Big Data Course

Dynamic network processes

Marc Santolini and Liubov Tupikina

27 Nov 2019

CRI Paris

CANCELED

Planning		
9	18 Dec (3 ho Anirudh, Felix	**nic Digital Drivers i Open** to the digi data form & database) to p project(s) ale, significance, ing the digital tools for 2 orking with the digital tools
10	8 Jan (3 hours) - Anirudh, Felix	**h** and peer-evaluation strong and 3 weak 's project) oject tutors and
11	15 Jan (Anirudh, Felix	**dual/Group arch Project** 1. Work session - gr ntinues.
12	22 Jan (3 hou - Anirudh, Felix	**Individual/Group Mental Health Research Project** Final presentations and discussion

OUTLINE: OTHER CHANGES

Planning		
N°	**Type (CM/TP/TD) & hourly rate**	Describe here the content of each session. (1. means first part of the course, 2. the second part, ~1h20 each)
1	16 October (3 hours) - Marc, Liubov, Anirudh, Felix	**Introduction to Networks and Big Data Efforts** 1. **Introduction** to the network part of the course "Why network science?" with presentation of **topics** covered 2. Introduction to the big data part, with a focus on big data for mental health and presentation of **topics** covered
2	23 October (3 hours) - Marc, Liubov	**Networks - basics of network analysis and visualisation** Network metrics and data analysis & create teams for personal projects 1. Theory (network construction, centralities, statistical significance, networkx, other packages) 2. Hands-on session working on a dataframe, creating a network, making some statistical analysis, network visualisation
3	6 Nov (3 hours) - Loic	**Big Data - Infrastructure of big data 1** • Introduction to data engineering, what is the day-to-day jobs and what skills are needed **Data engineering**
4	13 Nov (3 hours) - Loic	**Big Data - Infrastructure of big data 2**

~~12 December~~ —> MOVED TO JANUARY

Planning		
5	20 Nov (3 hours) - Marc, Liubov, Raphael	**Networks - mobility, web-based data** 1. Hands-on Liuba mobility / geographically embedded networks 2. Hands-on How to get data from Youtube/Twitter/etc APIs for data analysis (Raphael Tackx)
6	27 Nov (3 hours) - Marc, Liubov	**Networks - spreading processes** 1. Dynamics of networks: network growth, network attack, hands on session 2. Dynamics on networks: information spreading (eg spread of fake news, softwares), hands on session
7	4 Dec (3 hours) - Marc, Liubov	**Networks - advanced topics** 1. Temporal networks, networks with attributes: multilayer, multiplex, simplicial, etc. 2. **Intro to statistics and data science with R**
8	11 Dec (3 hours) - Marc, Liubov	**Networks EVALUATION part 1** Reverse classroom: 10 minutes presentation of a contribution to Wikipedia about an advanced topic / a paper related to network science. **Show slides including modifications of wikipedia page, max 10 slides (1 min per slide)**
9	18 Dec (3 hours) - Marc, Liubov	**Network EVALUATION part 2** **10 min project presentation** (visualisation and descriptive analysis of network data) through an **interactive demonstration of a project notebook including images**

TEXTBOOK: CH. 3-5 AND 8

http://networksciencebook.com/

Network Science

by Albert-László Barabási

TODAY'S GOAL

1. Understand how networks can change in time

2. Exercises to handle dynamic processes on networks

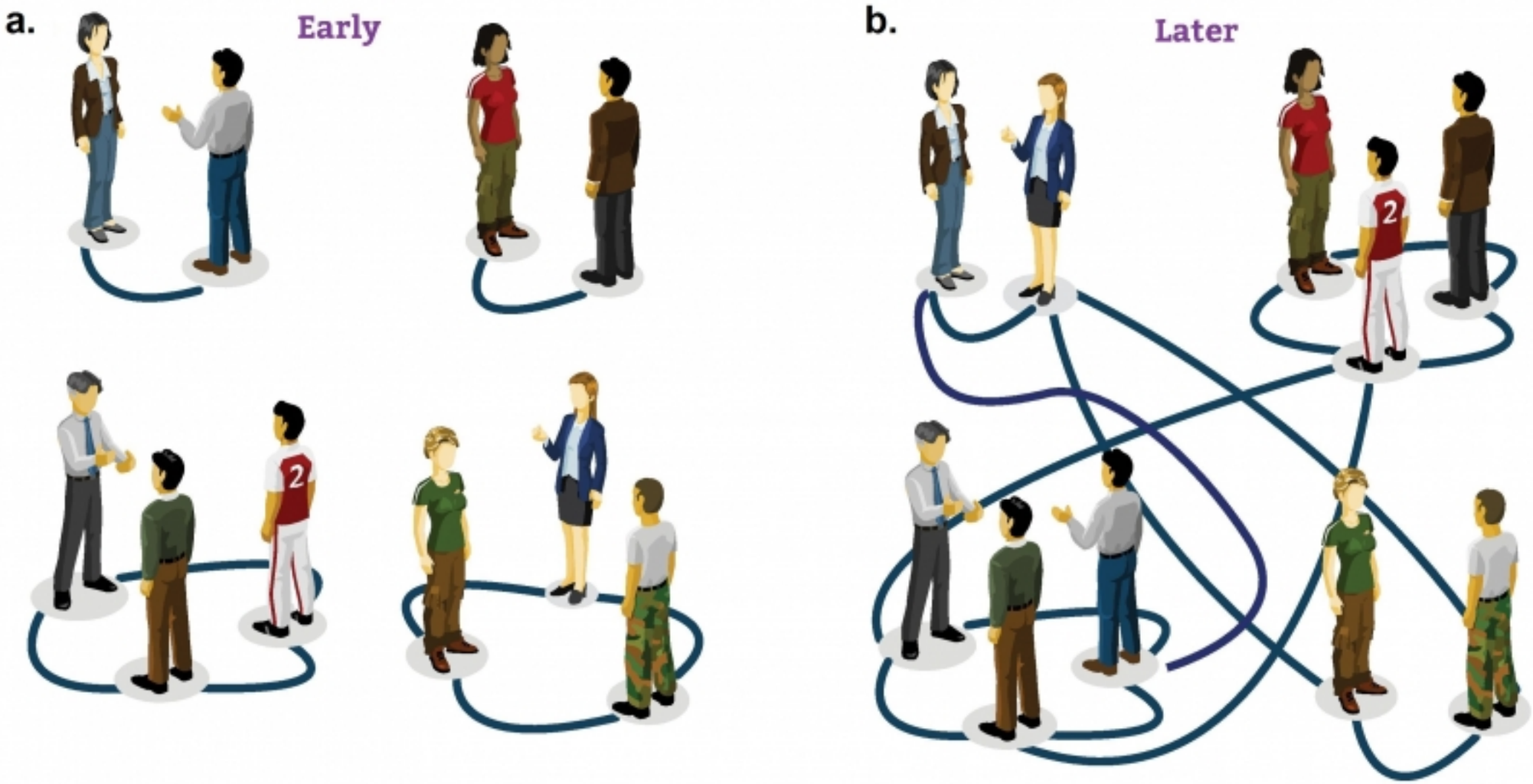
b

Before duplication

How to grow a network

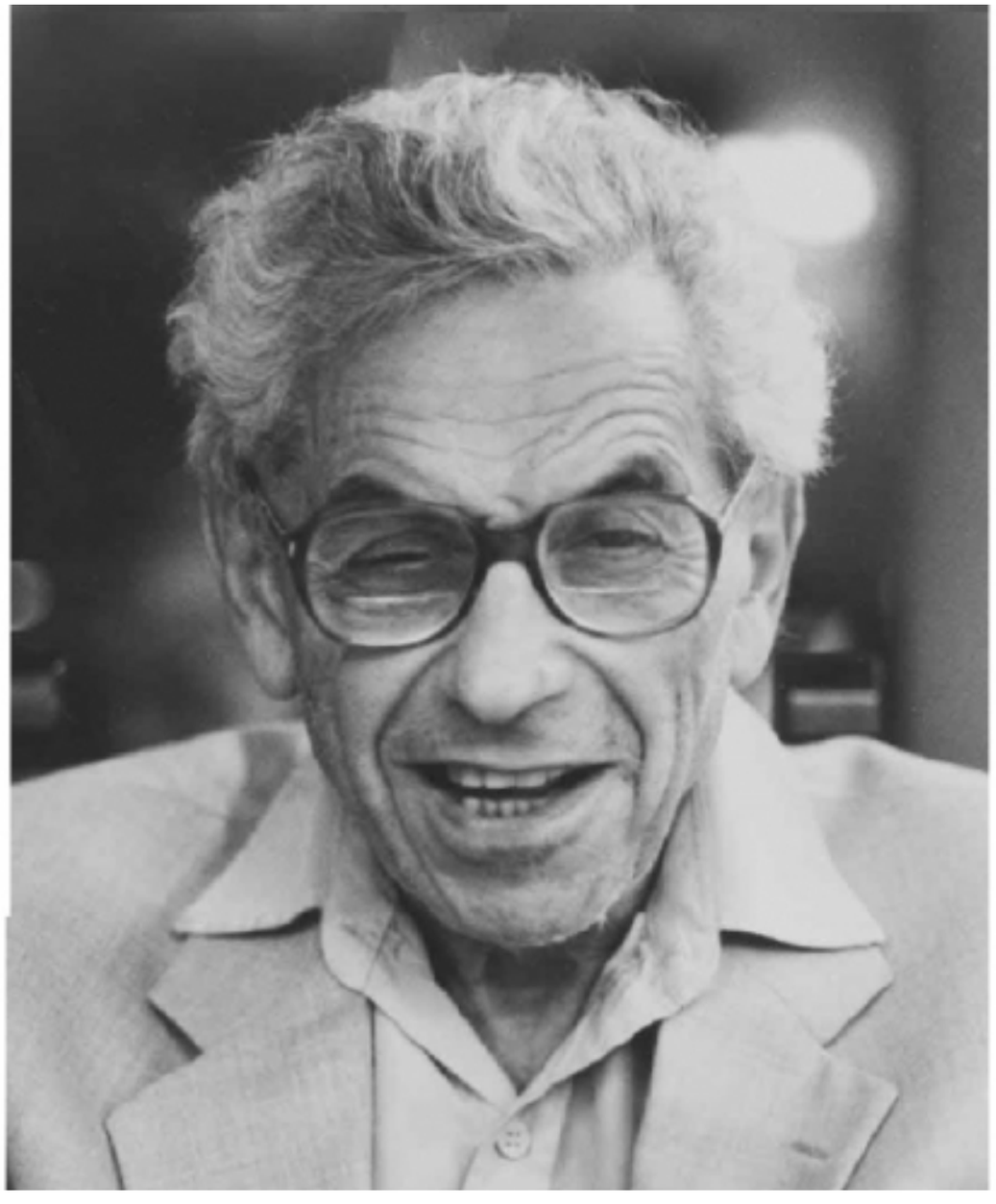
From random to preferential attachment

After duplication


a.
Early
b.
Later
2
2

Erdös-Renyi, 1959-1968

a.



Pál Erdős (1913-1996)

500+ co-authors

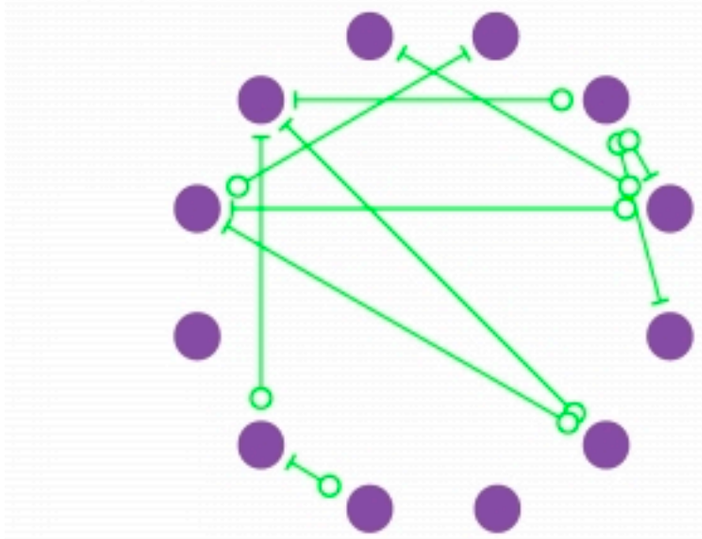
b.

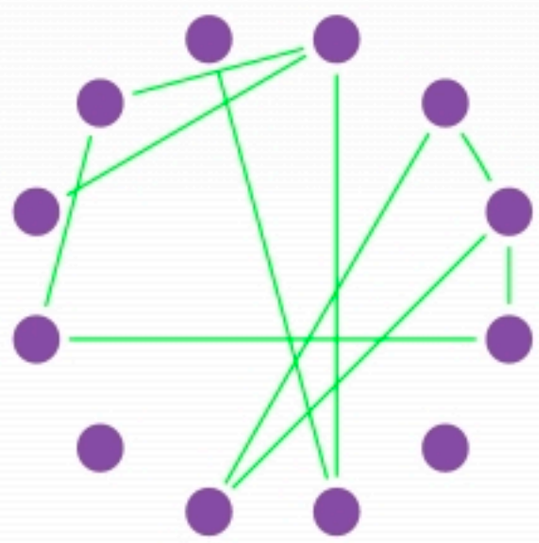
Alfréd Rényi (1921-1970)

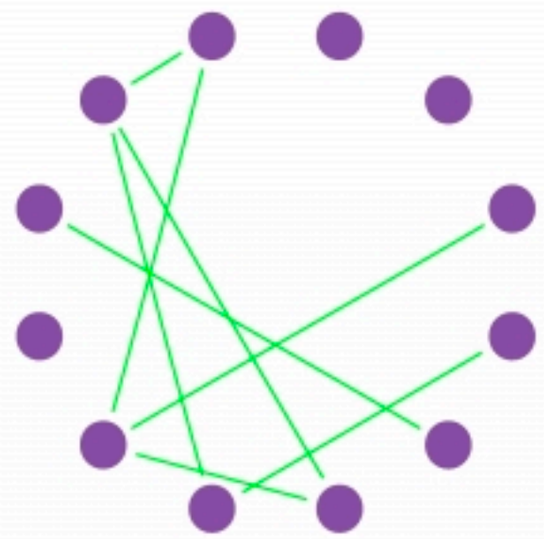
“A mathematician is a device for turning coffee into theorems”

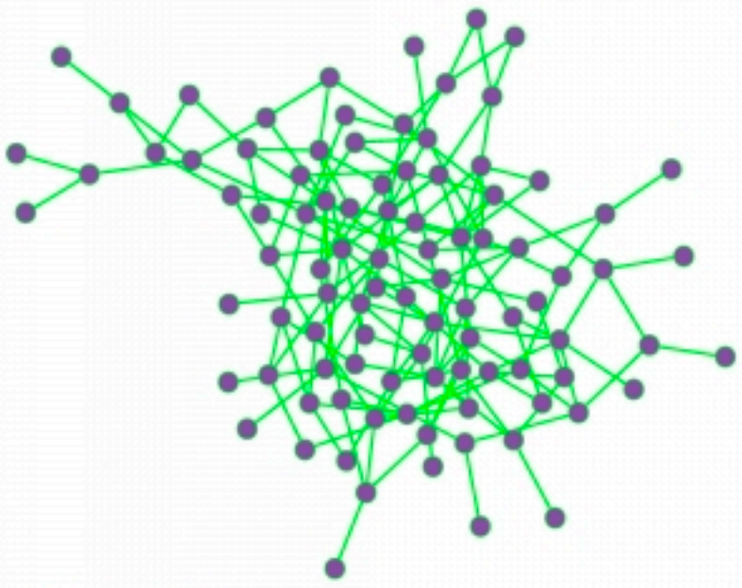
ER RANDOM PROCESS

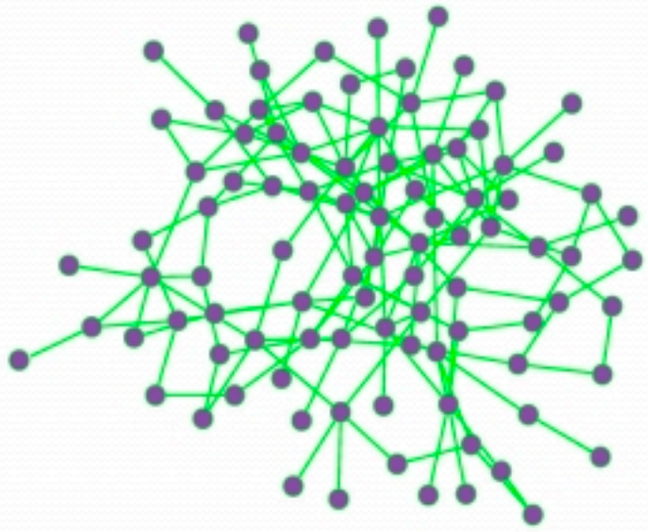
1. Create empty network with N isolated nodes
2. Select a node pair and generate a random number between 0 and 1. If the number exceeds p, connect the selected node pair with a link, otherwise leave them disconnected.
3. Repeat step (2) for each of the N(N-1)/2 node pairs.

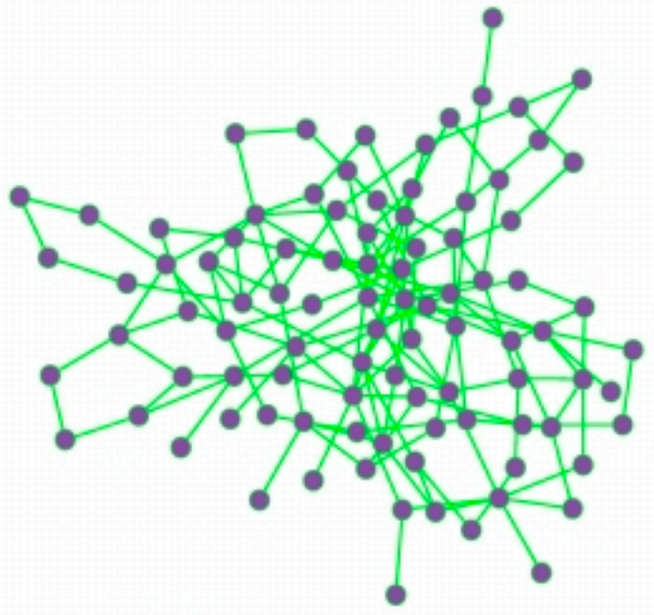




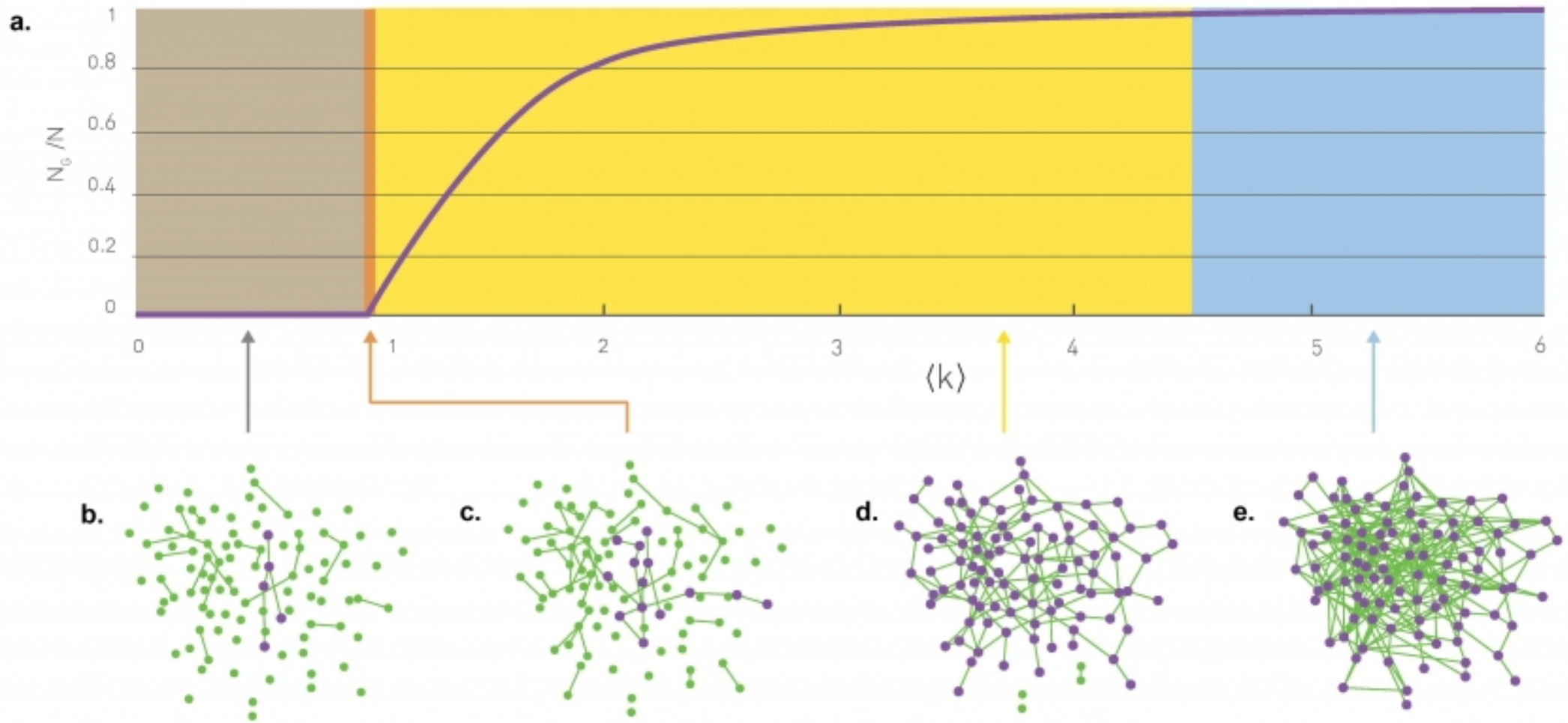








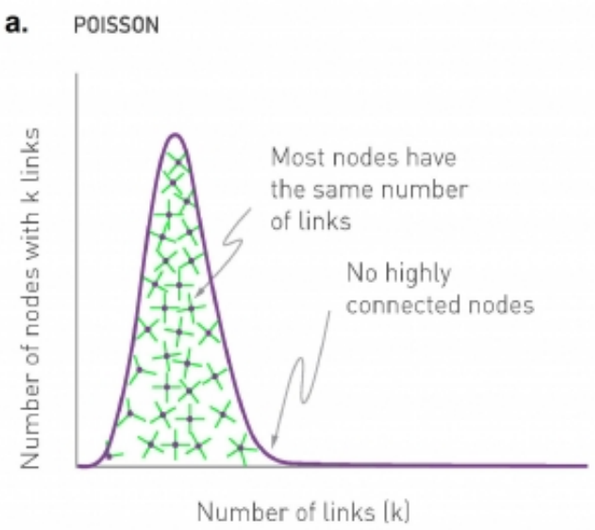
PERCOLATION TRANSITION

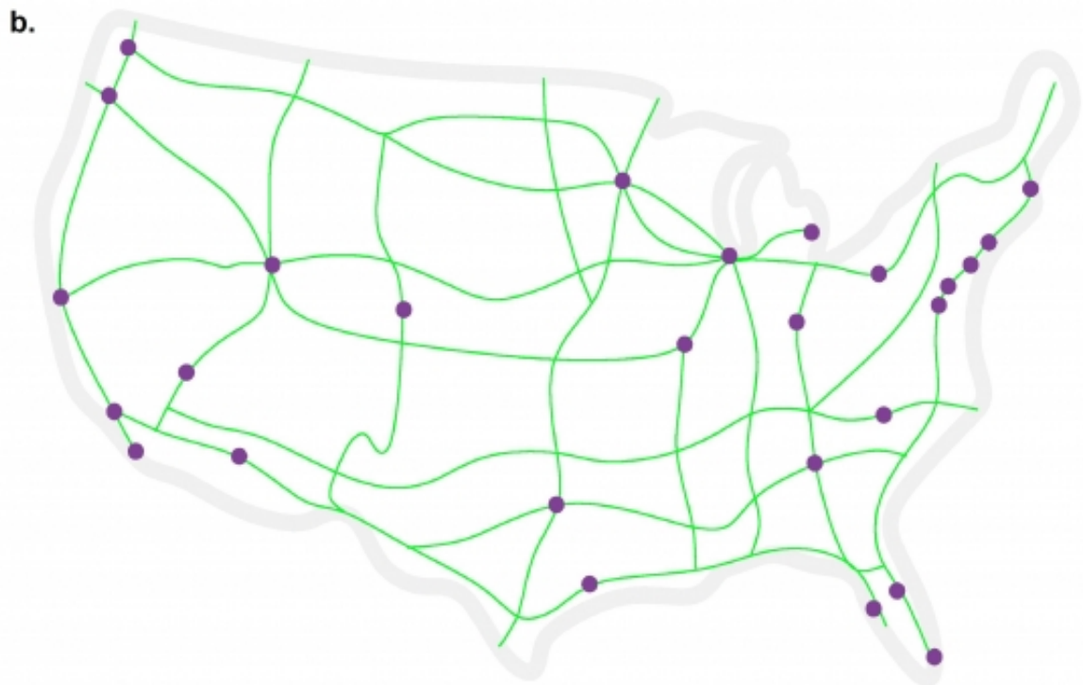


emergence of a giant component for growing density:
percolation transition at average degree $< k > = 1$

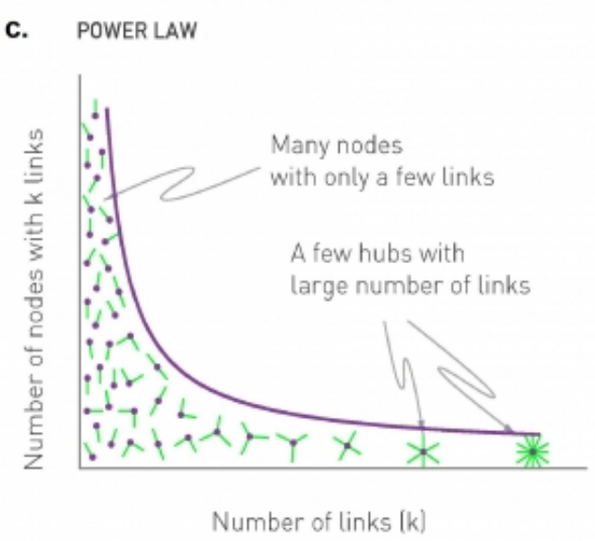
REMINDER: SCALE-FREE NETWORKS

Erdos-Renyi network

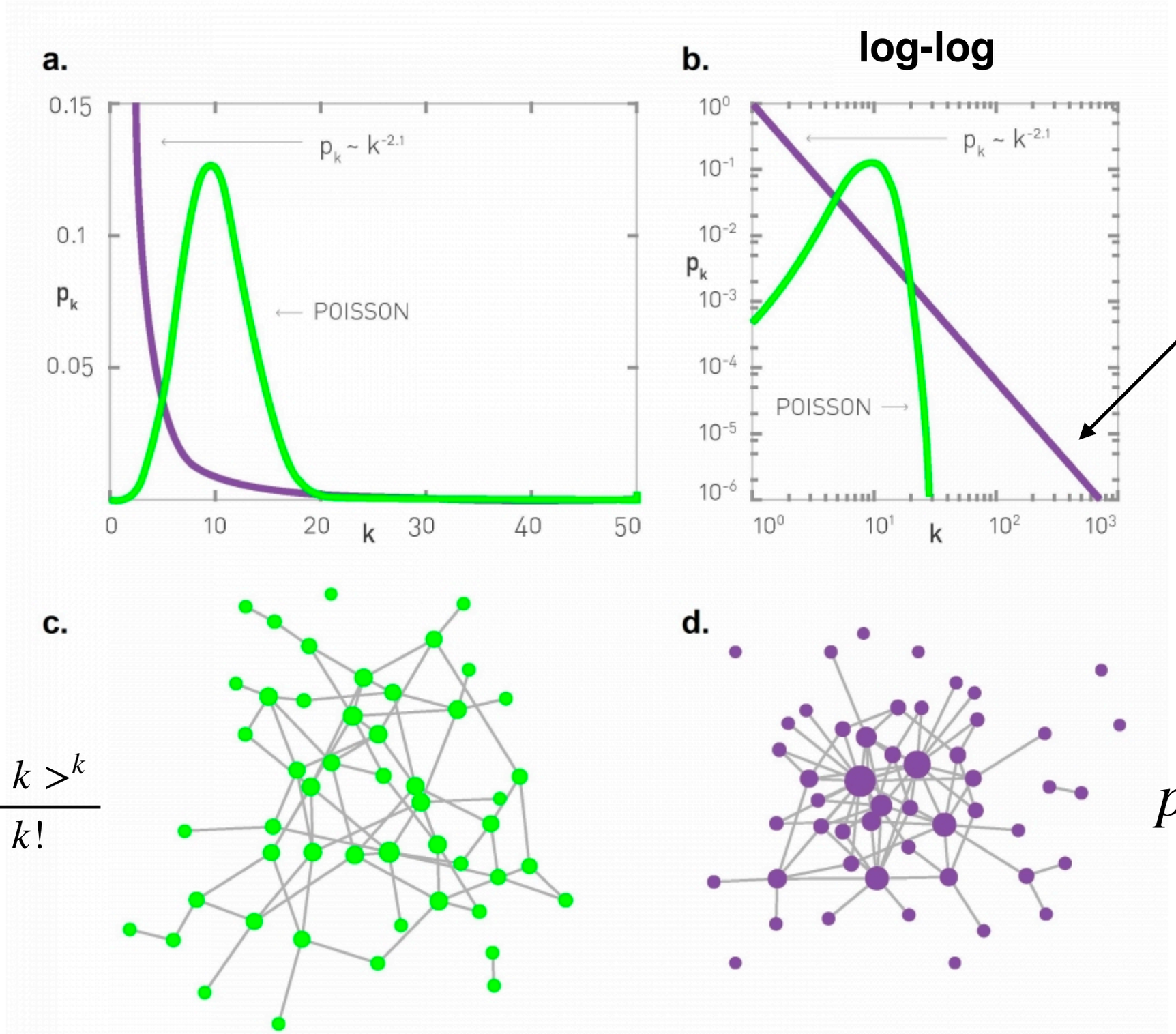




scale-free network

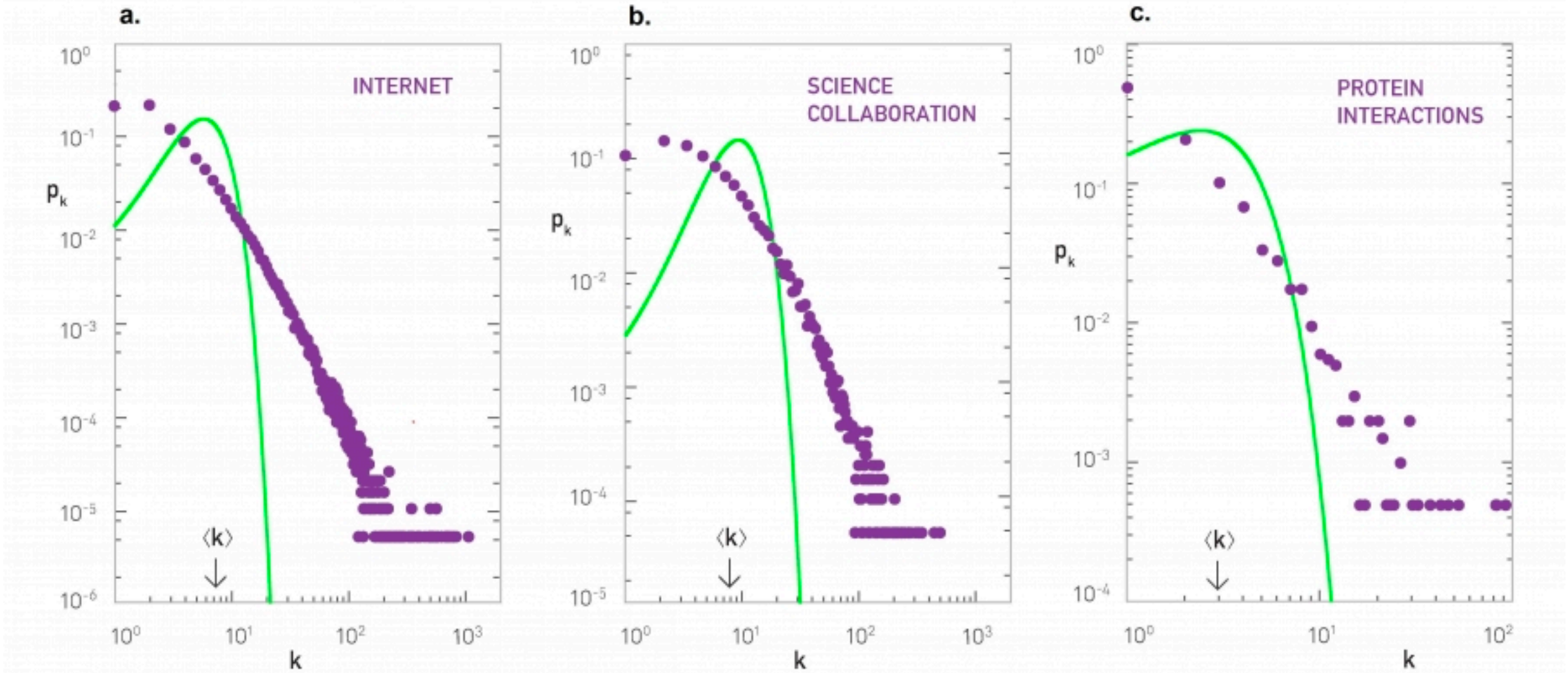


HUBS



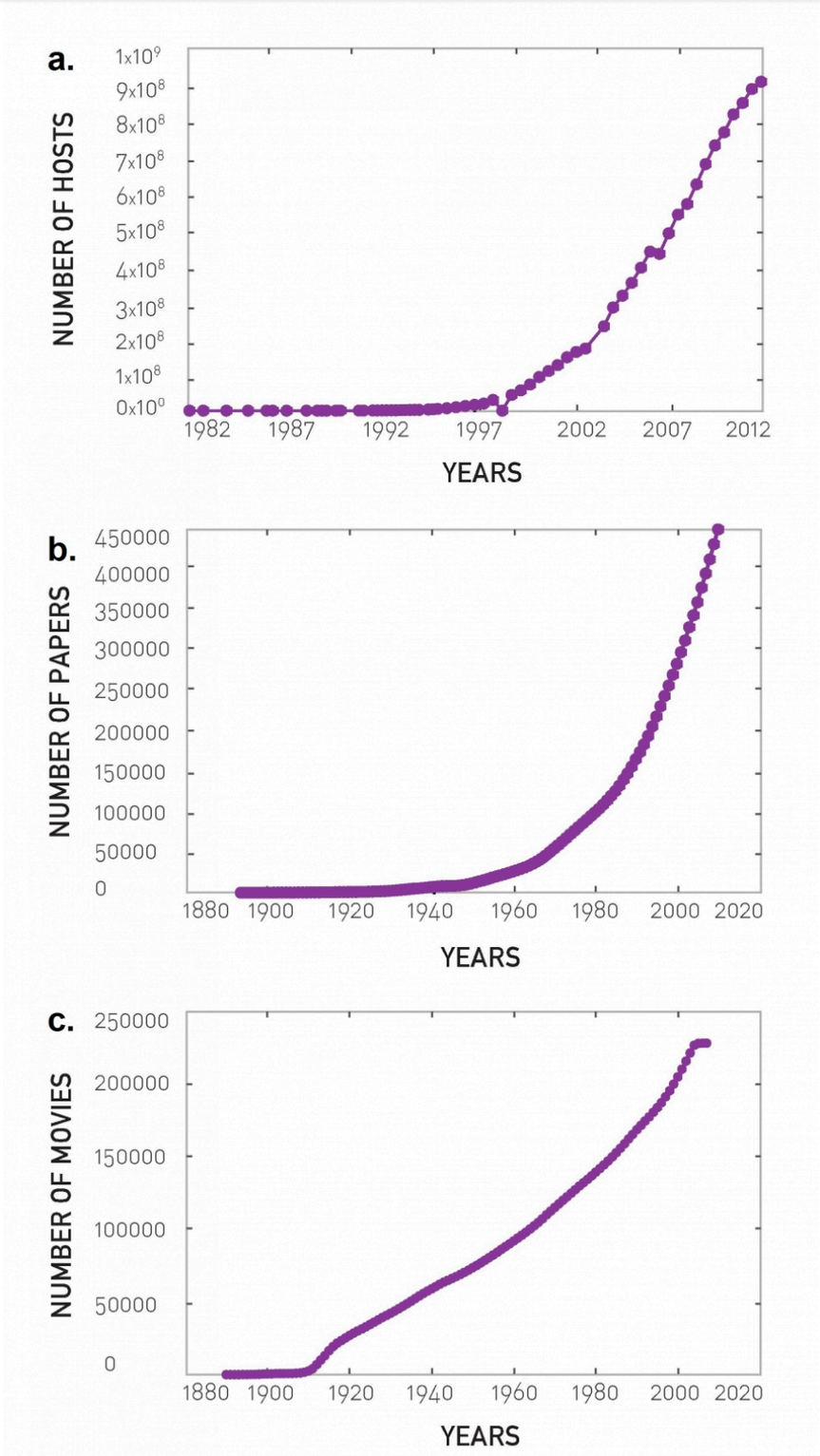
$$p_k \sim e^{-<k>} \frac{< k >^k}{k!}$$

$$p(k) \sim k^{-\gamma}$$


a.
INTERNET
b.
SCIENCE
COLLABORATION
c.
PROTEIN
INTERACTIONS
p_k
k
⟨k⟩
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^1
10^2
10^3

REAL NETWORKS GROW IN TIME

number of nodes increases not just density

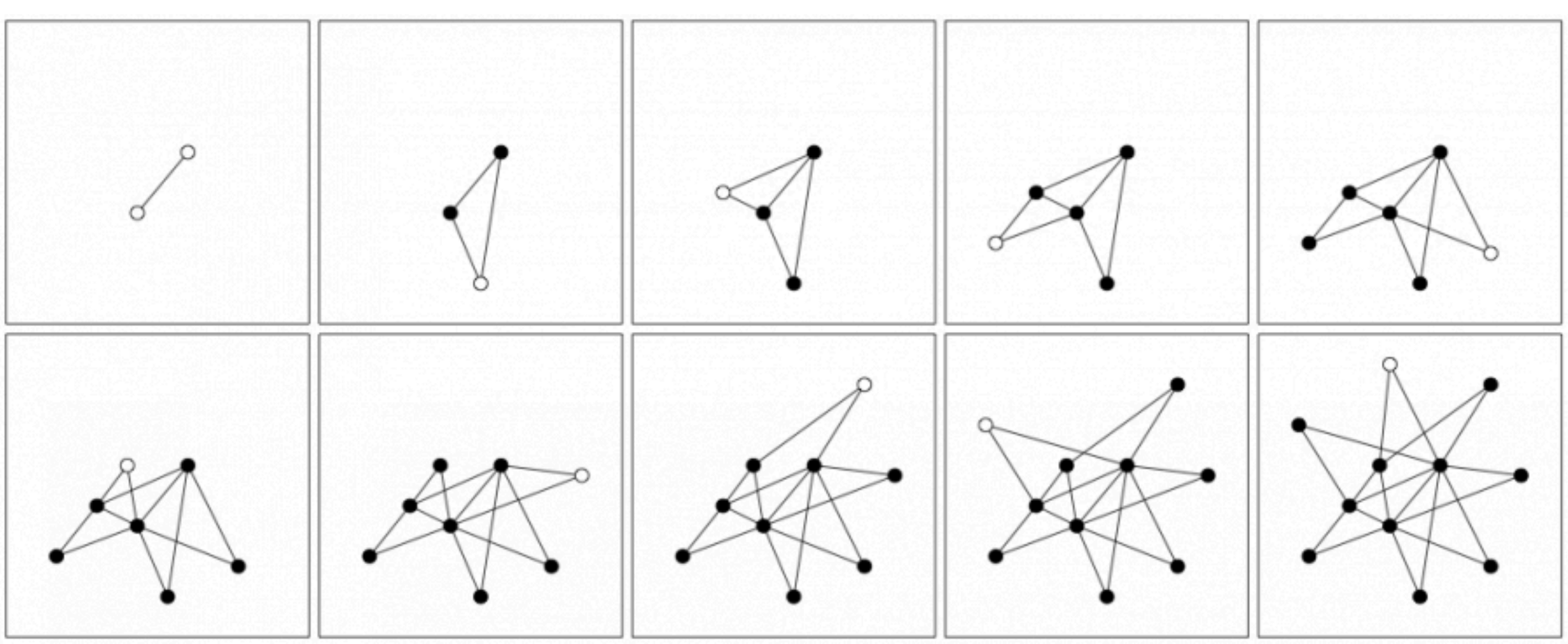


Barabasi-Albert model 1999

Preferential attachement

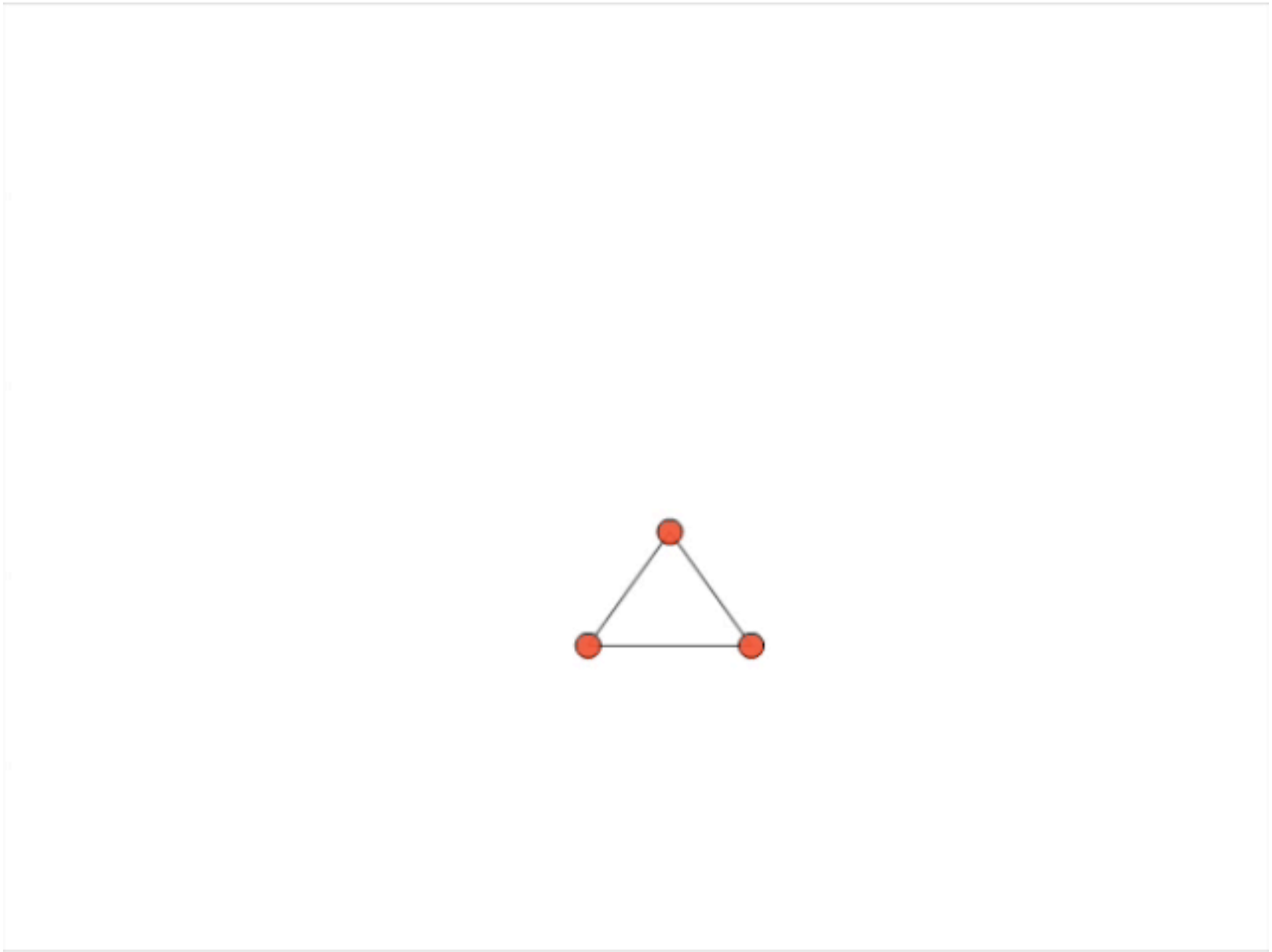
Probability $\Pi(k)$ that a link from a new node connects to node i:

$$\Pi(k_i) \sim \frac{k_i}{\sum_j k_k}$$

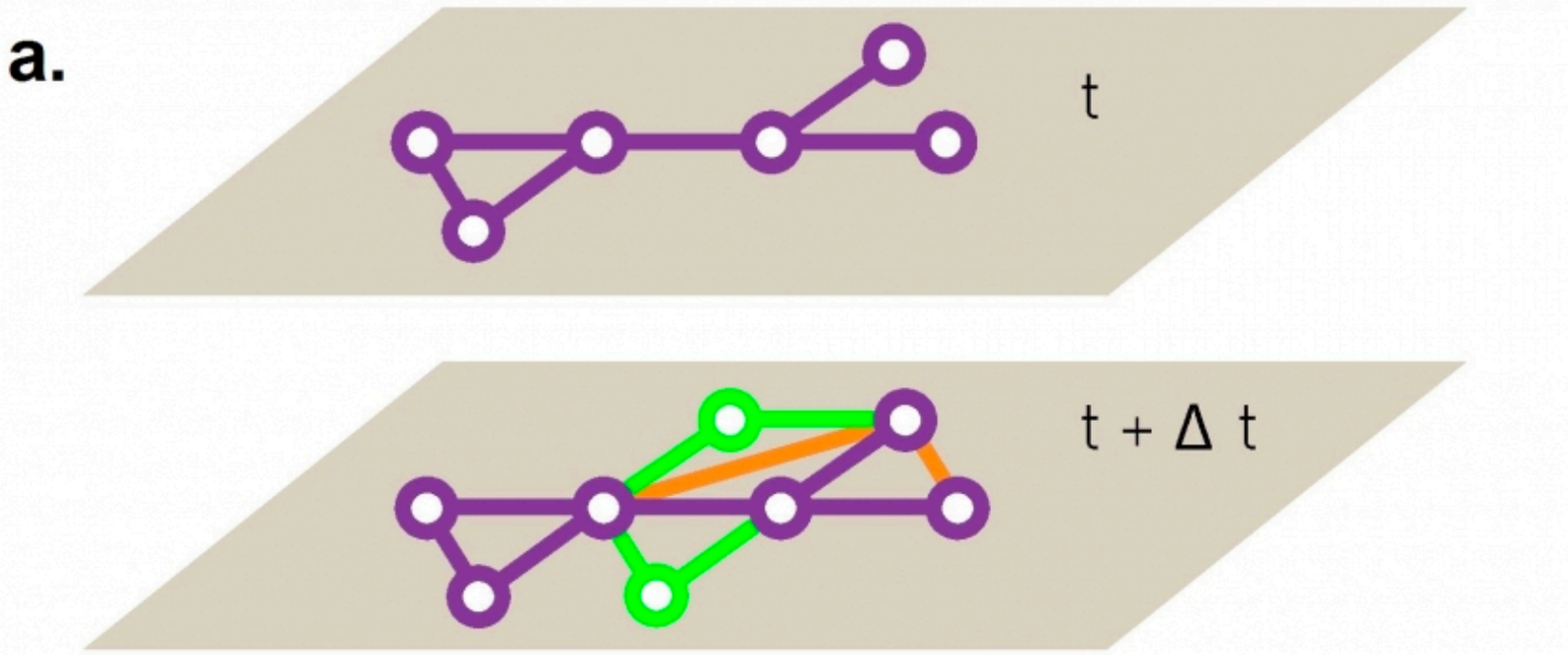


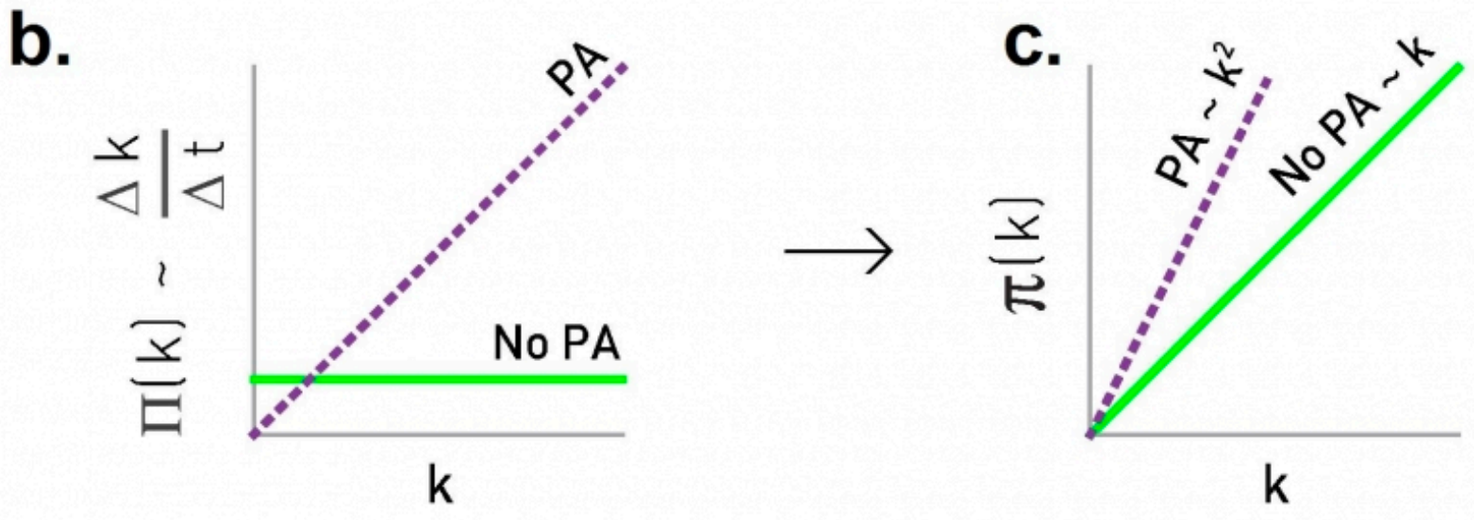
Barabasi-Albert model 1999

Preferential attachement

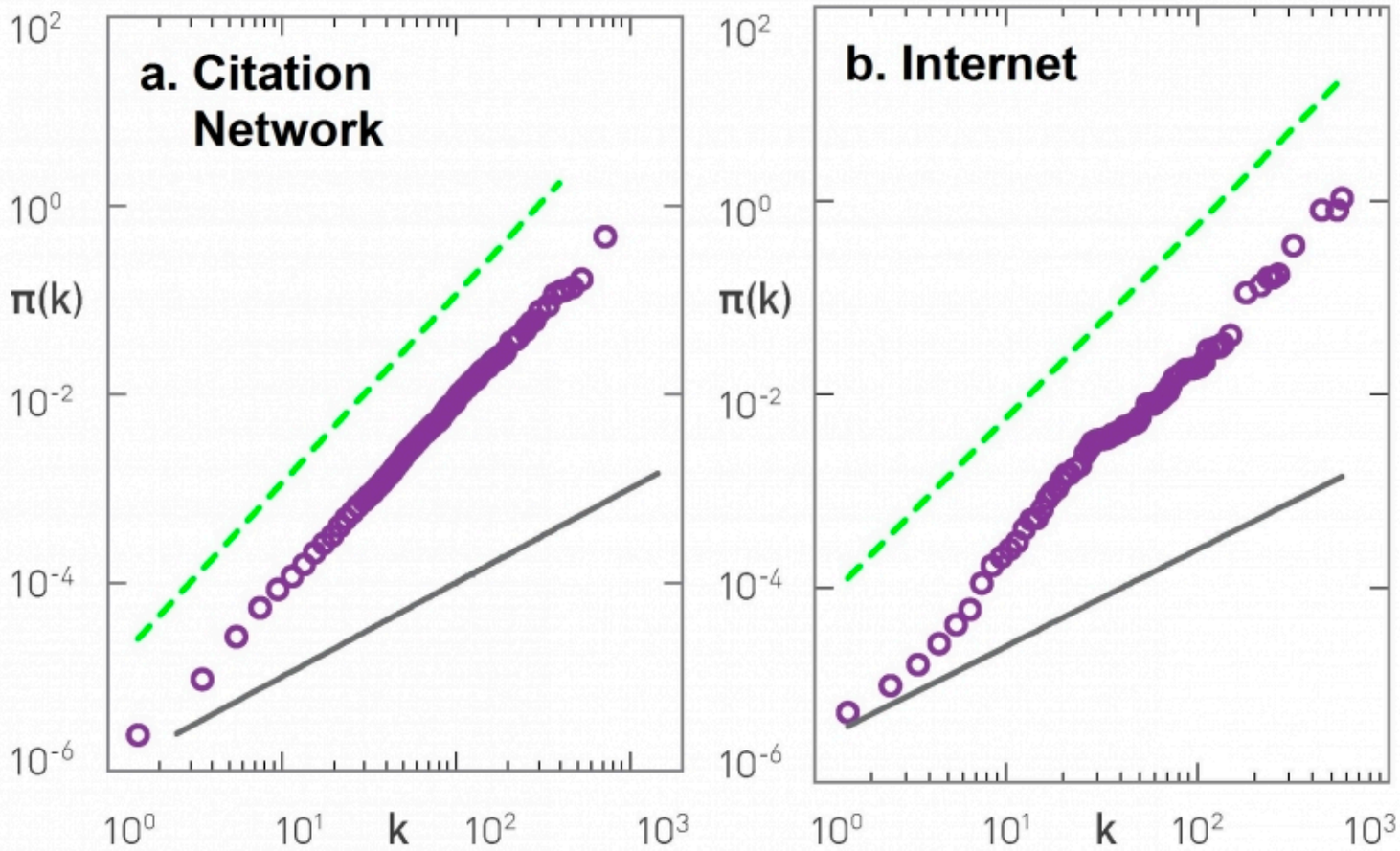


Measuring Preferential Attachement (PA)





$$\pi(k) = \sum_{k_i=0}^{k} \Pi(k_i)$$

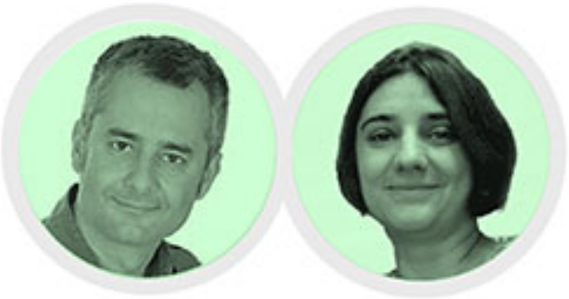

a. Citation Network
b. Internet
c. Collaboration Network
d. Actor Network
π(k)
k
10^2
10^0
10^-2
10^-4
10^-6
10^1
10^3
10^4

György Pólya
PÓLYA PROCESS
MATHEMATICIAN

George Kinsley Zipf
WEALTH DISTRIBUTION
ECONOMIST

Herbert Alexander Simon
MASTER EQUATION
POLITICAL SCIENTIST

Robert Merton
MATTHEW EFFECT
SOCIOLOGIST



Albert-László Barabási & Réka Albert
PREFERENTIAL ATTACHMENT
NETWORK SCIENTISTS

George Udmy Yule
YULE PROCESS
STATISTICIAN

Robert Gibrat
PROPORTIONAL GROWTH
ECONOMIST

Derek de Solla Price
CUMULATIVE ADVANTAGE
PHYSICIST

MILESTONES

PUBLICATION DATE

1923 1925 1931 1935 1941 1945 1950 1955 1960 1968 1970 1976 1980 1985 1990 1995 1999 2000 2005 2010

XXI

György Pólya (1887-1985)
Preferential attachment made its first appearance in 1923 in the celebrated urn model of the Hungarian mathematician György Pólya [2]. Hence, in mathematics preferential attachment is often called a **Pólya process**.

George Udmy Yule (1871-1951)
used preferential attachment to explain the power-law distribution of the number of species per genus of flowering plants [3]. Hence, in statistics preferential attachment is often called a **Yule process**.

Robert Gibrat (1904-1980)
proposed that the size and the growth rate of a firm are independent. Hence, larger firms grow faster [4]. Called **proportional growth**, this is a form of preferential attachment.

George Kinsley Zipf (1902-1950)
used preferential attachment to explain the fat tailed distribution of wealth in the society [5].

Herbert Alexander Simon (1916-2001)
used preferential attachment to explain the fat-tailed nature of the distributions describing city sizes, word frequencies, or the number of papers published by scientists [6].

Derek de Solla Price (1922-1983)
used preferential attachment to explain the citation statistics of scientific publications, calling it **cumulative advantage** [7].

Robert Merton (1910-2003)
In sociology preferential attachment is often called the **Matthew effect**, named by Merton [8] after a passage in the Gospel of Matthew.

Barabási (1967) & **Albert** (1972)
introduce the term **preferential attachment** to explain the origin of scale-free networks [1].

EXAMPLE IN BIOLOGY

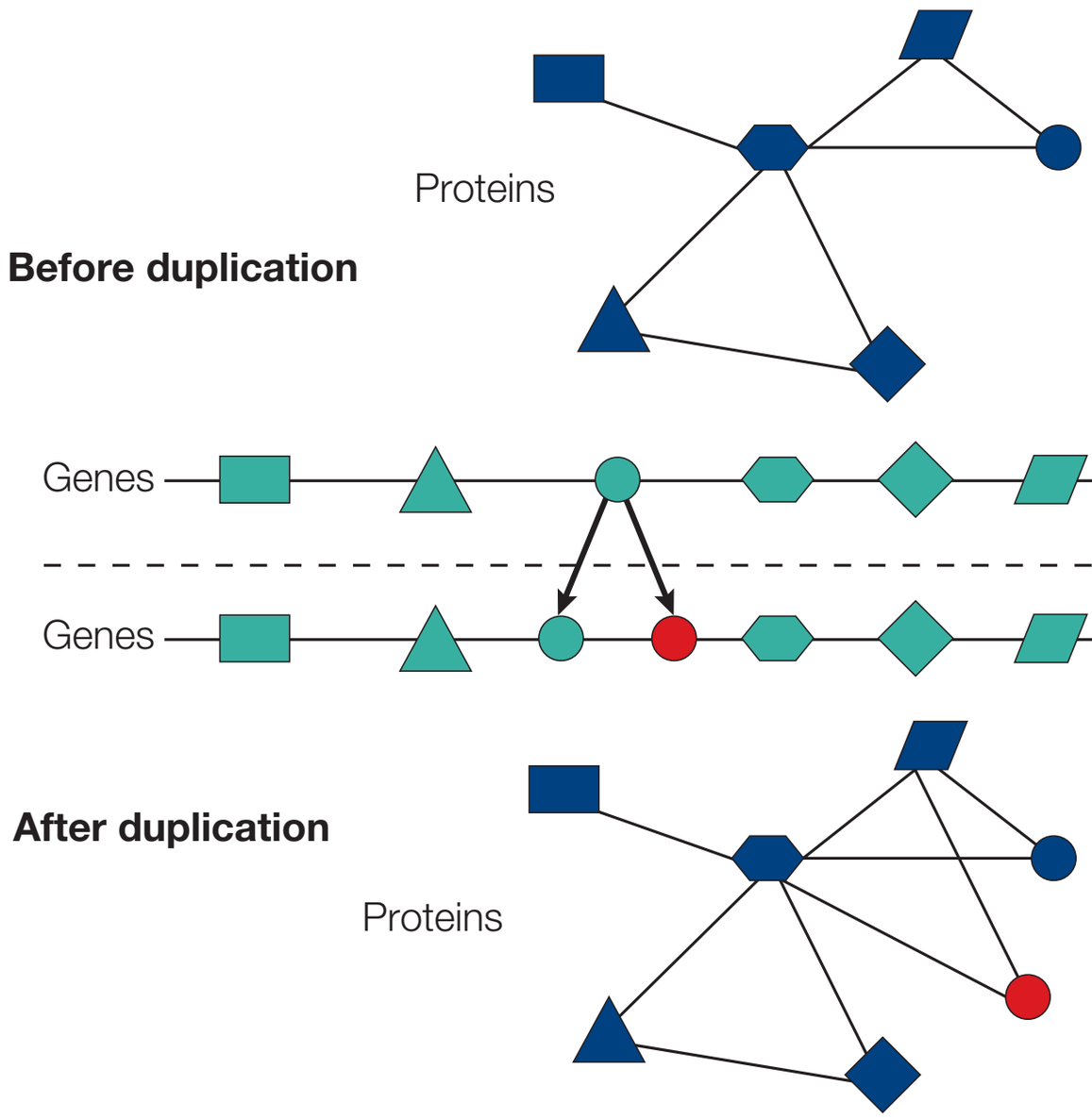
NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION

Albert-László Barabási & Zoltán N. Oltvai‡*

A key aim of postgenomic biomedical research is to systematically catalogue all molecules and their interactions within a living cell. There is a clear need to understand how these molecules and the interactions between them determine the function of this enormously complex machinery, both in isolation and when surrounded by other cells. Rapid advances in network biology indicate that cellular networks are governed by universal laws and offer a new conceptual framework that could potentially revolutionize our view of biology and disease pathologies in the twenty-first century.

Barabasi et al., Nat Rev Genet 2004

preferential attachment: highly connected proteins have higher chance to be connected to a duplicated protein



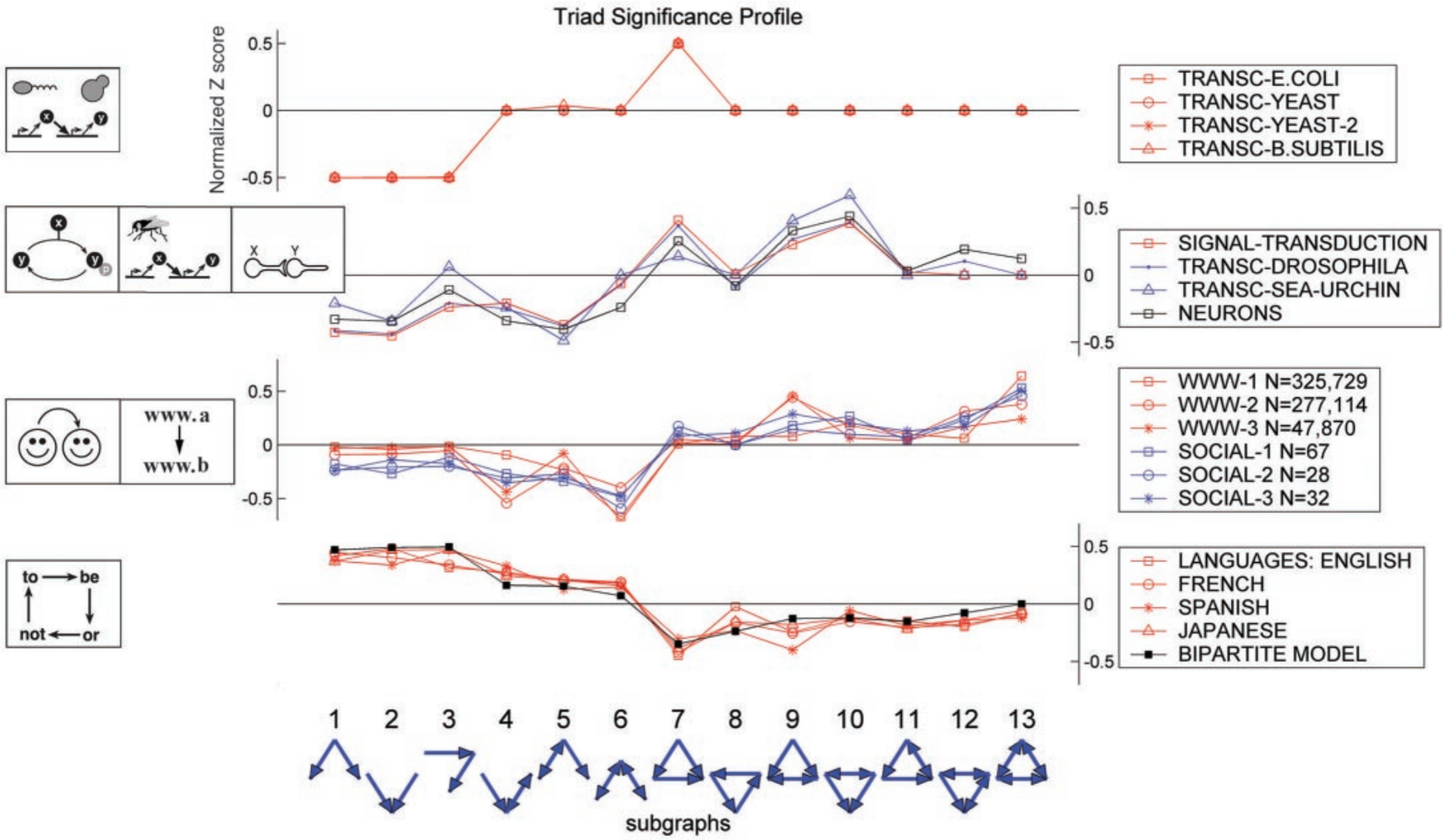
HIGHER ORDERS: TRIADIC CLOSURE

Superfamilies of Evolved and Designed Networks

Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, Uri Alon*

5 MARCH 2004 VOL 303 SCIENCE

in real networks, there is also a preferential attachment to form certain **network motifs** (like the triadic "feed-forward loop")



ADVANCED METHODS

Exponential Random Graph Models (ERGMs)

(eg in R https://statnet.github.io/Workshops/ergm_tutorial.html)

$$p(X = x) = \frac{\exp\{\theta' z(x)\}}{\kappa(\theta)}$$

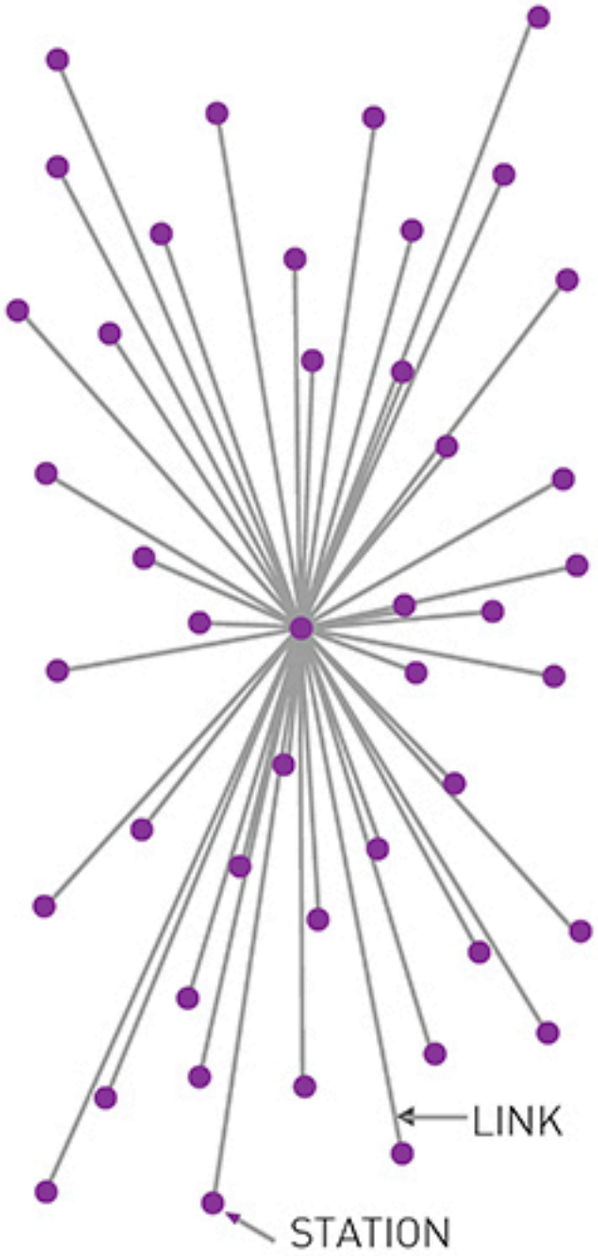
Where:
X is a random network on n nodes
x is the observed network
θ is a vector of parameters (like regression coefficients)
z(x) is a vector of network statistics
κ is a normalizing constant, to ensure the probabilities sum to 1:
$\kappa(\theta) = \sum_{\substack{all possible \\ graphs \\ x}} \exp\{\theta' z(x)\}$

Typical terms in the model

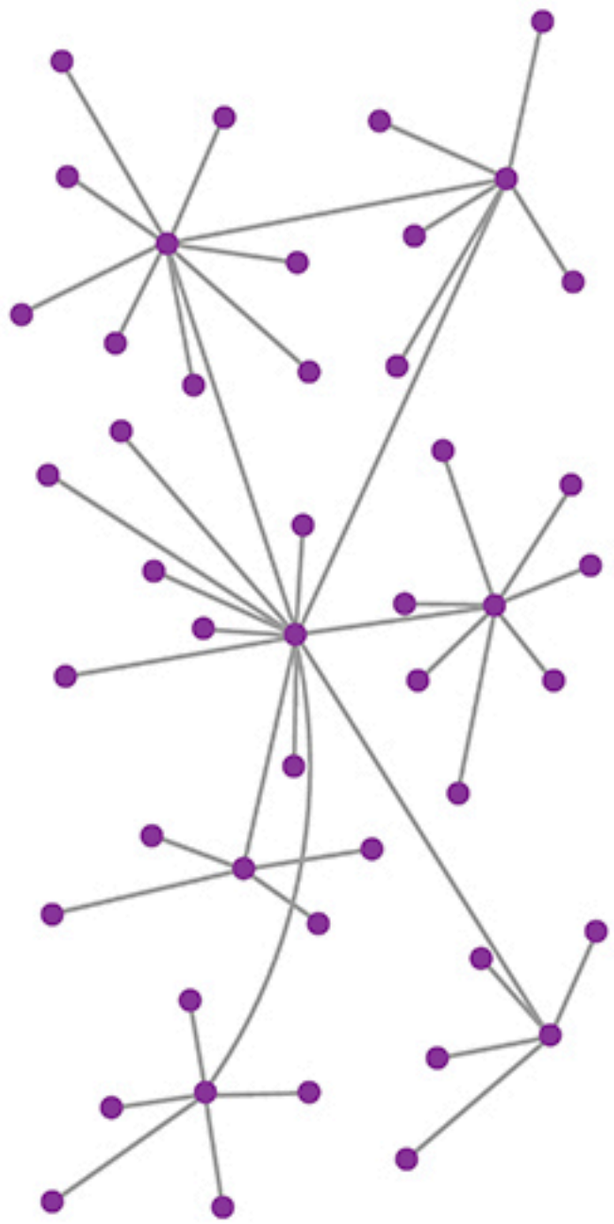
- Edges
- Mutuality
- Homophily
- Degree distribution
- Triangles
- …

How to attack a network

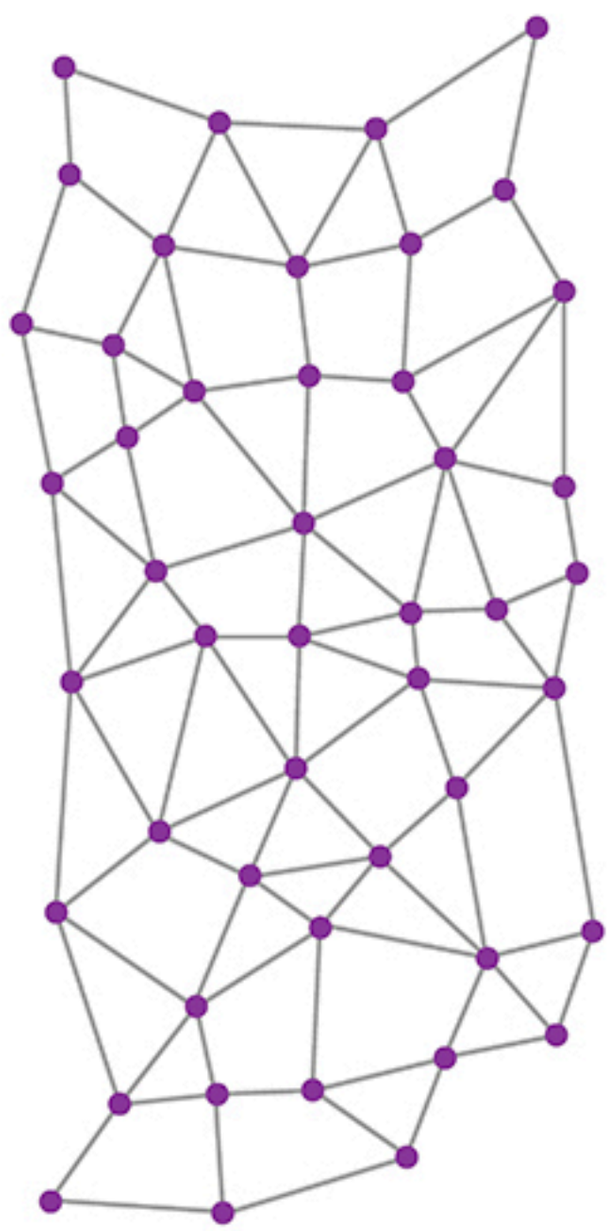
Robustness and fragility



a. CENTRALIZED



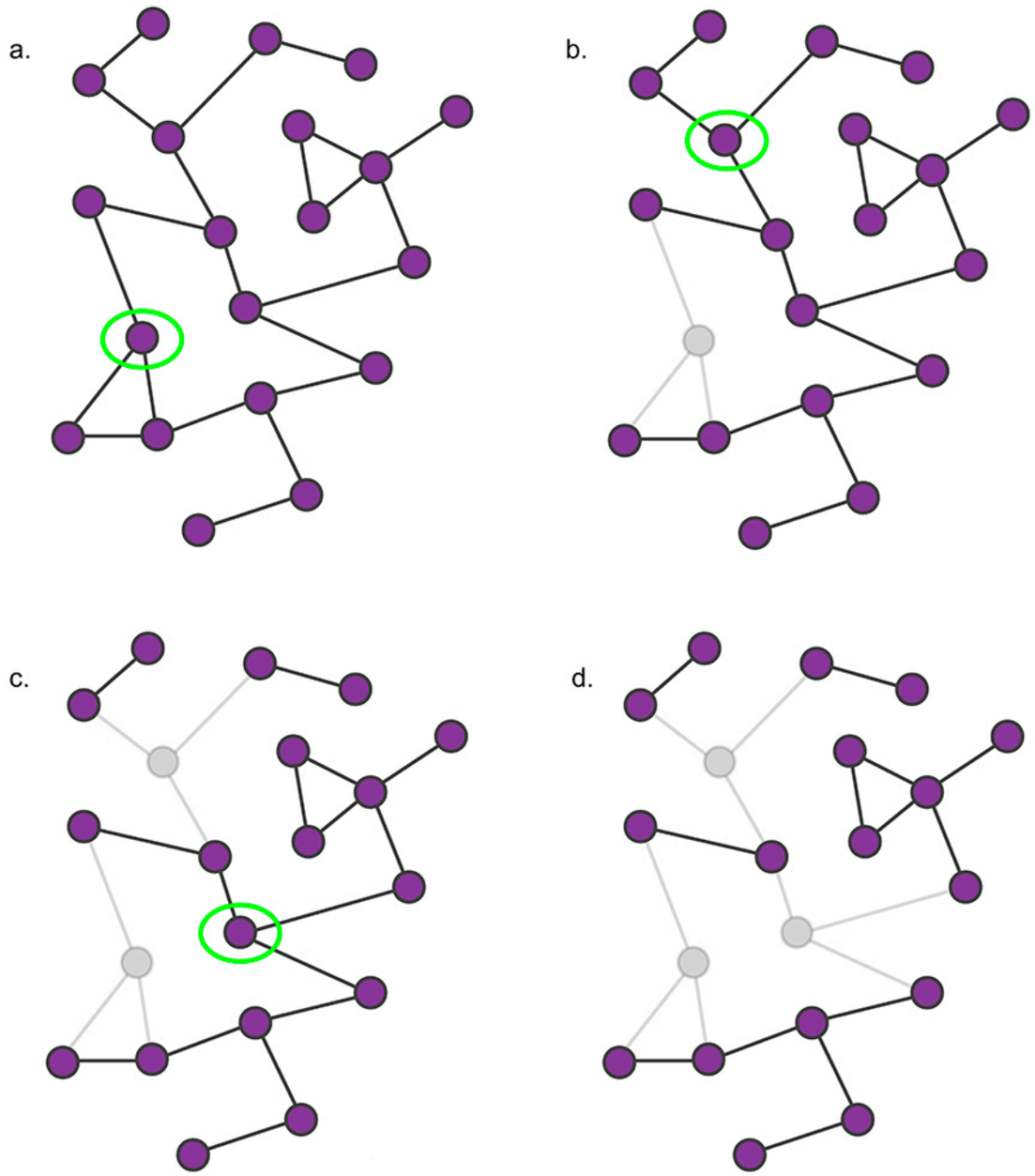
b. DECENTRALIZED



c. DISTRIBUTED

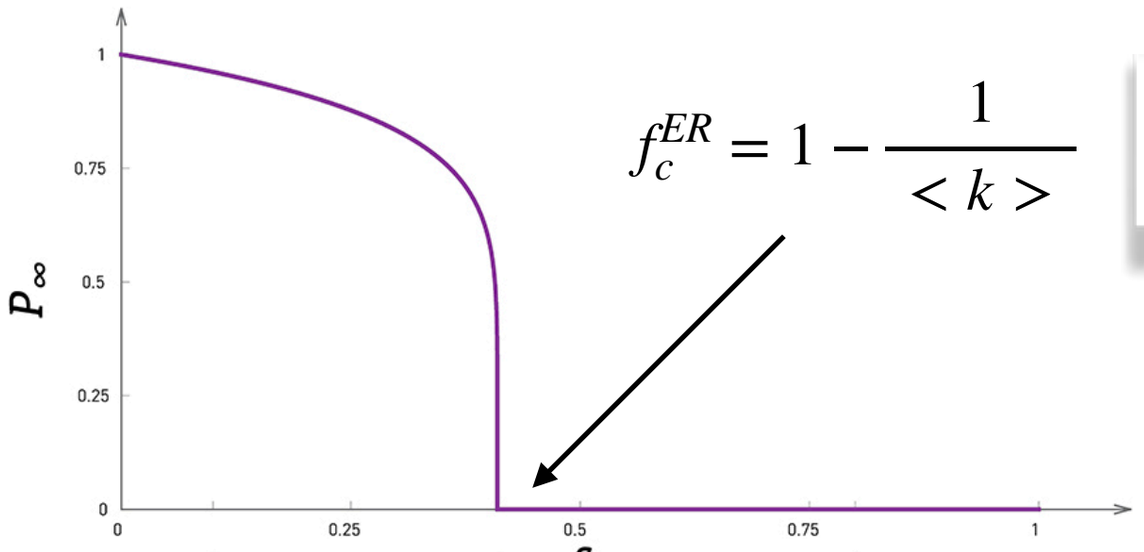
RANDOM FAILURES

At each step, remove a node at random along with its links



INVERSE PERCOLATION

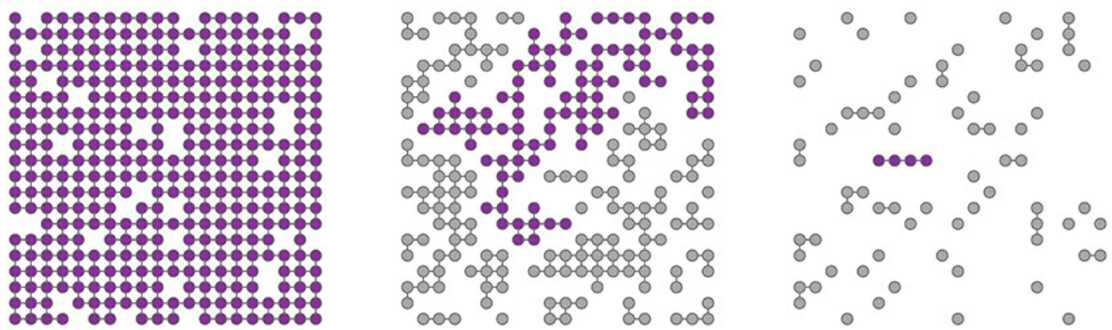
Proportion of nodes in the “giant” (largest connected) component



$$f_c^{ER} = 1 - \frac{1}{< k >}$$

for a random network, f_c is finite

square lattice (each node connected to neighbors)



$0 < f < f_c$:

There is a giant component.

$P_\infty \sim |f - f_c|^\beta$

$f = f_c$:

The giant component vanishes.

$f > f_c$:

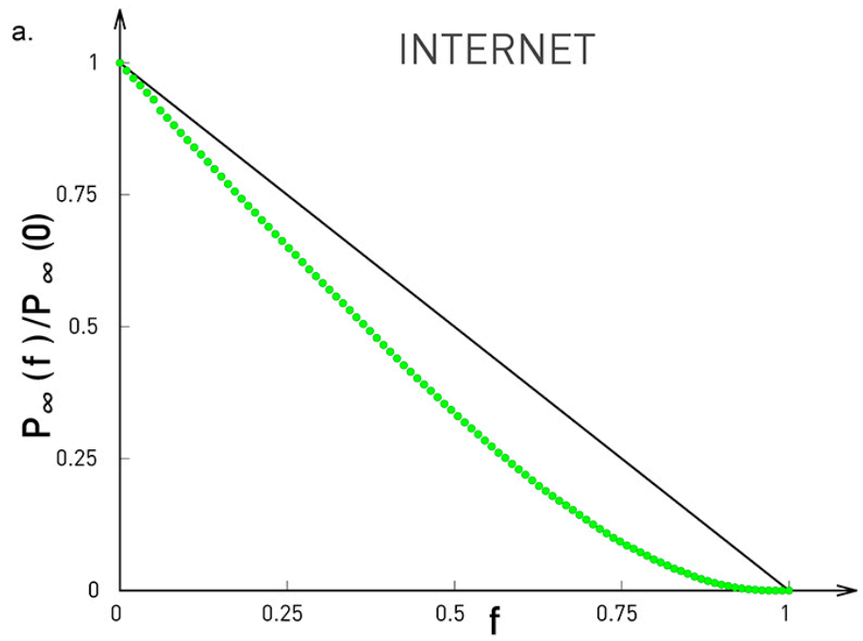
The lattice breaks into many tiny components.

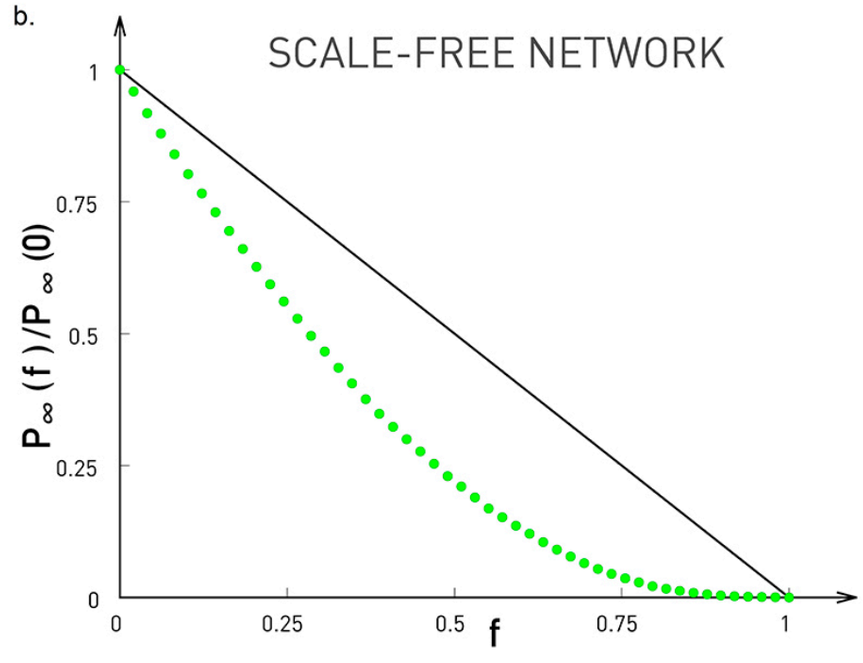
ROBUSTNESS OF SCALE FREE NETWORKS

general solution:

$$fc = 1 - \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1}$$

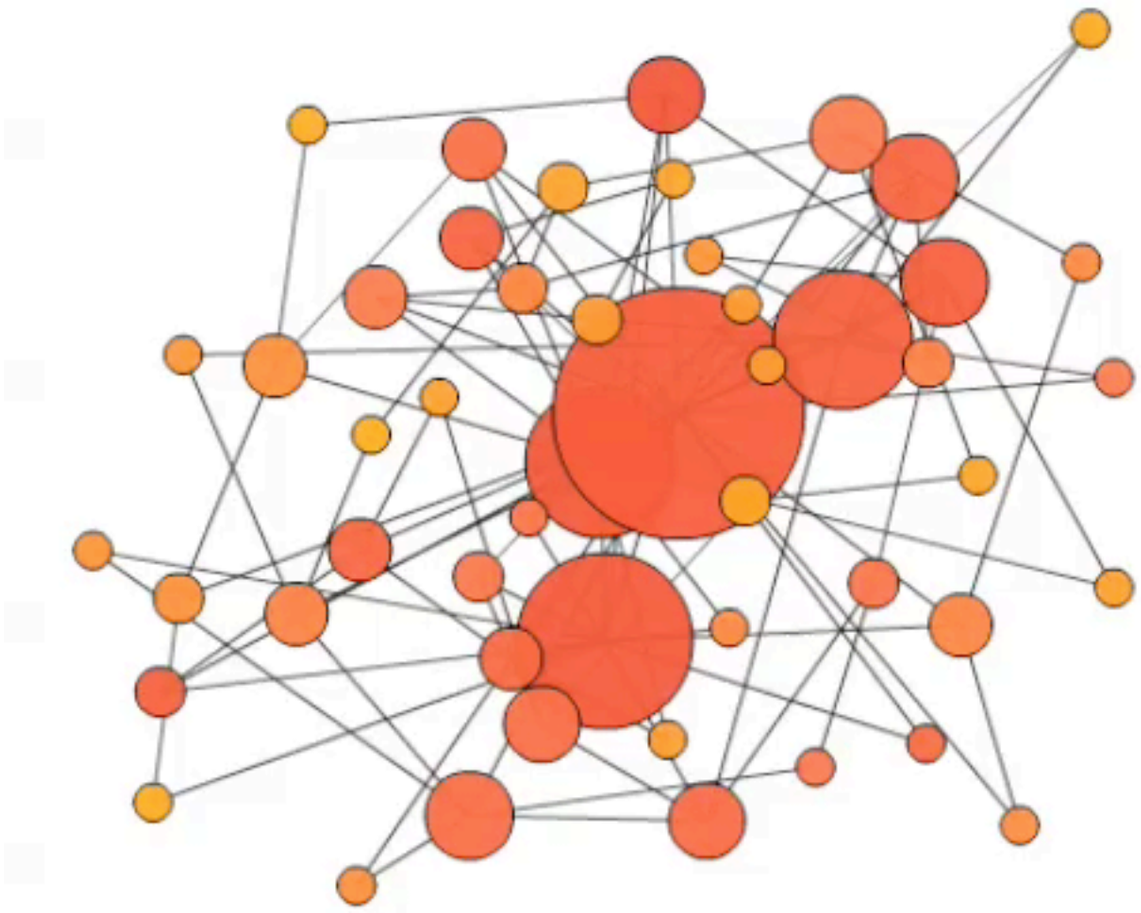
for many scale-free networks, $< k^2 > \to \infty$ to that $f_c = 1$ **and the network is absolutely robust!**





Scale-free Network Under Node Failures

random node removal



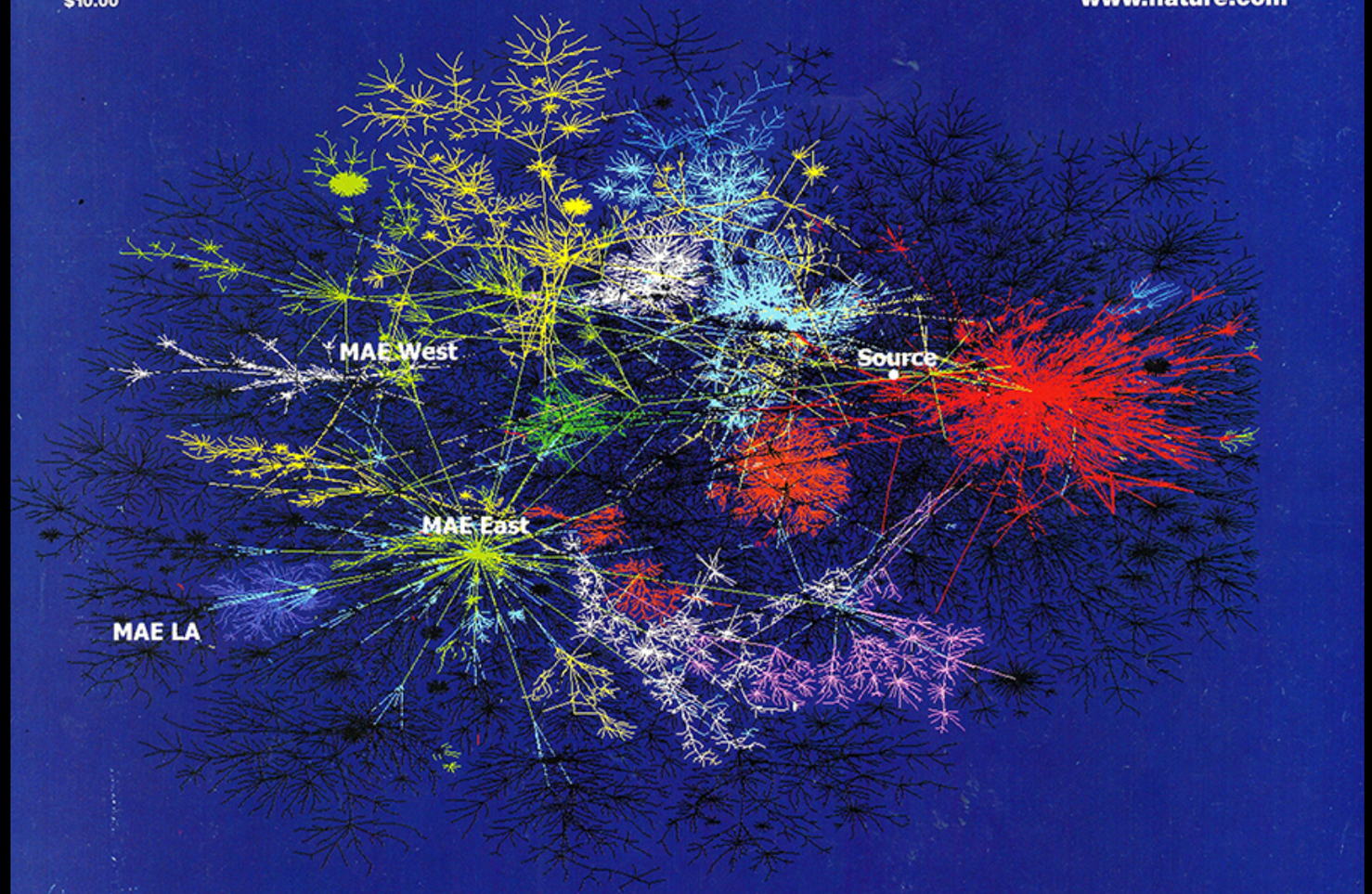
27 July 2000

International weekly journal of science

nature

$10.00

www.nature.com



Achilles' heel of the Internet

Obesity Mice that eat more but weigh less

Ocean anoxic events Not all at sea

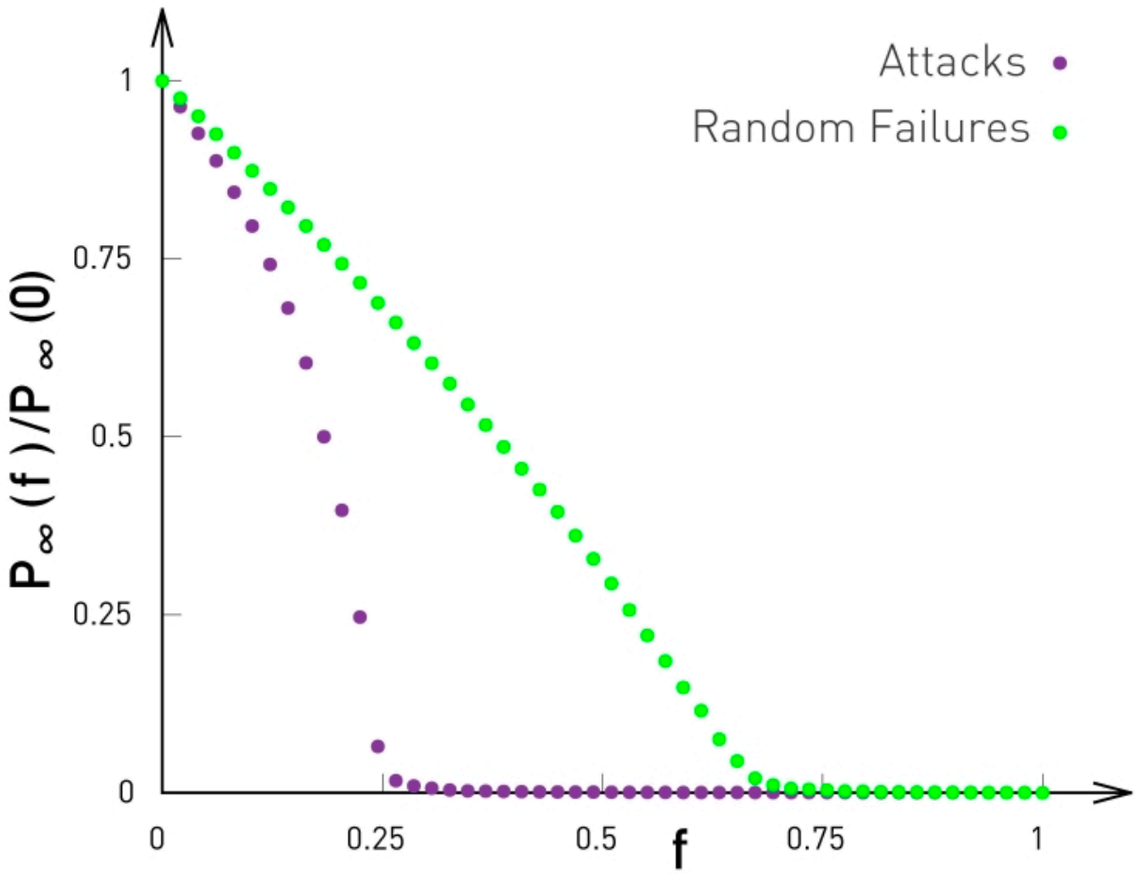
Cell signalling Fringe sweetens Notch

new on the market
oligonucleotides

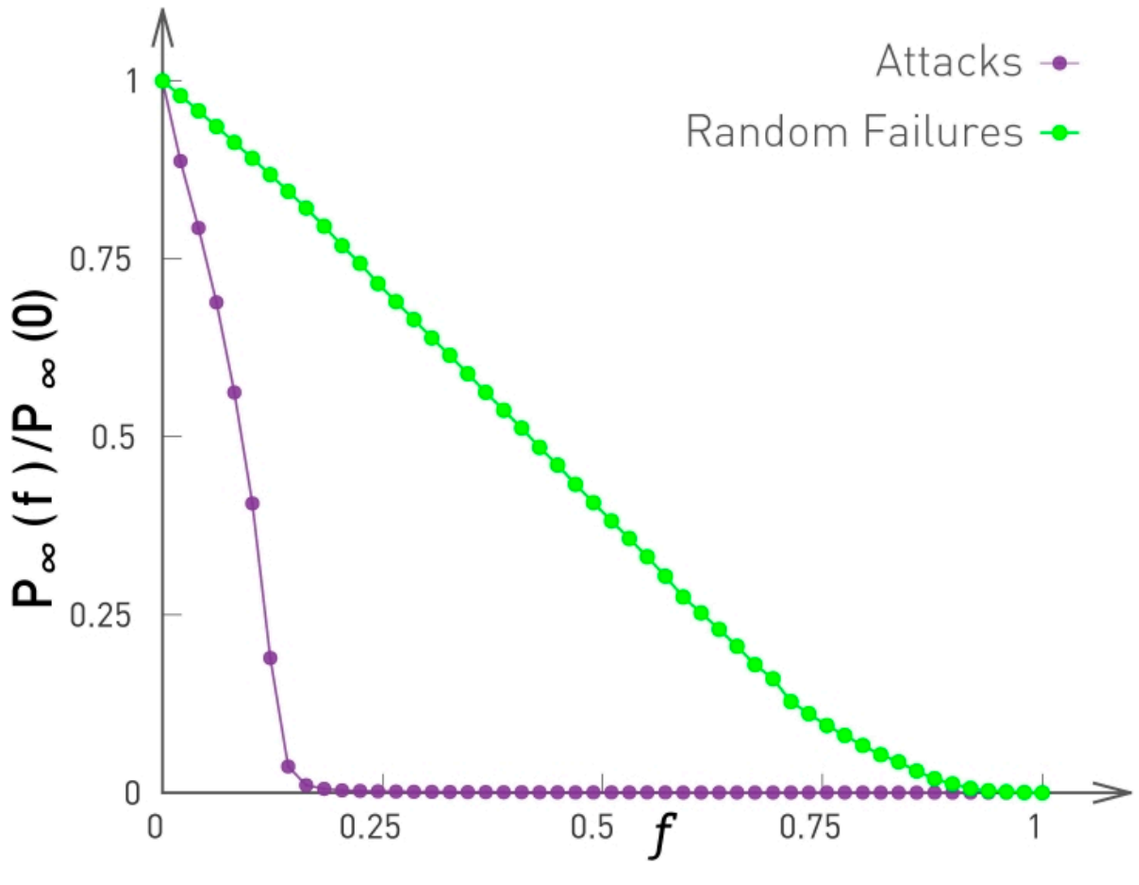
Albert,R., Jeong,H. and Barabási,A. (2000) Error and attack tolerance of complex networks. Nature, 406, 378–382.

Networks Under Attack

Random (ER) network

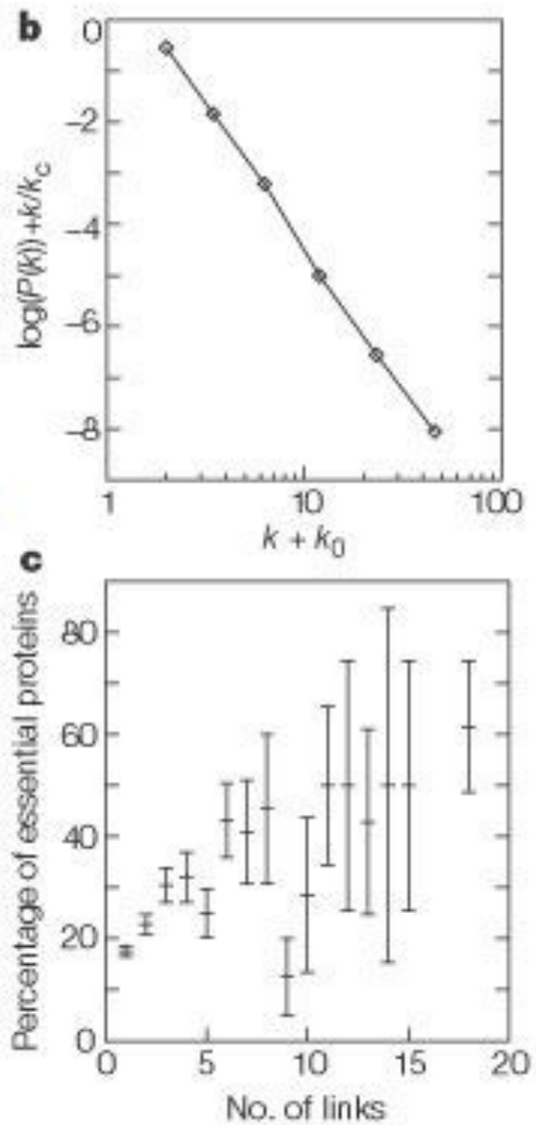


Scale-free (BA) network



while very robust against random attacks, scale-free networks are very fragile against targeted attacks on hubs!

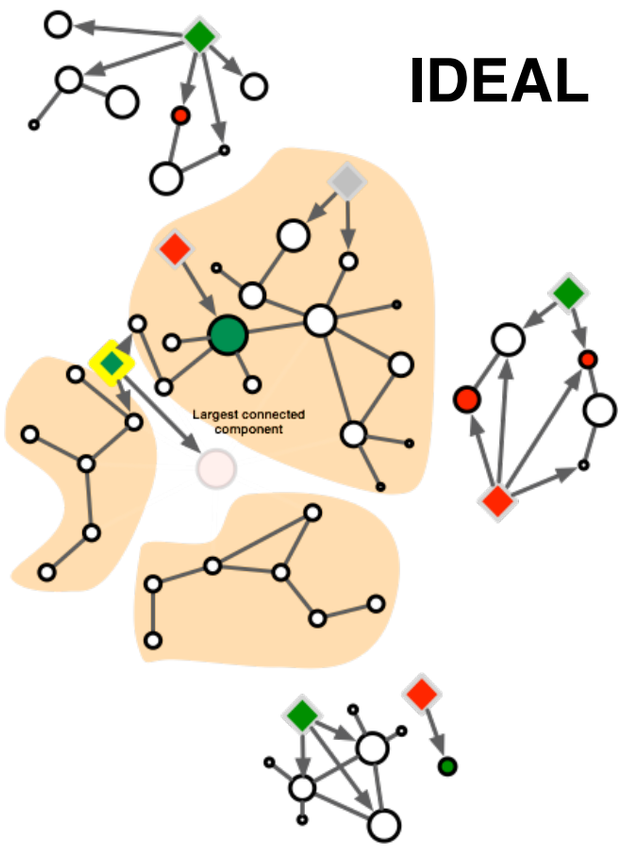
EXAMPLE: CENTRALITY-LETHALITY



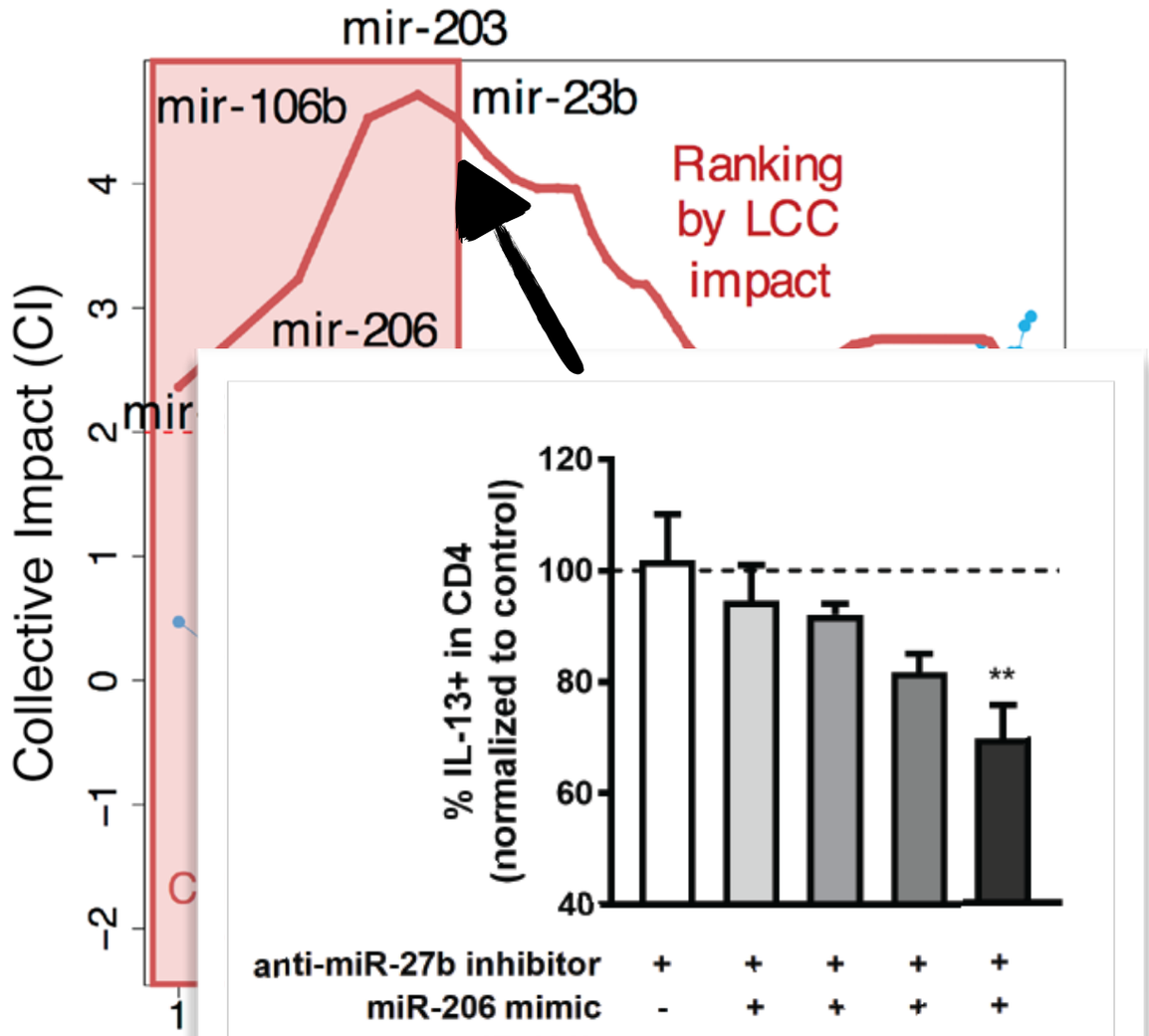
Jeong et al., “Lethality and centrality in protein networks“ Nature 2001

EXAMPLE: MICRO-RNAS ATTACK

Network impact of miRNAs



Kilic et al, JCI insight 2018

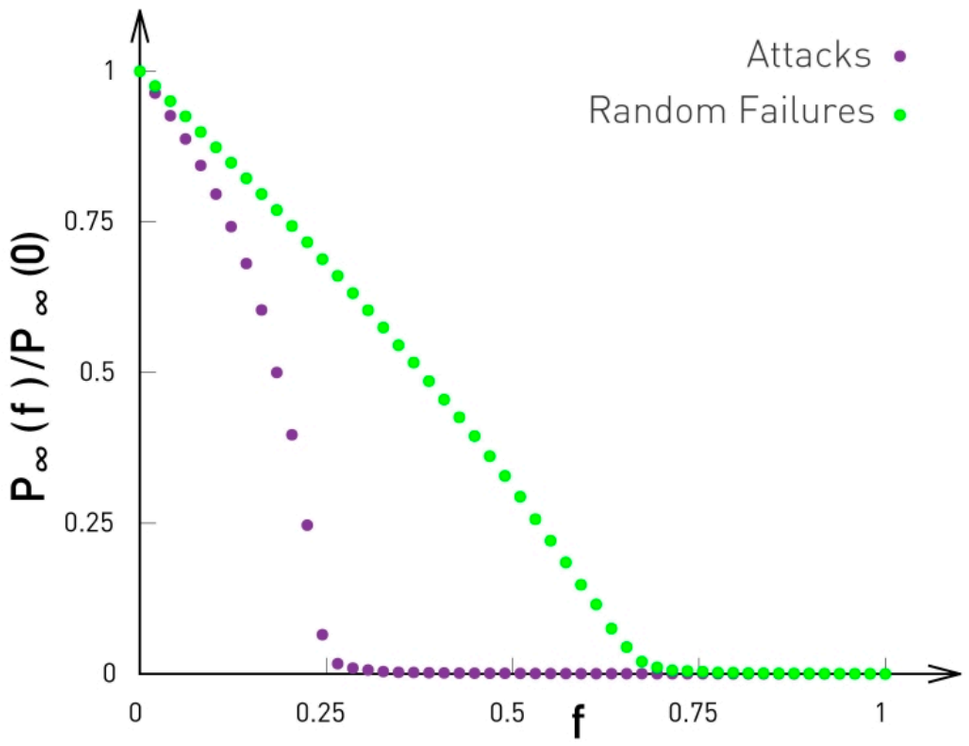


➡Network impact predicts functional miRNAs

Exercise: using python and networkx, reproduce the plots below. Display the network at 3 stages (initial, middle, end)

file: resources Python /robustness_network.ipynb

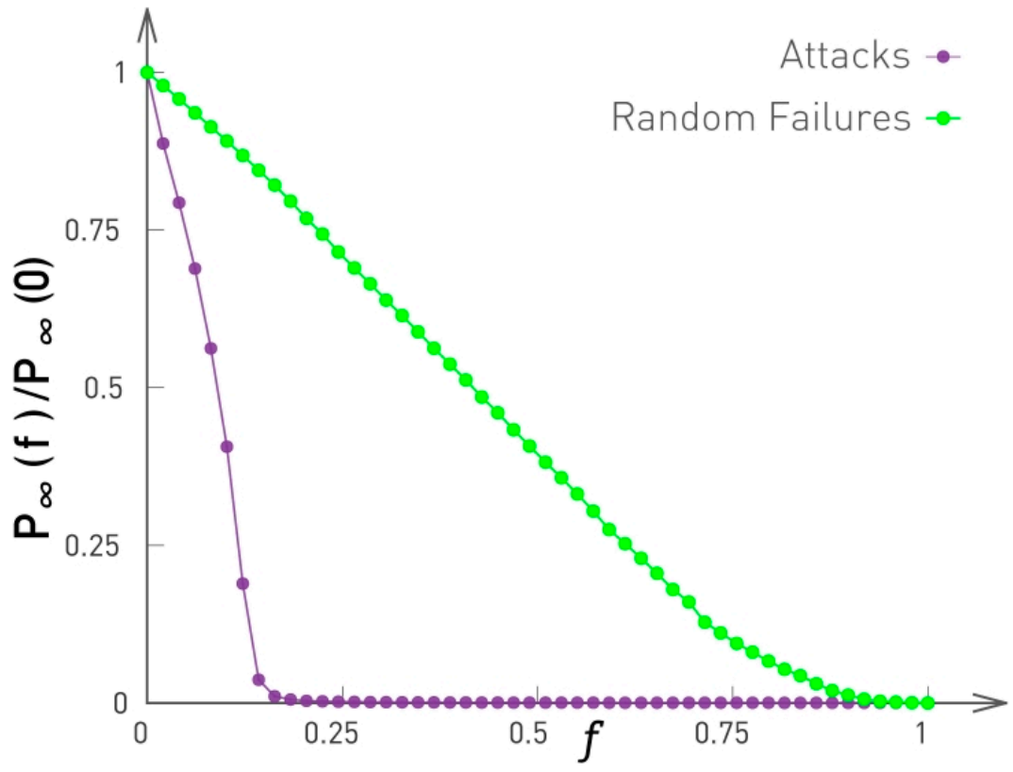
Erdös-Renyi network



Parameters

- N = 10,000
- <k> = 3

Barabasi-Albert network



Parameters

- degree exponent γ = 2.5
- kmin = 2
- N = 10,000