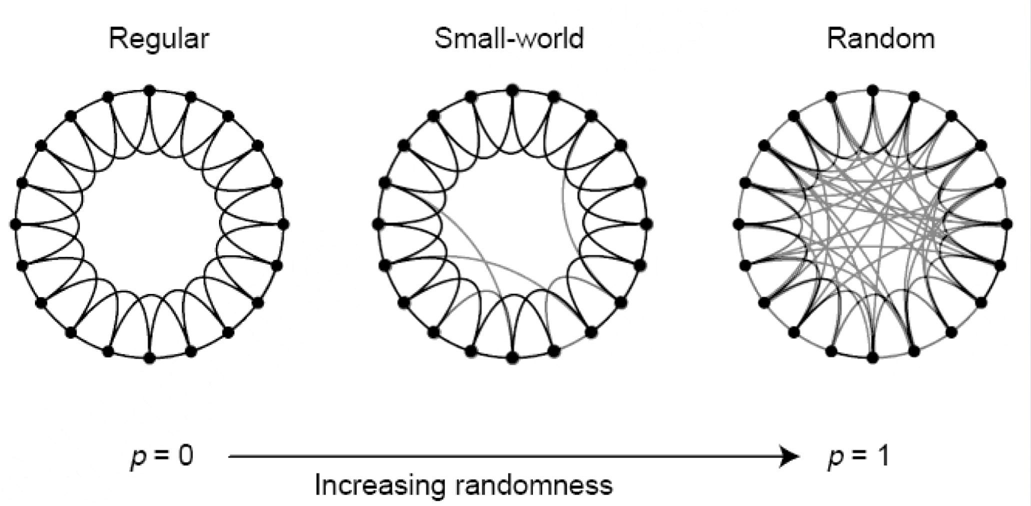
D.J. Watts and S. Strogatz, Nature (1998)

- Algorithm:
    1. Start with a ring lattice with n nodes in which every node is connected to its first K neighbours ($K/2$ on either side).
    2. Randomly rewire each edge of the lattice with probability p such that self-connections and duplicate edges are excluded.

By varying p the network can be transformed from a completely ordered ($p=0$) to a completely random ($p=1$) structure

The Watts-Strogatz model

(Global) Clustering coefficient (Definition 2)



- *p=0* - regular ring with constant clustering: $C = \frac{3(K-2)}{4(K-1)}$
  - *0 ≤ C ≤ 3/4*
  - Independent of *n*

- *p>0* - we can count triangles and tuples

Global clustering coefficient

$$C = \frac{\frac{1}{4}NK(\frac{1}{2}K-1)\times 3}{\frac{1}{2}NK(K-1)+NK^2p+\frac{1}{2}NK^2p^2} = \frac{3(K-2)}{4(K-1)+8Kp+4Kp^2}$$

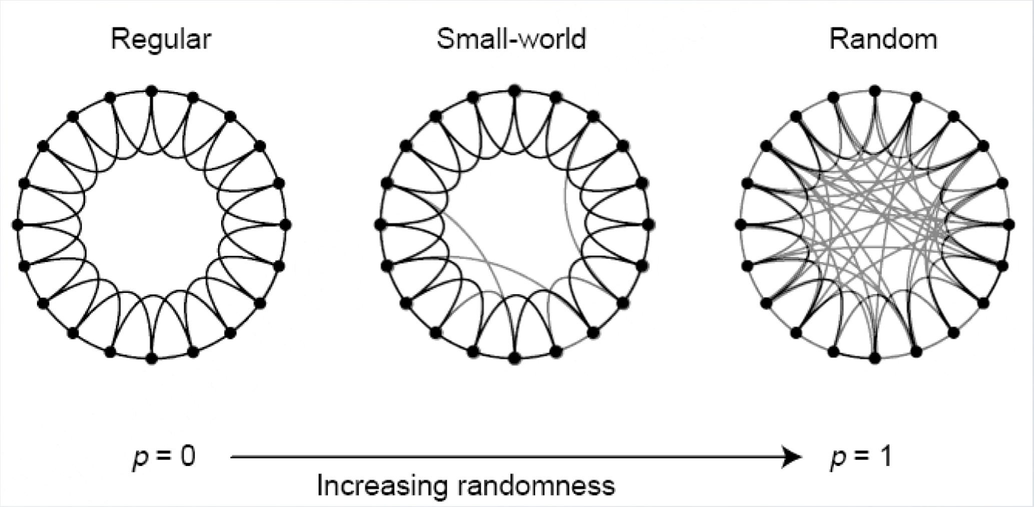
- Independent of n
- if p→0 it recovers the ring value
- if p→1 , small

For details, see: http://tuvalu.santafe.edu/~aaronc/courses/5352/fall2013/csci5352_2013_L12.pdf

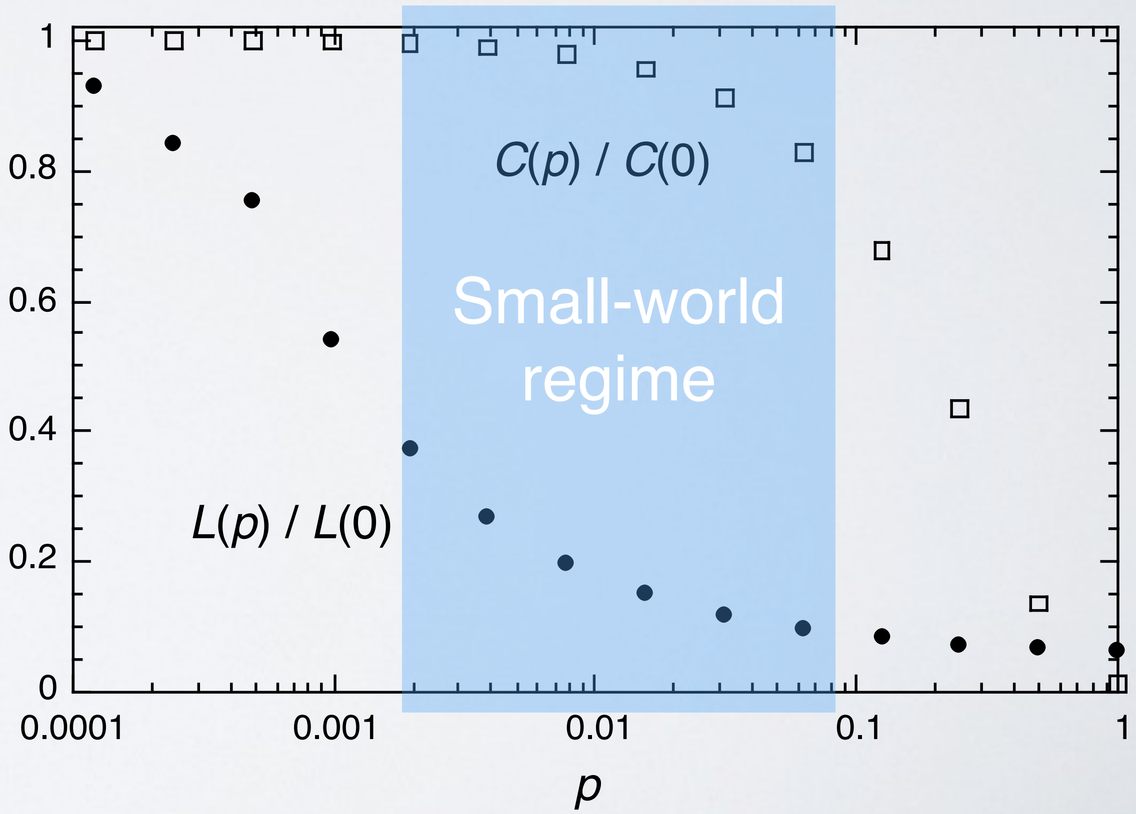
The Watts-Strogatz model

Average path length (Definition 2)

- No closed form solution



- From numerical simulations:



- See Newman, M. E. (2000). Models of the small world. *Journal of Statistical Physics*, *101*(3-4), 819-841.

for details

L=avg path length

ER Random Network - catch up

Network	Degree distribution	Path length	Clustering coefficient
Real world networks	broad	short	large
Regular lattices	constant	long	large
ER random networks	Poissonian	short	small
Configuration Model	Custom, can be broad	short	small
Watts & Strogatz (in SW regime)	Poissonian	short	large

Scale-free networks

Scale-free networks

A network is called ***Scale-free*** **when its degree distribution follows (to some extent) a Power-law distribution**

Power-law distribution:

(PDF)

$$P(k) \sim Ck^{-\alpha} = C\frac{1}{k^{\alpha}}$$

α (sometimes γ) called the **exponent** of the distribution

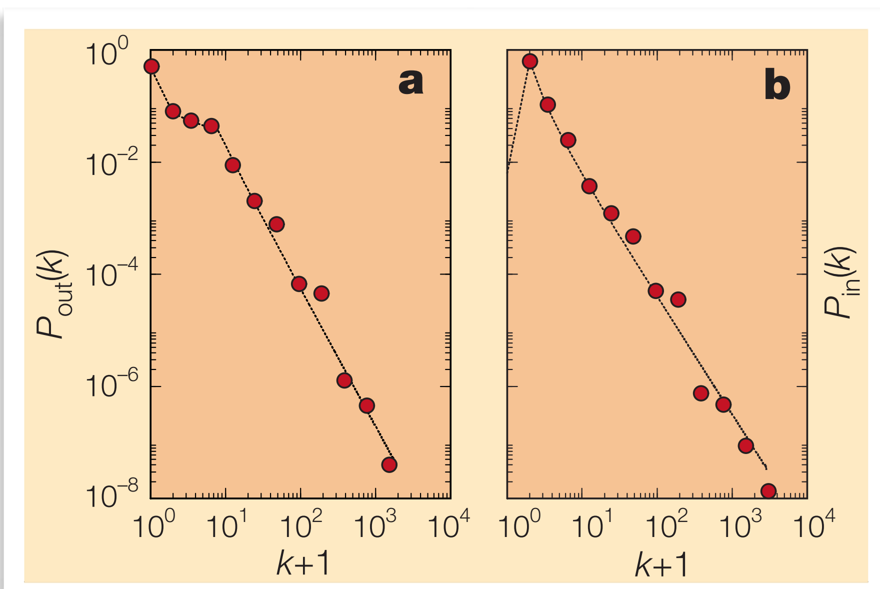
Positive values

Here, defined as continuous (approximation)

Scale-free networks - first observations

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

Diameter of the world wide web

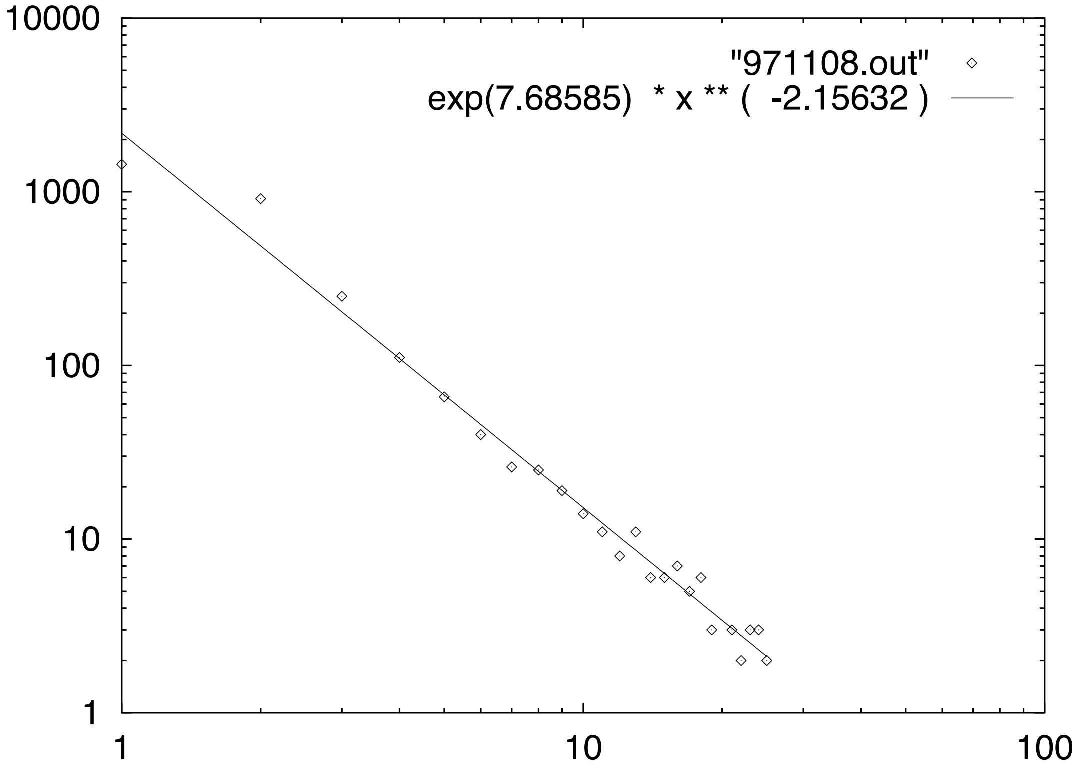


Power law
Appear as a line
On a log-log plot

Scale-free networks - other examples

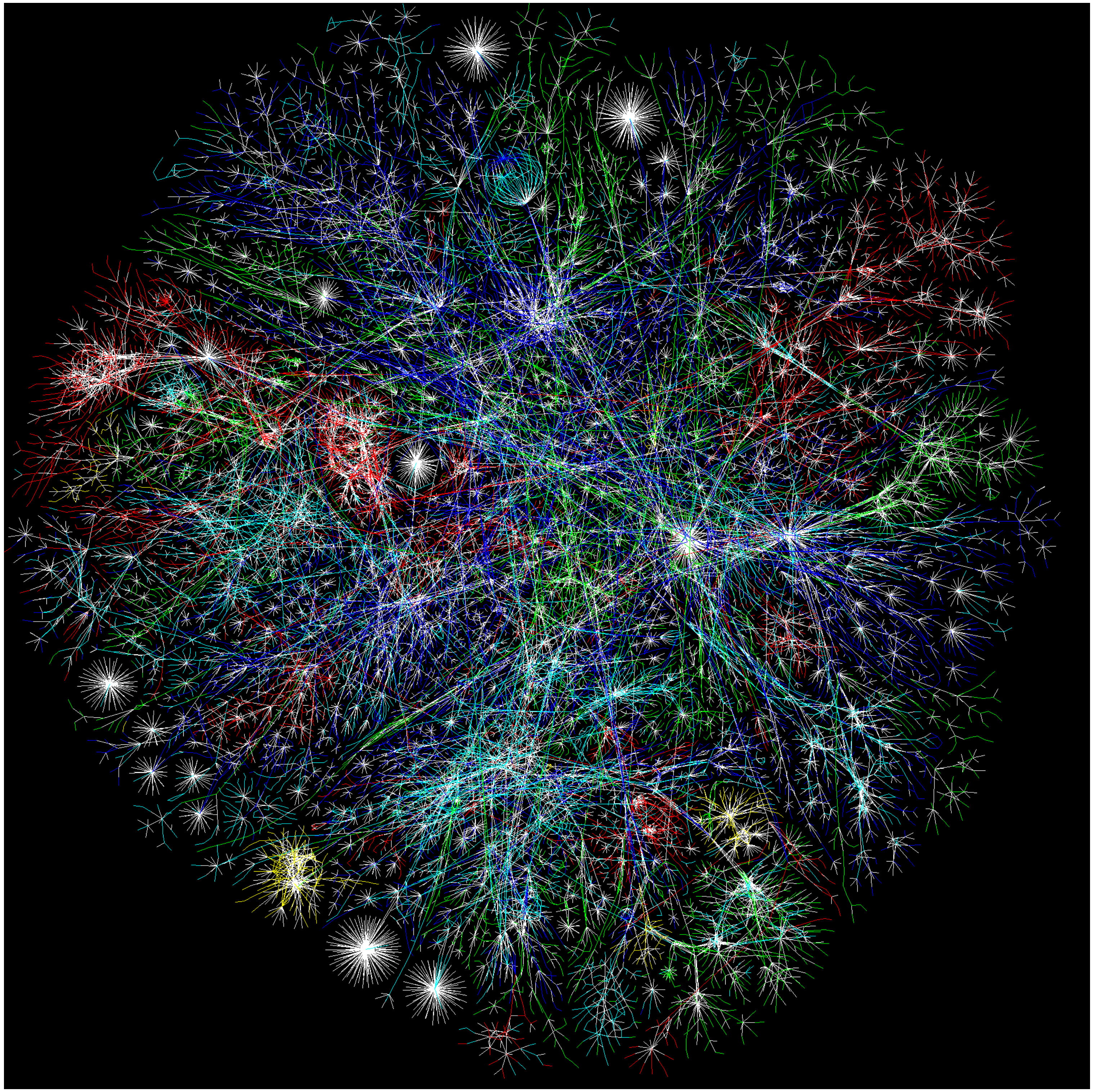
The internet

- Nodes: routers
- Links: Physical wires



(a) Int-11-97

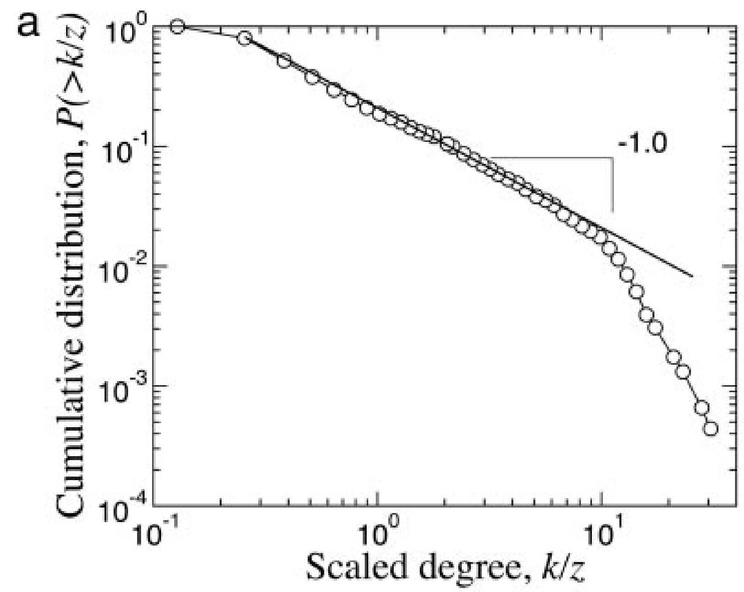
Faloutsos, Faloutsos and Faloutsos (1999)



Scale-free networks - other examples

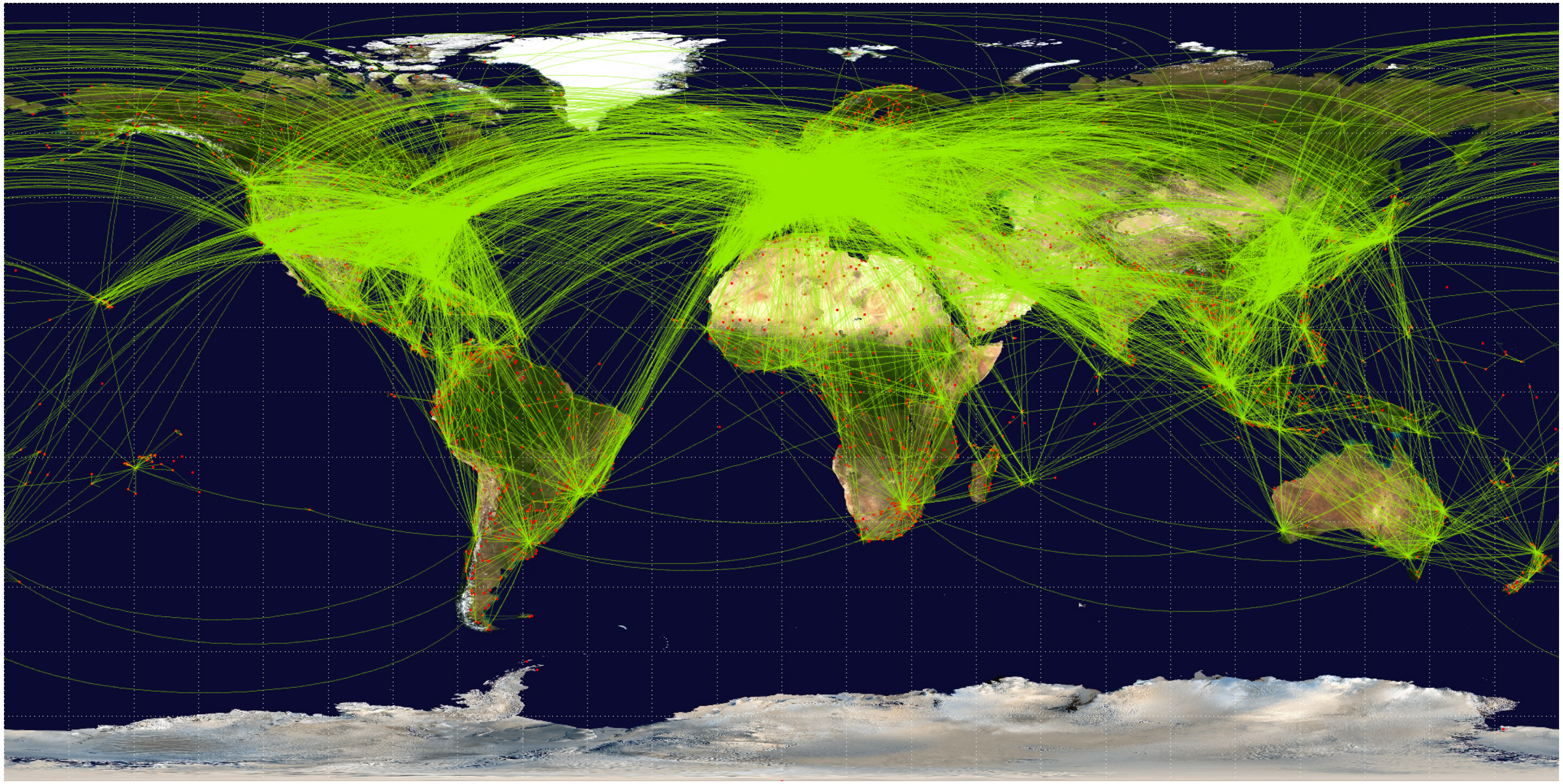
Airline route map network

- Nodes: airports
- Links: airplane connections



Guimera et.al. (2004)

Note: the cumulative distribution of a power law is also a line on a log-log plot

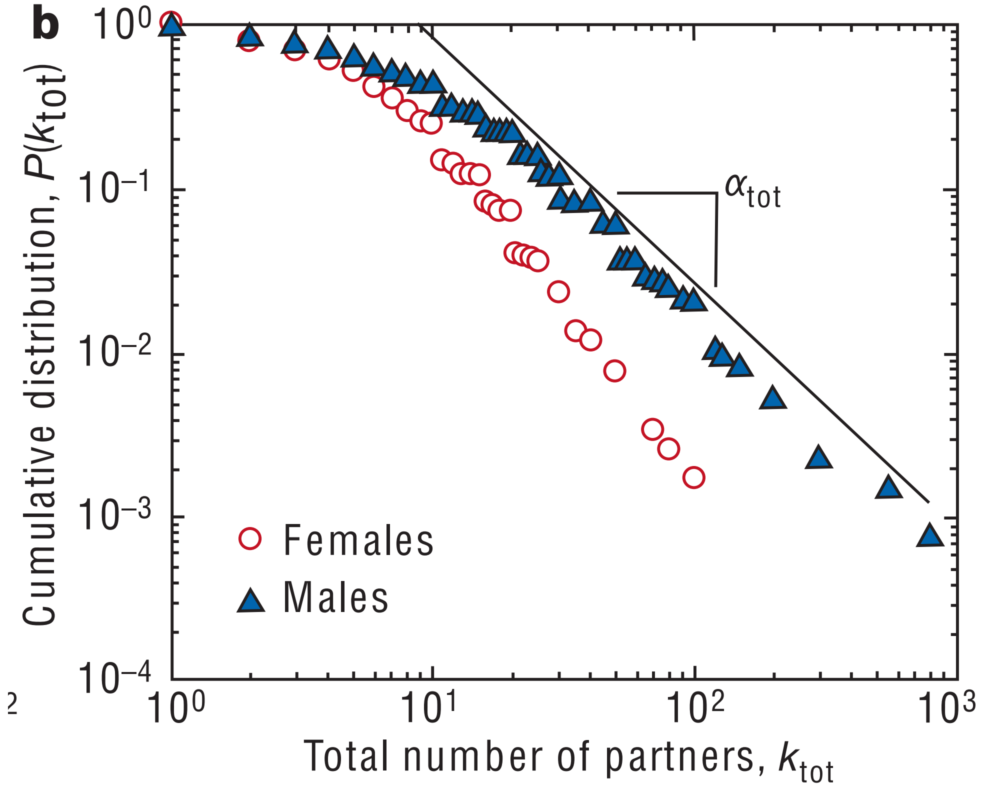


Scale-free networks - other examples

Sexual-interaction networks

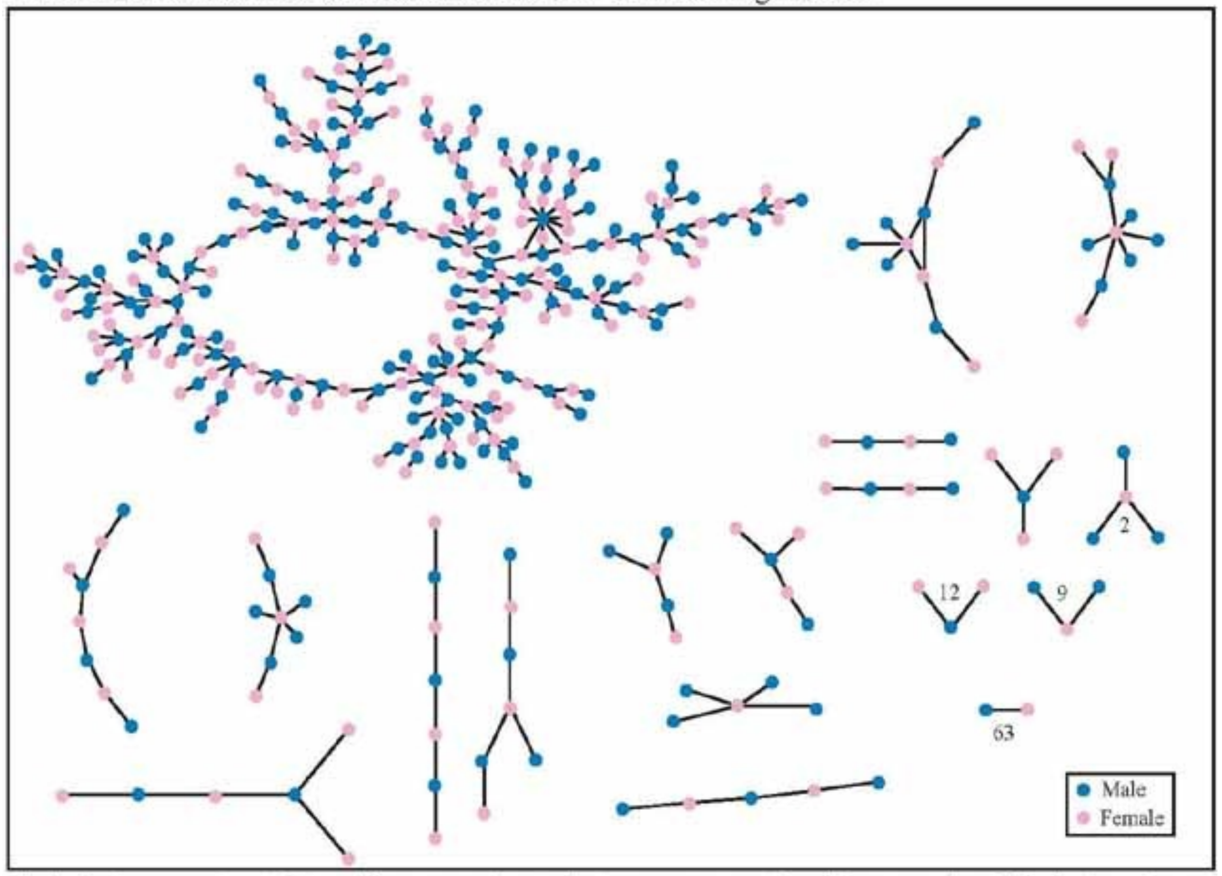
- Nodes: individuals
- Links: sexual incursion

Liljeros et.al. (2001)



Bearman et.al. (2004)

The Structure of Romantic and Sexual Relations at "Jefferson High School"

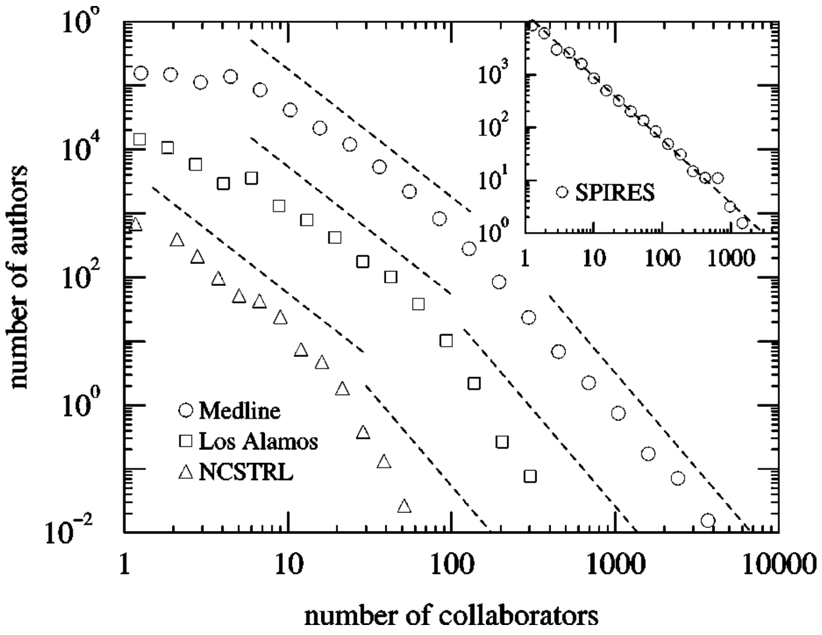


Each circle represents a student and lines connecting students represent romantic relations occuring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e. we found 63 pairs unconnected to anyone else).

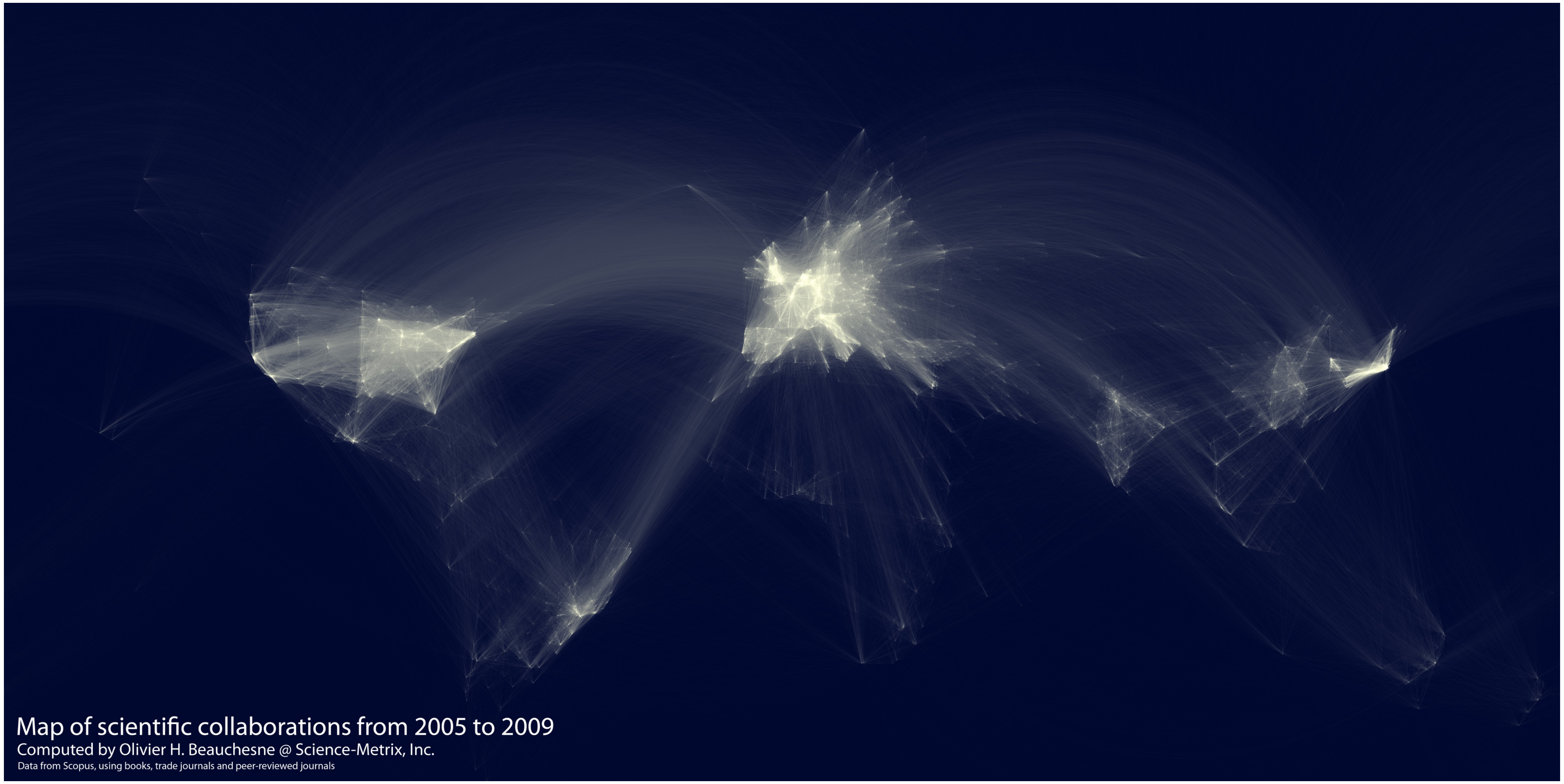
Scale-free networks - other examples

Scientific collaborations

- Nodes: scientists (here geo-localised)
- Links: common papers



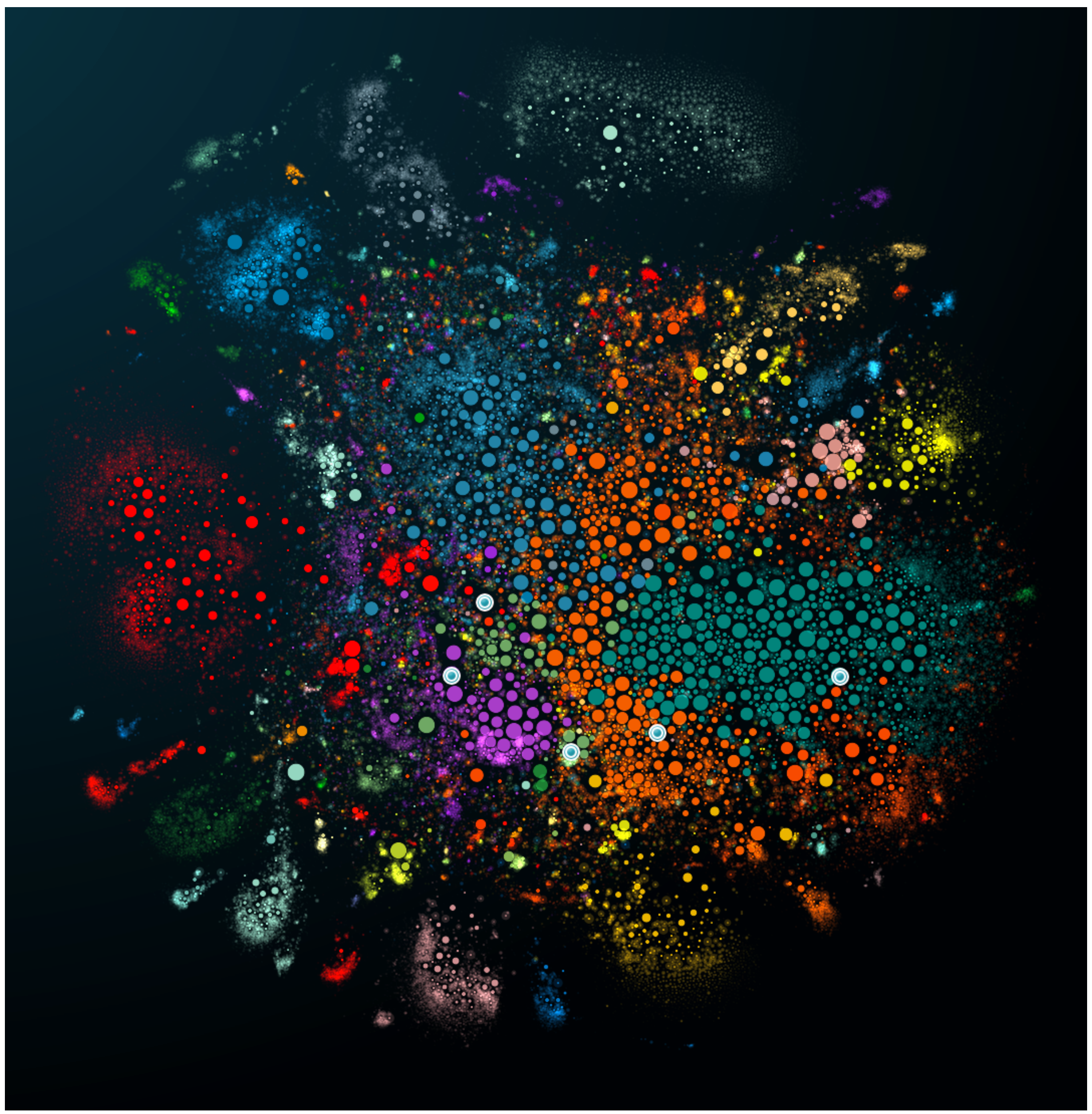
Newman (2001)



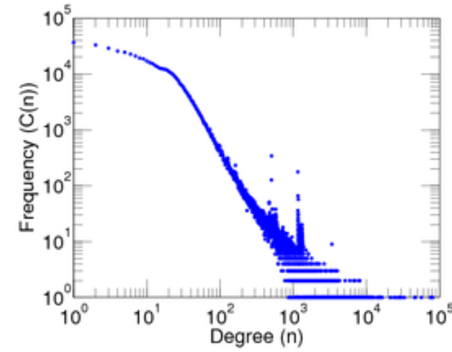
Scale-free networks - other examples

Online social networks

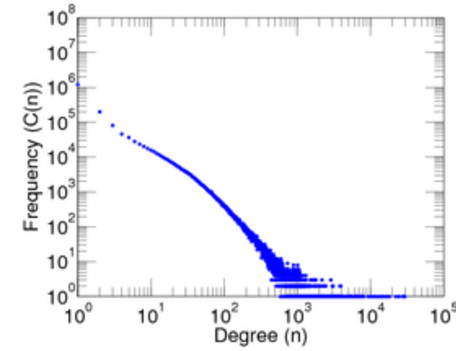
- Nodes: individuals
- Links: online interactions



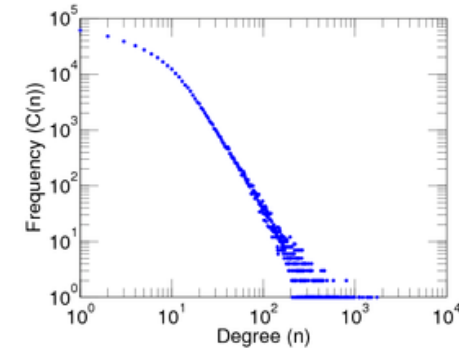
Social network of Steam
http://85.25.226.110/mapper



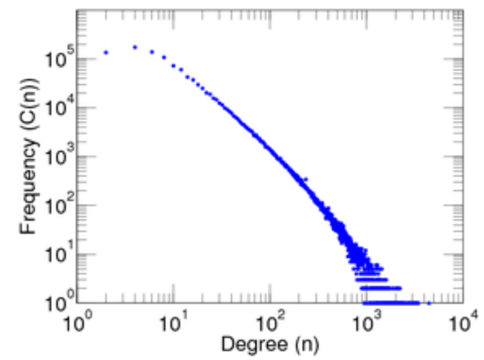
Catster/Dogster Familylinks/Friendships



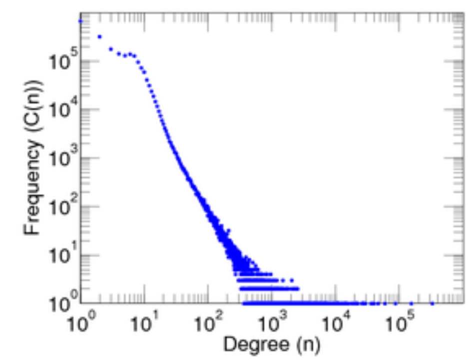
Chinese Wikipedia internal links



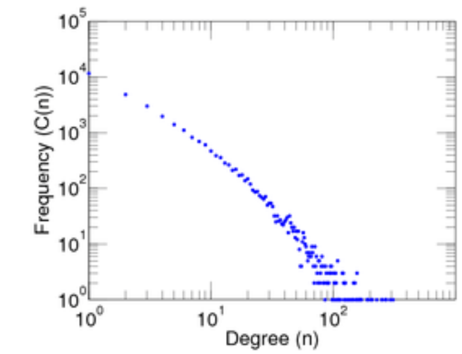
CiteSeer



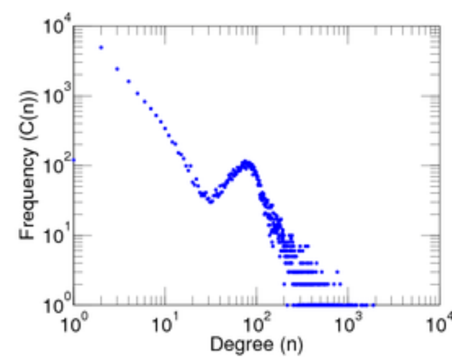
DBLP



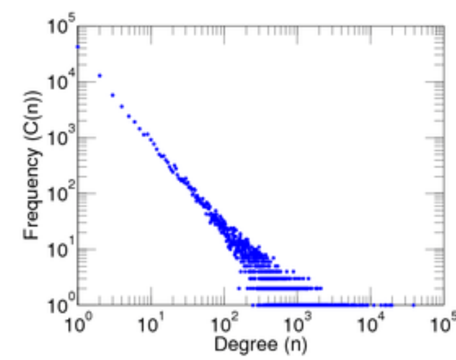
DBpedia



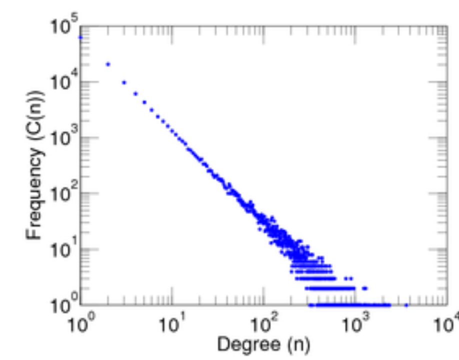
Digg



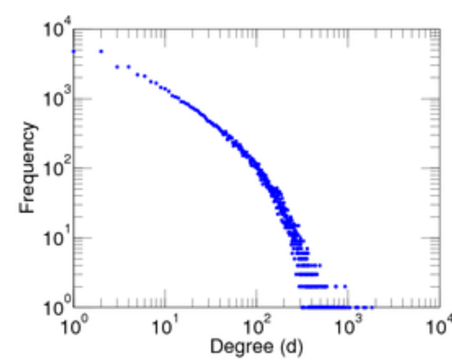
Edinburgh Associative Thesaurus



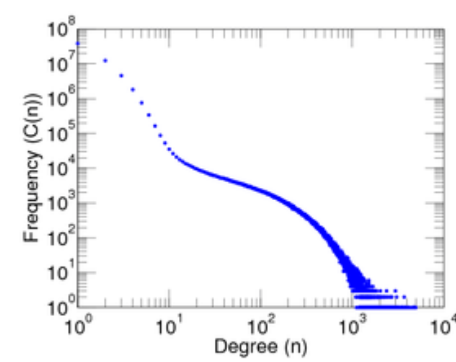
Enron



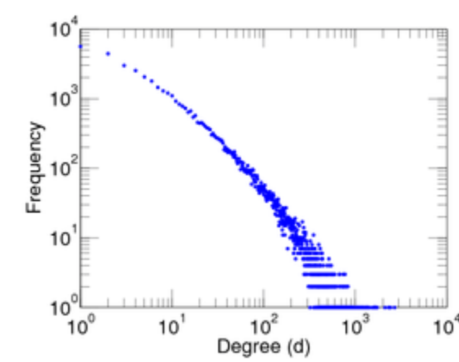
Epinions trust



Facebook friendships



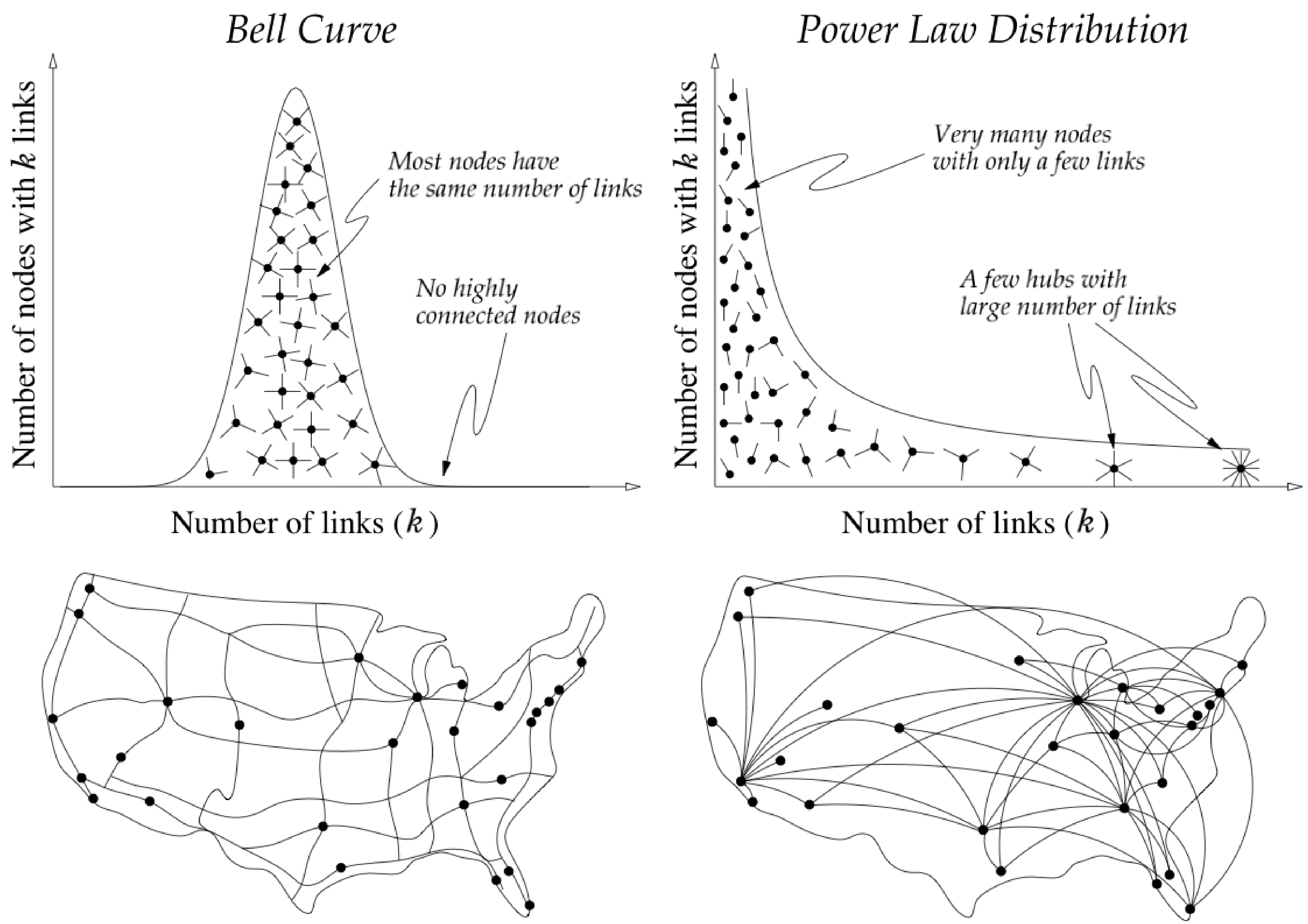
Facebook social graph



Facebook wall posts

Scale-free distribution

What does it mean?

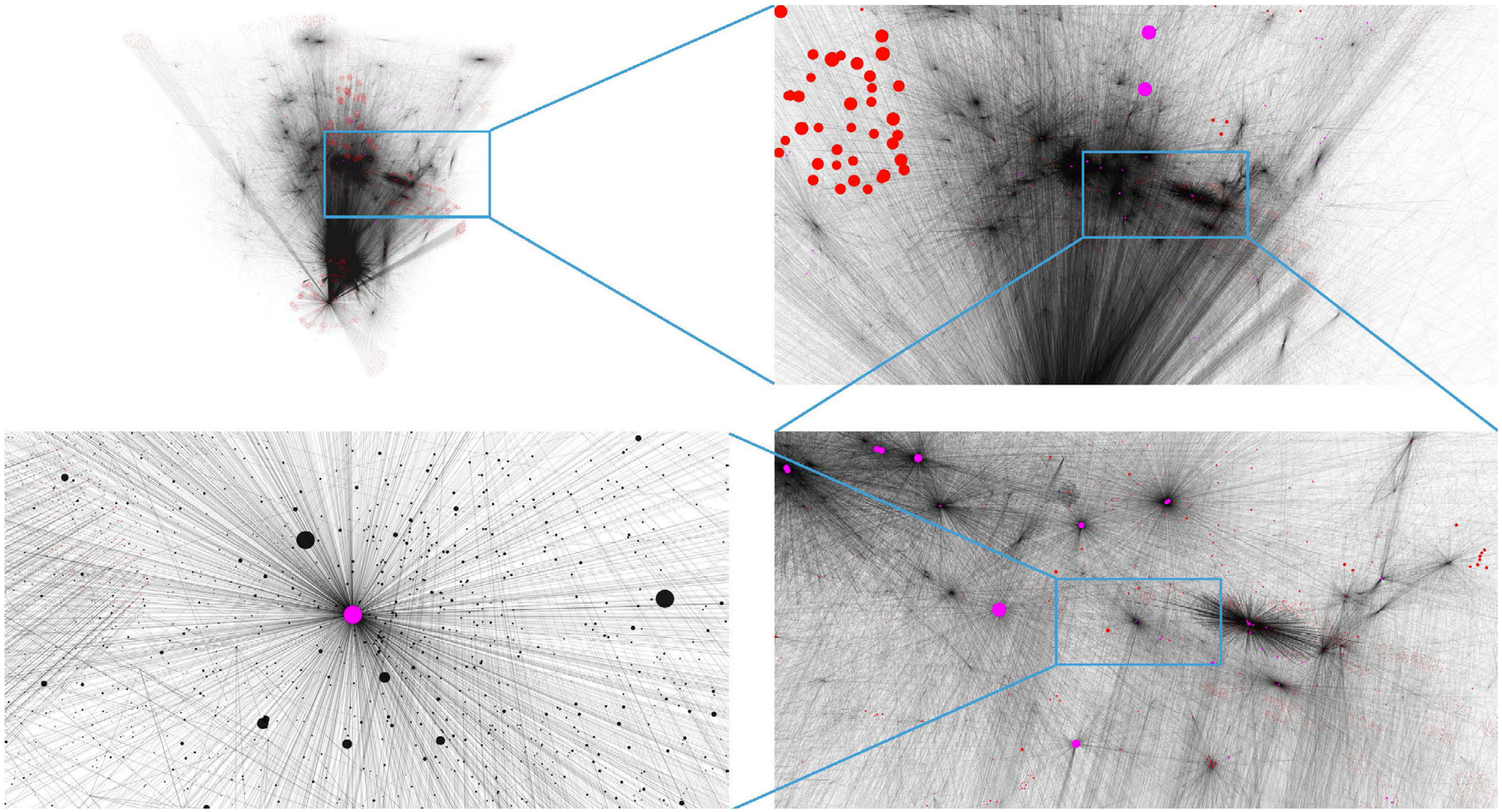


AL. Barabási, Linked (2002)

Degree fluctuations have no characteristic scale (scale invariant)

Scale-free networks

Idea of *scale free*



AL Barabási, Network Science Book (2013)

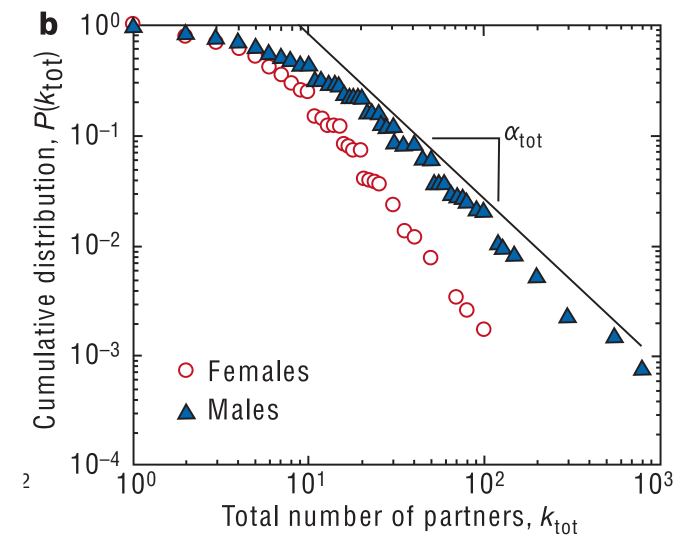
Proper definition

$$P(k) \sim Ck^{-\alpha}$$

$$C = \frac{1}{\int_{k_{\min}}^{\infty} k^{-\alpha} dk} = (\alpha - 1)k_{\min}^{\alpha-1}$$

$$P(k) = (\alpha - 1)k_{\min}^{\alpha-1}k^{-\alpha}$$

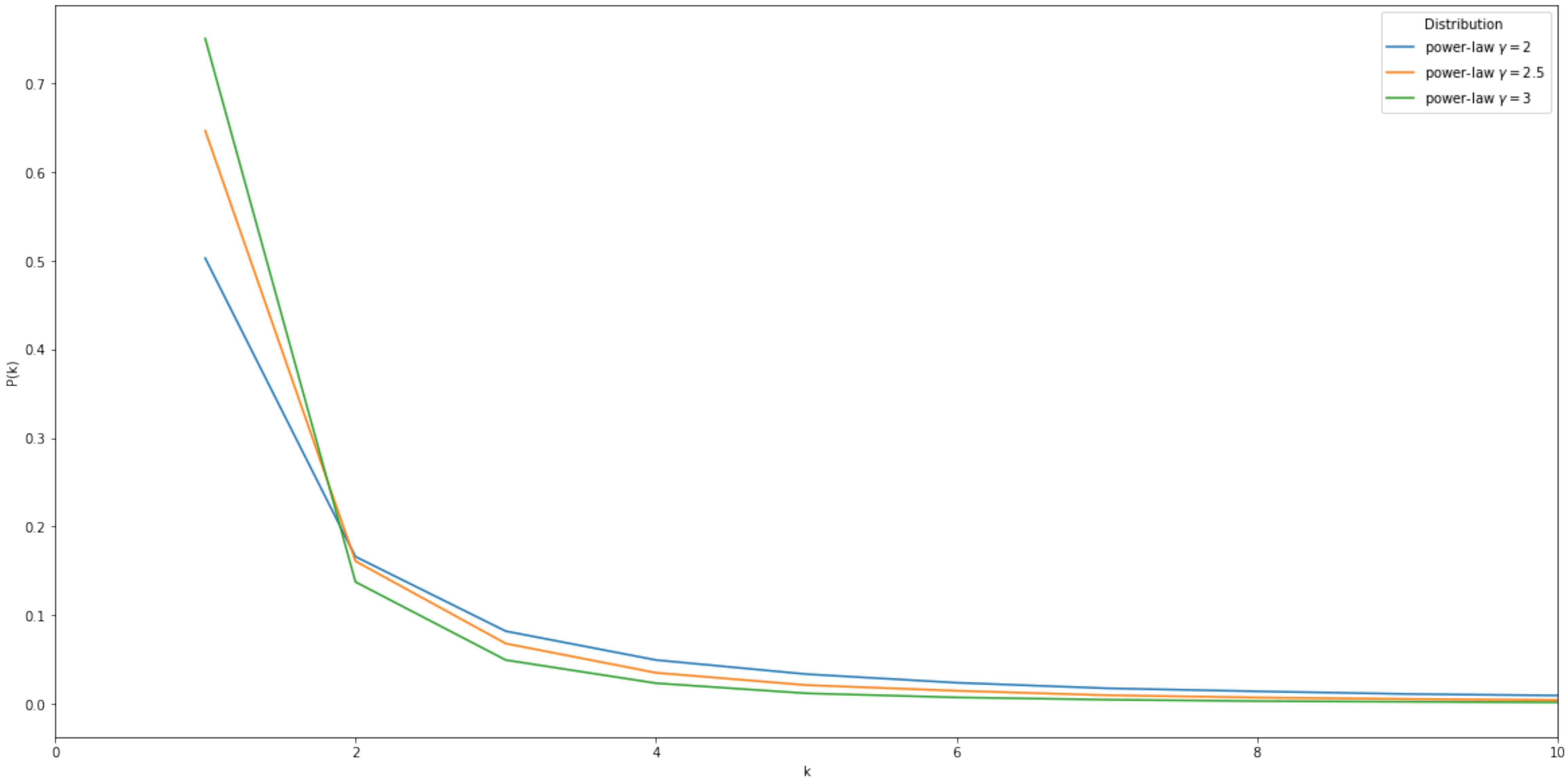
$$P(k) = \frac{\alpha - 1}{k_{\min}}\left(\frac{k}{k_{\min}}\right)^{-\alpha}$$



Scale-free networks

Power law plotted with a linear scale, for k<=10

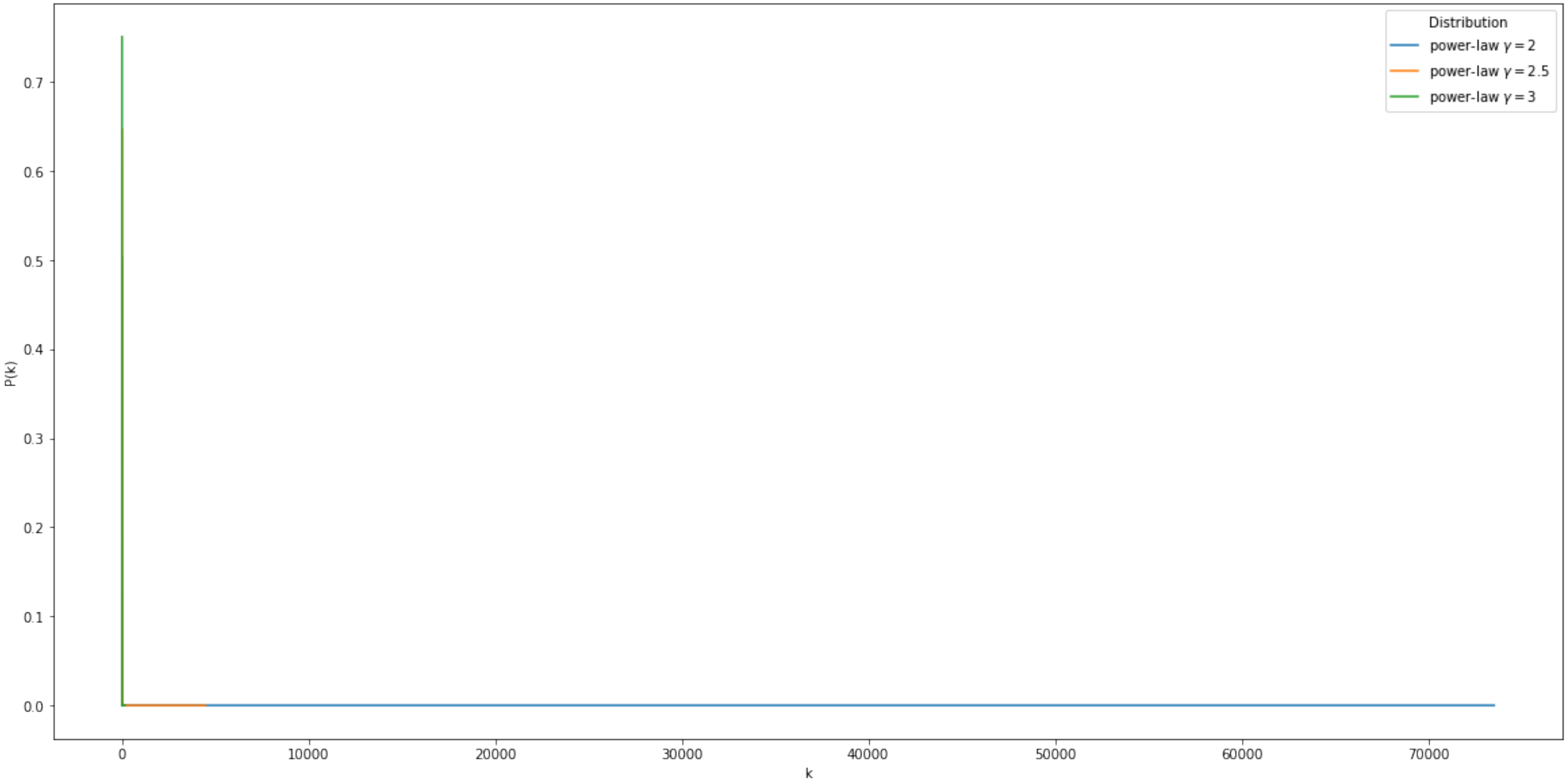
(100 000 samples)



Scale-free networks

Power law plotted with a linear scale, for k<100000

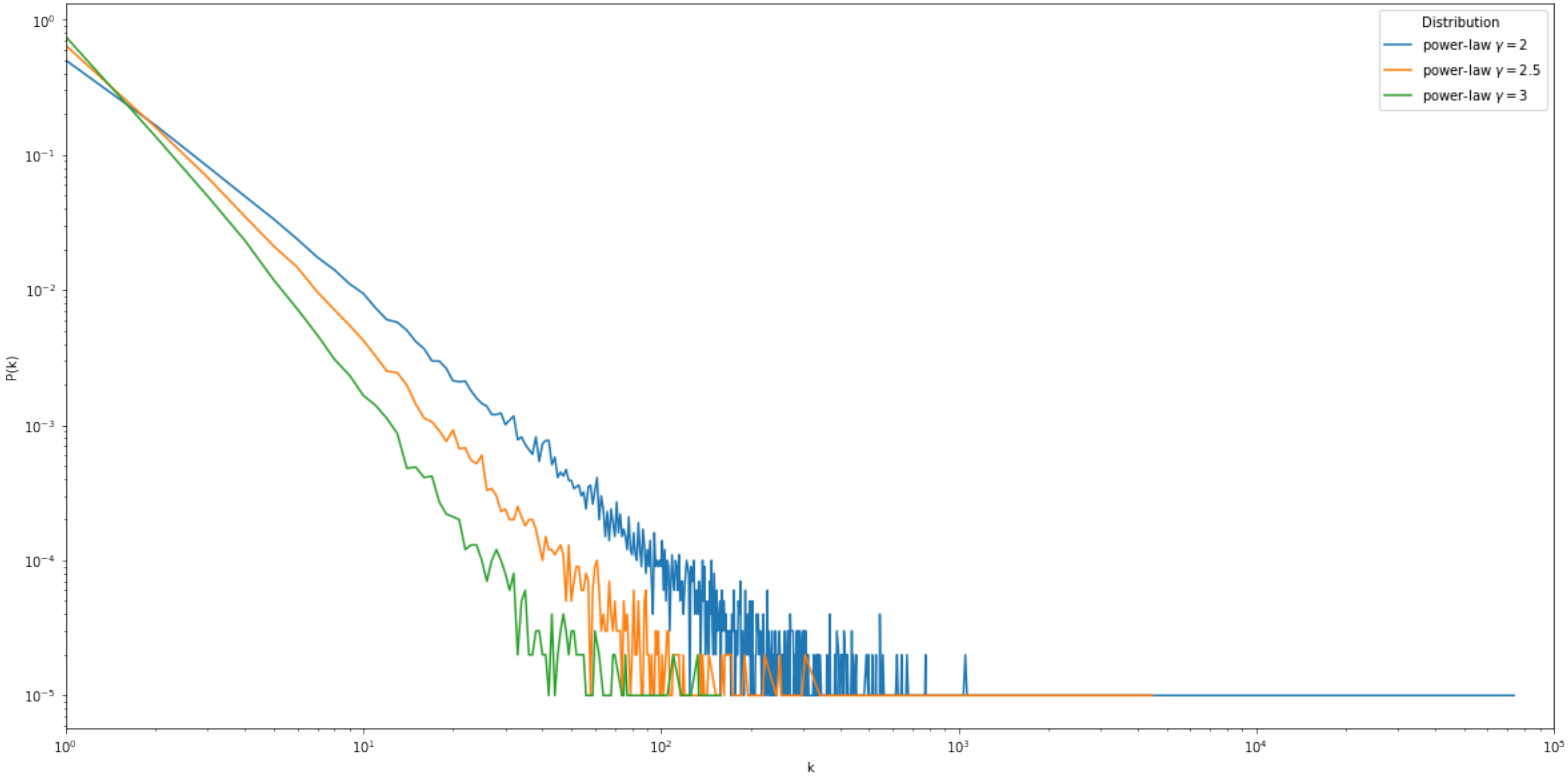
(100 000 samples)



Scale-free networks

Power law plotted with a log-log scale, for k<100000

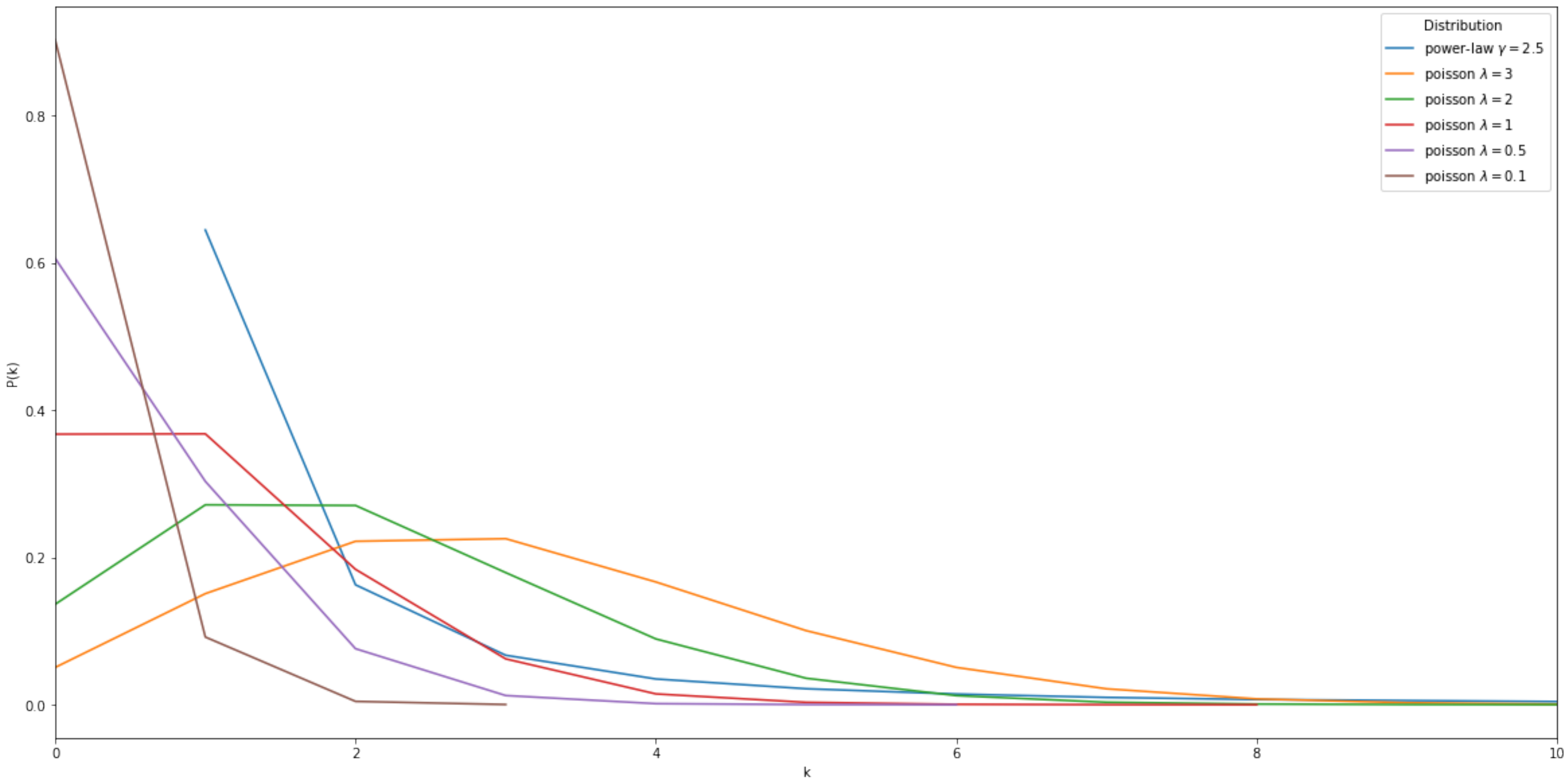
(100 000 samples)



Scale-free networks

Comparing a poisson distribution and a power law

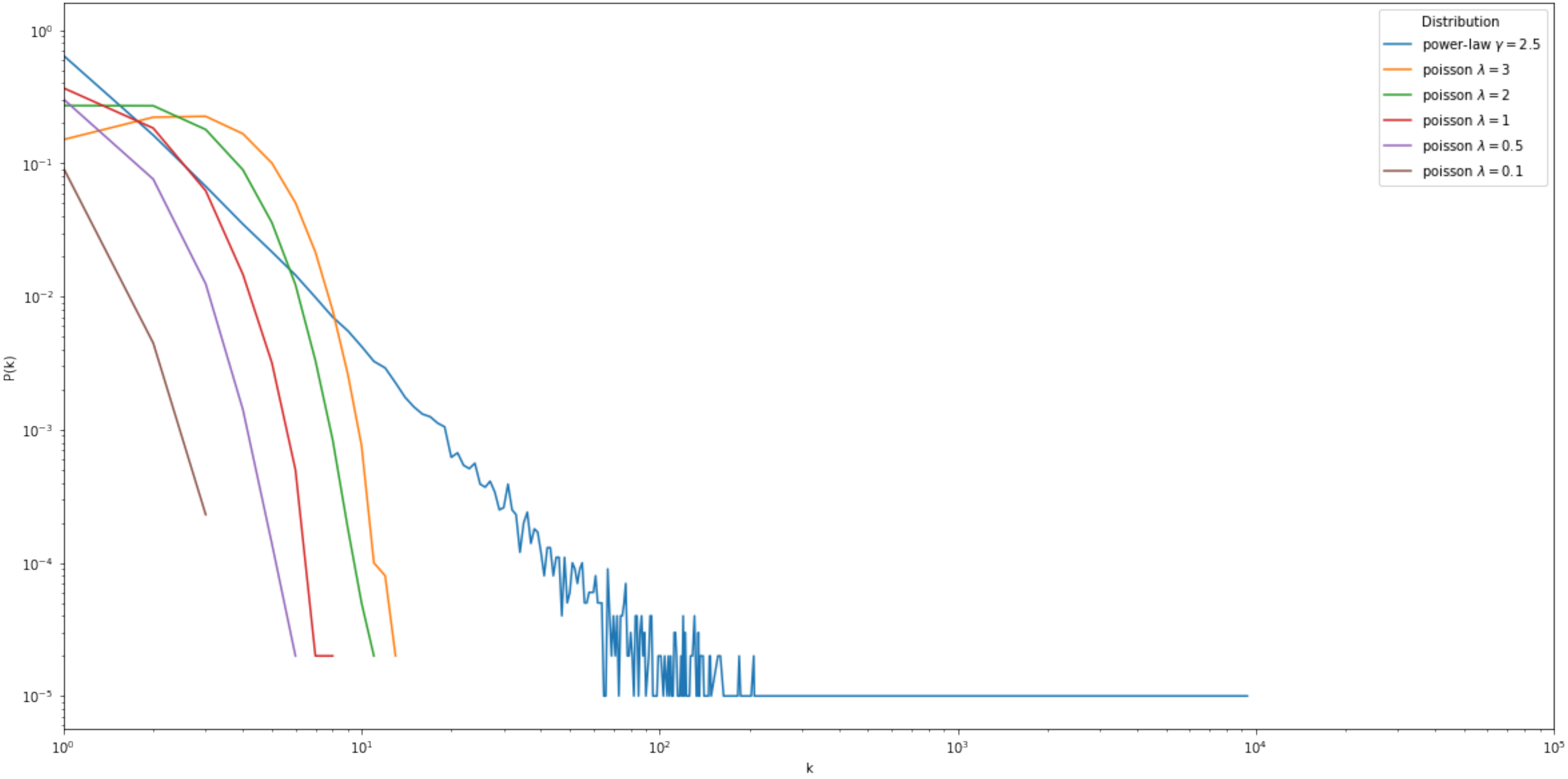
$$\frac{\lambda^k e^{-\lambda}}{k!}$$



Scale-free networks

Comparing a poisson distribution and a power law

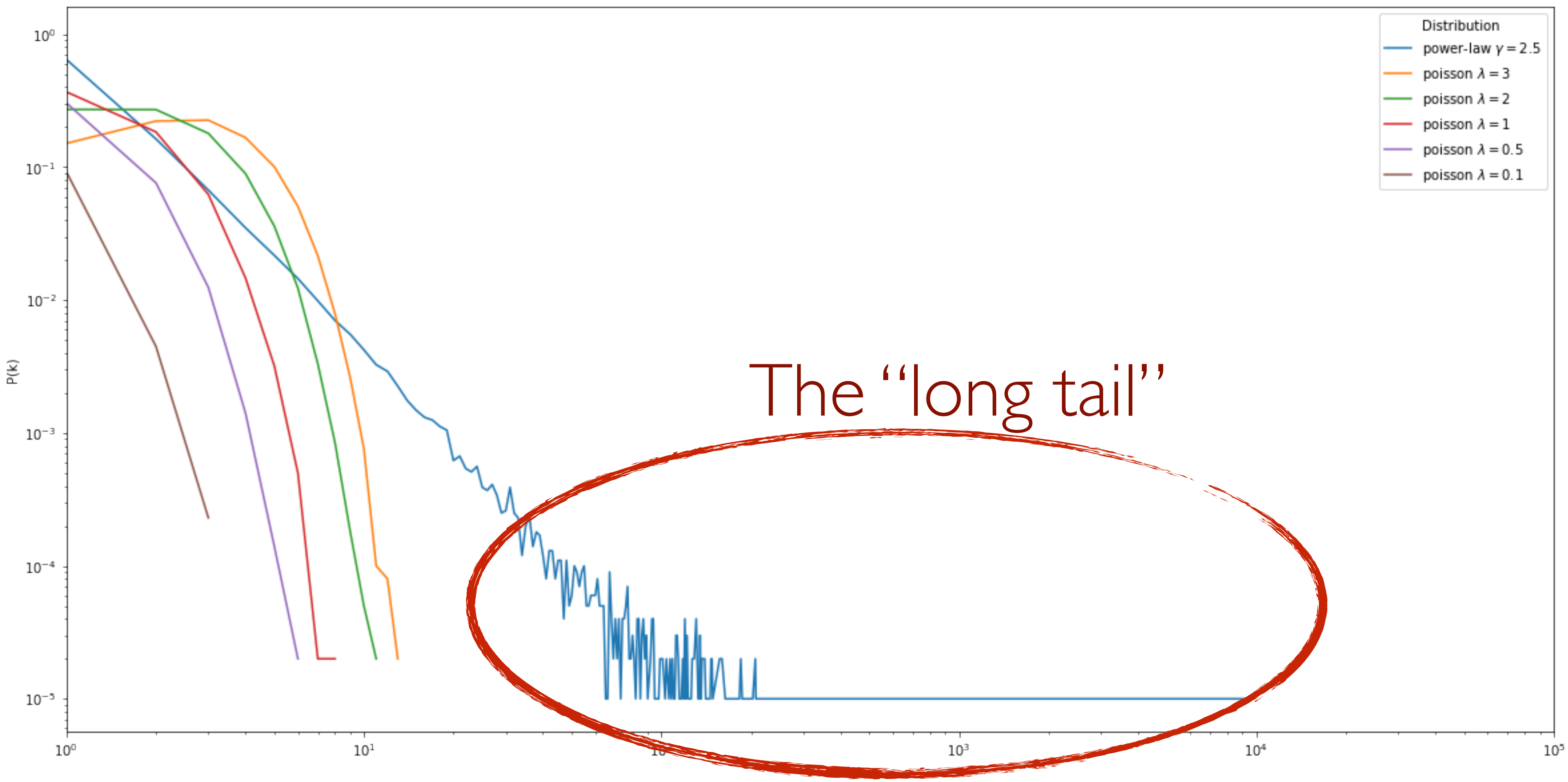
$$\frac{\lambda^k e^{-\lambda}}{k!}$$



Scale-free networks

Comparing a poisson distribution and a power law

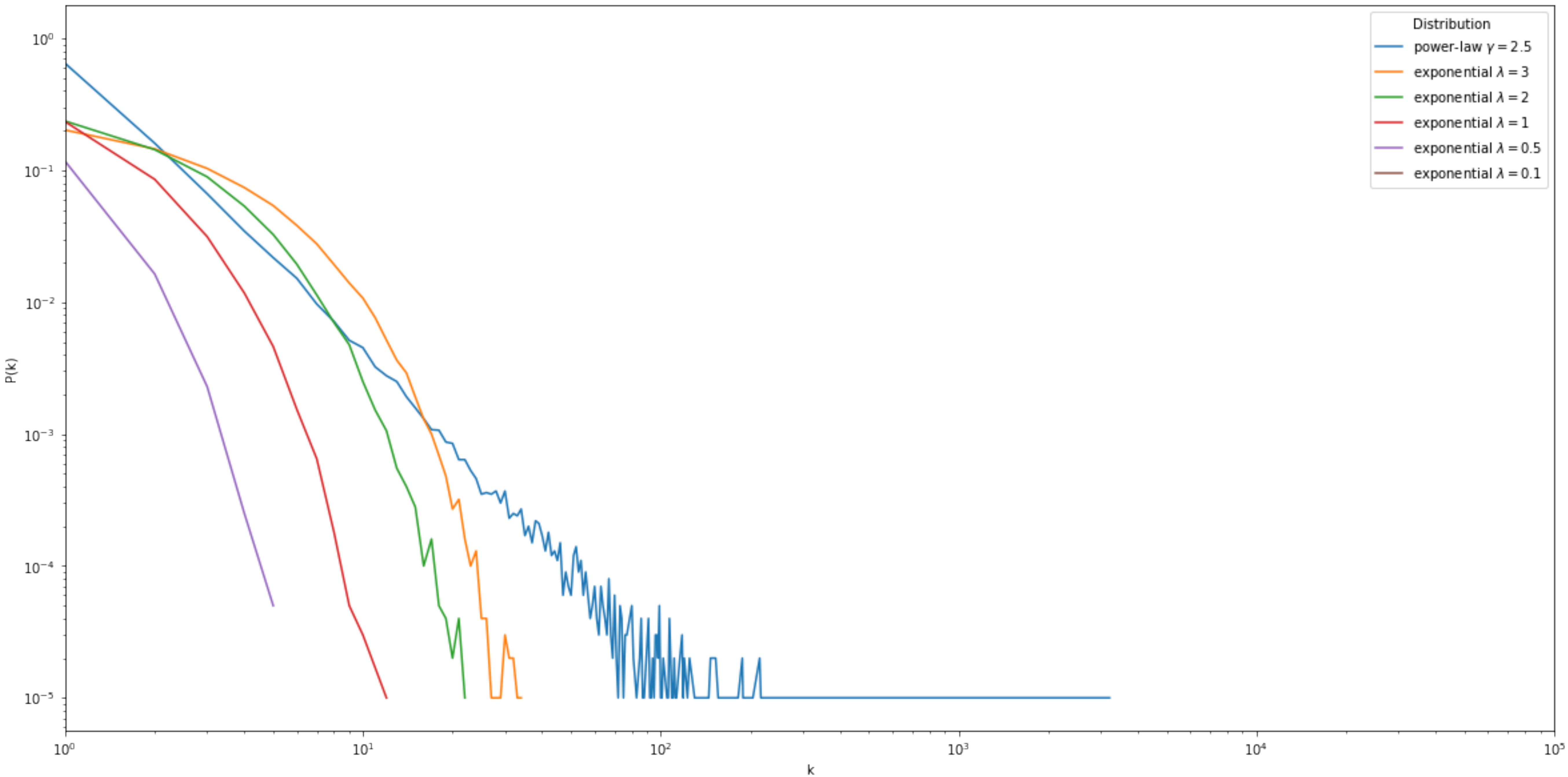
$$\frac{\lambda^k e^{-\lambda}}{k!}$$



Scale-free networks

Comparing an exponential distribution and a power law

$$\begin{cases} \lambda e^{-\lambda k} & k \geq 0, \\ 0 & k < 0. \end{cases}$$



Moments

Distribution: $$P(k) = (\alpha - 1)k_{\min}^{\alpha-1}k^{-\alpha}$$

Moments: $$\langle k^m \rangle = \int_{k_{\min}}^{\infty} k^m p(k) dk$$

With:

$\langle k^1 \rangle$ Average

$\langle k^2 \rangle$ Variance (converge like)

$\langle k^3 \rangle$ Skewness (converge like)

…

Moments

Distribution: $P(k) = (\alpha - 1) k_{\min}^{\alpha-1} k^{-\alpha}$

Moments:

$$\langle k^m \rangle = \int_{k_{\min}}^{\infty} k^m p(k) dk$$

$$\langle k^m \rangle = (\alpha - 1) k_{\min}^{\alpha-1} \int_{k_{\min}}^{\infty} k^{-\alpha+m} dk$$

$$\langle k^m \rangle = k_{\min}^{m} \left(\frac{\alpha - 1}{\alpha - 1 - m} \right)$$

Defined for $\alpha > m + 1$,
Otherwise diverge (+inf)

$$\int x^n dx = \frac{1}{n+1} x^{n+1} + c, \; n \neq -1$$

http://tuvalu.santafe.edu/~aaronc/courses/7000/csci7000-001_2011_L2.pdf

Moments

Moments: $$\langle k^m \rangle = k_{\min}^m \left(\frac{\alpha - 1}{\alpha - 1 - m} \right)$$ Defined for $\alpha > m + 1$, Otherwise diverge (+inf)

=> Mean: $$\langle k \rangle = \frac{\alpha - 1}{\alpha - 2} k_{\min}$$ (But diverges for $\alpha \leq 2$)

$$\langle k^2 \rangle = \frac{\alpha - 1}{\alpha - 3} k_{\min}^2$$ (But diverges for $\alpha \leq 3$)

Moments

What does divergence means in practice ?

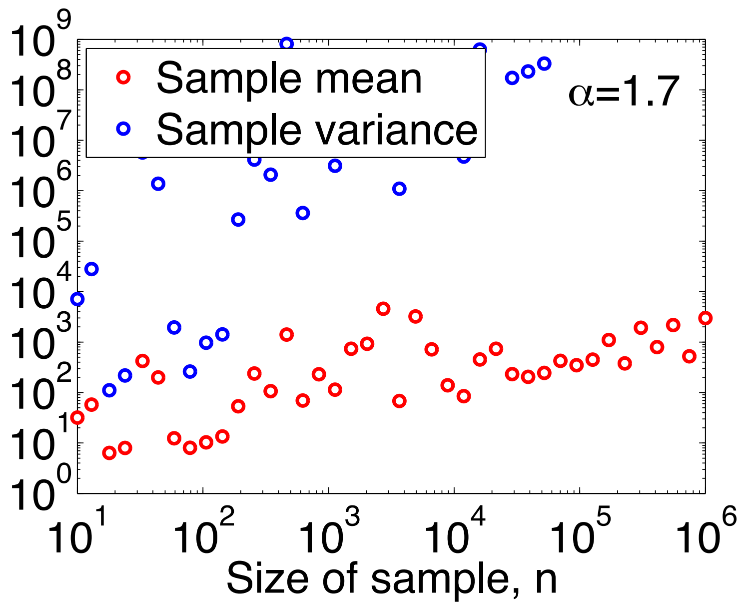
We can always *compute* the mean and variance, given samples of a distribution (e.g., an observe degree distribution)

=>The value computed depends on the **size** of the sample, it is not a characteristic of the distribution.

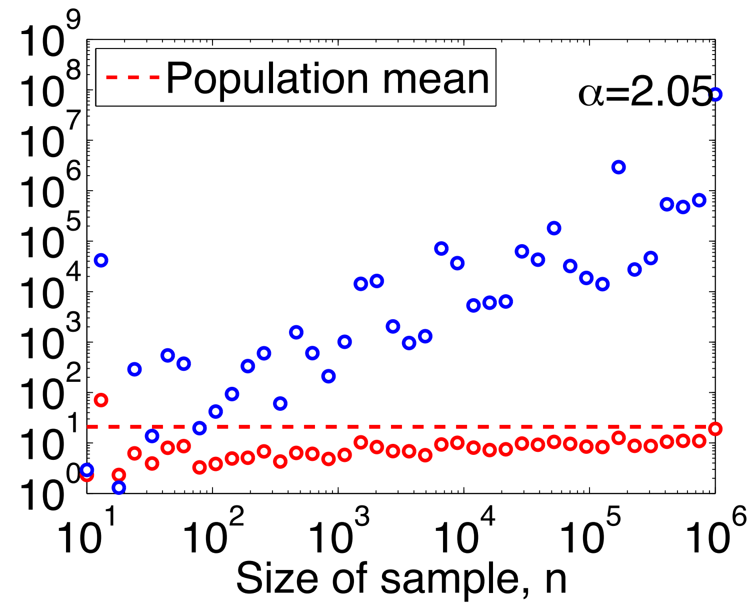
Moments are *dominated* by elements in the long tail. Some events are rare, but they have so large values, that if observed, they are strong enough to modify substantially the corresponding moment. And they appear frequently enough so that the mean will continue to shift when increasing the sample size

Moments

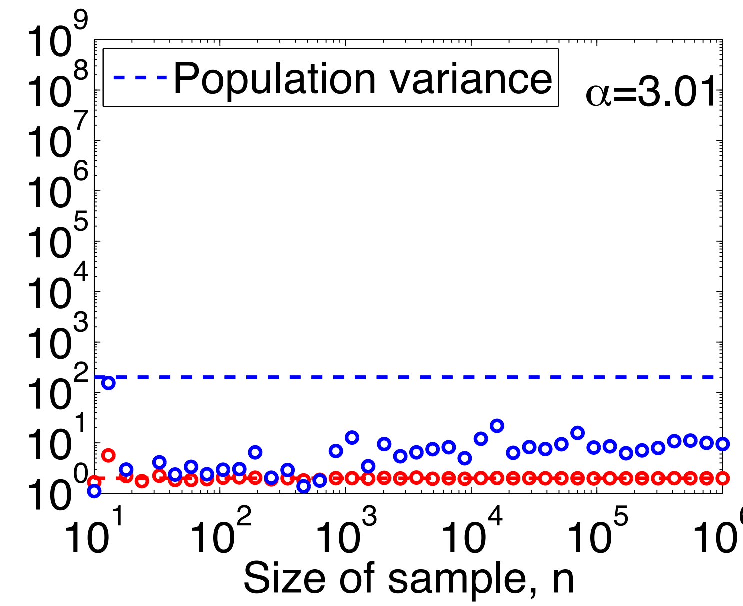
$\alpha < 2$
Mean diverge



$2 < \alpha < 3$
Mean well defined, Variance diverge



$\alpha > 3$
Mean and variance defined



=> Even when well defined, **moments converge very slowly**

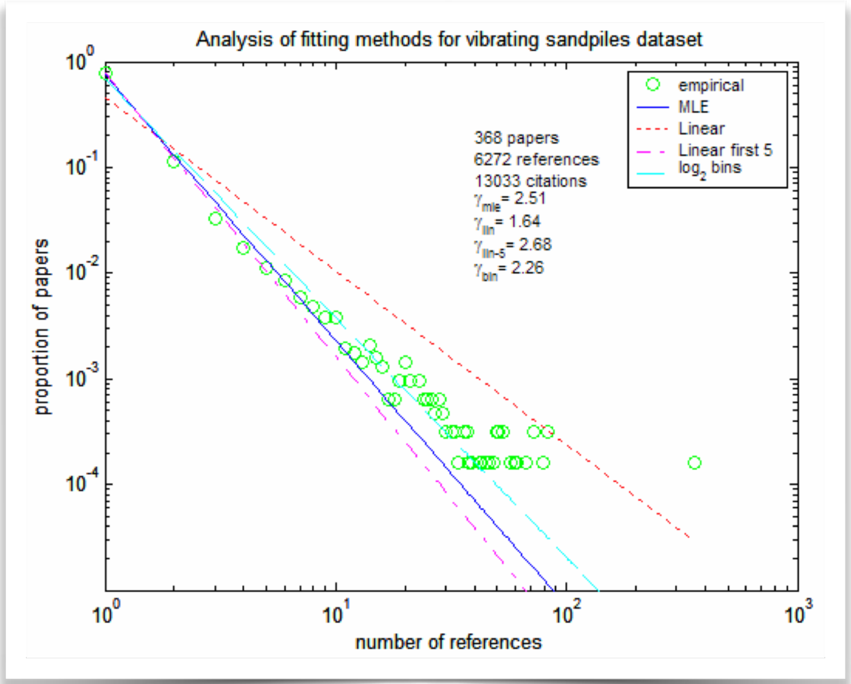
More details at [http://tuvalu.santafe.edu/~aaronc/courses/5352/fall2013/csci5352_2013_L11.pdf]

Computing the exponent of an observed network

Method 1: find the slope of the line of the log-log plot

Problem: most of data is on first values, so we *overfit* based on a few values in the long tail

Better approach
Maximum Likelihood Estimation (MLE)
Find the parameters of the distribution maximizing the probability to generate observations



[Fitting to the Power-Law Distribution, Goldstein et al.]
https://arxiv.org/vc/cond-mat/papers/0402/0402322v1.pdf

Scale-free networks

Network	Size	$\langle k \rangle$	κ	γ_{out}	γ_{in}
WWW	325 729	4.51	900	2.45	2.1
WWW	4×10^7	7		2.38	2.1
WWW	2×10^8	7.5	4000	2.72	2.1
WWW, site	260 000				1.94
Internet, domain*	3015–4389	3.42–3.76	30–40	2.1–2.2	2.1–2.2
Internet, router*	3888	2.57	30	2.48	2.48
Internet, router*	150 000	2.66	60	2.4	2.4
Movie actors*	212 250	28.78	900	2.3	2.3
Co-authors, SPIRES*	56 627	173	1100	1.2	1.2
Co-authors, neuro.*	209 293	11.54	400	2.1	2.1
Co-authors, math.*	70 975	3.9	120	2.5	2.5
Sexual contacts*	2810			3.4	3.4
Metabolic, *E. coli*	778	7.4	110	2.2	2.2
Protein, *S. cerev.**	1870	2.39		2.4	2.4
Ythan estuary*	134	8.7	35	1.05	1.05
Silwood Park*	154	4.75	27	1.13	1.13
Citation	783 339	8.57			3
Phone call	53×10^6	3.16		2.1	2.1
Words, co-occurrence*	460 902	70.13		2.7	2.7
Words, synonyms*	22 311	13.48		2.8	2.8

Exponent: γ_{out}, γ_{in}

Albert, R. et.al. Rev. Mod. Phy. (2002)

- Average values are not reliable since the convergence is very slow
- Furthermore, average values are meaningless since the fluctuations are infinitely large (diverging variance)

Exponents of real-world networks are usually between 2 and 3

Scale-free networks

Why do most of the real networks have degree exponent between 2 and 3?

- If the exponent is smaller than 2, the distribution is so skewed that we expect to find nodes with a degree *larger* than the size of the network => not possible in finite networks

Scale-free networks

Why do most of the real networks have degree exponent between 2 and 3?

- To detect a scale-free network its degree distribution needs to span through several (at least 2-3) orders of magnitude ⇒ $K_{max} \sim 10^3$
- If the exponent is large (>3), large degrees become so rare that the size of the sample (i.e., size of observed network) must be enormous to indeed observe such a node
- Example: let's choose $\gamma = 5$, $K_{min} = 1$ and $K_{max} \sim 10^3$

$$K_{\max} = K_{\min} N^{\frac{1}{\gamma - 1}}$$

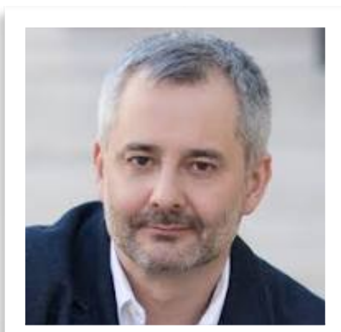
$$N = \left(\frac{K_{\max}}{K_{\min}} \right)^{\gamma - 1} \approx 10^{12}$$

We need to observe 10^{12} nodes to observe a node of degree 1000 for exponent=5

=> Forget about (single planet) social networks…

Scale-free networks

- Are real networks really Scale Free ?
- In most real networks, the scale free stands only for *a range* of degrees, i.e., between a minimum degree and maximum degree different than those observed (cut-offs)
- Some other distributions, in particular log-normal distributions, might “look like” power-law



Albert-László Barabási

Aaron Clauset

Emergence of scaling in random networks (1999)

Scale-free networks are rare (2018)

Love is All You Need - Clauset's fruitless search for scale-free networks (2018)

Petter Holme

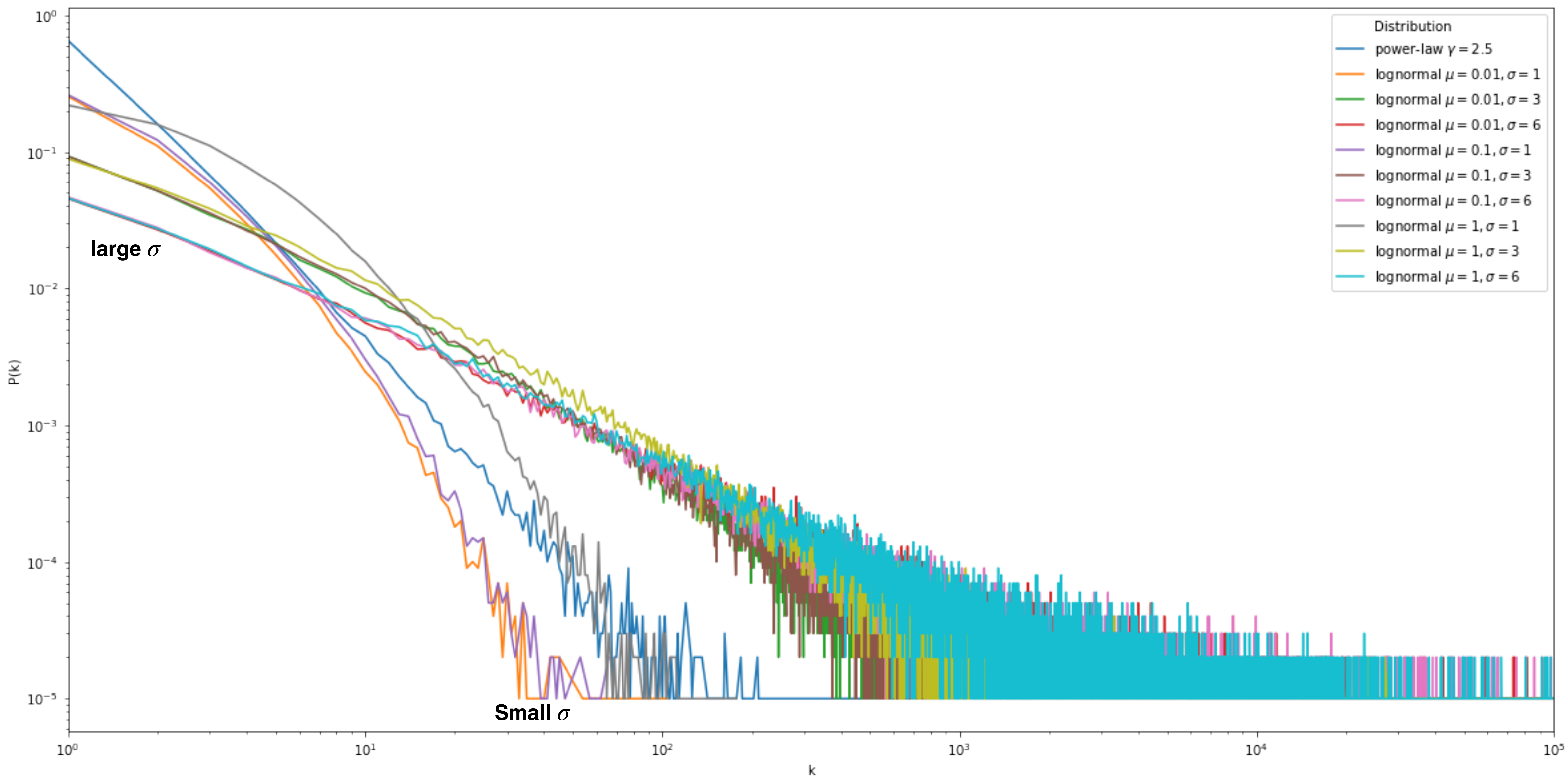
Rare and everywhere: Perspectives on scale-free networks (2019)

Scale-free networks

Comparing a log-normal distribution and a power law

Log-normal distribution = Probability distribution of a random variable whose logarithm is normally distributed

$$\frac{1}{k\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln k-\mu)^2}{2\sigma^2}\right) \qquad k^{-\alpha}$$



μ σ Mean, std of the log of the variable

Scale-free networks

Albert-László Barabási
@barabasi

@aaronclauset Every 5 years someone is shocked to re-discover that a pure power law does not fit many networks. True: Real networks have predictable deviations. Hence forcing a pure power law on these is like...fitting a sphere to the cow. Sooner or later the hoof will stick out.

4.13.4 Systematic Fitting Issues

The procedure described above may offer the impression that determining the degree exponent is a cumbersome but straightforward process. In reality, these fitting methods have some well-known limitations: **Network Science, Chapter 4, pg 159**

1. A pure power law is an idealized distribution that emerges in its form (4.1) only in simple models (**Chapter 5**). In reality, a whole range of processes contribute to the topology of real networks, affecting the precise shape of the degree distribution. These processes will be discussed in **Chapter 6**. If p_k does not follow a pure power law, the methods described above, designed to fit a power law to the data, will inevitably fail to detect statistical significance. While this finding can mean that the network is not scale-free, it most often means that we have not yet gained a proper understanding of the precise form of the degree distribution. Hence we are fitting the wrong functional form of p_k to the dataset.

Albert-László Barabási @barabasi · Jan 15, 2018
Replying to @barabasi
Chapter 6 in Network Science networksciencebook.com/chapter/6 discusses what you should be fitting to the degree distribution of *real* scale-free networks. You are right: Pure power laws are predictably rare. Scale-free networks are not.
1 · 21 · 45

Aaron Clauset @aaronclauset · Jan 15, 2018
Replying to @barabasi
Yes, science is hard and real data often messy. But it is worrying how criticisms of harsh statistical evaluations can be interpreted as a belief that "disagreement with data" (as Feynman would put it) should not be held against a favored theory or model.
3 · 5 · 18

Albert-László Barabási @barabasi · Jan 15, 2018
We are on the same page. The question is, what you test and what you conclude. There are multiple processes that contribute to the degree distribution that modify the power law. Hence testing for power laws only you are ignoring them all, leading to misleading takeway message.
2 · 4 · 10

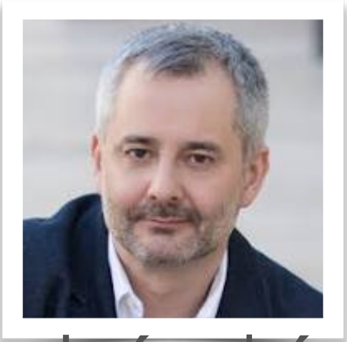
Aaron Clauset @aaronclauset · Jan 15, 2018
Perhaps. I feel good about the accuracy of our conclusions: we used rigorous statistical methods, tested 5 distributions, considered 5 levels of evidence, across nearly 1000 network datasets. The goal was to be thorough and to treat the SF hypothesis as falsifiable.
1 · 3 · 14

Albert-László Barabási @barabasi · Jan 15, 2018
The effort is amazing. The conclusions are less so. The feather falls slower than the rock, yet gravitation is not wrong. We add friction. You need to fit for each system the Pk that is right for it. That is hard, I know. Otherwise you ignore 20 year of work by hundreds.
2 · 4 · 6

Aaron Clauset @aaronclauset · Jan 15, 2018
It seems easy to get confused here: an empirical power-law degree distribution is evidence for SF structure, but no deviation from the power law can be evidence against SF structure? It is reasonable to believe a fundamental phenomena would require less customized detective work.

Scale-free networks

Aaron Clauset



Albert-László Barabási

-Rigorous **statistical tests** show that observed degree distributions are not compatible with a power law distribution (high p-values)

-Compared with different distributions, in particular log-normal, most degree distributions are more likely to be generated by something else than power laws

-Networks are real objects, not mathematical abstraction, therefore they are sensible to **noise** (real life limits…)

-Power law is a **good, simple model** of degree distributions of a class of networks

-20 years of **fruitful research** based on this model

A whole scientific article dedicated to the controversy:
Jacomy, M. (2020). Epistemic clashes in network science: Mapping the tensions between idiographic and nomothetic subcultures. *Big Data & Society*, *7*(2), 2053951720949577.

The Barabási-Albert model

of scale-free networks

Emergence of hubs

What did we miss with the earlier network models?

1. Networks are evolving
   - Networks are not static but growing in time as new nodes are entering the system

2. Preferential attachement
   - Nodes are not connected randomly but tends to link to more attractive nodes

- Pólya urn model (1923)
- Yule process (1925)
- Zipf's law (1941)
- Cumulative advantage (1968)
- Preferential attachement (1999)
- Pareto's law - 80/20 rule
- The rich get richer phenomena
- etc.

AL Barabási, Network Science Book (2013)

The Barabási-Albert model

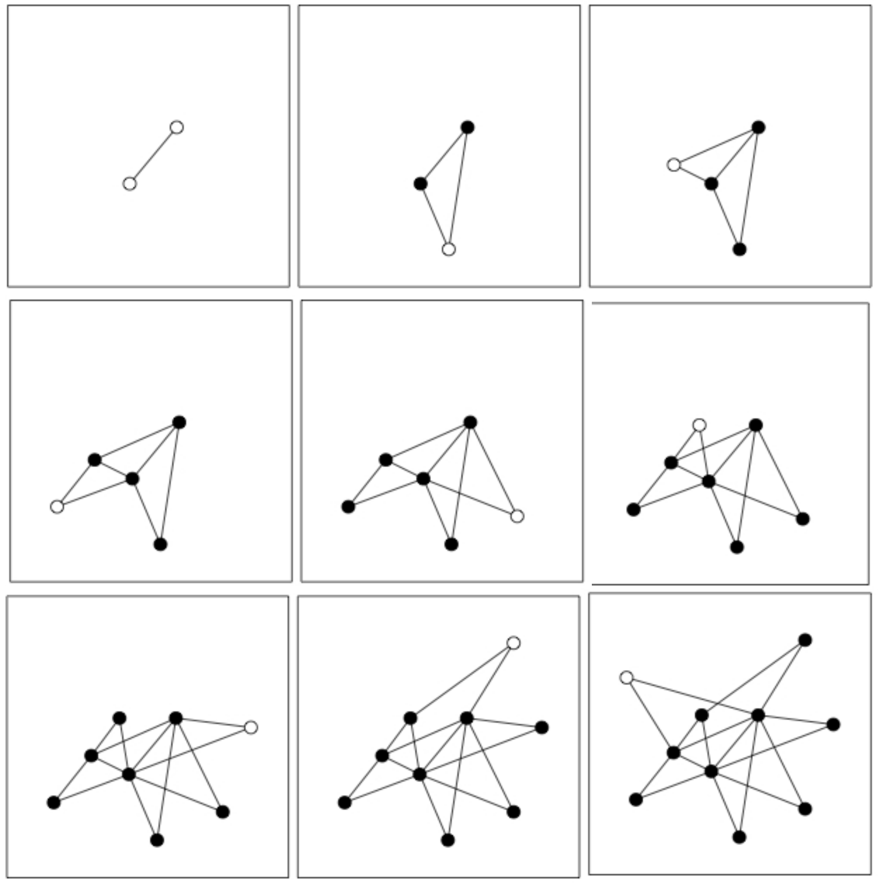
1. Start with m_0 connected nodes

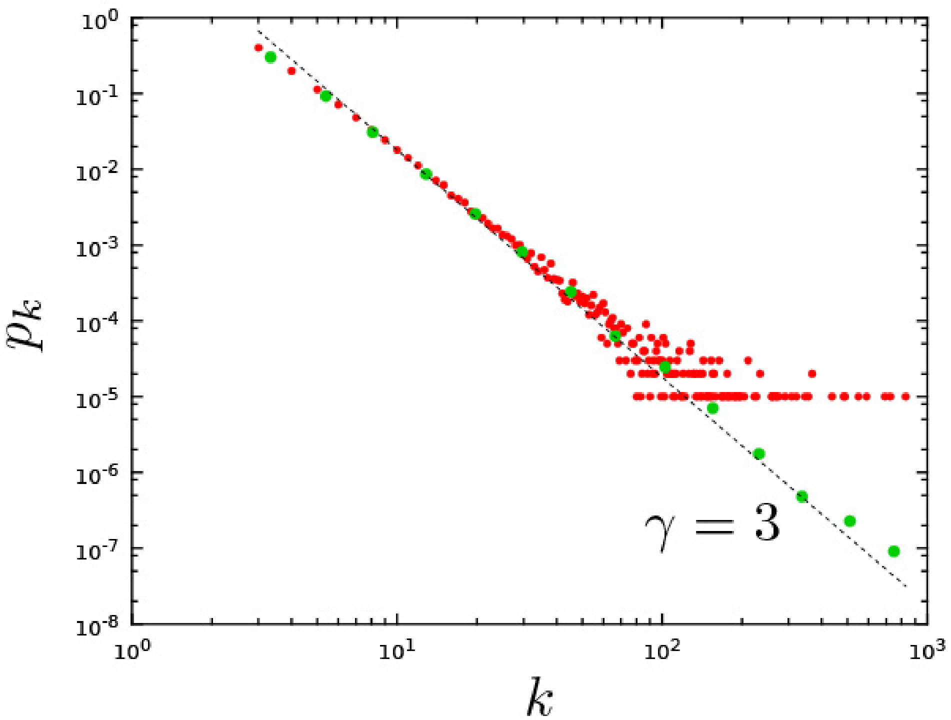
2. At each timestep we add a new node with m ($\leq m_0$) links that connect the new node to m nodes already in the network.

3. The probability $\pi(k)$ that one of the links of the new node connects to node i depends on the degree k_i of node i as

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

- The emerging network will be scale-free with degree exponent $\gamma=3$ independently from the choice of m_0 and m





AL Barabási, Network Science Book (2013)

ER Random Network - catch up

Network	Degree distribution	Path length	Clustering coefficient
Real world networks	broad	short	large
Regular lattices	constant	long	large
ER random networks	Poissonian	short	small
WS small-world networks	exponential	short	large
BA scale-free networks	power-law	short	Rather small

(some) Other random models

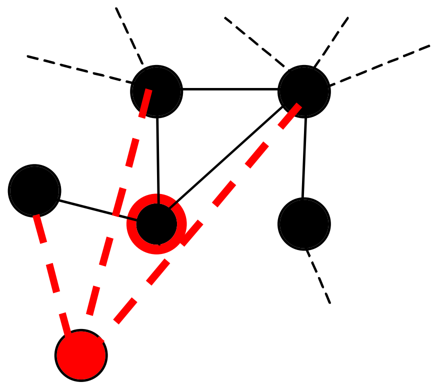
Other scale-free models

The vertex-copying model

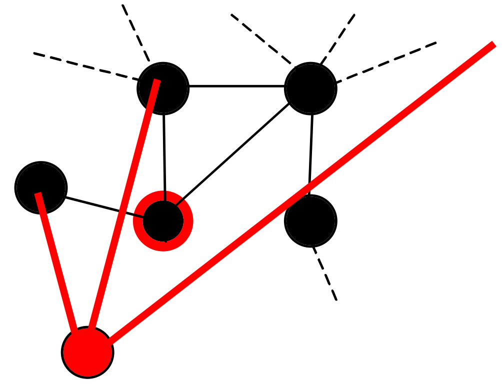
- Motivation:
  - Citations network or WWW where links are often copied
  - Local explanation to preferential attachement

1. Take a small seed network
2. Pick a random vertex
3. Make a copy of it
4. With probability p, move each edge of the copy to point to a random vertex
5. Repeat 2.-4. until the network has grown to desired size of N vertices

1. copy a vertex



2. rewire edges with *p*



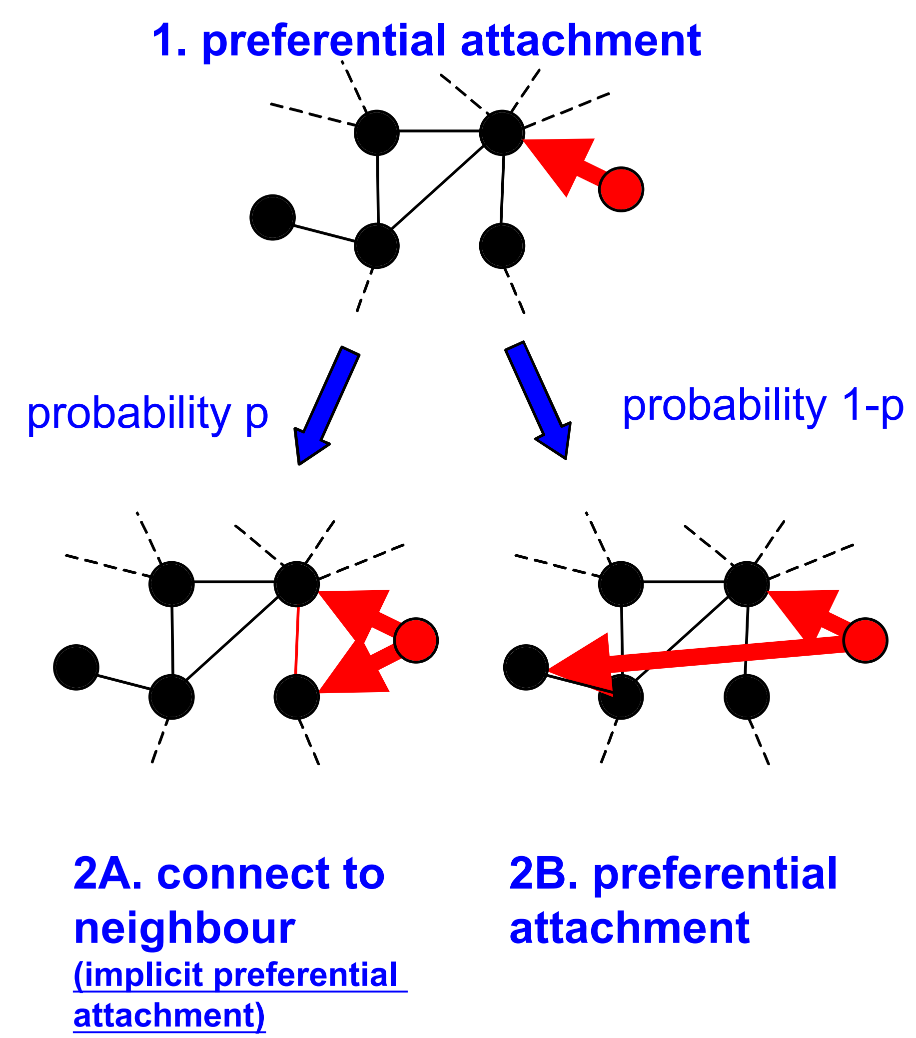
- Asymptotically scale-free with exponent $\gamma \geq 3$

Other scale-free models

The Holme-Kim model

- Motivation: more realistic clustering coefficient

1. Take a small seed network
2. Create a new vertex with m edges
3. Connect the first of the m edges to existing vertices with a probability proportional to their degree k (just like BA)
4. With probability p, connect the next edge to a random neighbour of the vertex of step 3., otherwise do 3. again
5. Repeat 2.-4. until the network has grown to desired size of N vertices



$$C(k) \propto \frac{1}{k}$$

for large N, ie clustering more realistic! This type of clustering is found in many real-world networks.

ER Random Network - catch up

Network	Degree distribution	Path length	Clustering coefficient
Real world networks	broad	short	large
Regular lattices	constant	long	large
ER random networks	Poissonian	short	small
WS small-world networks	exponential	short	large
BA scale-free networks	power-law	short	Rather small
Other models	power-law	short	Large

HETEROGENEOUS NODES

- Presented models assume that nodes are interchangeable globally
- Other models preserve some node properties
  - Spatial models: nodes have a fix position in space. Edge probability depends on node distance
  - Block models: nodes belong to a node group (block). Edge probability depends on blocks belonging
    - More during Community detection class
  -

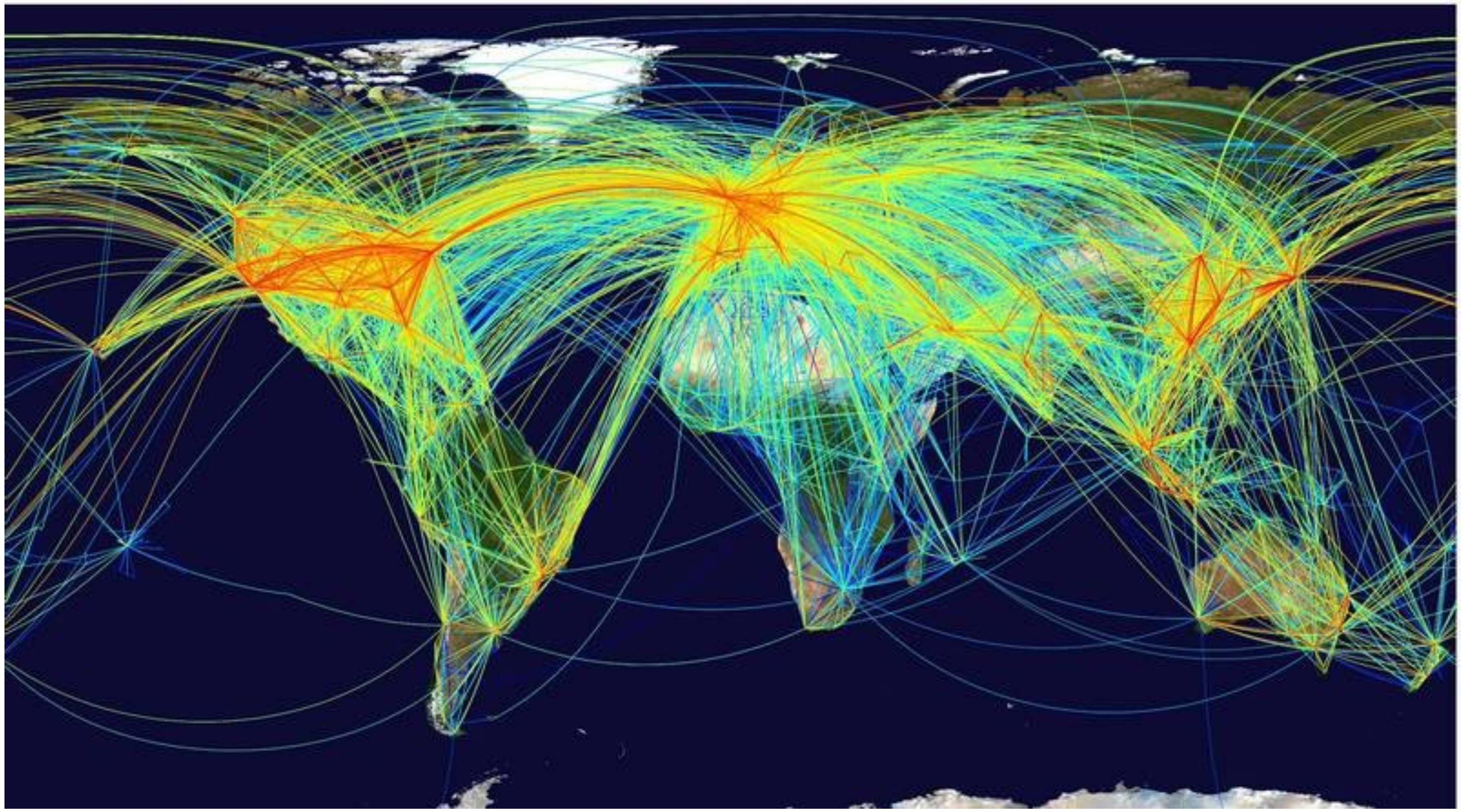
SPATIAL NETWORK MODELS

Spatial networks

A network is said spatial if the distance between nodes affect the probability of observing edges between them

Distance

- Physical distance
- Economical distance
- Social distance
- Difference in professional categories
- …



Geisel et.al. (2011)

Literature

Physics Reports 499 (2011) 1–101

Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

ELSEVIER

Spatial networks

Marc Barthélemy *

Institut de Physique Théorique, CEA, IPhT CNRS, URA 2306, F-91191 Gif-sur-Yvette, France
Centre d'Analyse et de Mathématique Sociales (CAMS, UMR 8557 CNRS-EHESS), Ecole des Hautes Etudes en Sciences Sociales, 54 bd. Raspail, F-75270 Paris Cedex 06, France

A R T I C L E   I N F O

Article history:
Accepted 8 November 2010
Available online 23 November 2010
editor: H. Orland

Keywords:

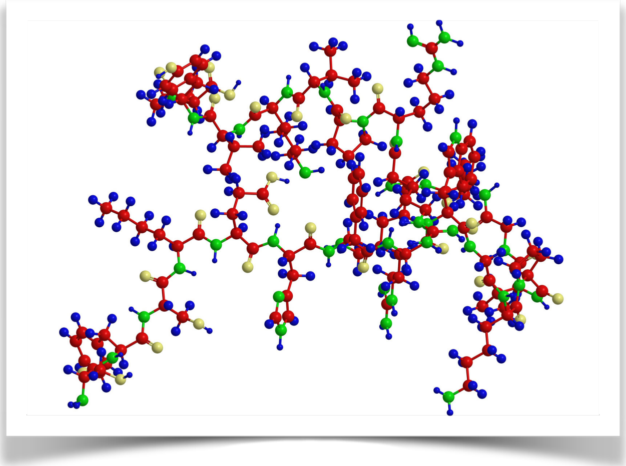
A B S T R A C T

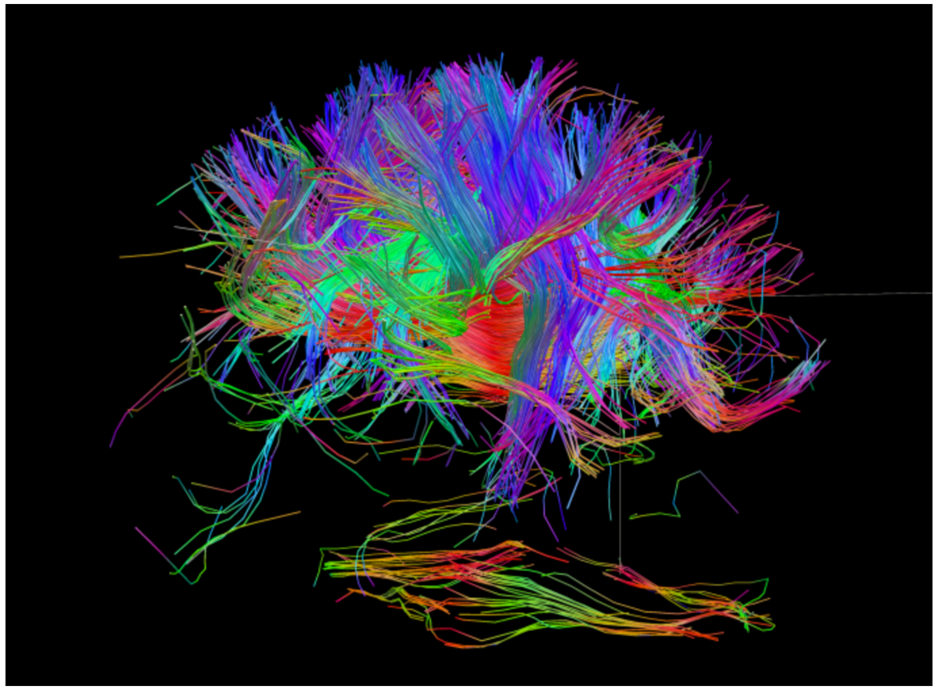
Complex systems are very often organized under the form of networks where nodes and edges are embedded in space. Transportation and mobility networks, Internet, mobile phone networks, power grids, social and contact networks, and neural networks, are all examples where space is relevant and where topology alone does not contain all the information. Characterizing and understanding the structure and the evolution of spatial networks is thus crucial for many different fields, ranging from urbanism to epidemiology.

Spatial networks

Types of spatial networks

- Transportation networks
  - Airline networks
  - Bus, subway, railway, and commuters
  - Cargo ship networks
- Infrastructure networks
  - Road and street networks
  - Power grids and water distribution networks
  - The internet
- Neural networks
- Protein networks
- Mobility networks
- Social networks
- …

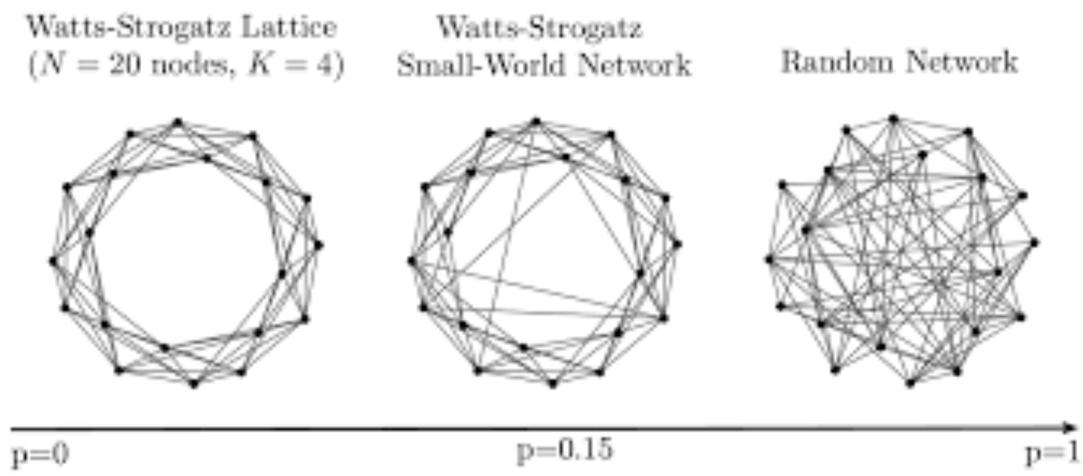


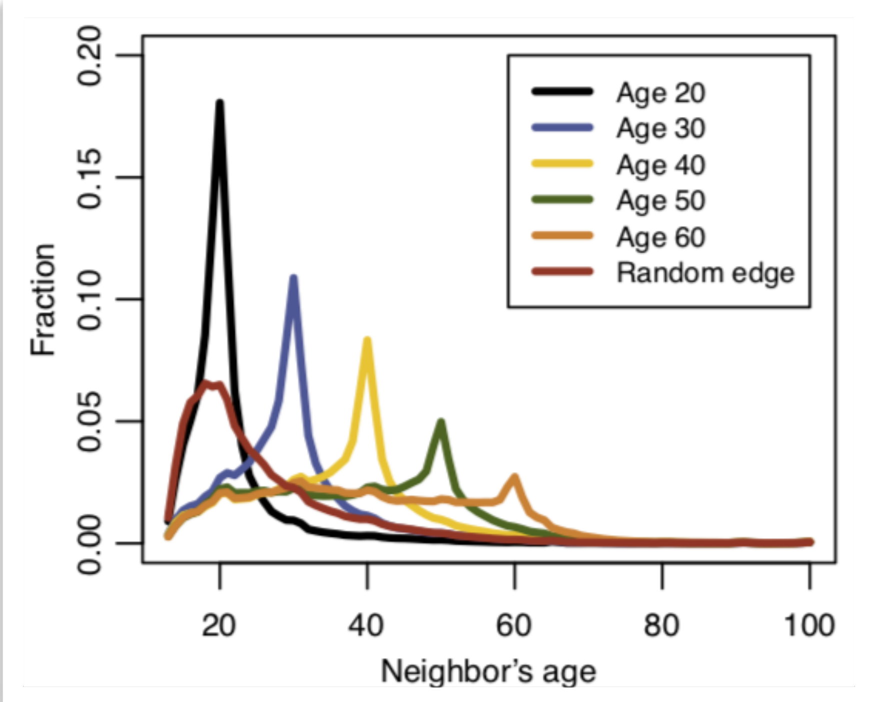


Spatial networks

Examples of 1D spaces

- The watts-Strogatz random graph is defined on a (circular) 1D space: each node is (initially) connected to its k closest nodes in this space.
- In social networks, users tend to be more connected with other users with similar age. We can consider age as a position on 1D space. The same is true about political opinions, if we consider a Left-Right spectrum.

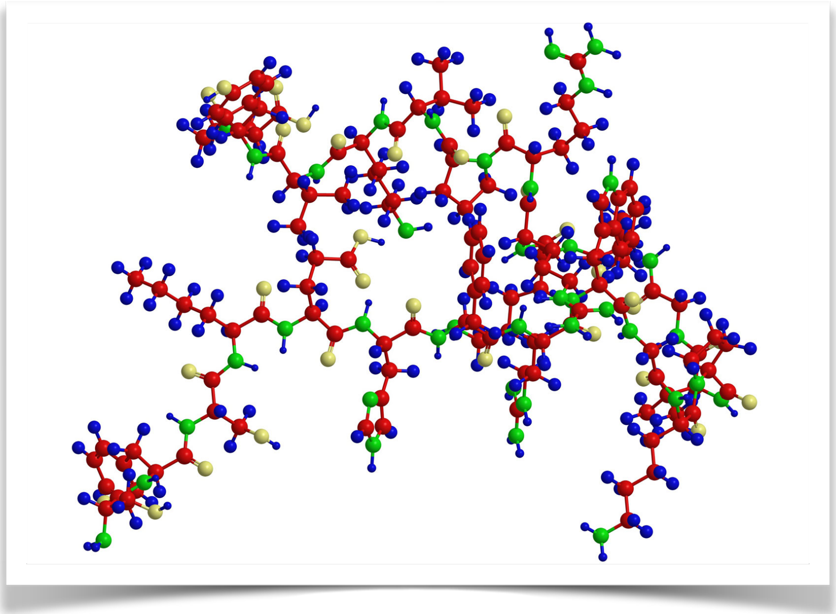


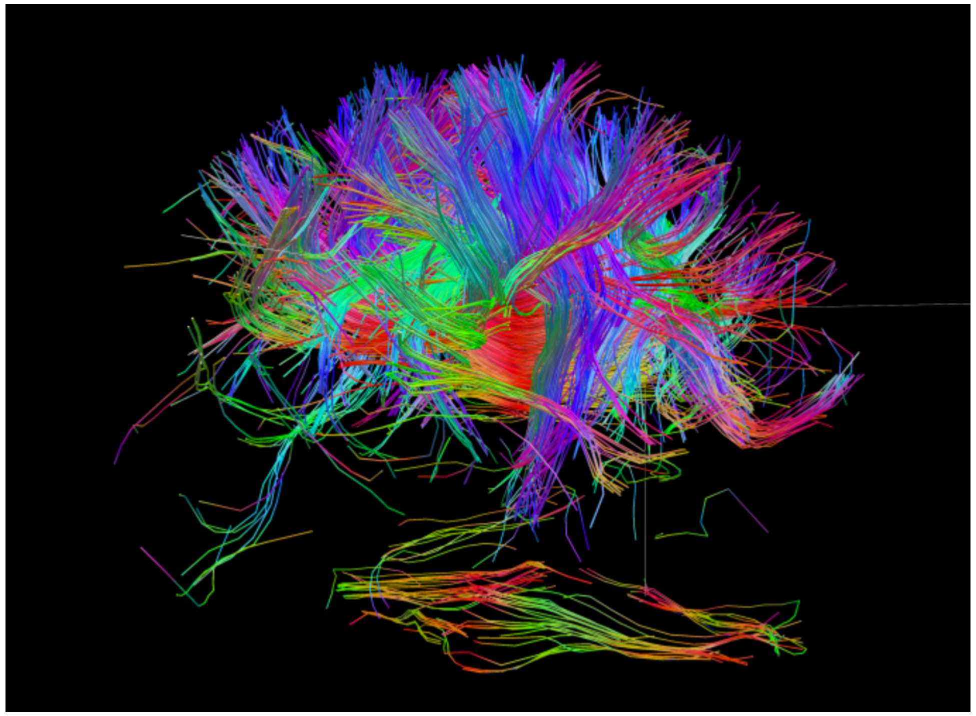


Spatial networks

Examples of 3+D spaces

- If we consider altitude, geographical networks are 3D spaces
- If we consider multiple nodes properties as dimensions, nodes can be located on high dimensional spaces, e.g., age, political opinion, revenue, geographical location, etc. Be careful however, that analyzing a spatial networks needs to define the *distance* between nodes, which can be tricky to define if dimensions are of different natures.
- Methods such as *graph embedding* assign locations in arbitrary large dimensions to nodes that summarize some of the network properties (see later class).





Spatial networks

Distances

The distance between each pair of nodes can be computed in different ways, depending on the nature of dimensions nodes are embedded in. The most common ones are:

- **Euclidean distance**, or $L^2 distance$ is the usual, straight line distance
- **Great-Circle distance** is used to measure the distance between points located on a sphere, typically the Earth for geographical data.
- **Dot product** and **Cosine Distance** are often used in high dimensions, in particular when it makes sense to multiply the location vectors.
- **Manhattan distance**, or $L^1 distance$, is sometimes used as a variant of Euclidean distance for high dimensional data (it is simply defined as the sum of differences in each of the dimensions.)
- Observed distance can sometimes be used, a typical example being **average time distance**: in datasets of trips or traffic, the time distance between dots might be only loosely proportional to geographical distance.

Simple models
of
spatial networks

Random geometric graphs

General definition:

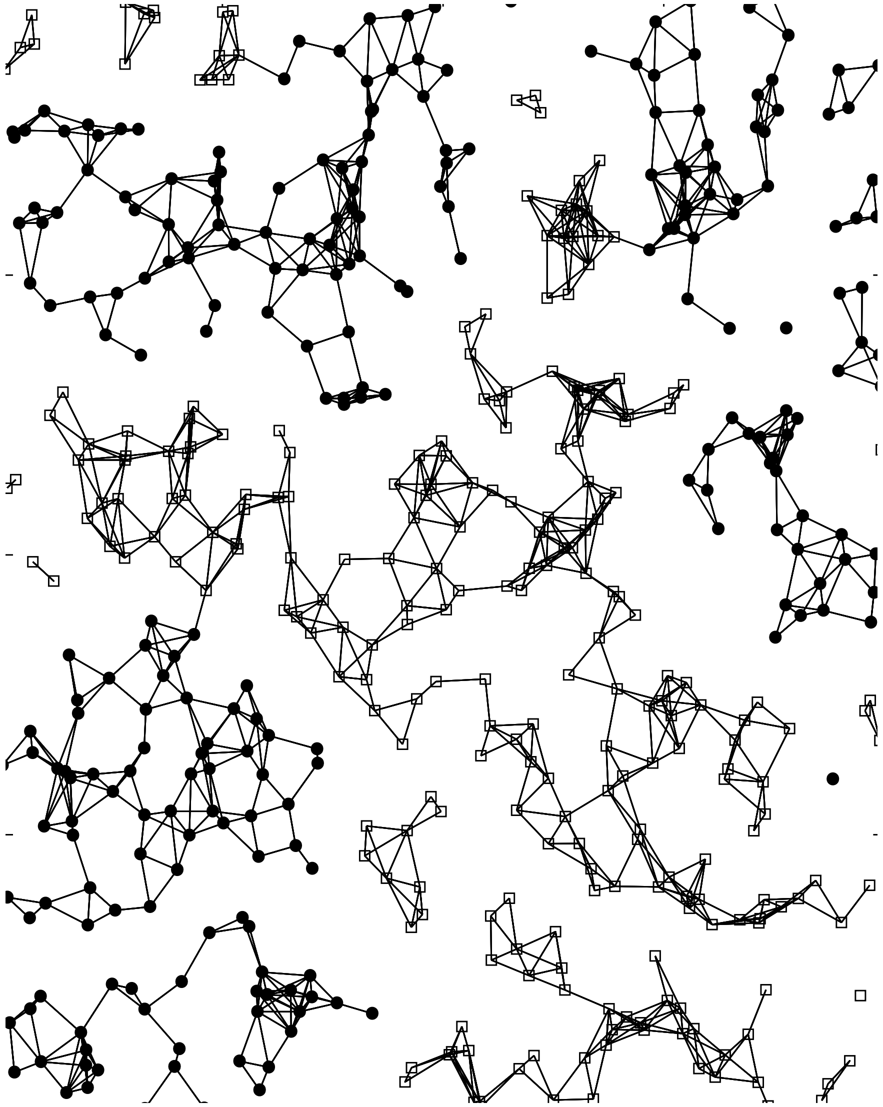
- Take a space and distribute nodes randomly
- Nodes are small spheres with radius r
- Two nodes are connected if their spheres overlap — separated with distance smaller than 2r
- Also called: disk-percolation

Degree distribution — Poisson distribution

Clustering coefficient (d=dimensions)

$$\langle C_d \rangle \sim 3\sqrt{\frac{2}{\pi d}} \left(\frac{3}{4}\right)^{\frac{d+1}{2}}$$

Independent of N contrary to random networks



Jesper Dall and Michael Christensen. "Random geometric graphs". In: *Physical review E* 66.1 (2002), p. 016121.

Soft RGG

Soft RGG (Waxman random graph)

Soft RGG, or **Waxman Random Graphs**[a], starts as the RGG by distributing nodes at random in a space, but instead of adding links between all nodes closer than a certain distance, it assign edges between nodes according to a **deterrence function** f, i.e., a function defining how distance affects the probability of observing edges between nodes.

The Soft RGG can model an ER random graph if f is constant function, $f(\Delta) = p$. It can model a classic RGG if f is a threshold function with:

$$f(d) = \begin{cases} 1 & \Delta \leq r \\ 0 & \Delta > r \end{cases}$$

[a] Waxman 1988.

Deterrence function

Deterrence function

A deterrence function defines how the distance affects the probability of observing an edge. It can be a probability (bounded on $[0, 1]$), or define a change ratio.

1. It can be defined *a priori*, usually as a classic monotonically decreasing function, e.g., Negative exponential($f(\Delta) = e^{-\alpha\Delta}$) or Negative power ($f(\Delta) = \Delta^{-\alpha}$), with α a parameter. A typical example of negative power in geographical data is when the probability of observing an edge decreases as the square of the distance, i.e., $f(\Delta) = \frac{1}{\Delta^2}$
2. It can also be learned from data, either by fitting parameters of a predefined function (e.g., the α parameter above), or by using an *Ad-Hoc deterrence function*.

The gravity law model

Formal description

Origin-destination matrix

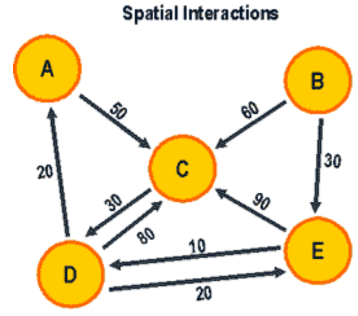
- Describe flow of individuals between locations
- Used since decades by geographers
- Definition:
  - divide the area of interest into zones (cells) labelled by $i=1...N$
  - count the number of individuals going from location i to location j

- directed
- weighted
- Beware:
  - strongly depends on the zone definition

$T(i,j)=$

O/D Matrix

	A	B	C	D	E	Ti
A	0	0	50	0	0	50
B	0	0	60	0	30	90
C	0	0	0	30	0	30
D	20	0	80	0	20	120
E	0	0	90	10	0	100
Tj	20	0	280	40	50	390



The gravity law

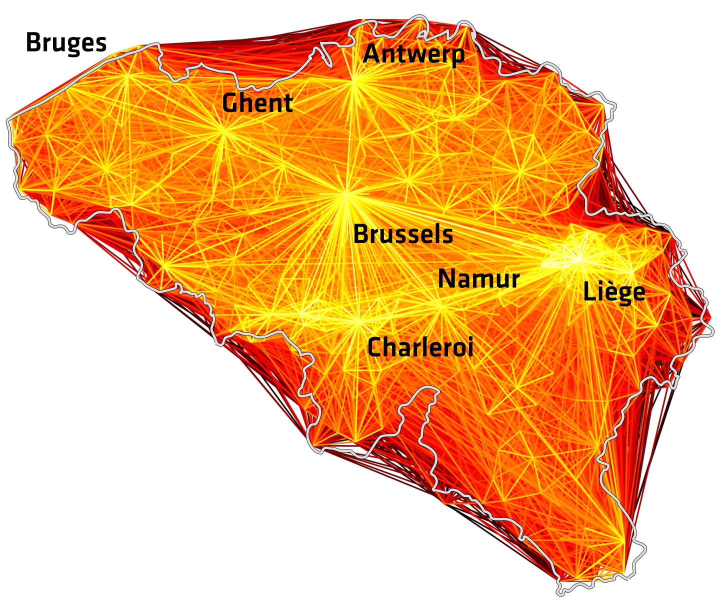
Number of trips from location i to location j is scaling as

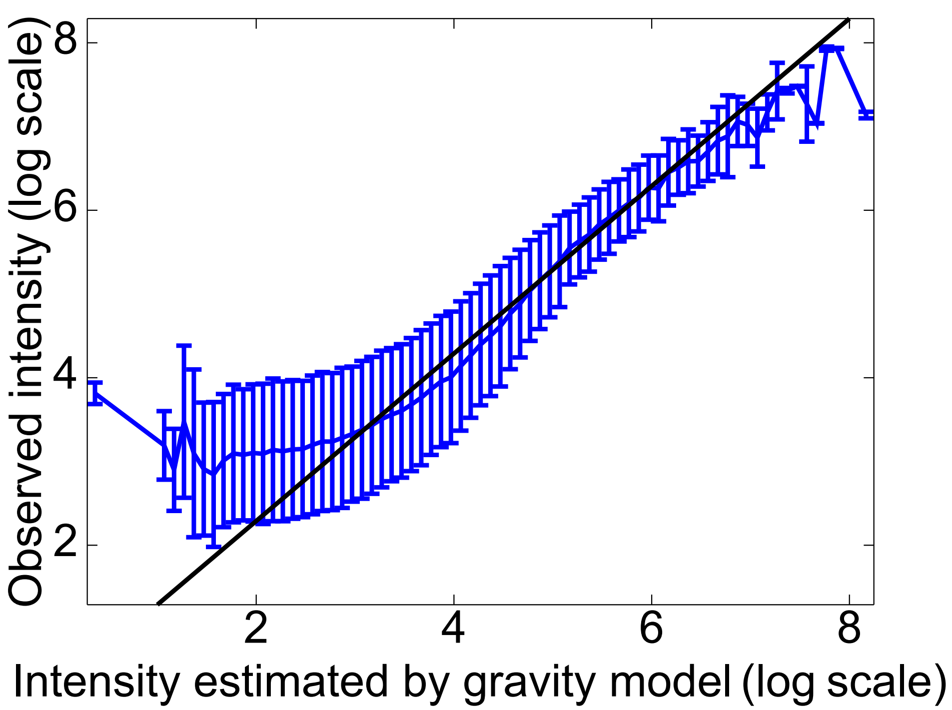
$$T_{ij} = K \frac{P_i P_j}{d_{ij}^{\sigma}}$$

- where $d_{ij} = d_E(i,j)$ is the distance between i and j
- $P_{i(j)}$ is the *population size* at location I(j)
- σ a parameter chosen or learned from data

Inter-city phone communication (Krings et.al.)

- mobile call communication intensity between Belgian cities





The gravity law

Number of trips from location i to location j is scaling as

$$T_{ij} = K\frac{P_iP_j}{d_{ij}^{\sigma}}$$

- where $d_{ij} = d_E(i,j)$ is the distance between i and j
- $P_{i(j)}$ is the *population size* at location $i(j)$

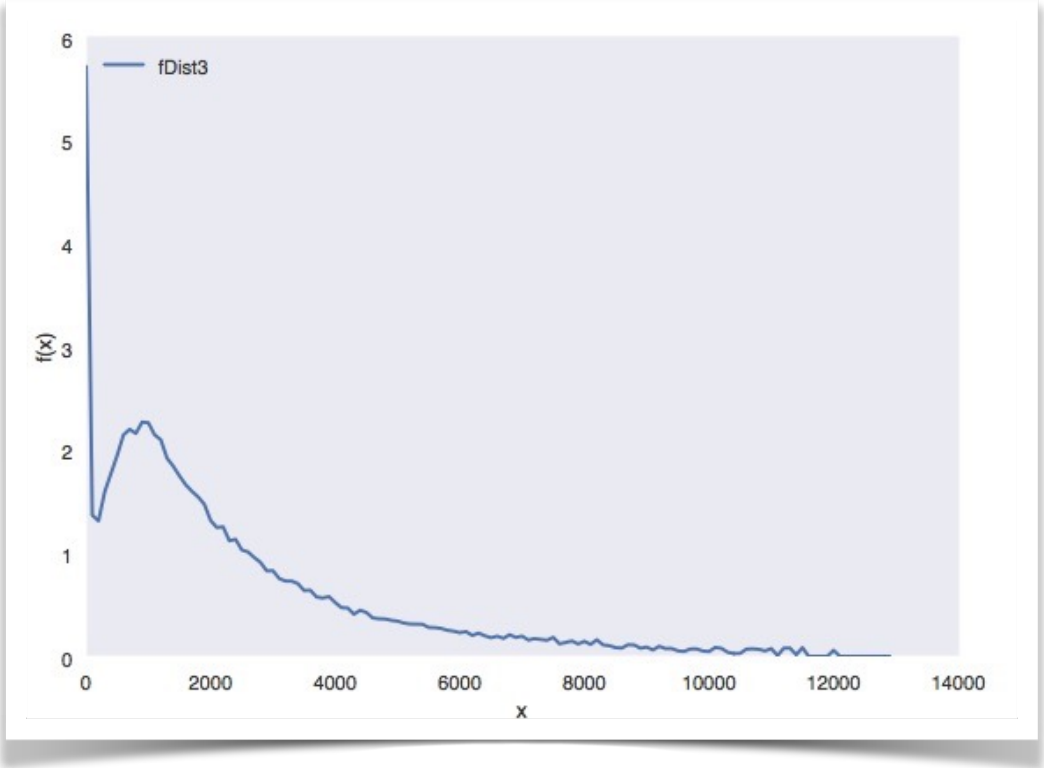
- In a general form: $T_{ij} \sim P_iP_jf(d(i,j))$
  - where $f(d(i,j))$ is the deterrence function describing the effect of space

Ad-hoc deterrence function

Agnostic deterrence function

- The influence of distance might be more complex than a power-law or an exponential. In particular, it is often non-monotonic (first increasing, then decreasing. Think of airplanes, bicycles, public transports… unlikely to use for short distances)
- A deterrence function can be learned from data
- Computed by comparing the number of trips observed at a given distance with the number of trip expected if distance has no effect (a configuration model)

$f(d)$



Distance d

The gravity law - as a network null model

Usage as a **network** null model

- Consider a spatial network (e.g., phone calls, trips, etc.)

- Fit a gravity model best explaining the observed network. If the population is unknown or not relevant, the degrees of nodes (in/out degrees in directed networks) can be used as a "*population*"

- =>Random model with a given edge probability for each pair of node

- The obtained network is a null model to which the observed network can be compared

The radiation law

The radiation law

Limitations of the gravity law

1. Requires previous data to fit

2. The number of travelers between destinations depends only on their populations and distances. In reality, this value depends probably of other *opportunities*

The radiation law

Intuition: Model how people move for jobs

1. Individuals look for job in all cities

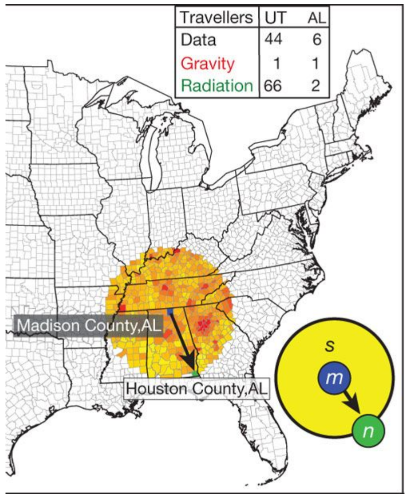
2. Each city has a number of job opportunities

    - Each job has a value of *interest*, considered random

3. What is the probability for a job-seeker to choose a job in city *c* located at distance *d?*

    - Depends only on how many jobs offered in cities at a distance equal or lower than *d* (probability to find a better job closer)

The model is parameter-free!



The radiation law

The model can be formulated in terms of radiation and absorption

- take locations i and j with populations (in-degree) m_i and n_j and at distance r_{ij}
- denote s_{ij} the total population in the circle with radius r_{ij} centered at i (excluding the source and destination population)
- P is the *power of attraction*, I.e., without other data, the degree.

Radiation Law of Spatial Interactions

The **Radiation Law**[a] is another random spatial model. Unlike previous ones, it does not depends on a deterrence function, and is parameter-free. It is based on the principle of relative opportunities: the probability of observing an interaction from i to j depends on P_i^{out}, P_j^{in}, and the sum of all P_k^{in} for $\Delta_{ik} < \Delta_{ik}$, i.e., other opportunities accessible at a shorter distance. More formally:

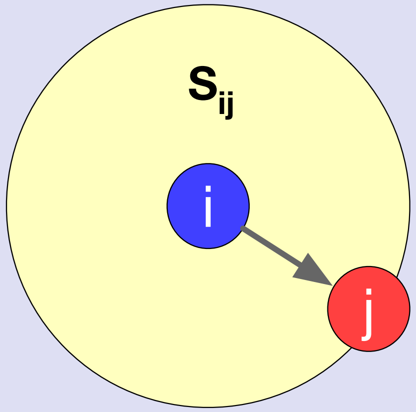
$$R_{ij} = k_i^{out} \frac{P_i^{out} P_j^{in}}{(P_i^{out} + s_{ij})(P_i^{out} + P_j^{in} + s_{ij})}$$

With $s_{ij} = \sum_{u \in V, \Delta_{iu} < \Delta_{ij}} P_u^{in}$ the sum of opportunities at a shorter distance than the target.

[a] Simini et al. 2012.

Radiation Law of Spatial Interactions

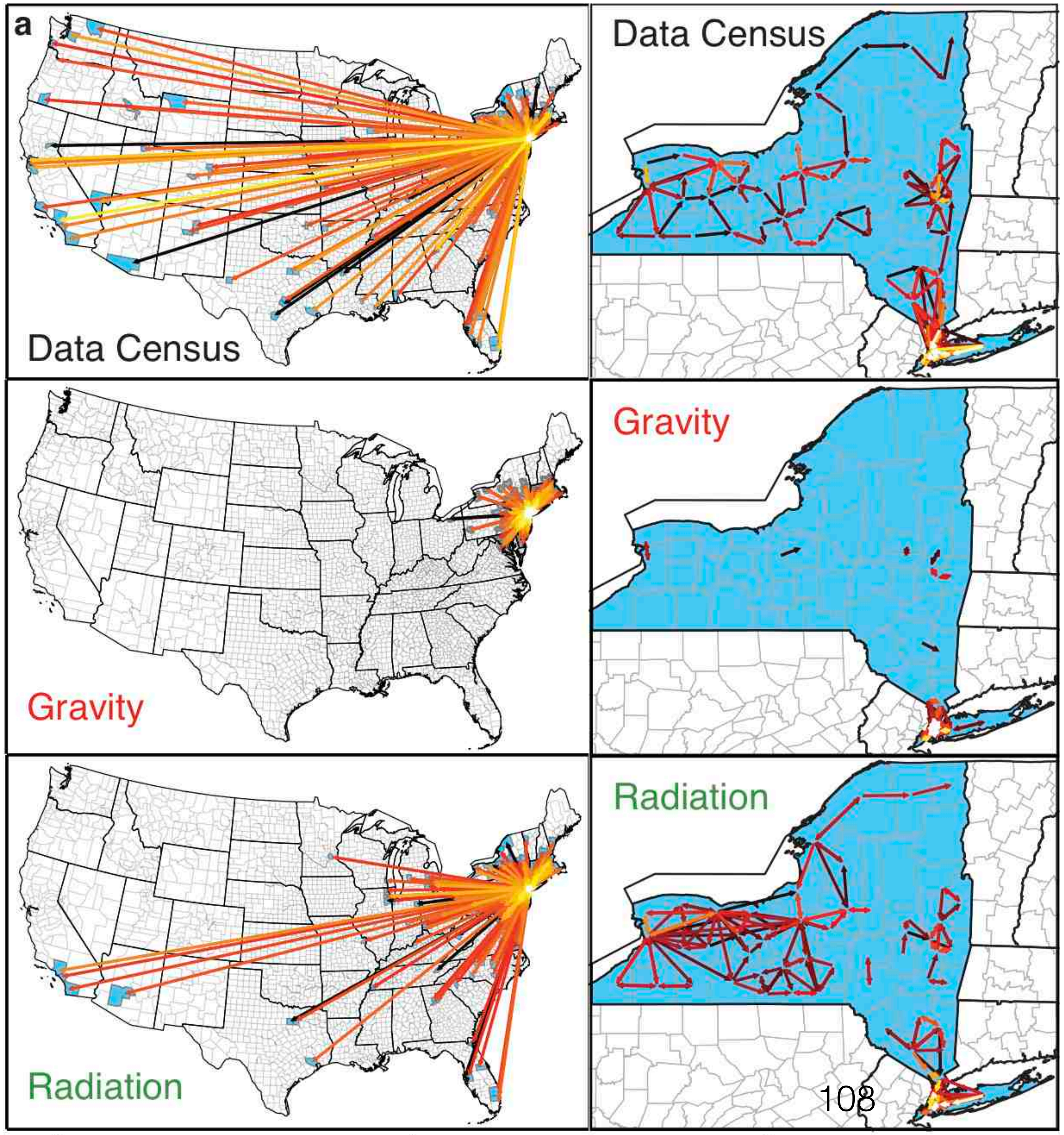
Illustration of the zone s_{ij} in which opportunities decrease the probability of interactions between i and j.

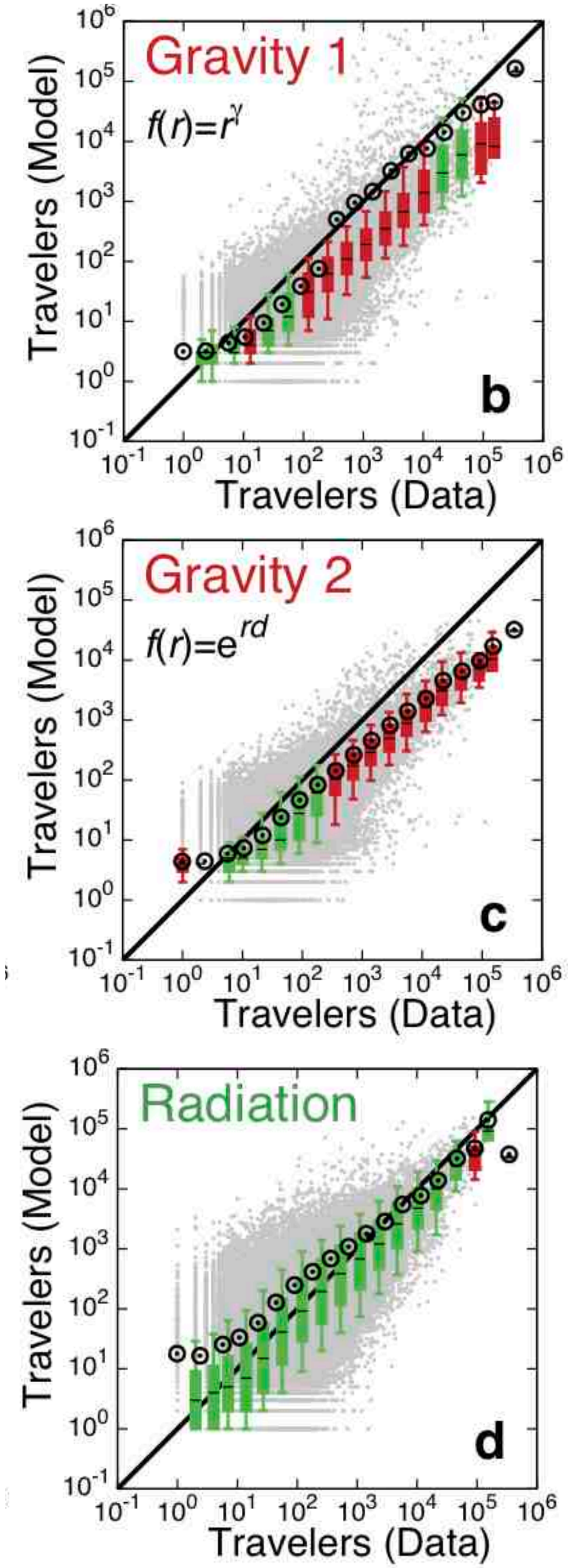


The radiation law

Comparison with census data and the gravity law predictions

Simini. et.al, Nature 2010





Radiation Law VS Gravity Law

+ Radiation:
  - No parameters
  - Two nodes of same degrees at similar distance can have different edge probability based on their location
+ Gravity:
  - Customizable deterrence function… The real world is complex !

End notes

- “All models are wrong, but some are useful”
- ER models, Configuration models, Gravity models are used as reference models in a large number of applications
- WS, BA models are more “making a point” type models: simple processes can explain some non-trivial properties of networks, unfound in random networks.
- Correlation is not causation. Are these simple processes the “cause” ? Maybe, maybe not, sometimes…