EFOP-3.4.3-16-2016-00009

A felsőfokú oktatás minőségének és hozzáférhetőségének együttes javítása a Pannon Egyetemen

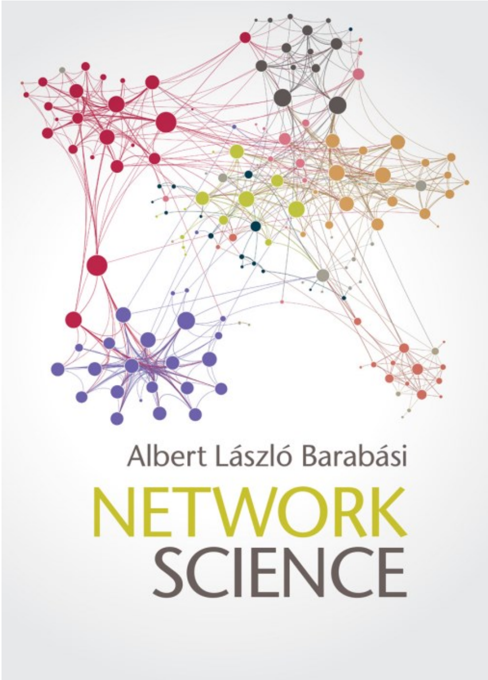
NETWORK ANALYSIS

Author: Dániel Leitold
Ágnes Vathy-Fogarassy

SZÉCHENYI 2020
MAGYARORSZÁG KORMÁNYA
Európai Unió
Európai Szociális Alap
BEFEKTETÉS A JÖVŐBE

Foreword

The slide series is created for the following textbook:



Albert László Barabási: Network Analysis

The book is online available: http://networksciencebook.com/

Course material

EFOP-3.4.3-16-2016-00009

A felsőfokú oktatás minőségének és hozzáférhetőségének együttes javítása a Pannon Egyetemen

Topics:

- 01 - Introduction
- 02 - Graph Theory
- 03 - Random Networks
- 04 - Scale-free Property
- 05 - BA Model
- 06 - Practice
- 07 - Evolving Networks
- 08 - Degree Correlations
- 09 - Network Robustness
- 10 - Communities
- 11 - Spreading Phenomena

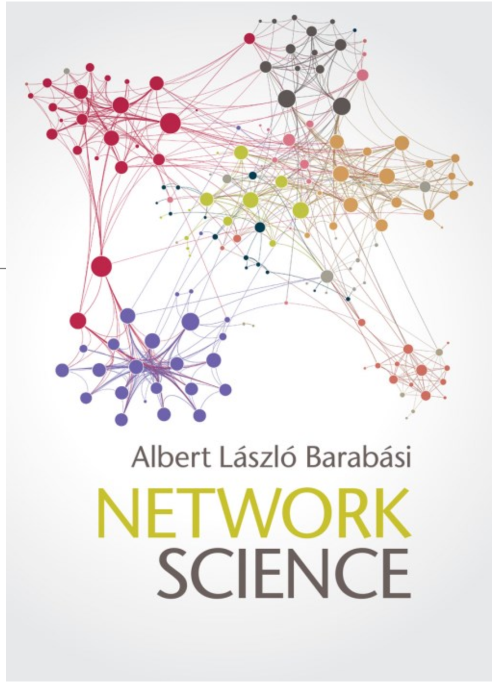
Network Analysis

01 – INTRODUCTION

Slides were created by: Daniel Leitold

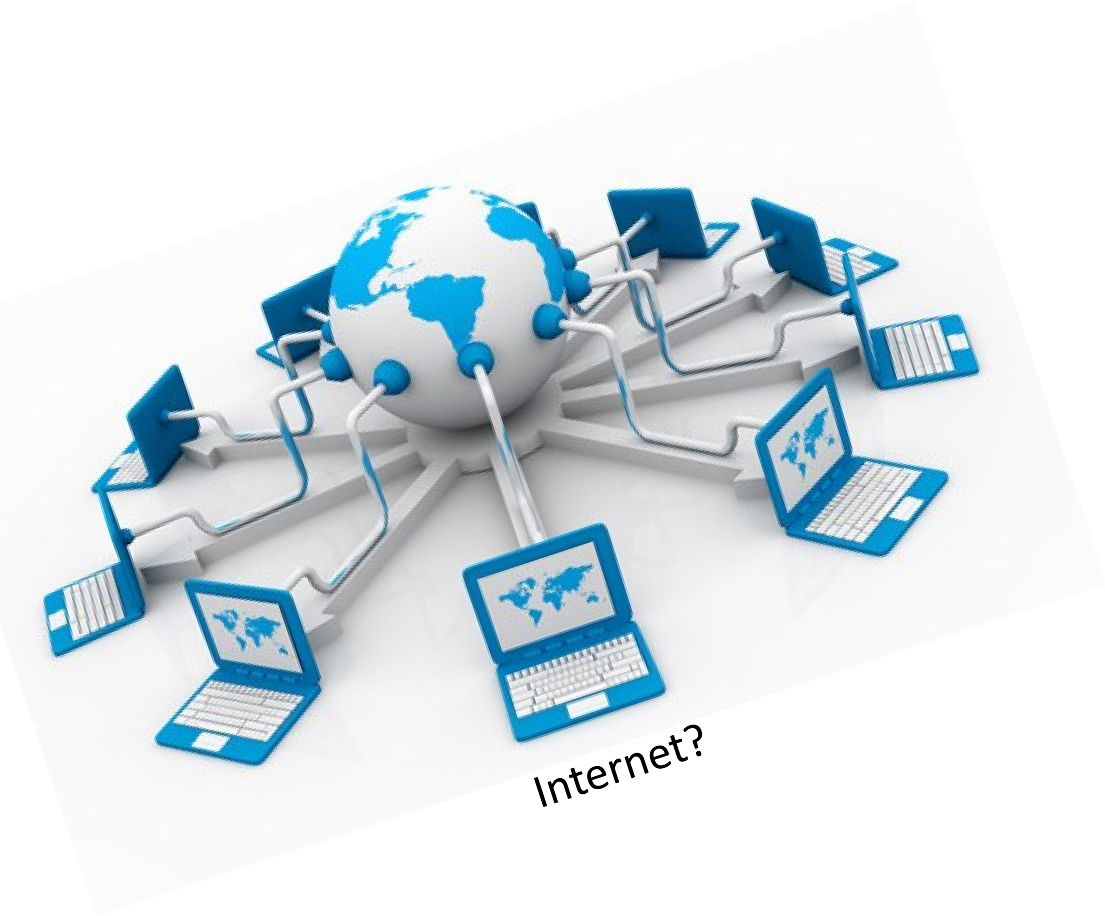
Network Science book (online)

Barabási, Albert-László. *Network Science*.
Cambridge University Press, 2016.

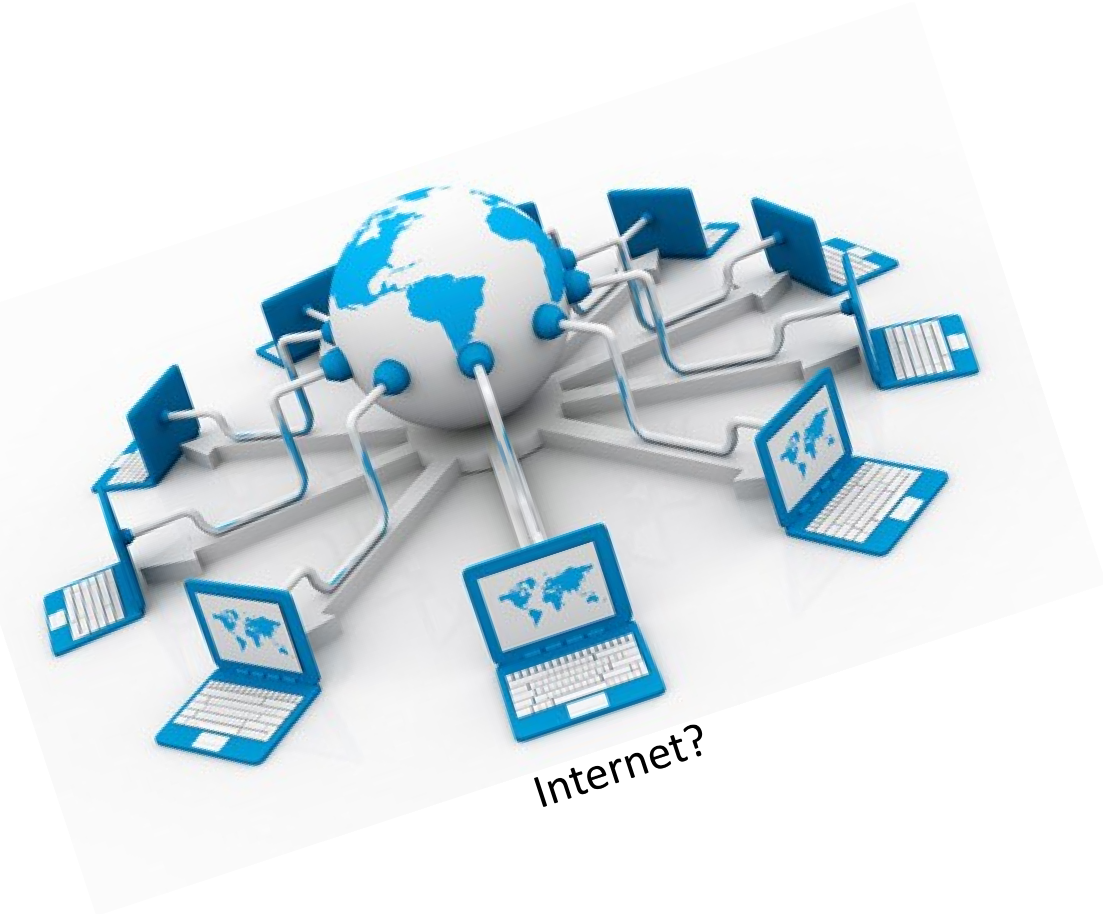


What is network science?

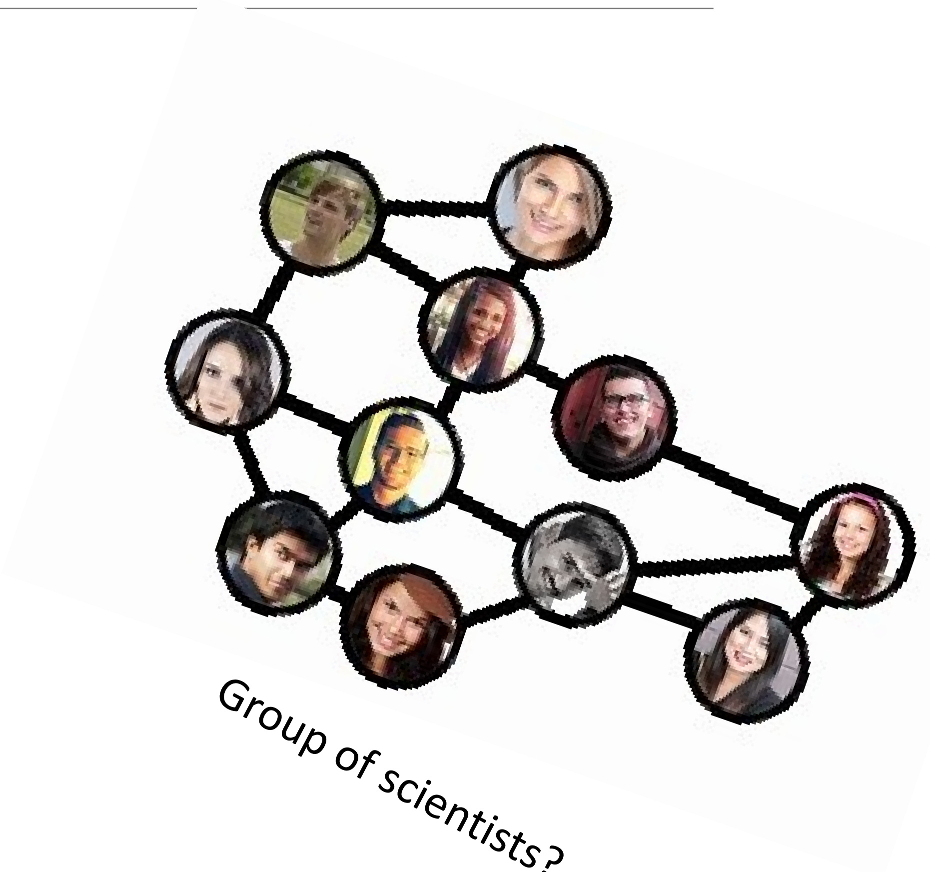
What is network science?



What is network science?



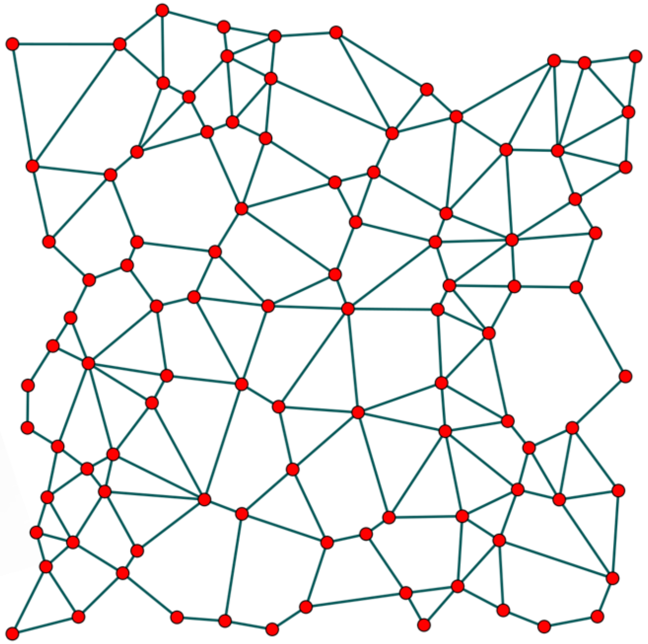
Internet?



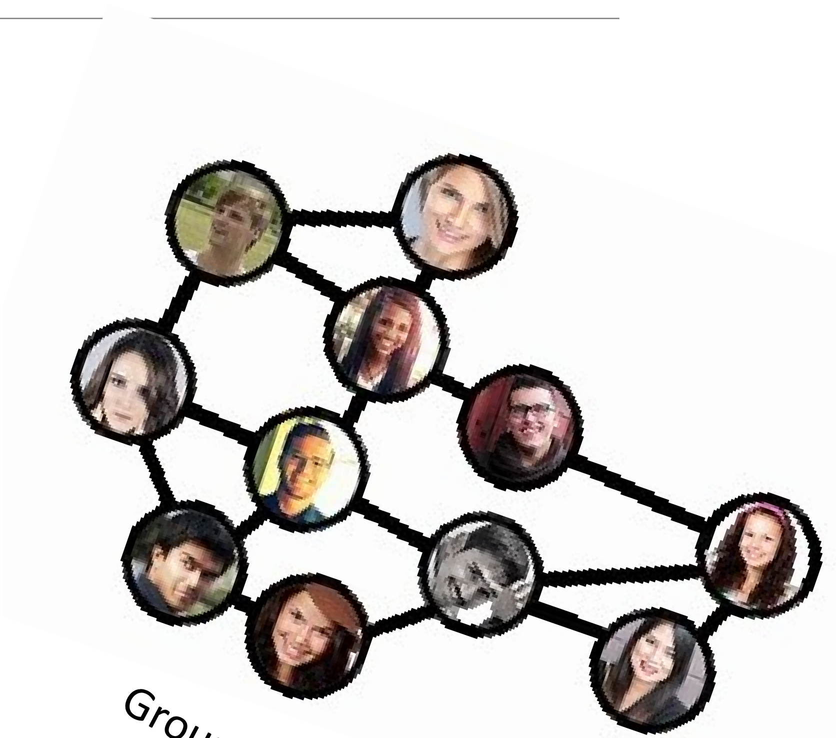
Group of scientists?

What is network science?

Internet?



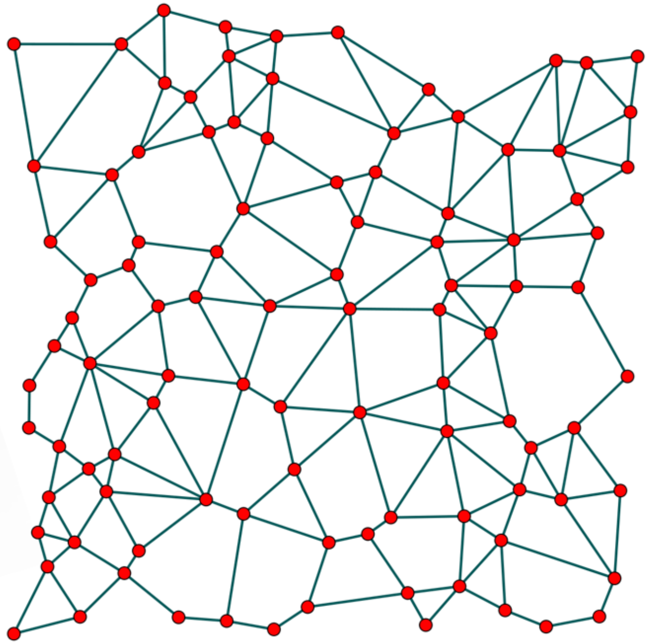
Graphs?



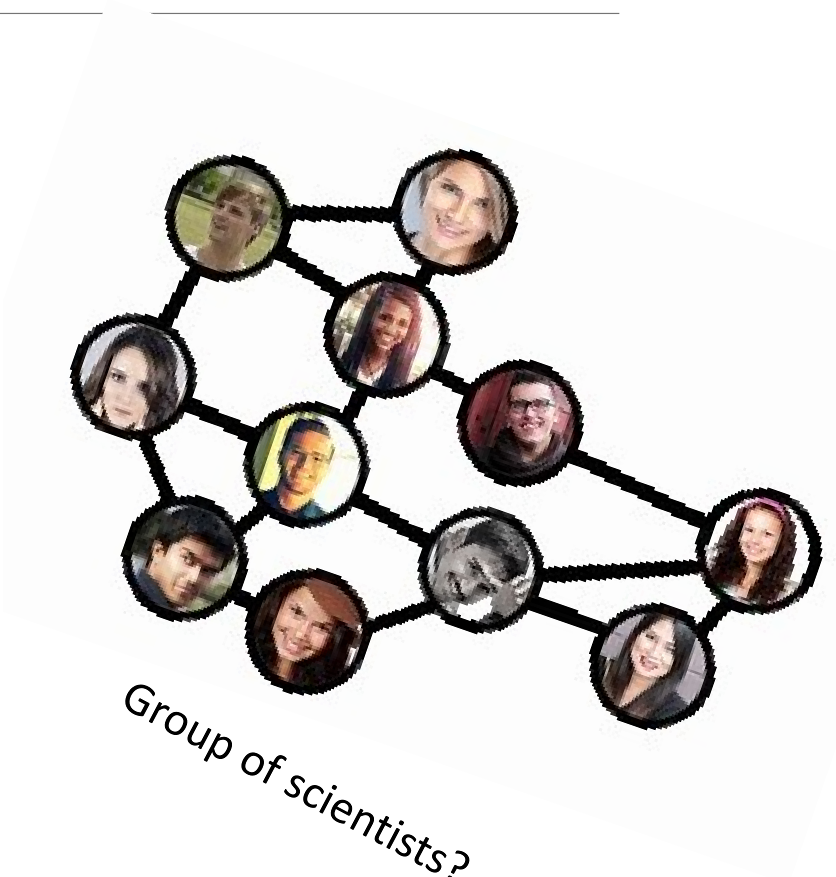
Group of scientists?

What is network science?

Internet?



Graphs?



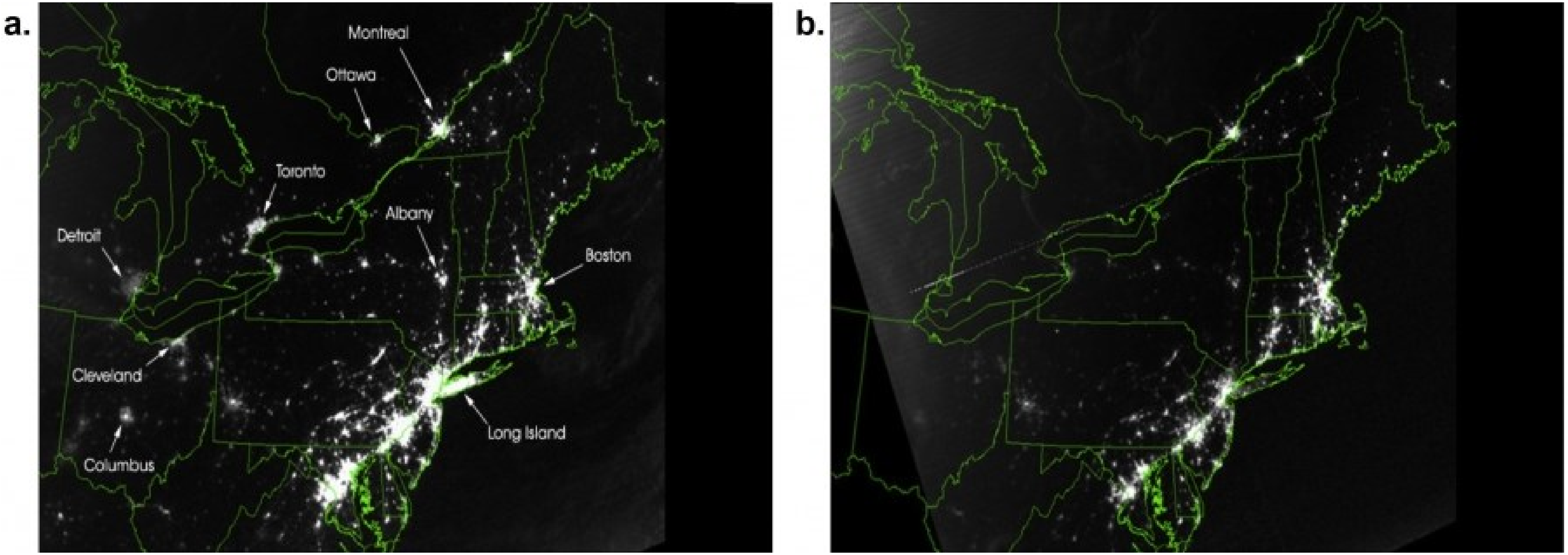
Group of scientists?

All together!

Example - 2003 North American Blackout

Toronto, Detroit, Cleveland, Columbus, Long Island are shining (a), and gone dark (b)

14th August 2003 – 45 million people in US and 10 million people in Ontario were left without power



Example - 2003 North American Blackout

Example - 2003 North American Blackout

Why is it important to us?

What is the network? What are the nodes and links?

How can we use network science to avoid cascading failures?

Could we have prevented the cascaded blackouts?

Example - 2003 North American Blackout

Why is it important to us?

A power grid is a complex system that can be analysed with engineering methods, but these methods cannot handle the complexity well derived from the interconnections.

What is the network? What are the nodes and links?

The network is the power grid itself. Nodes are the power plants and the links are the wires between the plants.

How can we use network science to avoid cascading failures?

With determining the overloaded plants, we can create a more robust network.

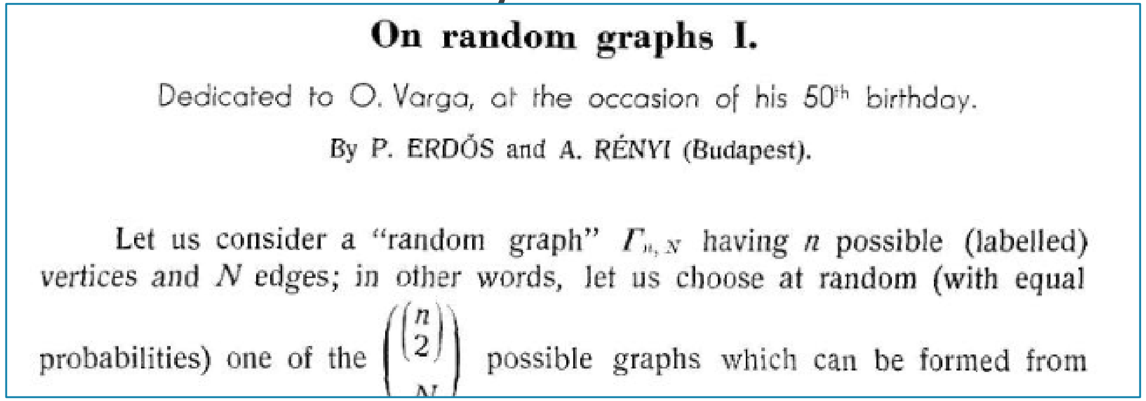
Could we have prevented the cascaded blackouts?

Probably yes.

When did network science start?

State 1: There are publications from Erdős-Rényi (1959) and Granovetter (1973).

State 2: There were social groups, trade routes and aqueduct in the ancient times already.

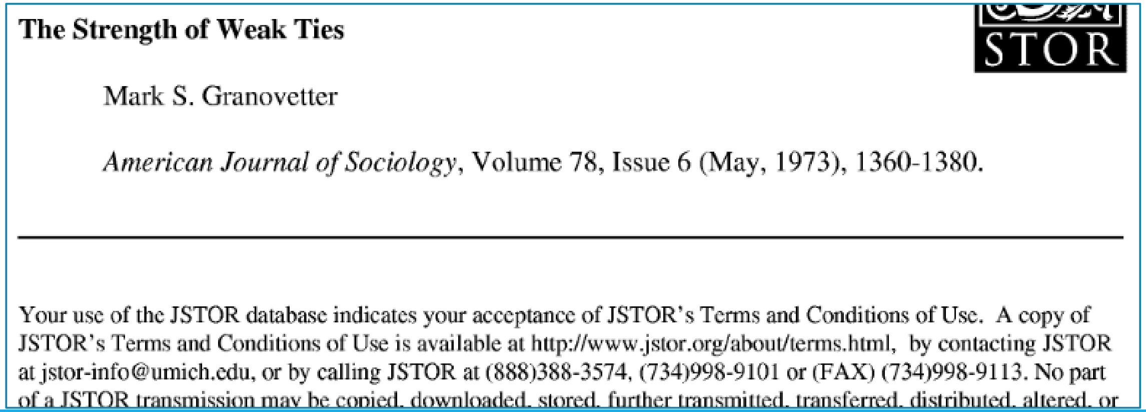


On random graphs I.

Dedicated to O. Varga, at the occasion of his 50th birthday.

By P. ERDŐS and A. RÉNYI (Budapest).

Let us consider a "random graph" $\Gamma_{n,N}$ having n possible (labelled) vertices and N edges; in other words, let us choose at random (with equal probabilities) one of the $\binom{\binom{n}{2}}{N}$ possible graphs which can be formed from

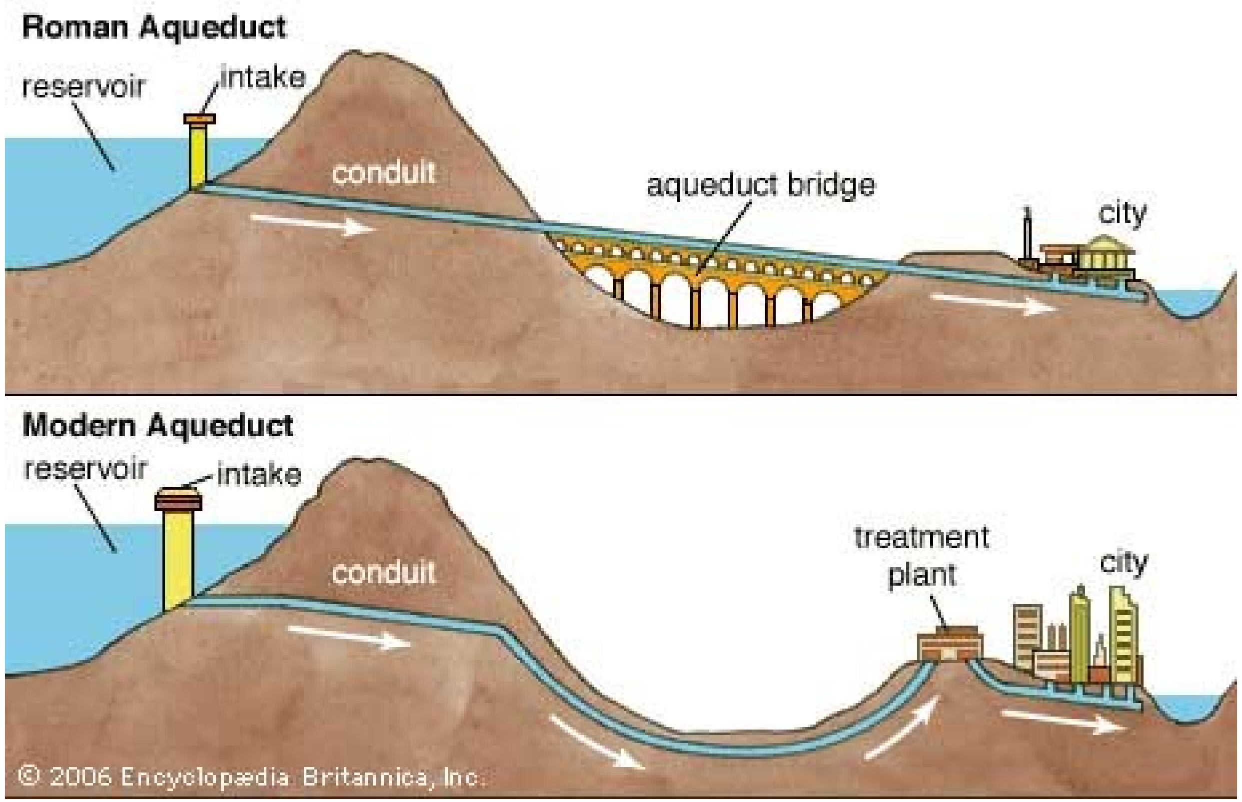


The Strength of Weak Ties

Mark S. Granovetter

American Journal of Sociology, Volume 78, Issue 6 (May, 1973), 1360-1380.

Your use of the JSTOR database indicates your acceptance of JSTOR's Terms and Conditions of Use. A copy of JSTOR's Terms and Conditions of Use is available at http://www.jstor.org/about/terms.html, by contacting JSTOR at jstor-info@umich.edu, or by calling JSTOR at (888)388-3574, (734)998-9101 or (FAX) (734)998-9113. No part of a JSTOR transmission may be copied, downloaded, stored, further transmitted, transferred, distributed, altered, or



When did network science start?

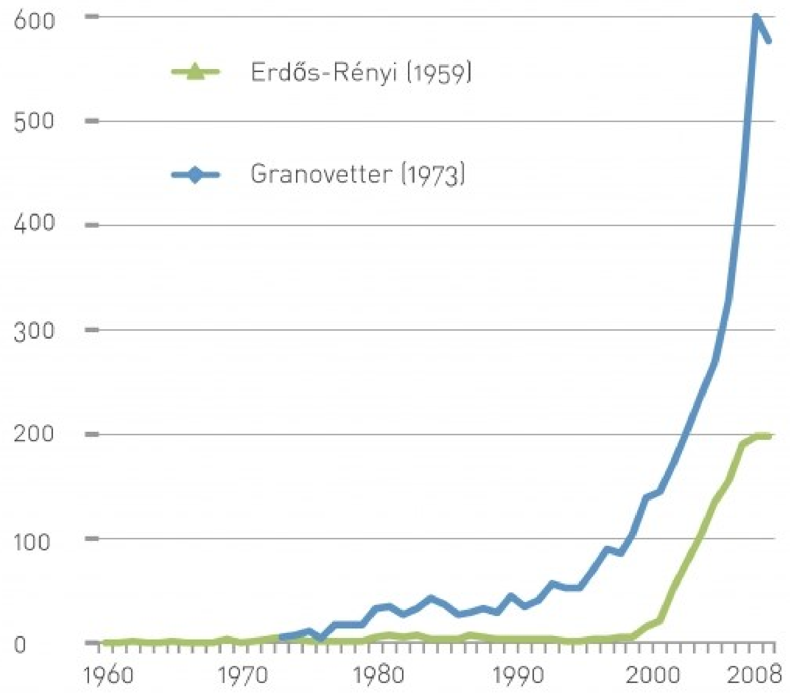
The network science is a new discipline. It became a separated discipline in the 21st century.

Citations for the previous two papers jump on 21st century.

Main author: Albert-László Barabási

Two main force of network science:

- Emergence of Network Maps
  - Internet
  - Hollywood
  - Chemical reactions
- Universality of Network Characteristics
  - Networks are different (nodes, links, how the links are appearing)
  - **BUT**, the structures of the different networks are similar



When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

Example:

Biological Research	Food web
Information Technologies	Co-purchases
Amazon	Protein reactions
Mother Nature	Wiring diagram

When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

Example:

Biological Research	Food web
Information Technologies	Co-purchases
Amazon	Protein reactions
Mother Nature	Wiring diagram

When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

Example:

Biological Research

Information Technologies

Amazon

Mother Nature

Food web

Co-purchases

Protein reactions

Wiring diagram

When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

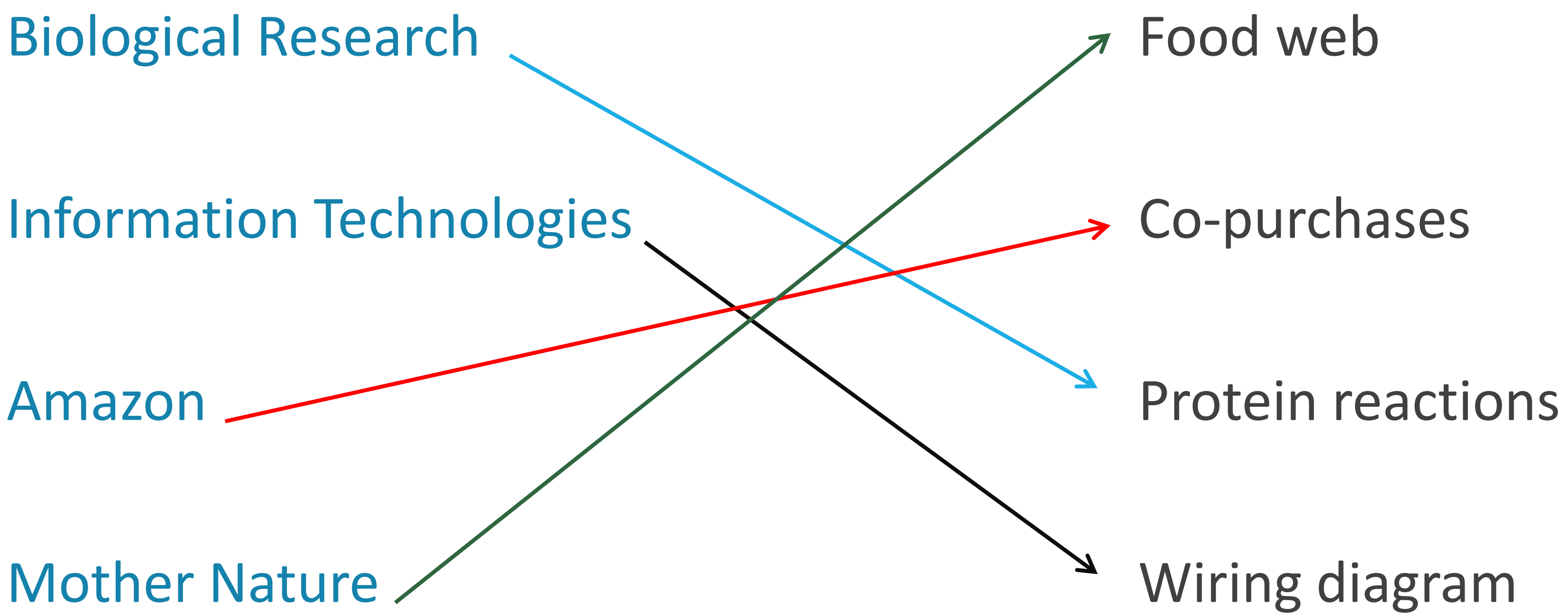
Example:

Biological Research	Food web
Information Technologies	Co-purchases
Amazon	Protein reactions
Mother Nature	Wiring diagram

When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

Example:



When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

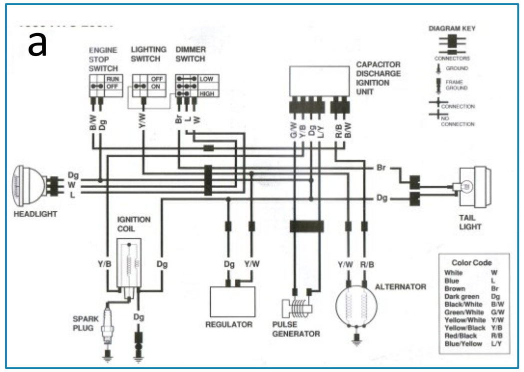
Example:

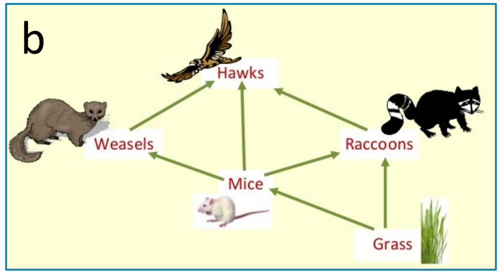
Biological Research

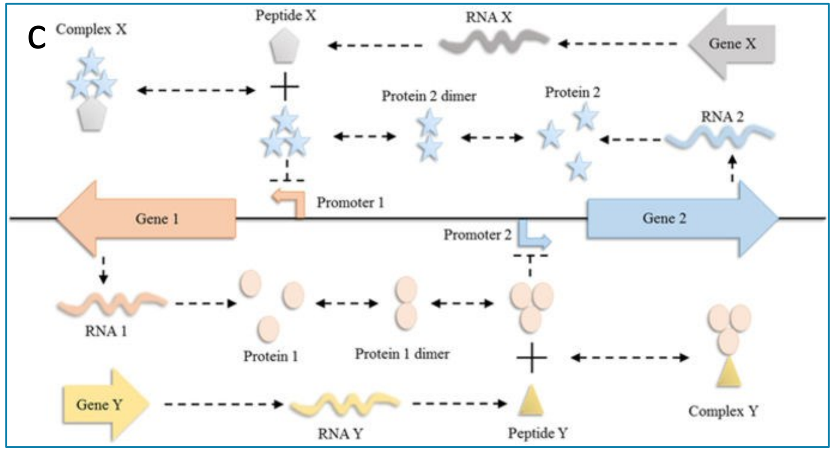
Information Technologies

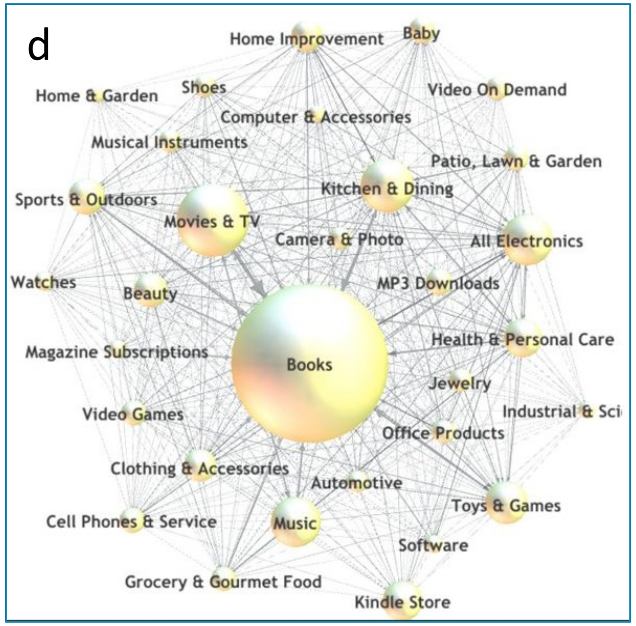
Amazon

Mother Nature









When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

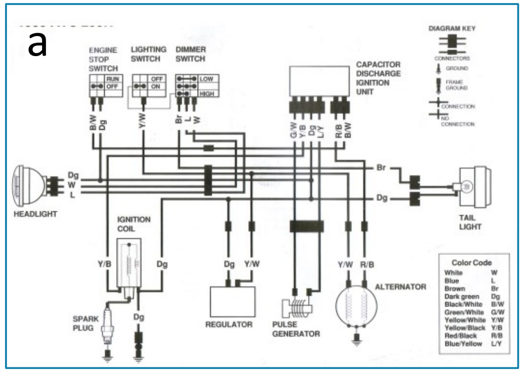
Example:

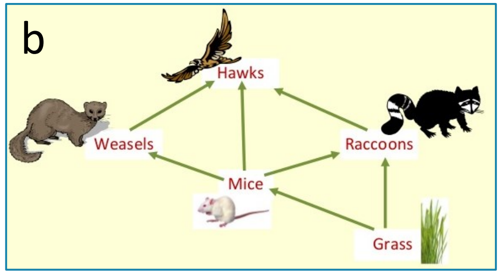
Biological Research - c

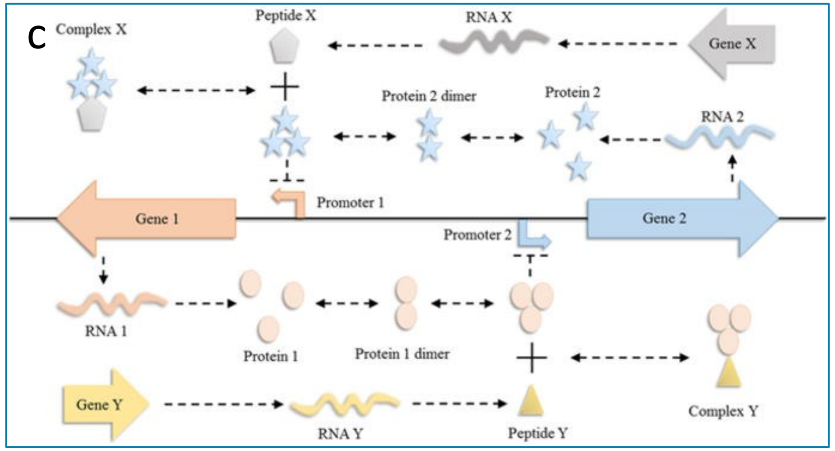
Information Technologies

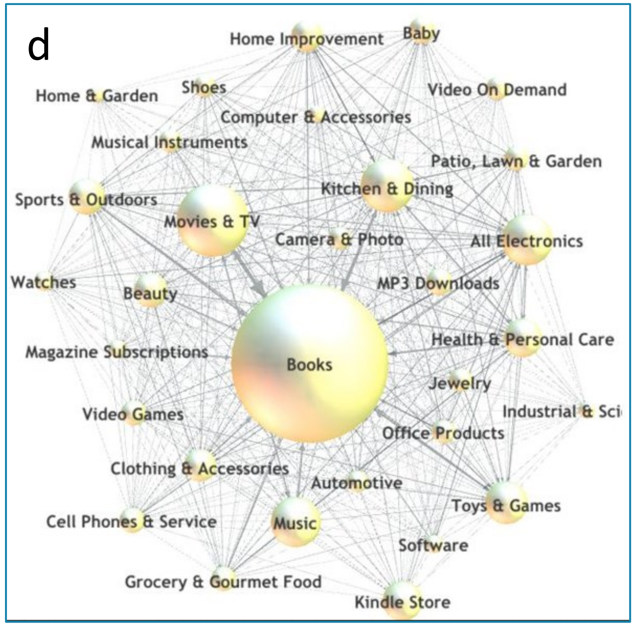
Amazon

Mother Nature









When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

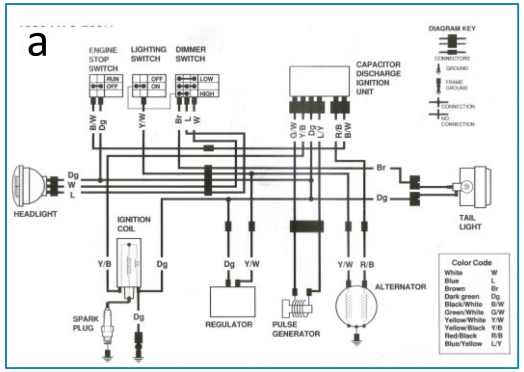
Example:

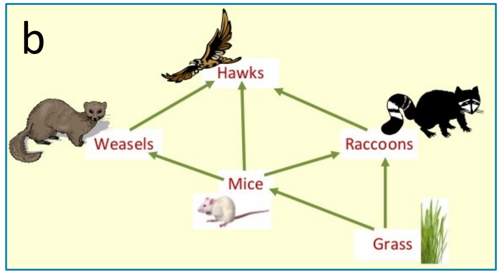
Biological Research - c

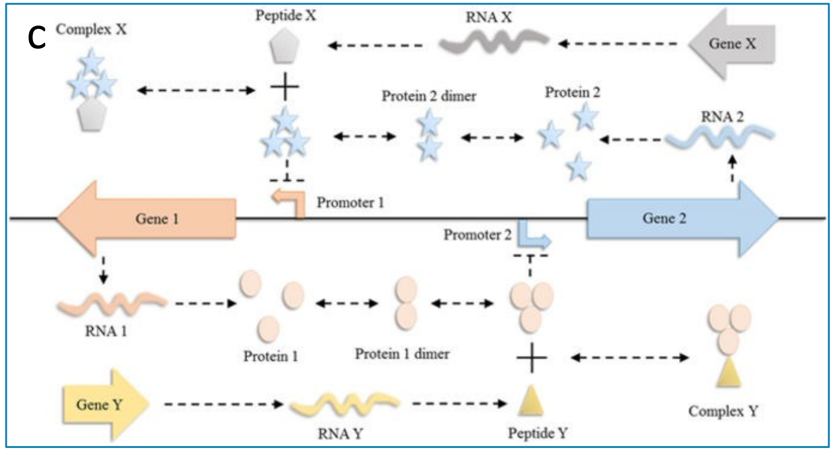
Information Technologies - a

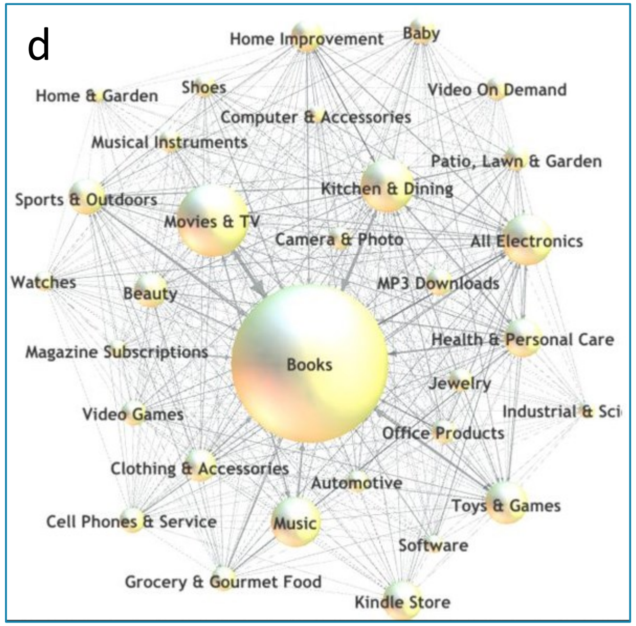
Amazon

Mother Nature





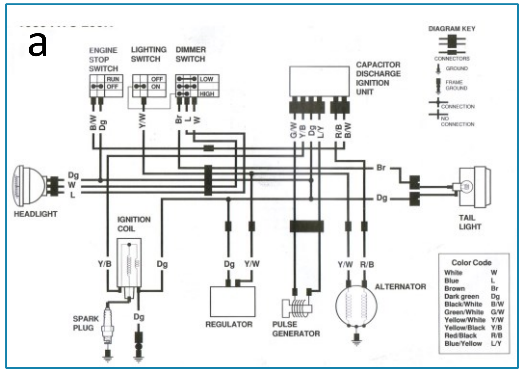


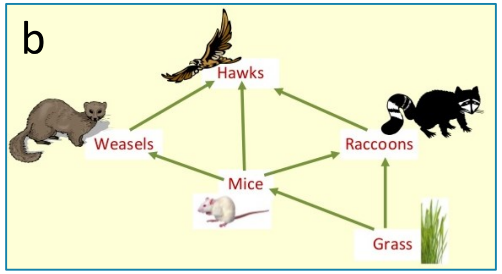


When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

Example:



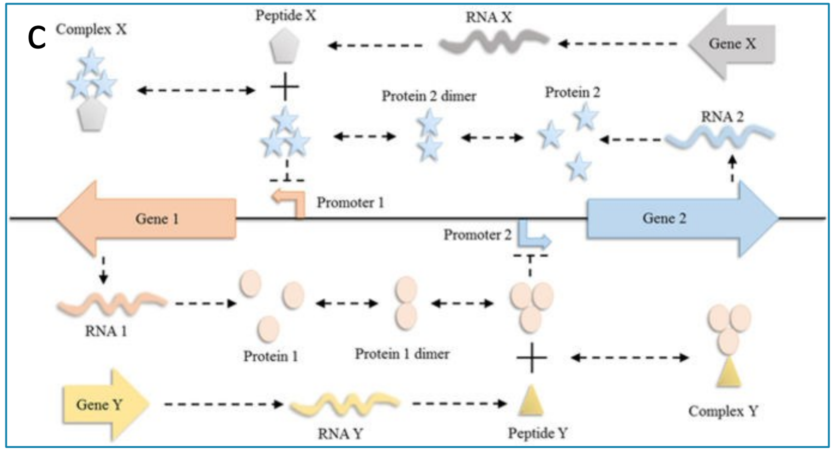


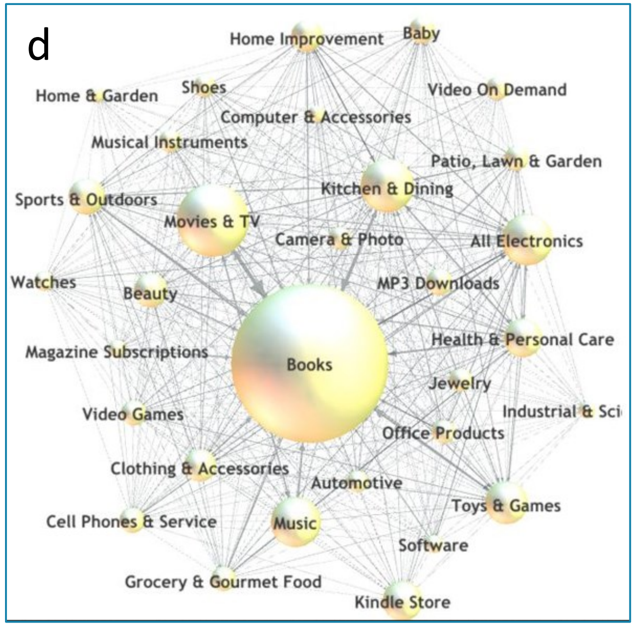
Biological Research - c

Information Technologies - a

Amazon - d

Mother Nature





When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

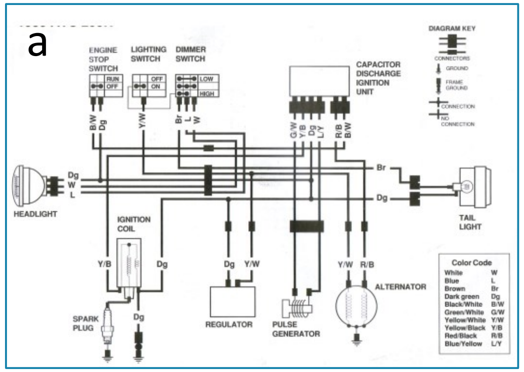
Example:

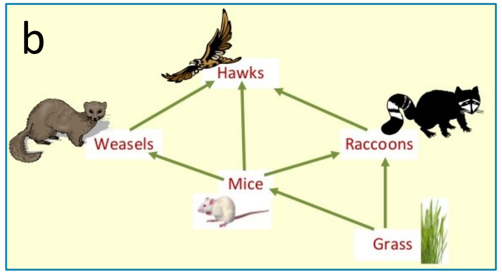
Biological Research - c

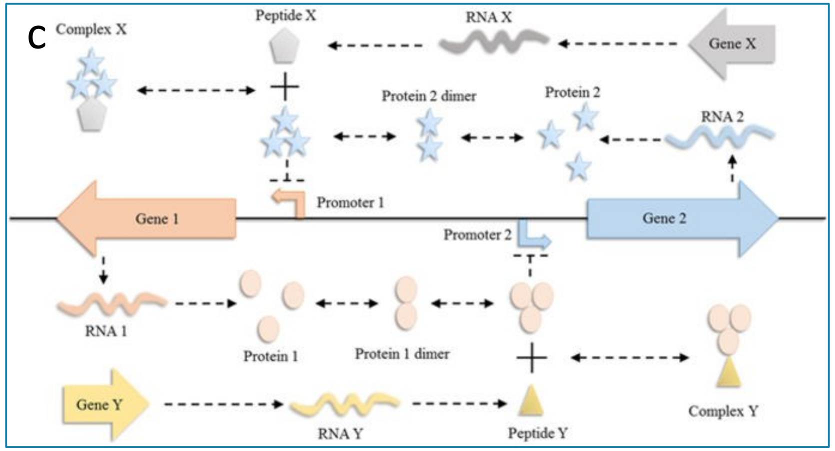
Information Technologies - a

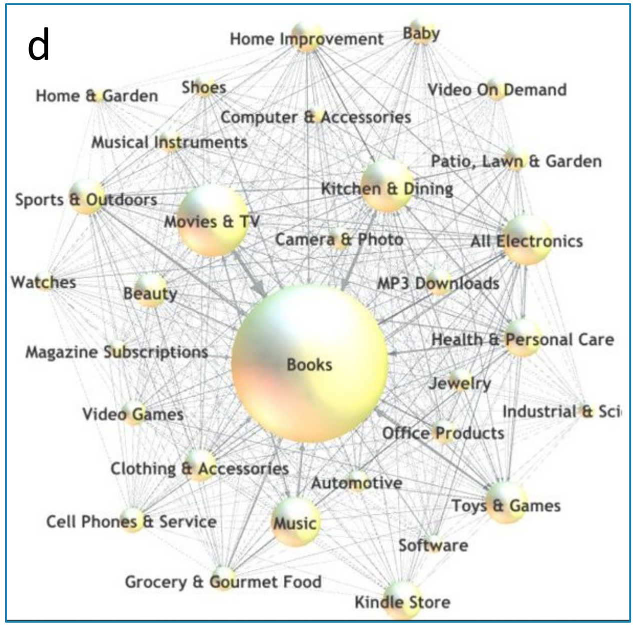
Amazon - d

Mother Nature - b









When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

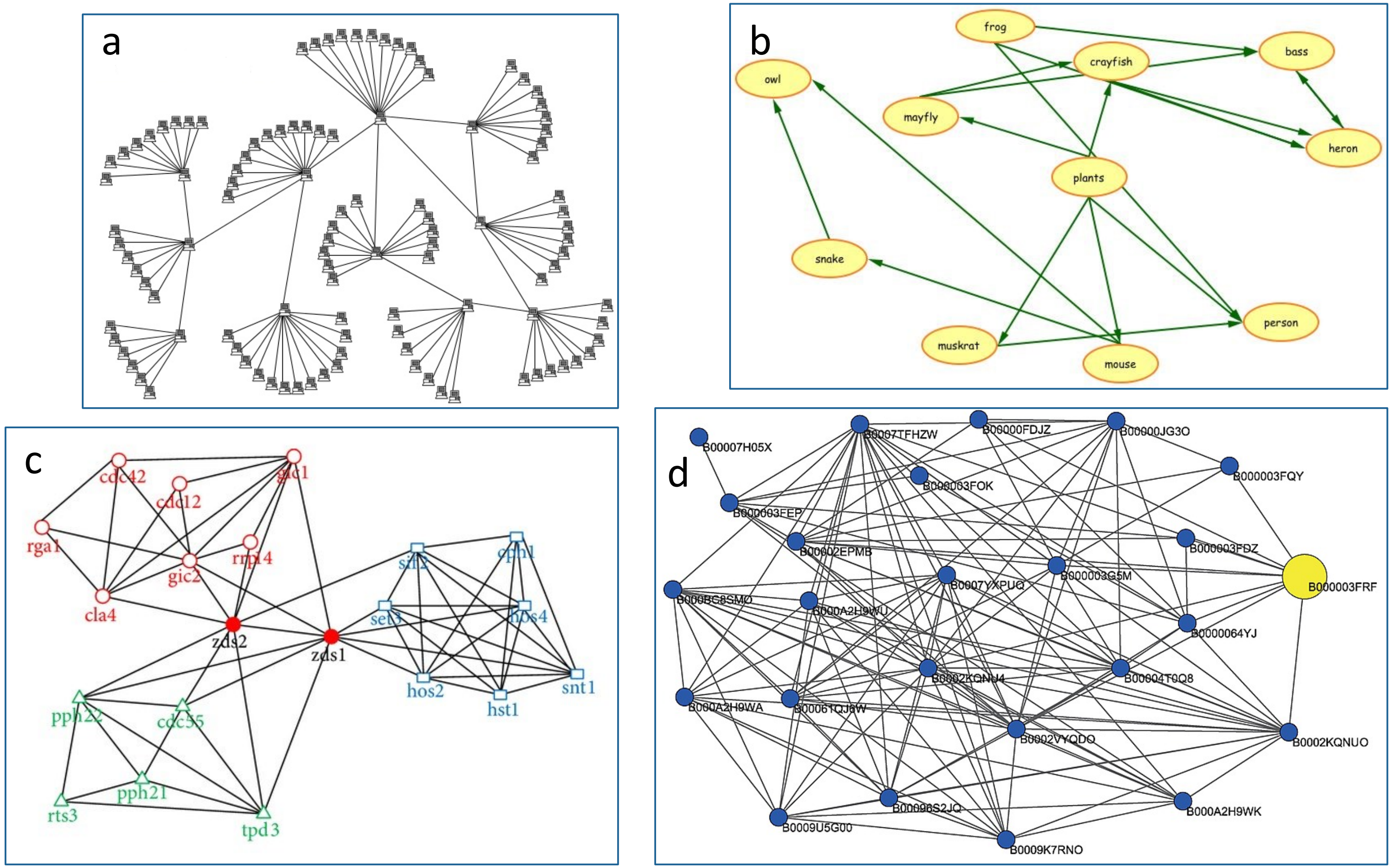
Example:

Biological Research

Information Technologies

Amazon

Mother Nature



When did network science start?

Why so late? The reason may be its interdisciplinary. What does it mean?

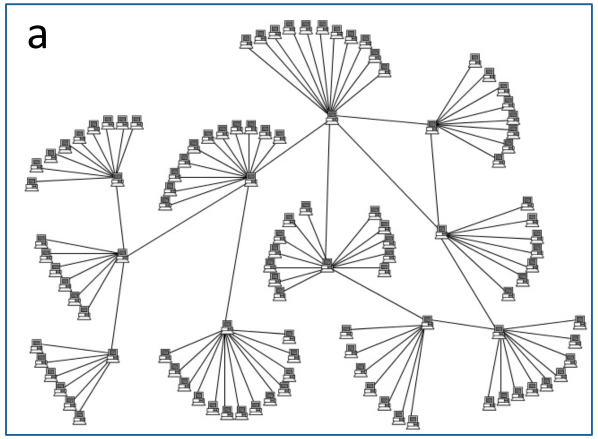
Example:

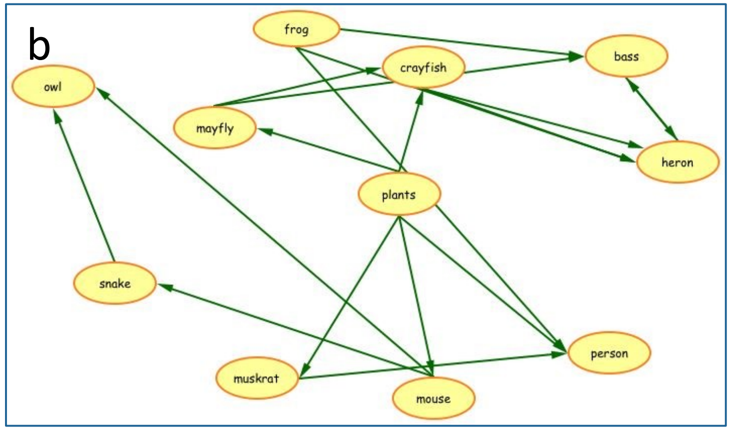
Biological Research - c

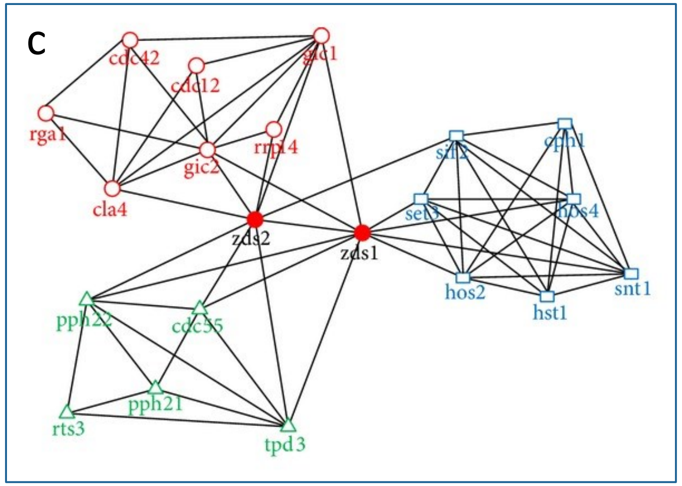
Information Technologies - d

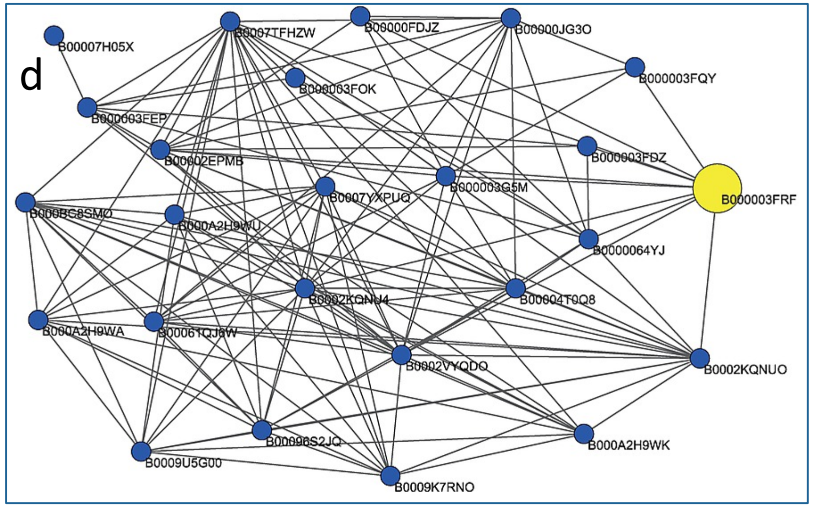
Amazon - a

Mother Nature - b









When did network science start?

Since each field had its own data representation, therefore network science-based researches were denied in the beginning.

BUT, network science demonstrates that science can cope with the challenge of complex systems.

Several key concepts of network science have their roots in graph theory.

What distinguishes network science from graph theory is its empirical nature, i.e. its focus on data, function and utility.

Network Science borrowed the followings:

- Formalism to deal with graph – from graph theory
- Dealing with randomness and universal principles – from statistical physics
- Dealing with control principles – from control and information theory
- Extracting information from incomplete and noisy data – from statistics

> *"I think the next century will be the century of complexity."*
> *Stephen Hawking, January 2000*

Is network science useful? – Societal Impacts

Economic Impact:

- Google search – PageRank measure for network.
- Facebook, LinkedIn, Twitter – advertising algorithms derived from network researcher.

Health:

- Gene networks: breakdown of molecular networks can cause human disease.
- Network pharmacology: cure disease without significant side effects (drug development).
- Network medicine: cellular interactions, drug targets in bacteria and humans.

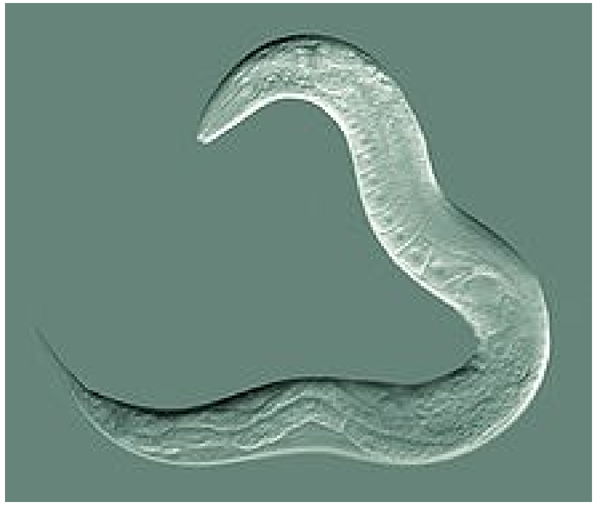
Security (fighting terrorism):

- Saddam Hussein was found by social network analysis.
- The perpetrator of the 11th March 2004 Madrid train bombings was found by the examination of the mobile call network.

Is network science useful? – Societal Impacts

Epidemics:

- In 2009, H1N1 pandemic was accurately predicted: Video.
- It helped to stop the spread of Ebola.
- In the autumn of 2010 in China, viruses, which spread through mobile phones, followed the predicted spreading scenario.



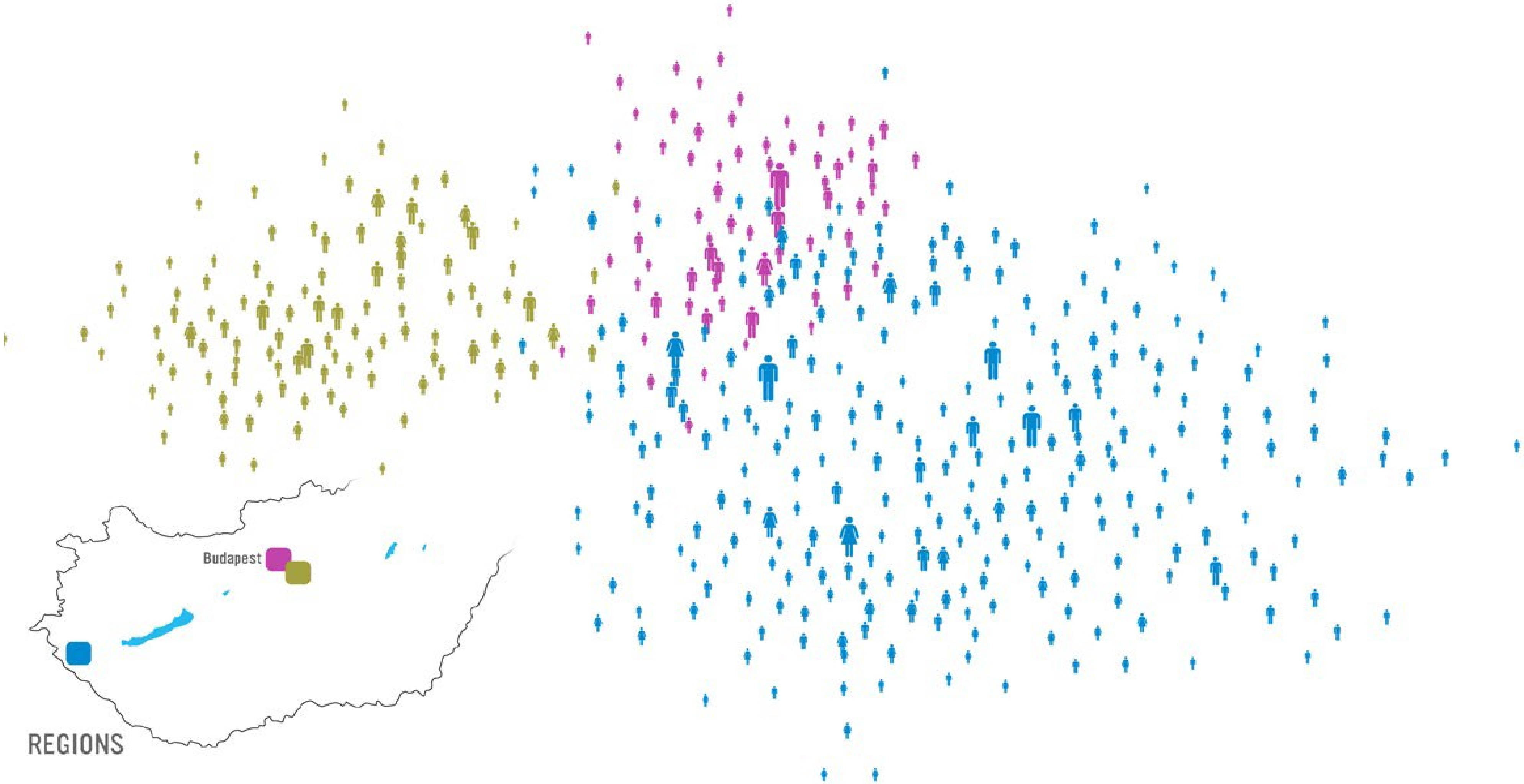
Neuroscience (mapping the brain):

- The human brain that consists of hundreds of billions of interlinked neurons is not understood.
- The only fully mapped brain available is that of the C. elegans worm, which consists of 302 neuron.

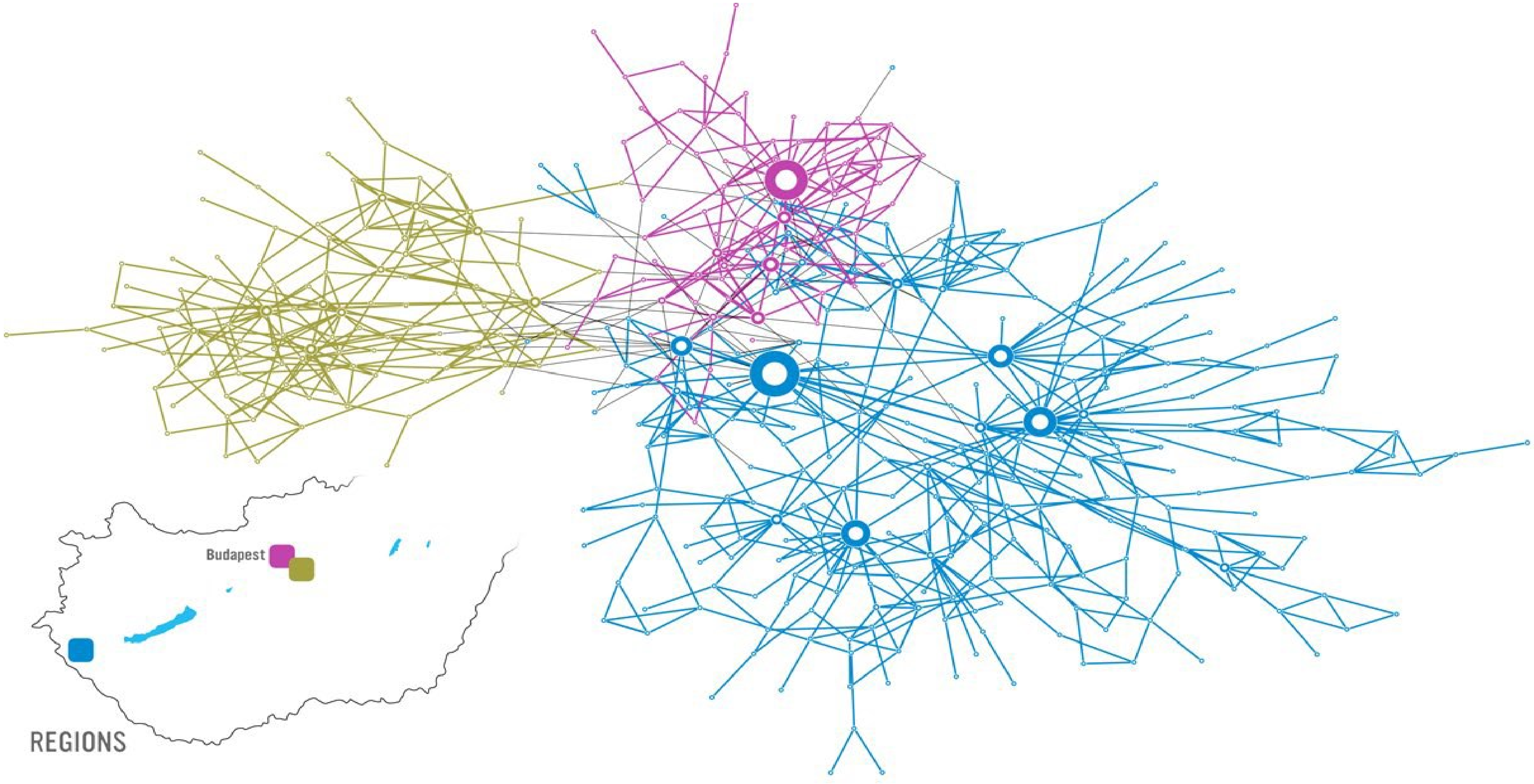
Organization management:

- The most important role in the success of an organization: the informal network, capturing who really communicates with whom.

Example – Organization management

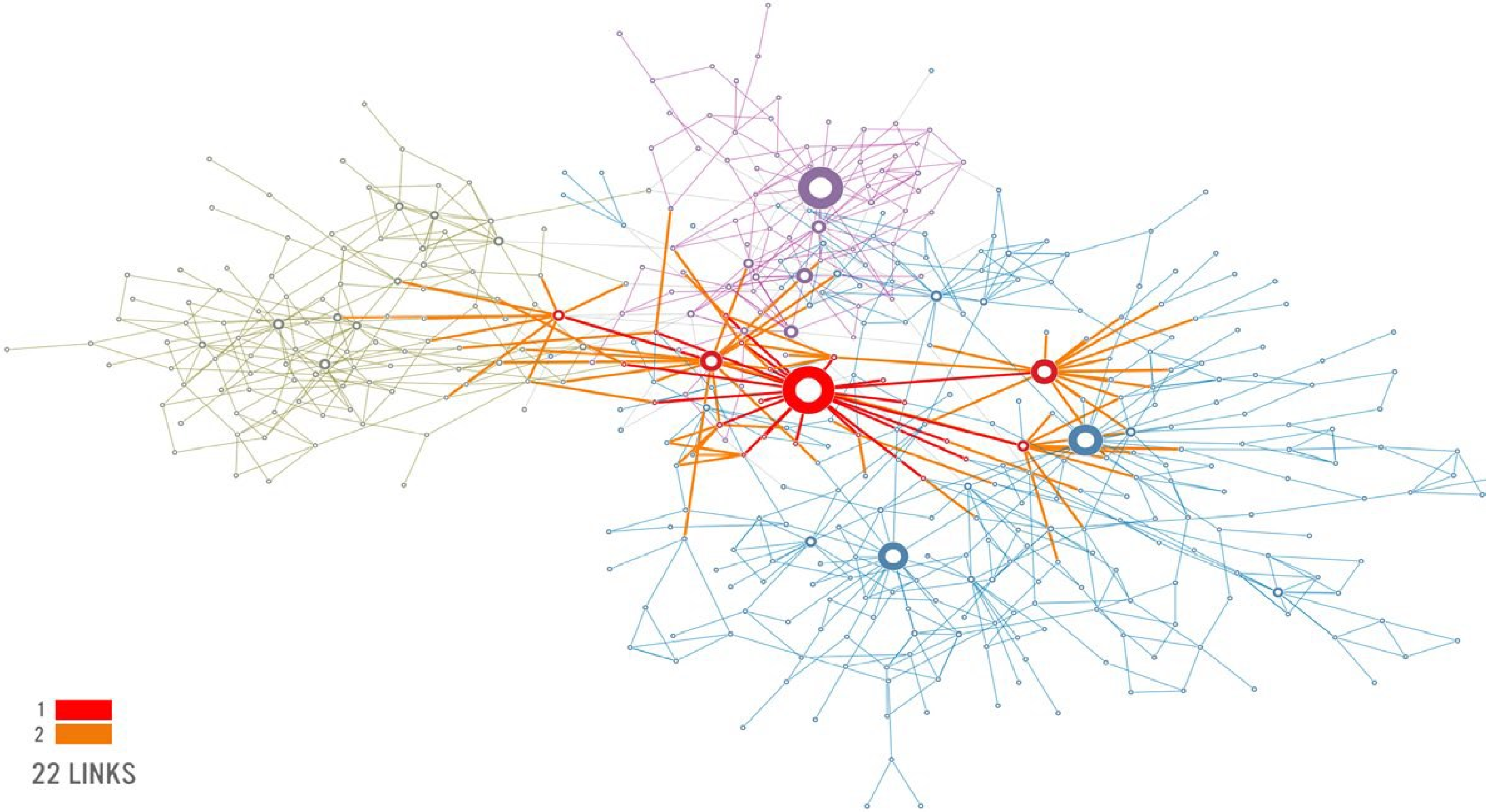


Example – Organization management



Example – Organization management

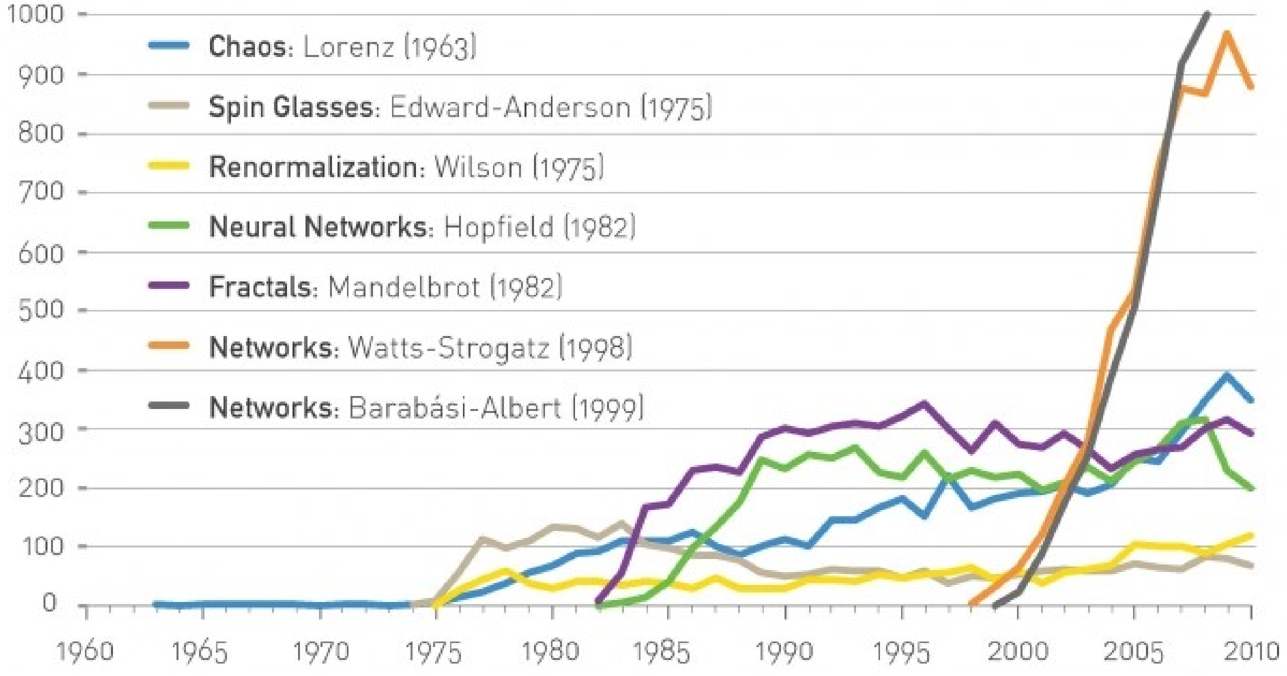
Example – Organization management



Is network science useful? – Scientific Impact

Nowhere is the impact of network science more evident than in the scientific community.

- Citation patterns of the most cited papers in the area of complex systems (each of them are citation classics such as the butterfly effect, fractals or neural networks).



Number of citations on the paper / year

Some other success:

- Network science courses on major universities.
- PhD programs in network science.
- Public excitement by books and movies like Linked, Nexus or Connected.
- and so on…

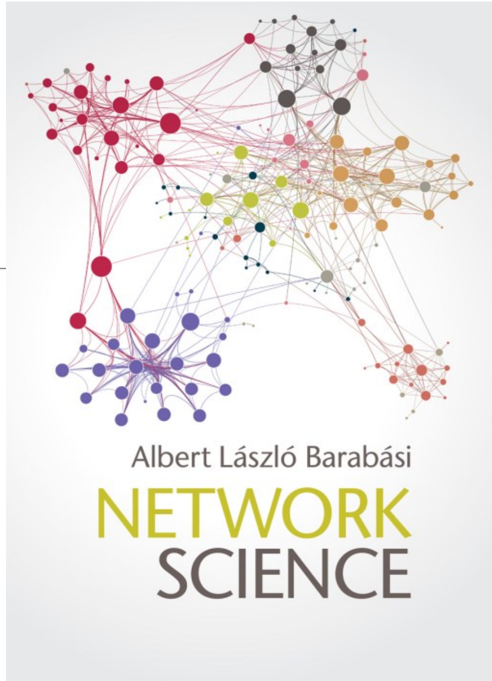
Network Analysis

02 – GRAPH THEORY

Slides were created by: Agnes Vathy-Fogarassy

Network Science book (online)

Barabási, Albert-László. *Network Science*.
Cambridge University Press, 2016.



The Bridges of Königsberg

Problem: How can one go through each bridge with using each only once?

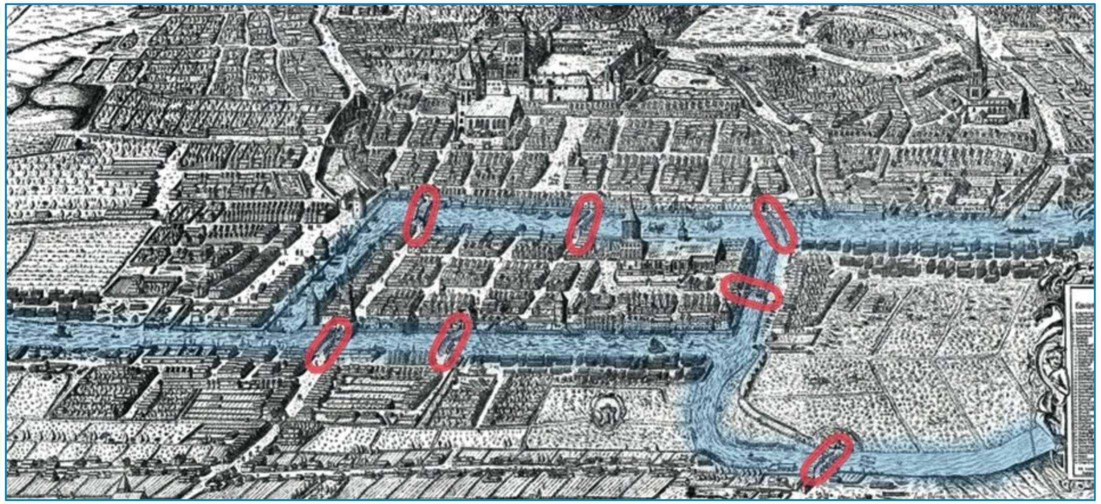
1735 – The beginning of graph theory.

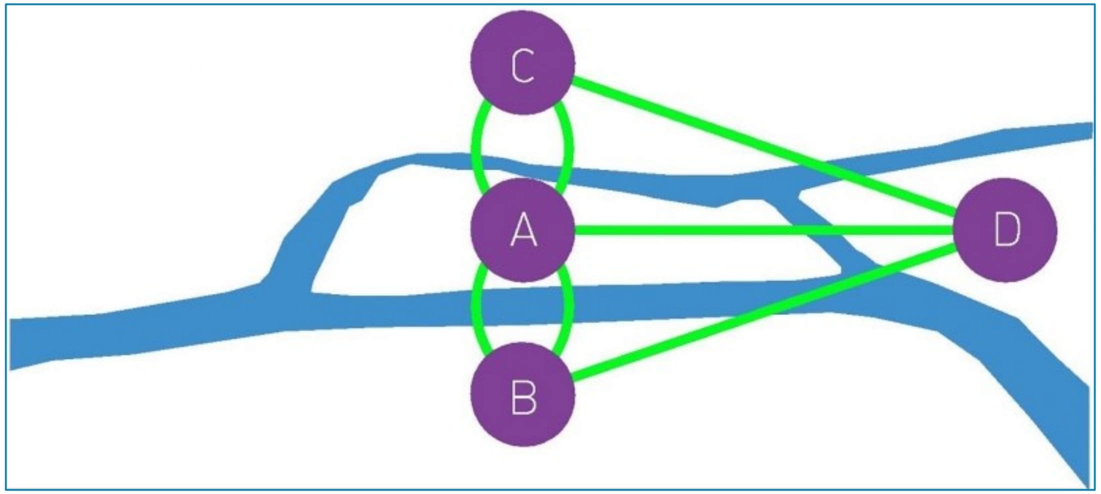
Euler's approach:

- Grounds are vertices.
- Bridges are edges.

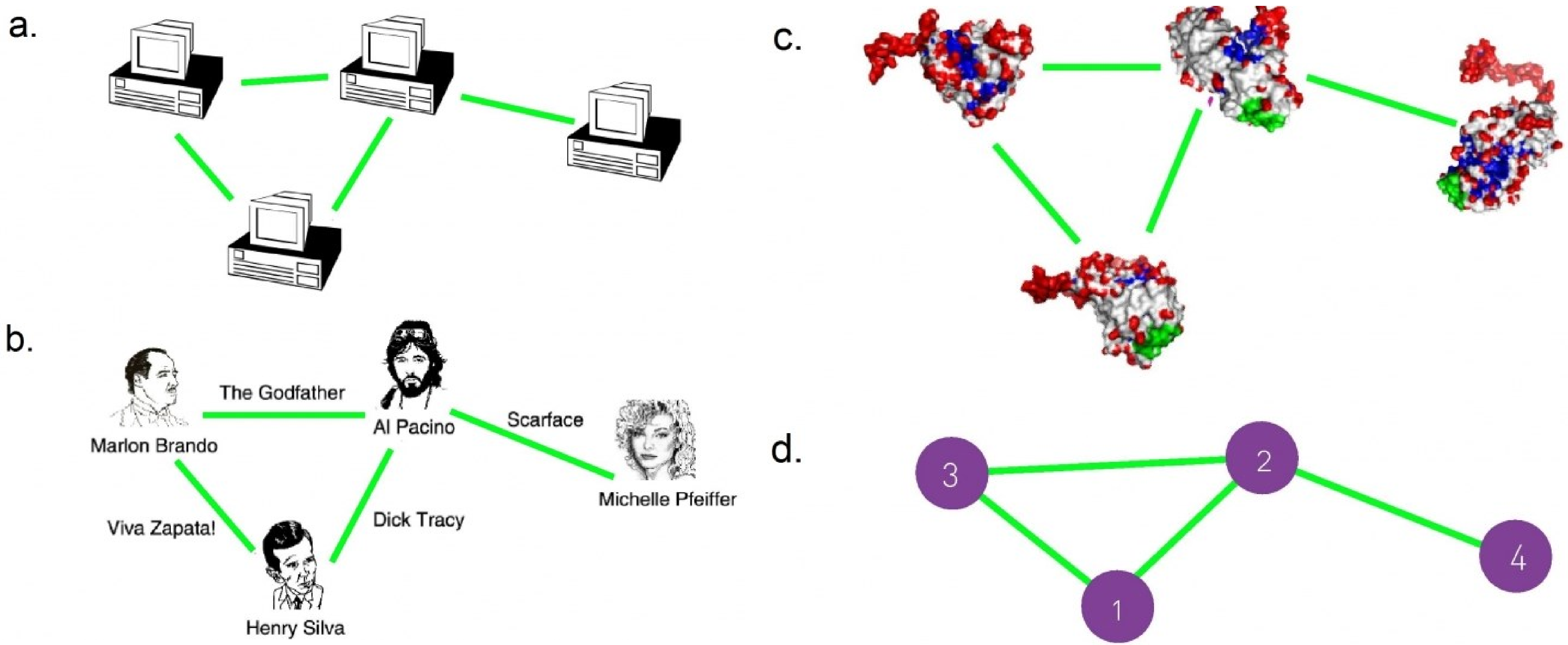
Solution: They build a new bridge between C and B (1875).

The Bridges of Königsberg (Video).





Networks and Graphs



Networks and Graphs

a – computer network

b – network of actors

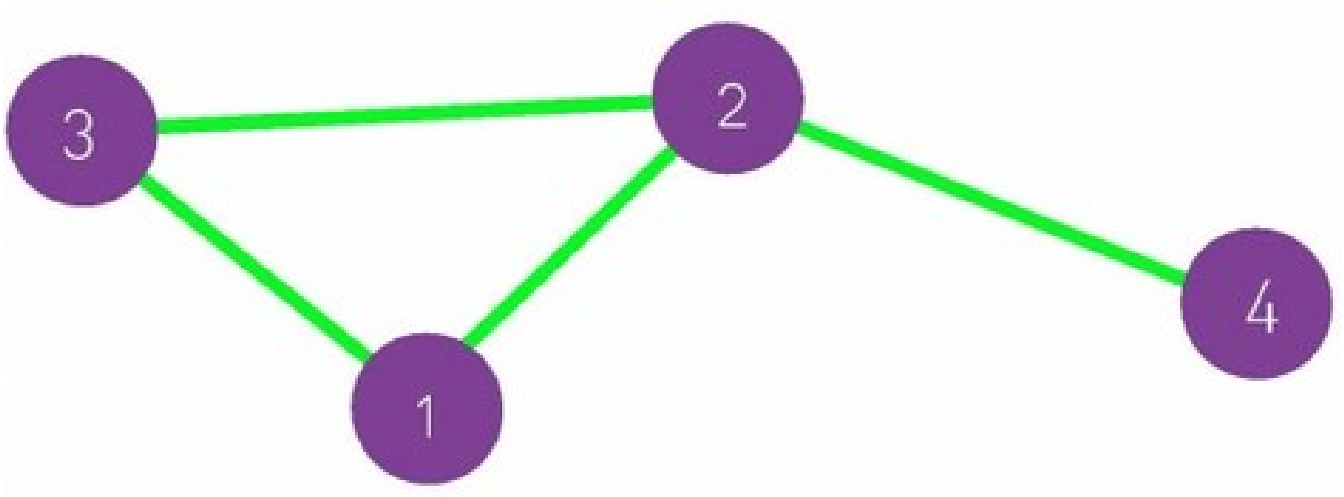
c – network of protein interactions

d – mathematical graph

Structurally these networks are the same.

Two important properties:

- Number of **nodes**:
  - N = 4
- Number of **links**:
  - L = 4



Degree and Average Degree

Questions: You have a social network from Facebook.

- What are the nodes and the links?
- Is it a directed or an undirected network?
- Who is the most well-known person?

Degree and Average Degree

You have a social network from Facebook.
Questions:

- What are the nodes and the links?
- Is it a directed or an undirected network?
- Who is the most well-known person?

Degree:

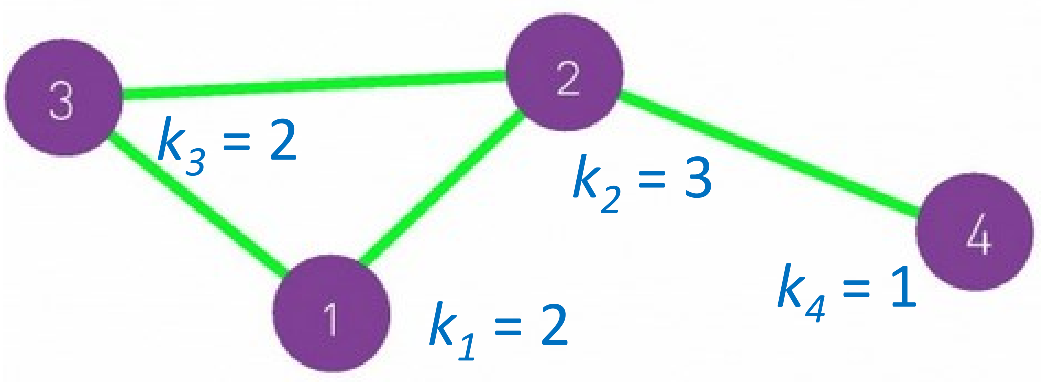
- k_i: degree of node i – the number of links belongs to node i

Total number of links in a network:

- $L = \frac{1}{2}\sum_{i=1}^{N} k_i$

Average degree:

- $\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}$



Degree and Average Degree – directed

Degree in directed case:

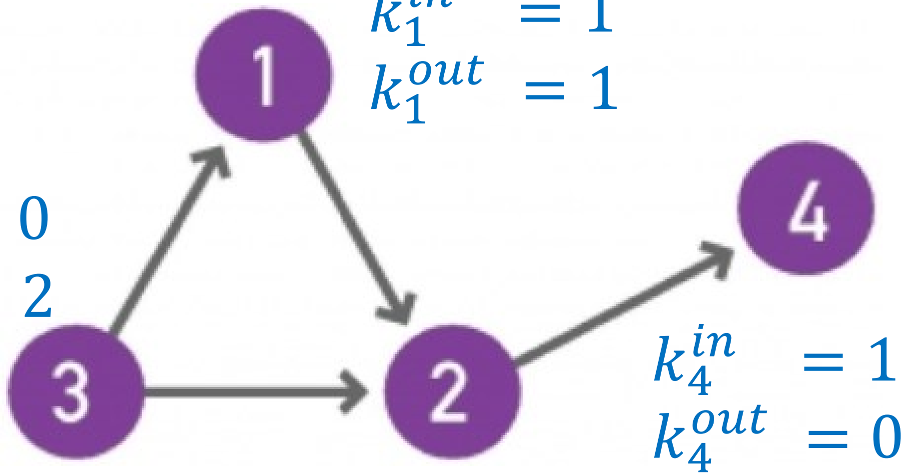
- Indigree (k_i^{in}): the number of links point to node i
- Outdegree (k_i^{out}): the number of links point from node i
- $k_i = k_i^{in} + k_i^{out}$

Total number of links in directed networks:

- $L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out}$

Average degree in directed networks:

- $\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{in}$
- $\langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{out}$
- $\langle k^{in} \rangle = \langle k^{out} \rangle = \frac{L}{N}$



$k_1^{in} = 1$, $k_1^{out} = 1$

$k_3^{in} = 0$, $k_3^{out} = 2$

$k_4^{in} = 1$, $k_4^{out} = 0$

$k_2^{in} = 2$, $k_2^{out} = 1$

Degree Distribution

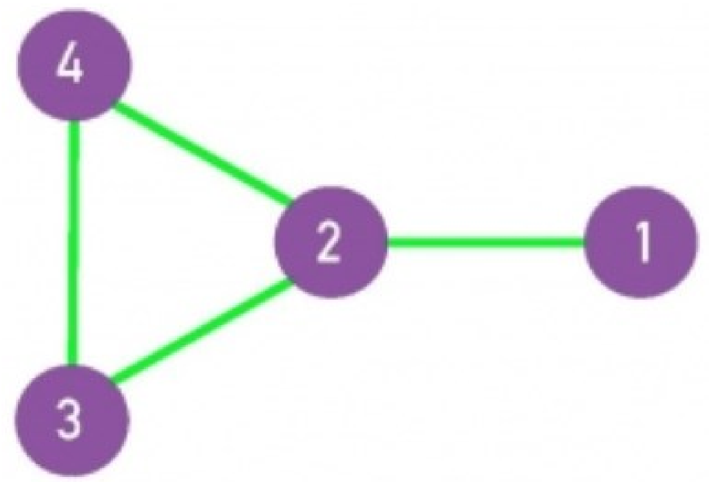
N_k: the number of nodes with degree k.

$p_k = \frac{N_k}{N}$: the probability that a randomly selected node has degree k.

Since p_k is a probability, it must be normalized: $\sum_{k=0}^{\infty} p_k = 1$.

Degree distribution had central role in discovering scale-free property.

Example 1:



Degree Distribution

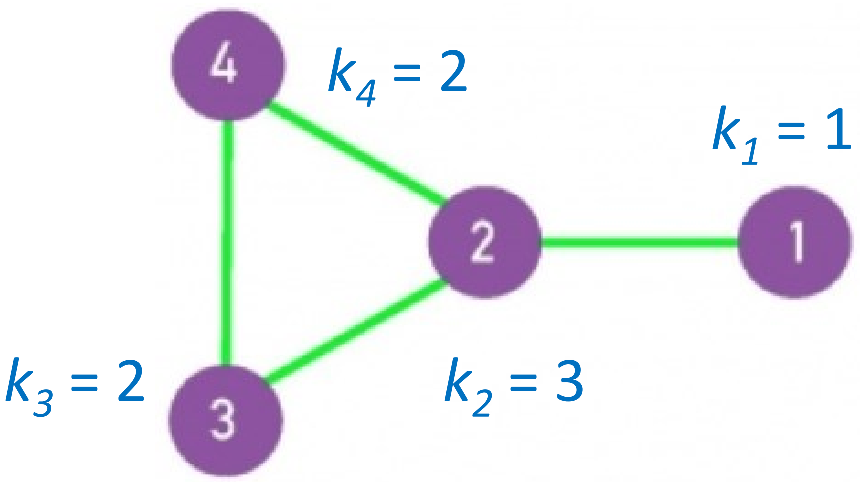
N_k: the number of nodes with degree k.

$p_k = \frac{N_k}{N}$: the probability that a randomly selected node has degree k.

Since p_k is a probability, it must be normalized: $\sum_{k=0}^{\infty} p_k = 1$.

Degree distribution had central role in discovering scale-free property.

Example 1:



Degree Distribution

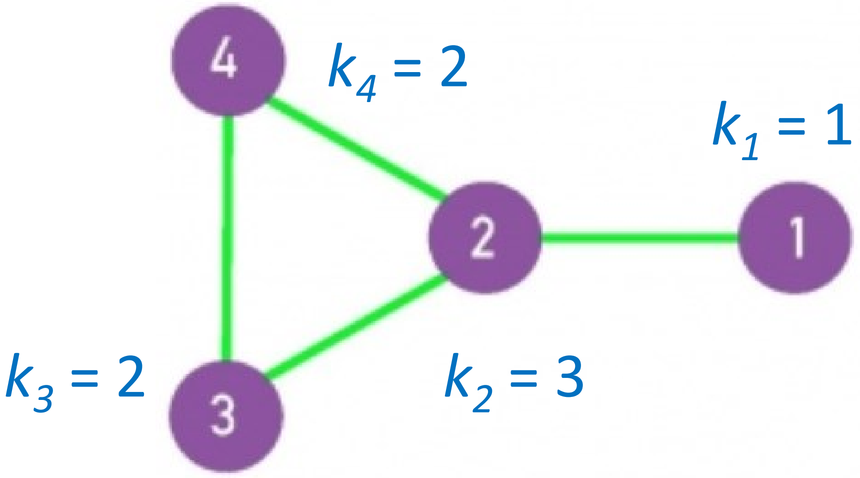
N_k: the number of nodes with degree k.

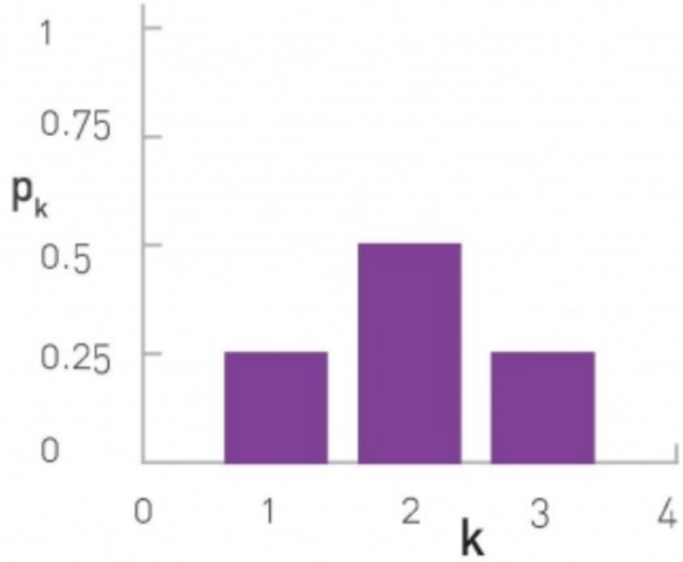
$p_k = \frac{N_k}{N}$: the probability that a randomly selected node has degree k.

Since p_k is a probability, it must be normalized: $\sum_{k=0}^{\infty} p_k = 1$.

Degree distribution had central role in discovering scale-free property.

Example 1:





Degree Distribution

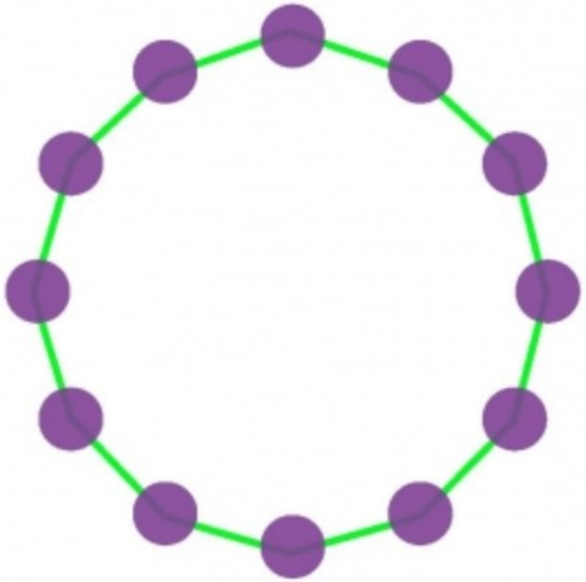
N_k: the number of nodes with degree k.

$p_k = \frac{N_k}{N}$: the probability that a randomly selected node has degree k.

Since p_k is a probability, it must be normalized: $\sum_{k=0}^{\infty} p_k = 1$.

Degree distribution had central role in discovering scale-free property.

Example 2:



Degree Distribution

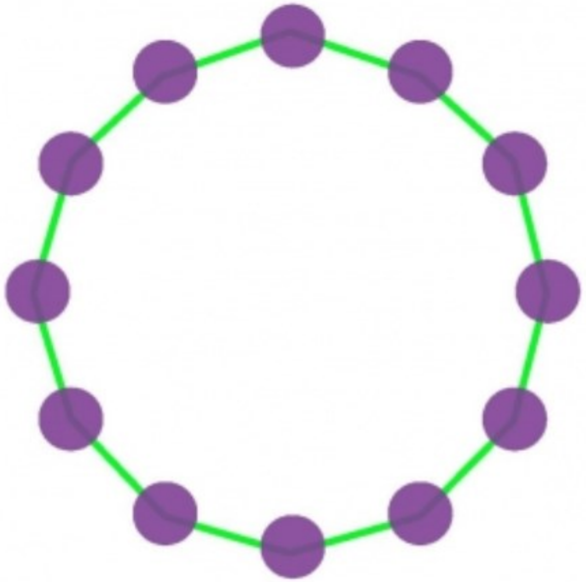
N_k: the number of nodes with degree k.

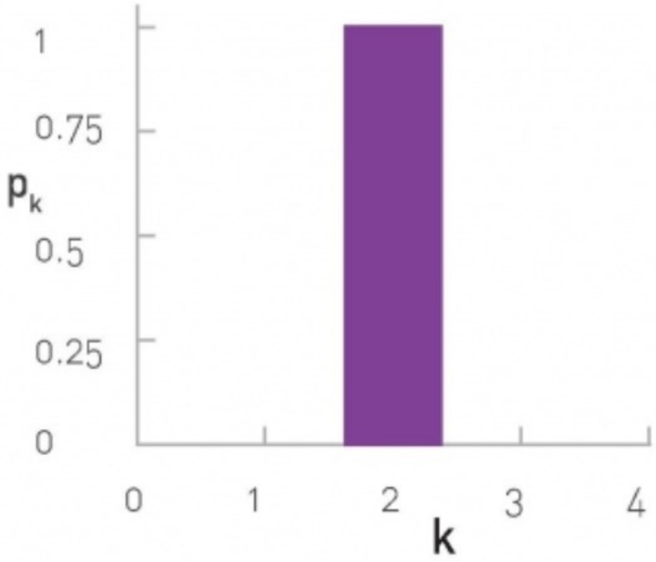
$p_k = \frac{N_k}{N}$: the probability that a randomly selected node has degree k.

Since p_k is a probability, it must be normalized: $\sum_{k=0}^{\infty} p_k = 1$.

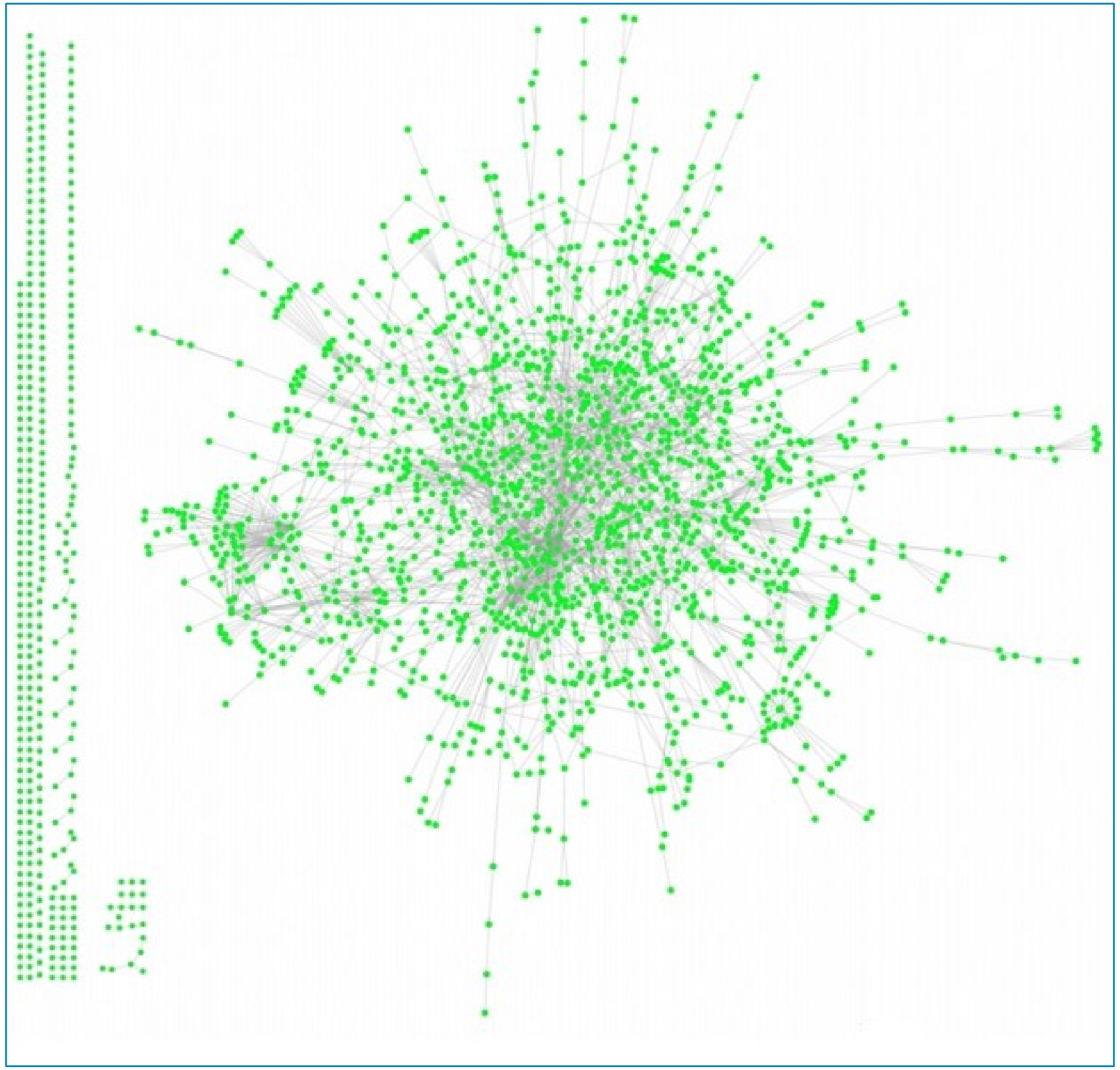
Degree distribution had central role in discovering scale-free property.

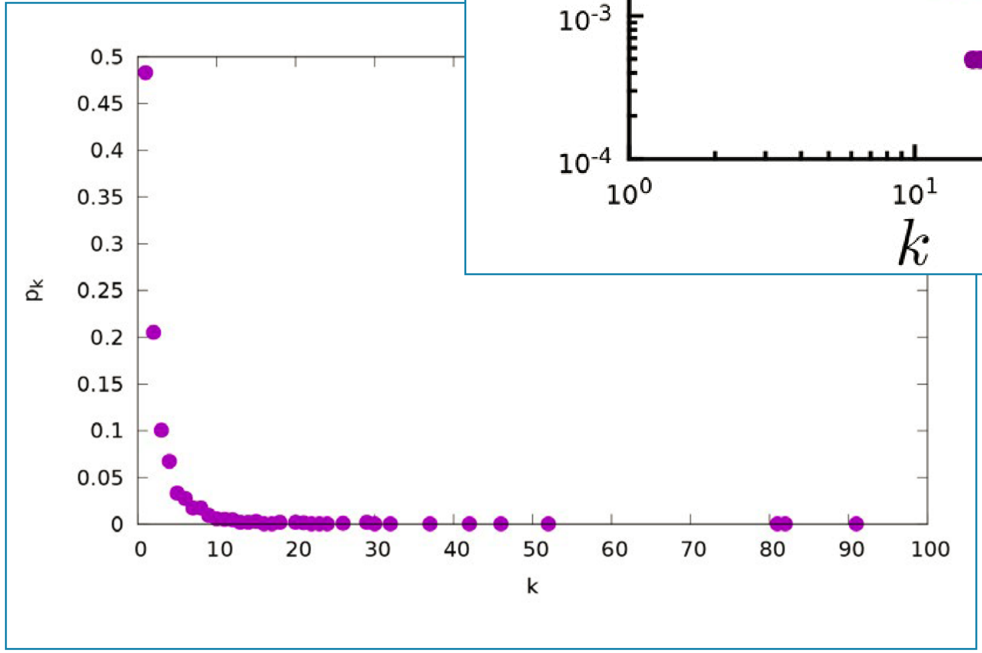
Example 2:

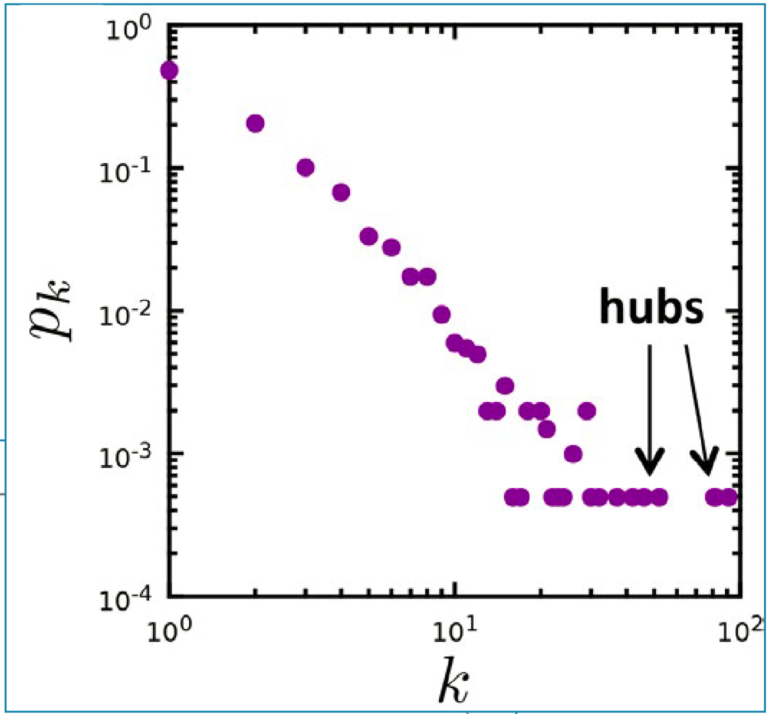




Degree Distribution – real example





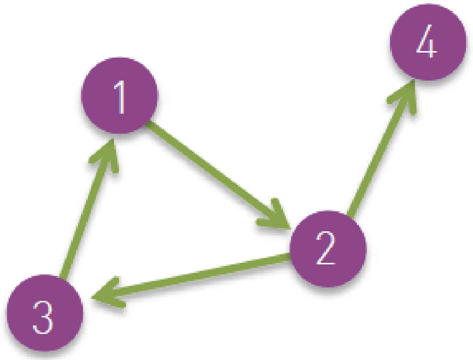


Adjacency Matrix

Mathematical description of a network: A

Directed case:

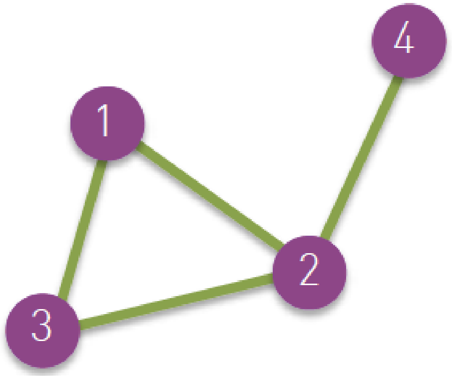
- $A_{ij} = 1$, if there is a link from node i to node j
- $A_{ij} = 0$, if there is no link from node i to node j



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Undirected case:

- $A_{ij} = A_{ji} = 1$, if there is a link between node i and j



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Real Networks are Sparse

The number of links in an undirected network can be between:

- $L_{min} = 0$
- $L_{max} = \binom{N}{2} = \frac{N(N-1)}{2}$.

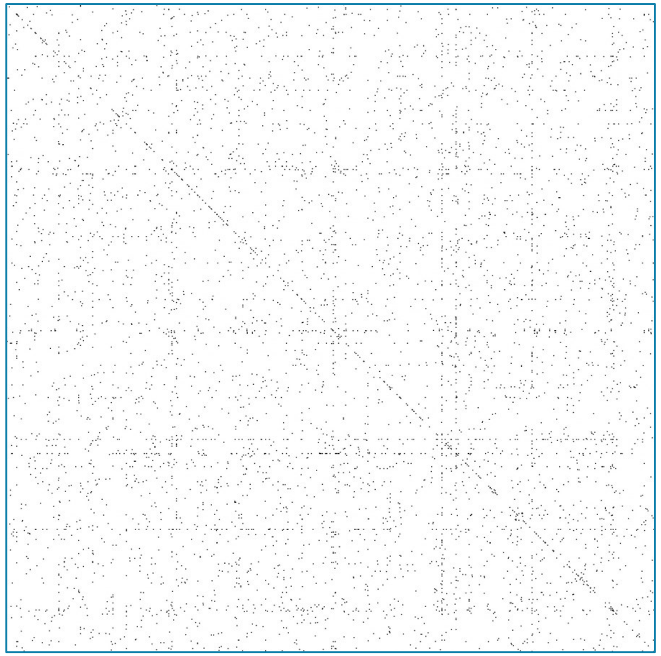
In reality $L \ll L_{max}$.

In yeast protein-protein interaction network:

- $N = 2018$
- $L = 2930$
- Theoretical maximum: $L_{\max} = 219\,853$
- Only 1.33% of possible connections

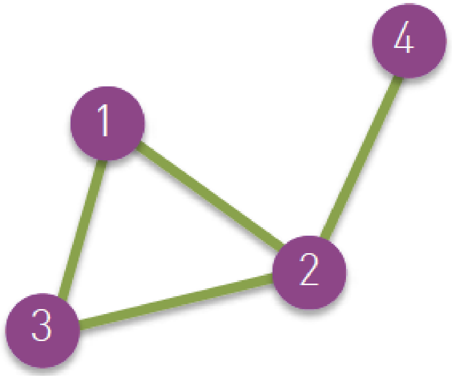
Solution:

- Edge list:



Edge list:

1 2
1 3
2 3
2 4



Weighted Networks

If we want to qualify the links, then we can associate weights for them.

For example:

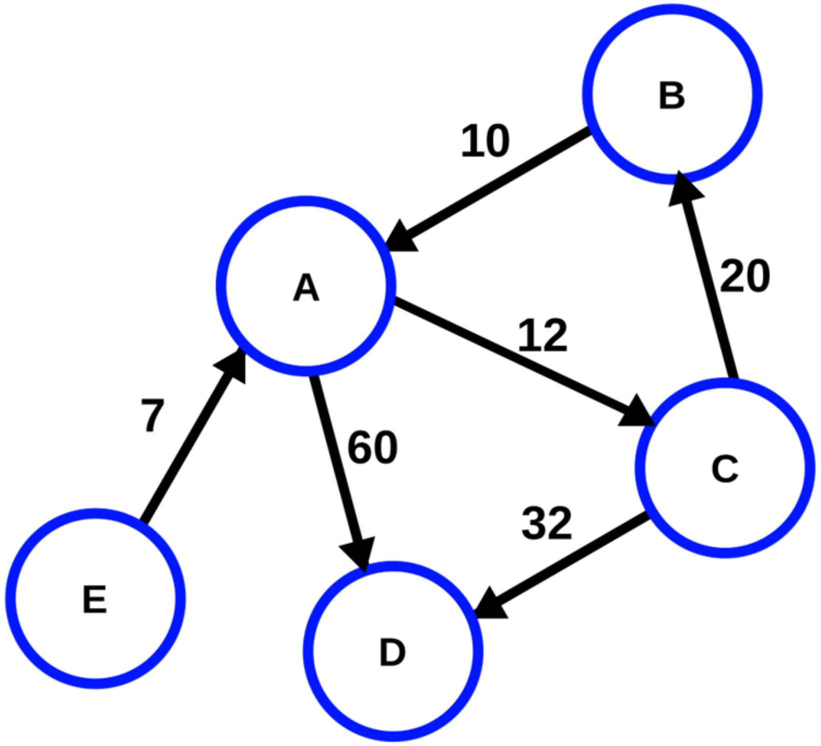
- Number of e-mails
- Length of phone call
- Distance between two cities
- …

In adjacency matrix:

- $A_{ij} = w_{ij}$

In edge list:

- From node, to node, weight
  - E.g. A, C, 12

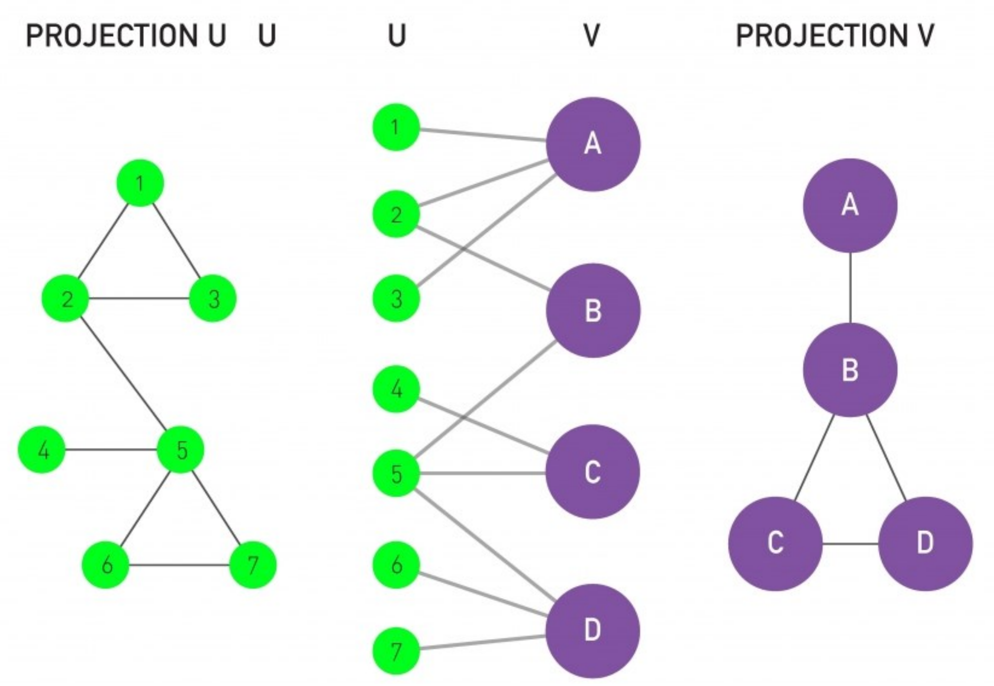


Bipartite Networks

Bigraph: a network whose nodes can be divided into two disjoint sets U and V such that each link connects a U-node to a V-node.

Projections:

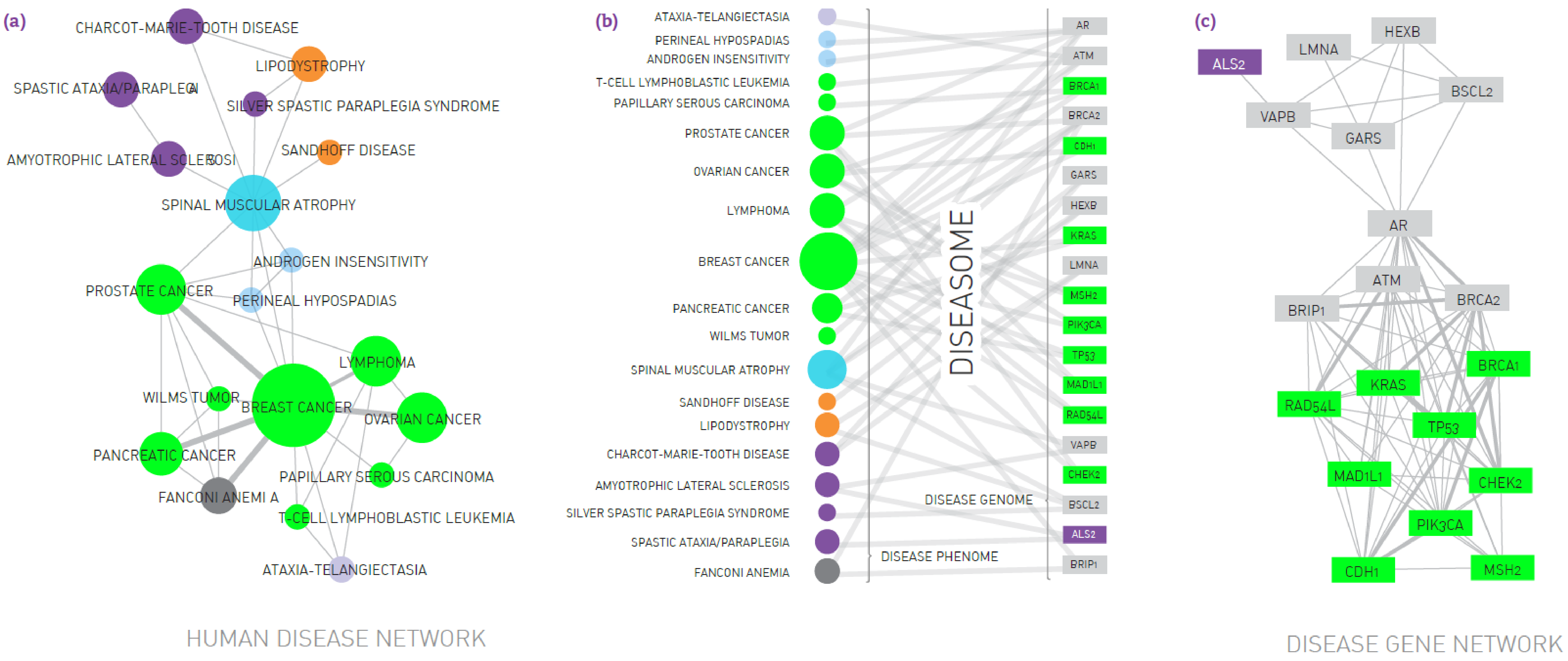
- 2 projections can be generated
- Projection U: two nodes are connected if they have at least one common neighbour from set V.
- Projection V: analogously



Example:

- Network of actors
- Network of diseases
- Network of recipe-ingredients

Bipartite Networks – Diseasome network



Paths and Distances

Path: Sequence of nodes such that each node is connected to the next one along the path by a link.

Shortest (Geodesic) path, d: The path with the shortest distance d between two nodes.

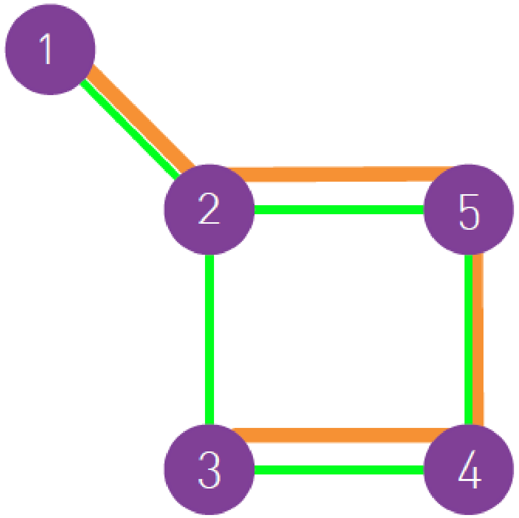
Network Diameter, $d_{\max}$: maximum shortest path in the network.

Average Path Length, $\langle d \rangle$: The average of the shortest paths between all pairs of nodes.

Cycle: A path with the same start and end node.

Eulerian Path: A path that traverses each link exactly once.

Hamiltonian Path: A path that visits each node exactly once.



Paths and Distances

Path: Sequence of nodes such that each node is connected to the next one along the path by a link.

Shortest (Geodesic) path, d: The path with the shortest distance d between two nodes.

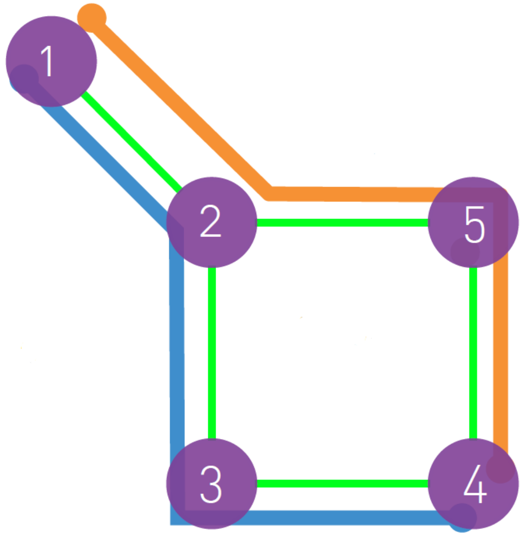
Network Diameter, $d_{\max}$: maximum shortest path in the network.

Average Path Length, $\langle d \rangle$: The average of the shortest paths between all pairs of nodes.

Cycle: A path with the same start and end node.

Eulerian Path: A path that traverses each link exactly once.

Hamiltonian Path: A path that visits each node exactly once.



Paths and Distances

Path: Sequence of nodes such that each node is connected to the next one along the path by a link.

Shortest (Geodesic) path, d: The path with the shortest distance d between two nodes.

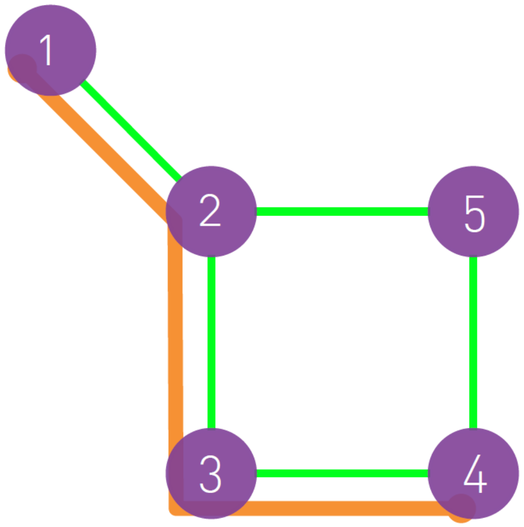
Network Diameter, $d_{\max}$: maximum shortest path in the network.

Average Path Length, $\langle d \rangle$: The average of the shortest paths between all pairs of nodes.

Cycle: A path with the same start and end node.

Eulerian Path: A path that traverses each link exactly once.

Hamiltonian Path: A path that visits each node exactly once.



Paths and Distances

Path: Sequence of nodes such that each node is connected to the next one along the path by a link.

Shortest (Geodesic) path, d: The path with the shortest distance d between two nodes.

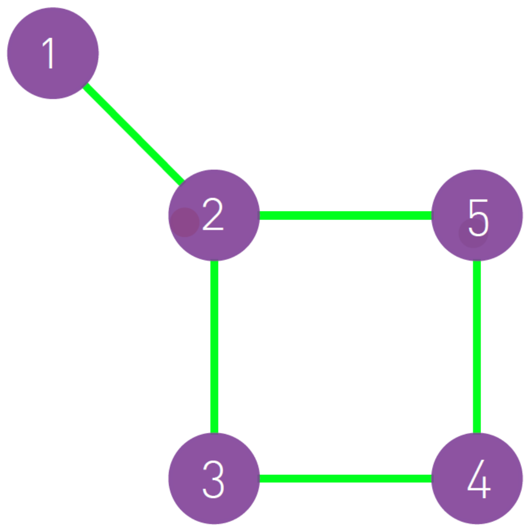
Network Diameter, $d_{\max}$: maximum shortest path in the network.

Average Path Length, $\langle d \rangle$: The average of the shortest paths between all pairs of nodes.

Cycle: A path with the same start and end node.

Eulerian Path: A path that traverses each link exactly once.

Hamiltonian Path: A path that visits each node exactly once.



Paths and Distances

Path: Sequence of nodes such that each node is connected to the next one along the path by a link.

Shortest (Geodesic) path, d: The path with the shortest distance d between two nodes.

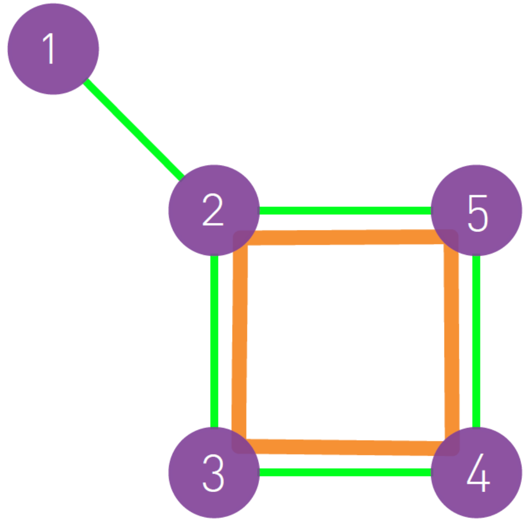
Network Diameter, $d_{\max}$: maximum shortest path in the network.

Average Path Length, $\langle d \rangle$: The average of the shortest paths between all pairs of nodes.

Cycle: A path with the same start and end node.

Eulerian Path: A path that traverses each link exactly once.

Hamiltonian Path: A path that visits each node exactly once.



Paths and Distances

Path: Sequence of nodes such that each node is connected to the next one along the path by a link.

Shortest (Geodesic) path, d: The path with the shortest distance d between two nodes.

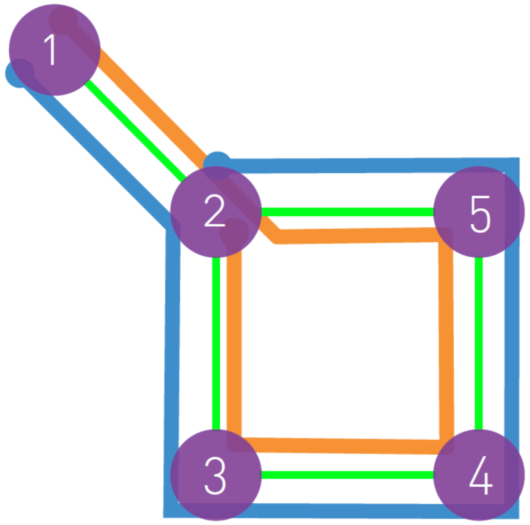
Network Diameter, $d_{\max}$: maximum shortest path in the network.

Average Path Length, $\langle d \rangle$: The average of the shortest paths between all pairs of nodes.

Cycle: A path with the same start and end node.

Eulerian Path: A path that traverses each link exactly once.

Hamiltonian Path: A path that visits each node exactly once.



Paths and Distances

Path: Sequence of nodes such that each node is connected to the next one along the path by a link.

Shortest (Geodesic) path, d: The path with the shortest distance d between two nodes.

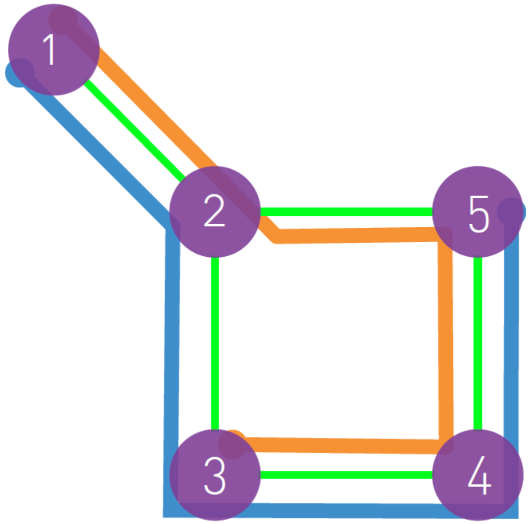
Network Diameter, $d_{\max}$: maximum shortest path in the network.

Average Path Length, $\langle d \rangle$: The average of the shortest paths between all pairs of nodes.

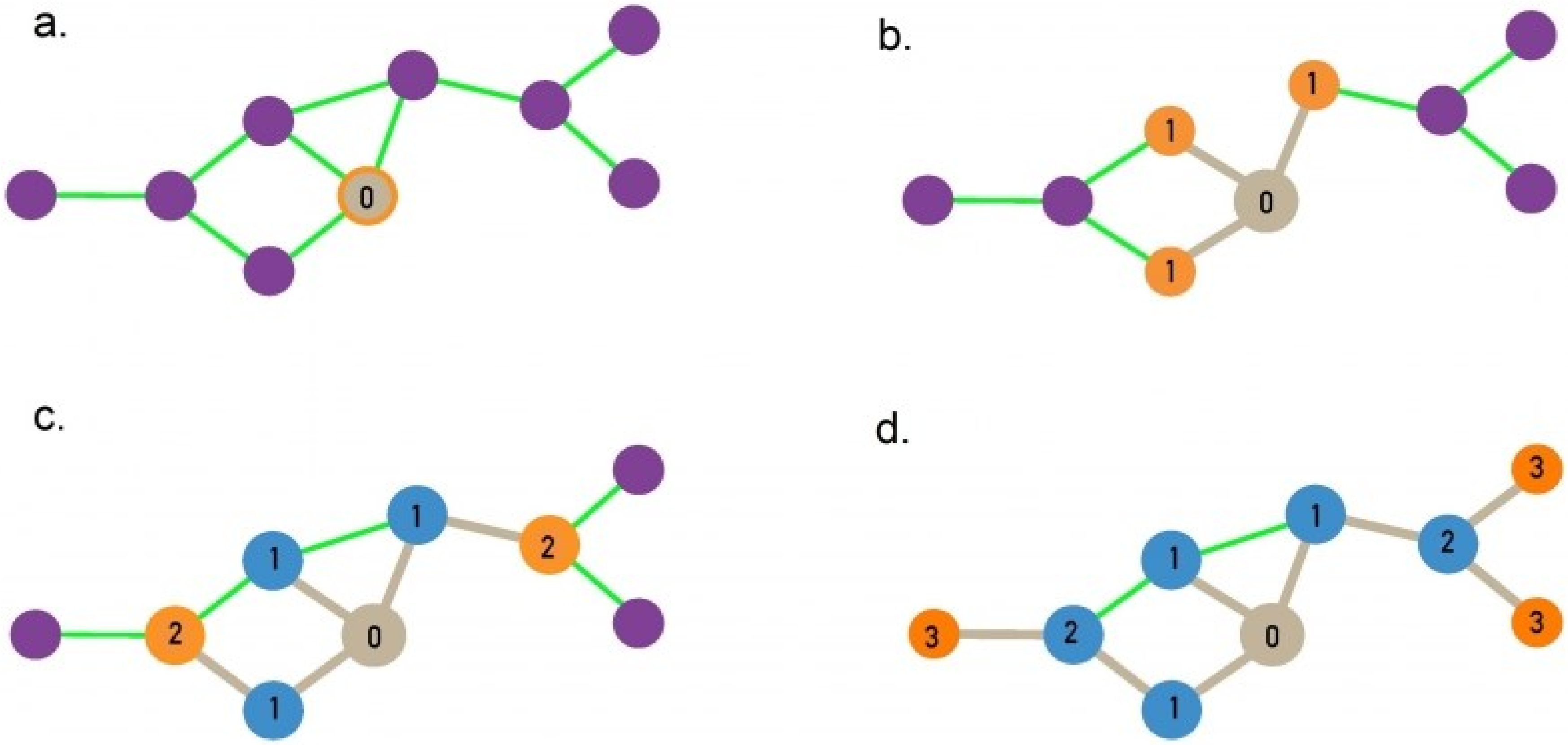
Cycle: A path with the same start and end node.

Eulerian Path: A path that traverses each link exactly once.

Hamiltonian Path: A path that visits each node exactly once.



Breadth-First Search (BFS) Algorithm



Connectedness

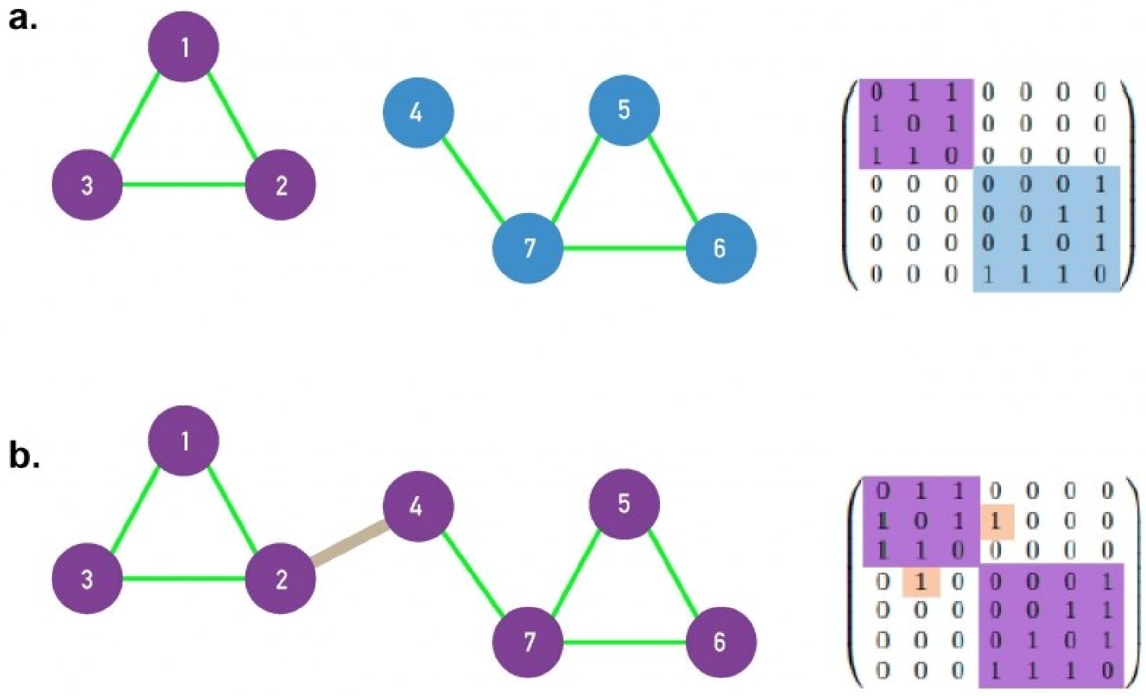
In an undirected network nodes i and j are *connected* if there is a path between them. They are *disconnected* if such a path does not exist, $d_{ij} = \infty$.

A network *is connected* if all pairs of nodes in the network are connected.

A network *is disconnected* if there is at least one pair of nodes with $d_{ij} = \infty$.

In a disconnected network we call its subnetworks *components* or *clusters*.

The link that connects two clusters is called *bridge*.



Clustering Coefficient (undirected case)

Cusltering Coefficient (C_i) measures the network's local link density.

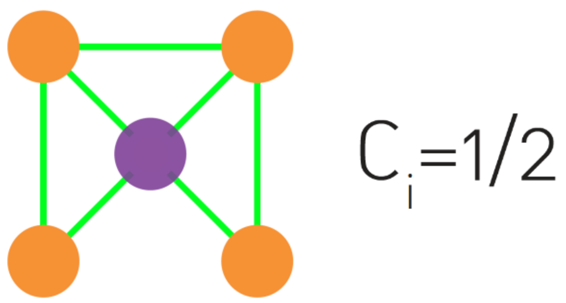
$$C_i = \frac{2L_i}{k_i(k_i-1)}$$

- L_i: number of links between the neighbours of node i

$C_i = 0$, if none of the neighbours of node i links to each other.

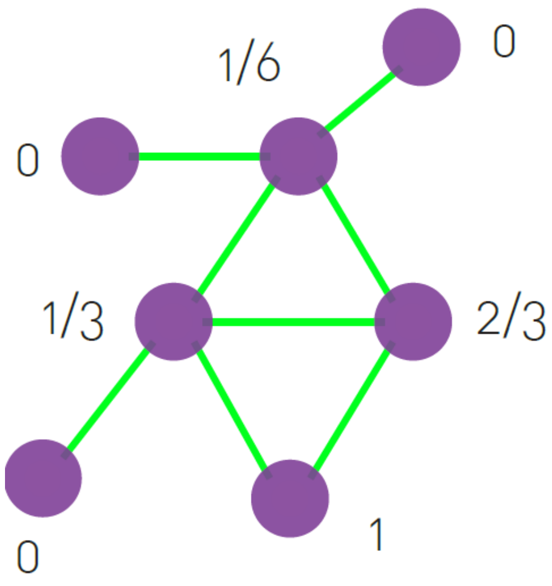
$C_i = 1$, if the neighbours of node i form a complete graph.

C_i is the probability that two neighbours of a node are connected to each other.



Average Clustering Coefficient ($\langle C \rangle$): degree of clustering of a whole network.

$$\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i$$



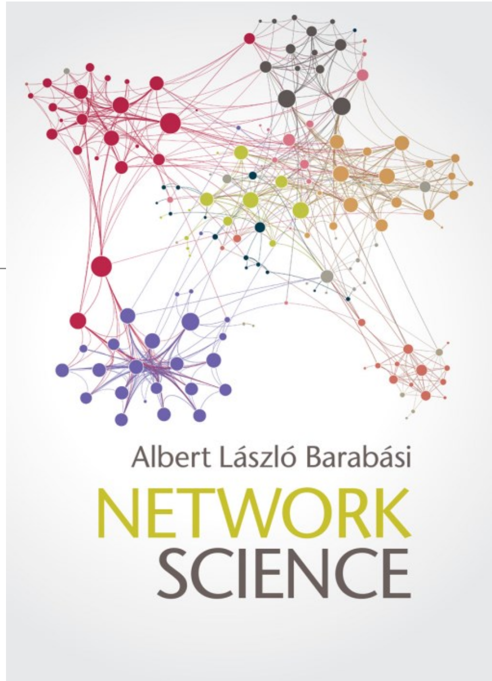
Network Analysis

03 – RANDOM NETWORKS

Slides were created by: Daniel Leitold

Network Science book (online)

Barabási, Albert-László. *Network Science*.
Cambridge University Press, 2016.



Party and wine

You invite 100 people for a party.

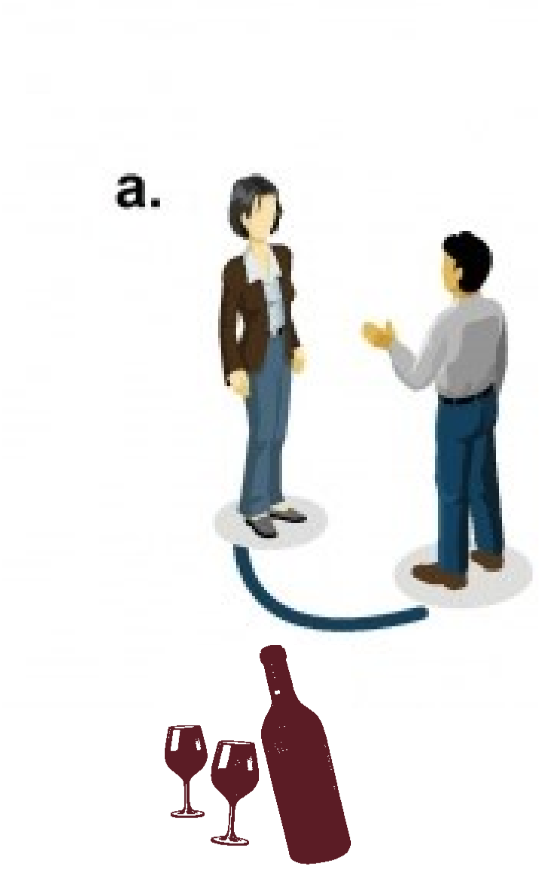
They do not know each other in the beginning.

Talking groups of 2 – 3 appear.

Then, you unfortunately said to Jane that the wine in unlabelled bottles is much better.

What happened?

Party and wine



She shares this information only with her acquaintances. If she talks just 5 minutes to each person, then to share this information with everyone takes 5*99 minutes that is more than 8 hours.

So can you calm down?

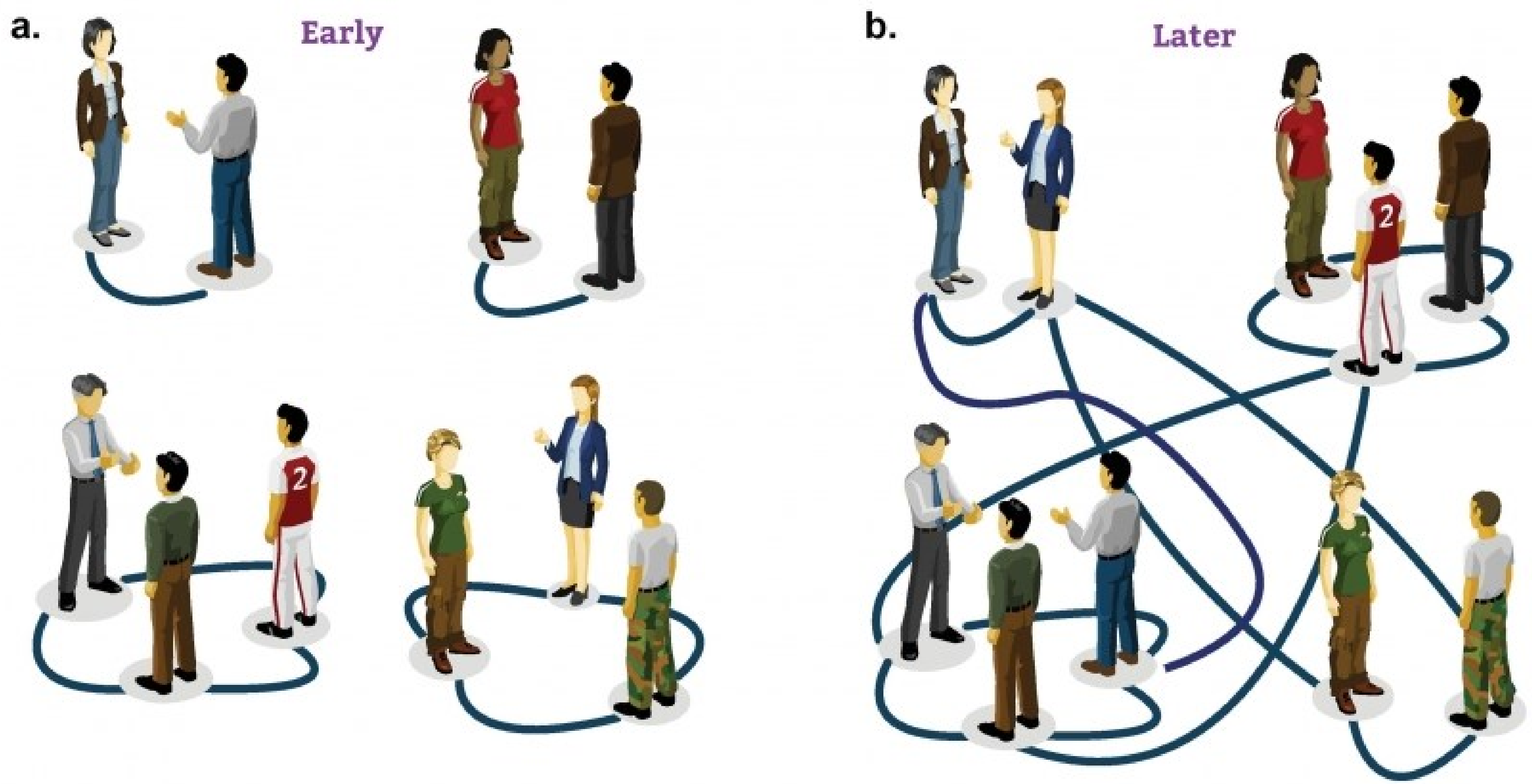
Party and wine

She shares this information only with her acquaintances. If she talks just 5 minutes to each person, then to share this information with everyone takes 5*99 minutes that is more than 8 hours.

So can you calm down?

NO!

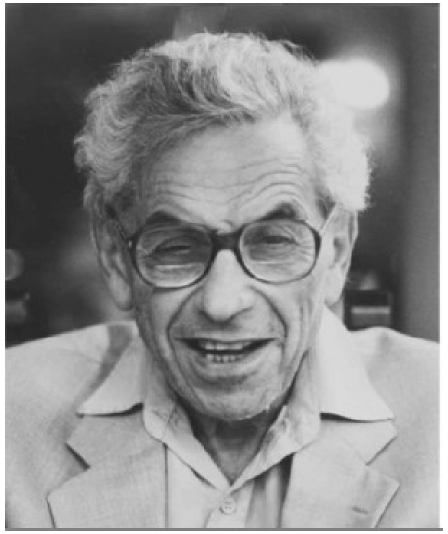
Party and wine



The Random Network Model

Two definitions:

- A random graph $G(N, p)$ is a graph of N nodes where each pair of nodes is connected by probability p. – Erős-Rényi model (ER model)
- A random graph $G(N, L)$ is a graph of N nodes that are connected by L randomly placed links.



Pál Erdős (1913-1996)

Alfréd Rényi (1921-1970)

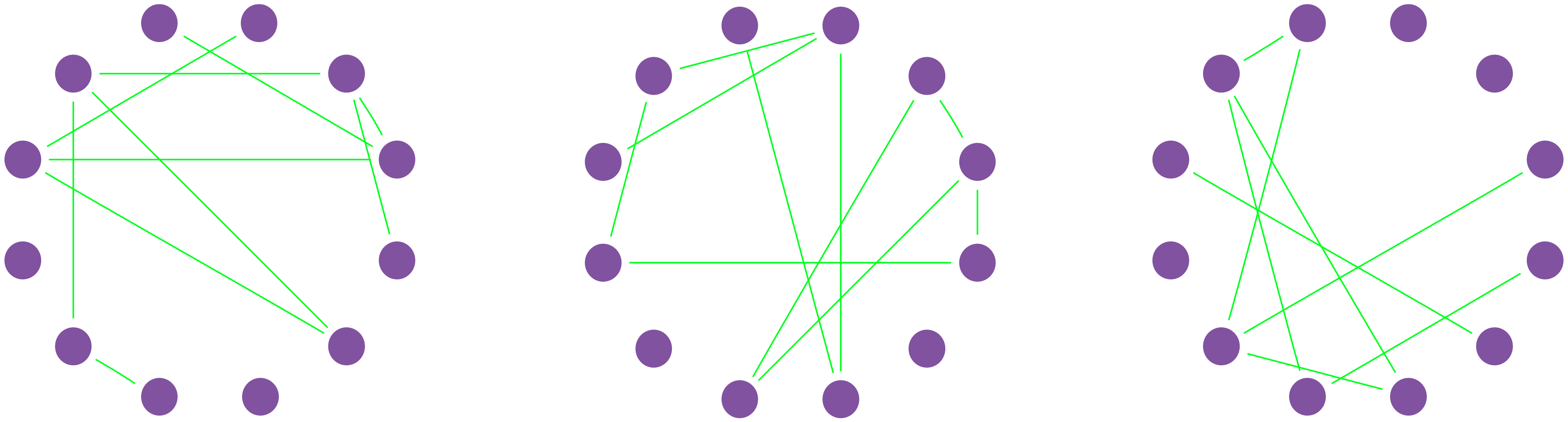
The Random Network Model

A random graph $G(N, p)$ is a graph of N nodes where each pair of nodes is connected by probability p.

A random graph $G(N, L)$ is a graph of N nodes that are connected by L randomly placed links.

$N = 12$

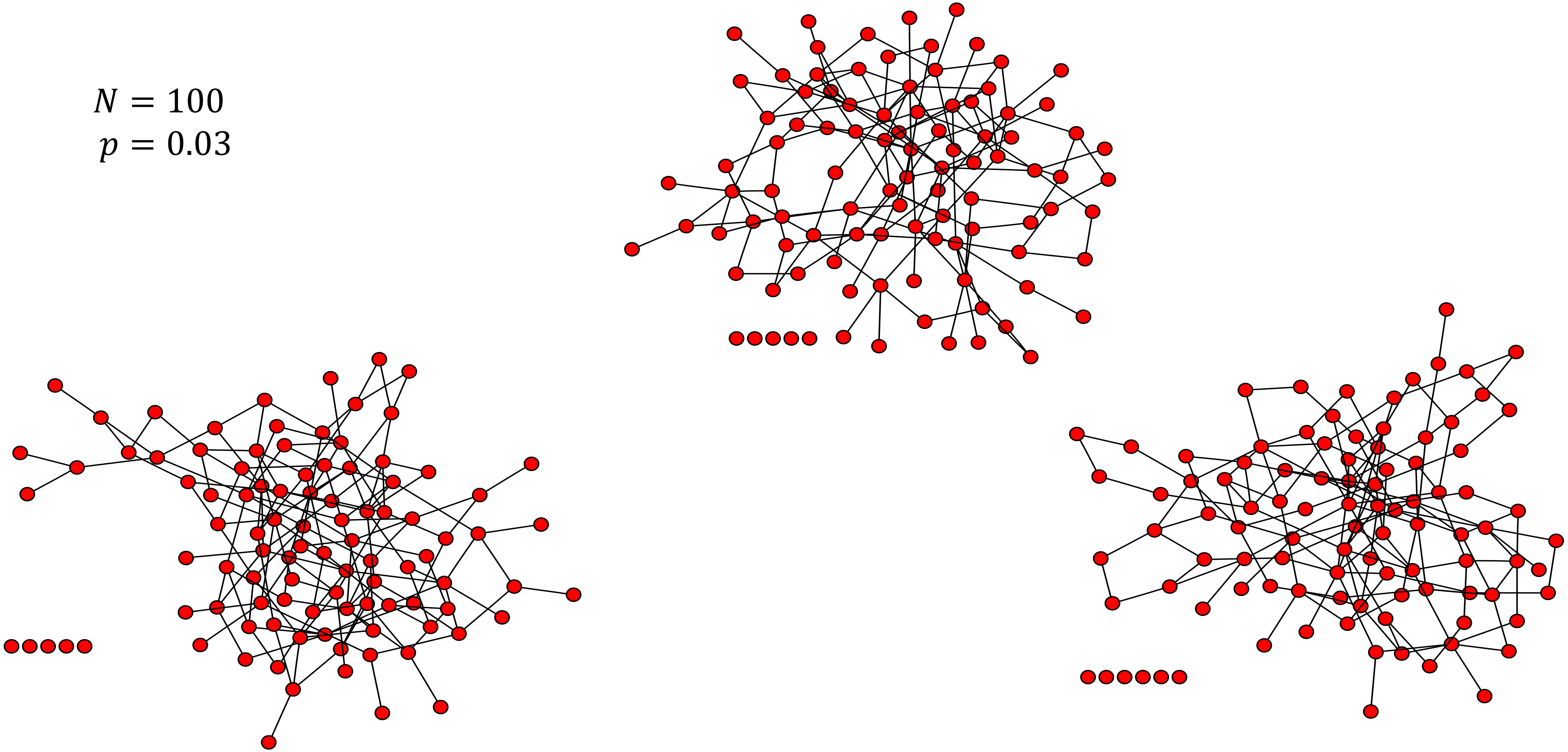
$p = \frac{1}{6}$



$L = 10$ $\quad$ $L = 10$ $\quad$ $L = 8$

The Random Network Model

$$N = 100$$
$$p = 0.03$$



Degree Distribution

The probability that a random node has exactly k links is the product of three terms:

Degree Distribution

The probability that a random node has exactly k links is the product of three terms:

- The probability that k links are connected to the node: p^k

Degree Distribution

The probability that a random node has exactly k links is the product of three terms:

- The probability that k links are connected to the node: p^k
- The probability that the remaining $(N - 1 - k)$ links are missing: $(1 - p)^{N-1-k}$

Degree Distribution

The probability that a random node has exactly k links is the product of three terms:

- The probability that k links are connected to the node: p^k
- The probability that the remaining $(N-1-k)$ links are missing: $(1-p)^{N-1-k}$
- A combinational factor: $\binom{N-1}{k}$

Degree Distribution

The probability that a random node has exactly k links is the product of three terms:

- The probability that k links are connected to the node: p^k
- The probability that the remaining $(N-1-k)$ links are missing: $(1-p)^{N-1-k}$
- A combinational factor: $\binom{N-1}{k}$

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

Degree Distribution

The probability that a random node has exactly k links is the product of three terms:

- The probability that k links are connected to the node: p^k
- The probability that the remaining $(N-1-k)$ links are missing: $(1-p)^{N-1-k}$
- A combinational factor: $\binom{N-1}{k}$

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

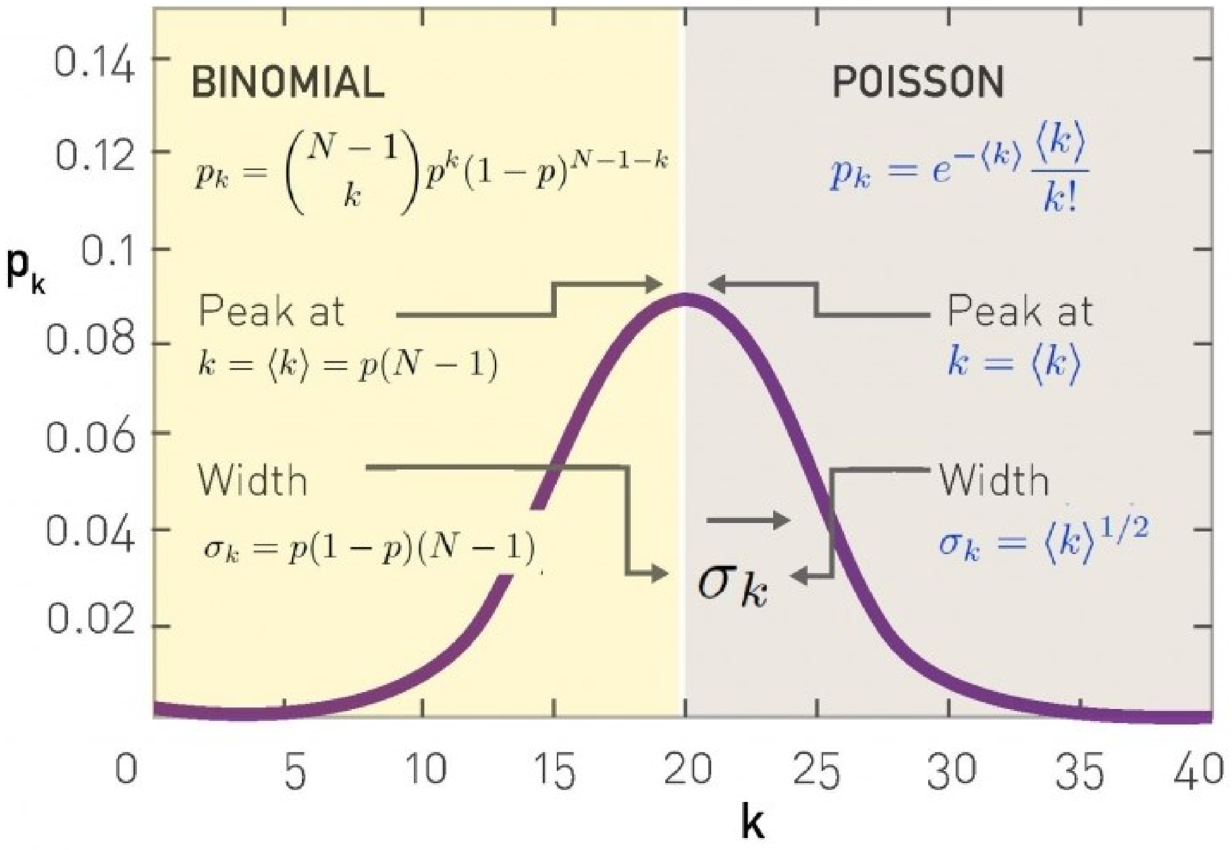
Binomial distribution

Degree Distribution

The most of real networks are sparse $\langle k \rangle \ll N$.

In this limit the degree distribution is well approximated by the Poisson distribution.

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$



Real Networks are Not Poisson

The human population is $N = 7 * 10^9$.

Sociologists estimate that a typical person knows about 1000 people.

According to Poisson distribution:

Real Networks are Not Poisson

The human population is $N = 7 * 10^9$.

Sociologists estimate that a typical person knows about 1000 people.

According to Poisson distribution:

- $k_{\max} = 1185$
- $\sigma_k = \langle k \rangle^{\frac{1}{2}} = 31.62$
- Usually: $\langle k \rangle \pm \sigma_k$
  between 968 and 1032

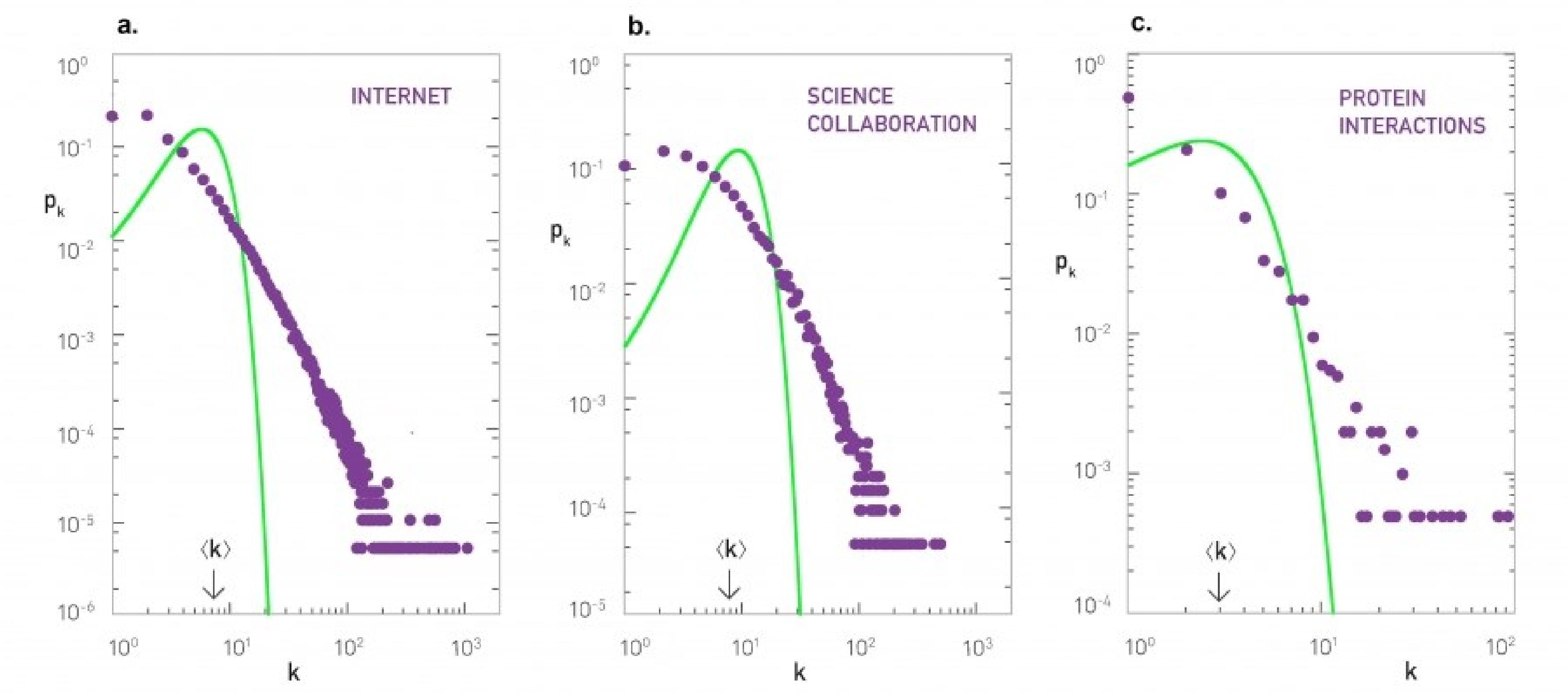
Real Networks are Not Poisson

The human population is $N = 7 * 10^9$.

Sociologists estimate that a typical person knows about 1000 people.

According to Poisson distribution:

- $k_{\max} = 1185$
- $\sigma_k = \langle k \rangle^{\frac{1}{2}} = 31.62$
- Usually: $\langle k \rangle \pm \sigma_k$ between 968 and 1032



The Evolution of a Random Network

The social network at the party is evolved by the new acquaintances.

This means a continuously changing p.

Firstly, how $\langle k \rangle$ influences the size of giant component

- Giant component (N_G): A significant connected portion of the network.

The Evolution of a Random Network

The social network at the party is evolved by the new acquaintances.

This means a continuously changing p.

Firstly, how $\langle k \rangle$ influences the size of giant component

- Giant component (N_G): A significant connected portion of the network.

Trivial cases:

- If $p = 0$, then $\langle k \rangle = 0$, $N_G = 1$, $\frac{N_G}{N} \to 0$
- If $p = 1$, then $\langle k \rangle = N - 1$, $N_G = N$, $\frac{N_G}{N} = 1$

The Evolution of a Random Network

The social network at the party is evolved by the new acquaintances.

This means a continuously changing p.

Firstly, how $\langle k \rangle$ influences the size of giant component

- Giant component (N_G): A significant connected portion of the network.

Trivial cases:

- If $p = 0$, then $\langle k \rangle = 0$, $N_G = 1$, $\frac{N_G}{N} \rightarrow 0$
- If $p = 1$, then $\langle k \rangle = N - 1$, $N_G = N$, $\frac{N_G}{N} = 1$

Suspicion:

- If $\langle k \rangle$ increases from $0 \rightarrow N - 1$, N_G grows gradually from $1 \rightarrow N$

The Evolution of a Random Network

The social network at the party is evolved by the new acquaintances.

This means a continuously changing p.

Firstly, how $\langle k \rangle$ influences the size of giant component

- Giant component (N_G): A significant connected portion of the network.

Trivial cases:

- If $p = 0$, then $\langle k \rangle = 0$, $N_G = 1$, $\frac{N_G}{N} \to 0$
- If $p = 1$, then $\langle k \rangle = N - 1$, $N_G = N$, $\frac{N_G}{N} = 1$

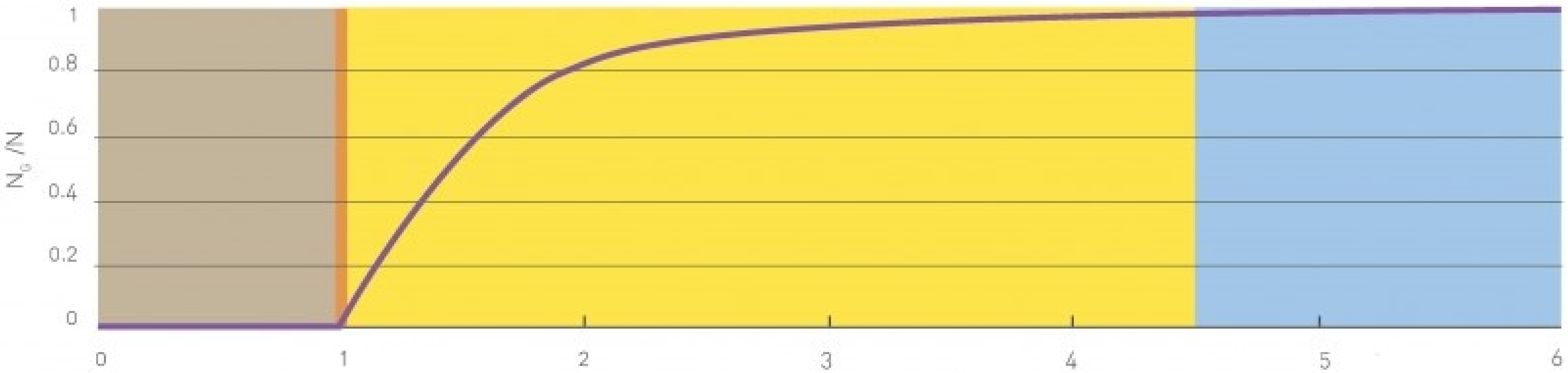
Suspicion:

- If $\langle k \rangle$ increases from $0 \to \ N - 1$, N_G grows gradually from $1 \to N$

Reality:

- $\frac{N_G}{N}$ increases rapidly, if $\langle k \rangle$ exceeds a critical value

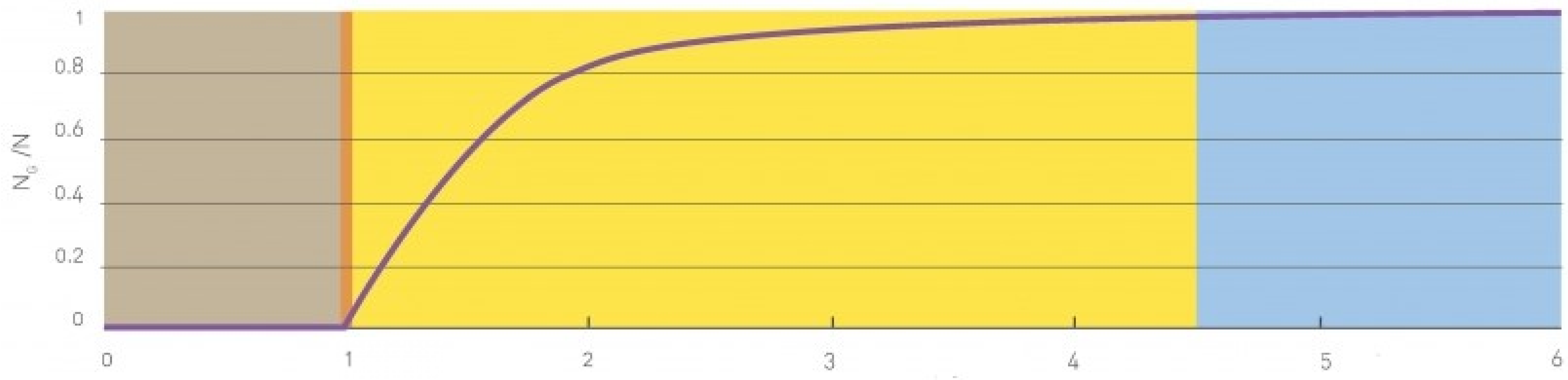
The Evolution of a Random Network



What is the critical value of $\langle k \rangle$?

Video

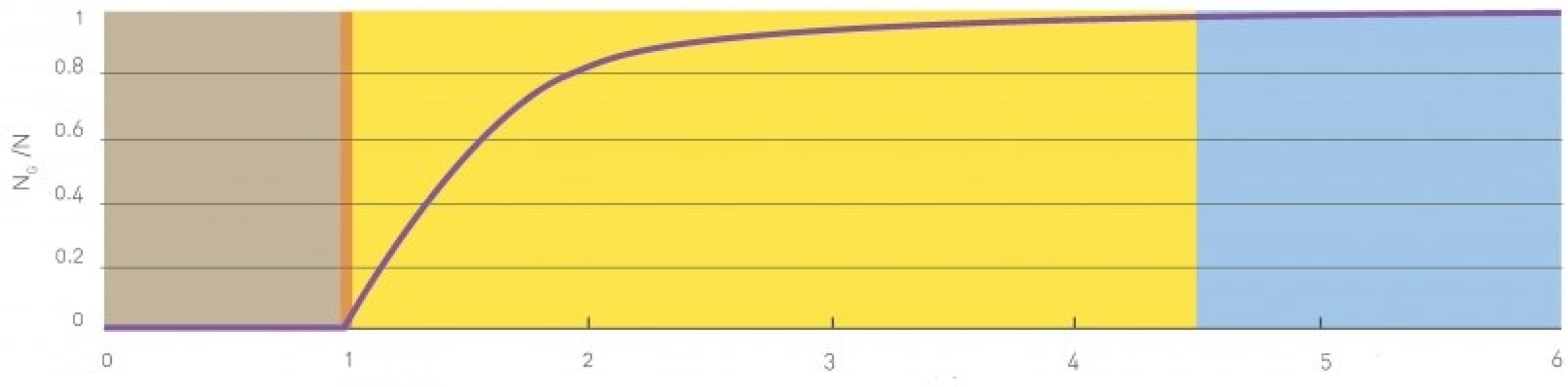
The Evolution of a Random Network



What is the critical value of $\langle k \rangle$? → 1

Video

The Evolution of a Random Network



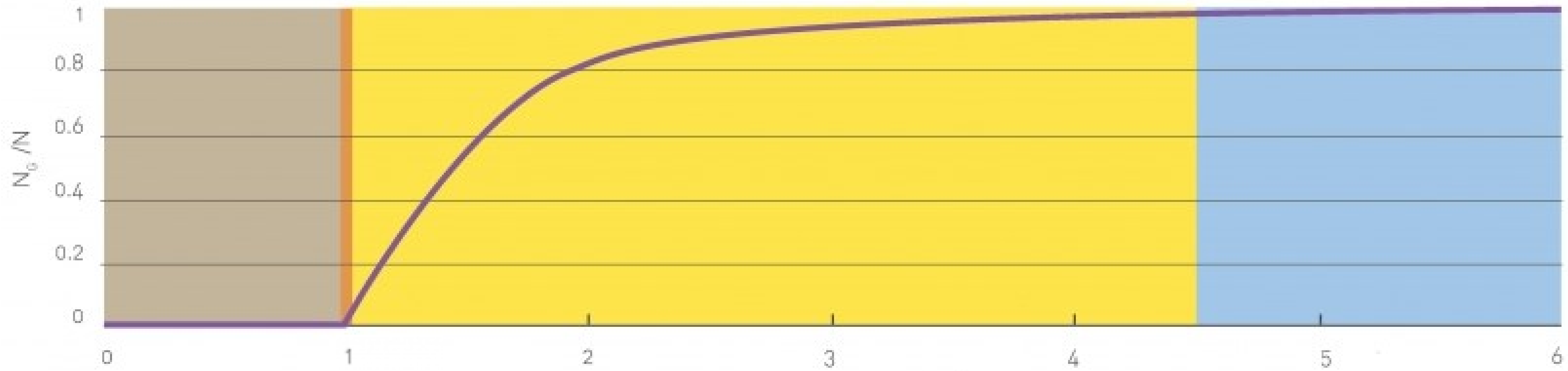
What is the critical value of $\langle k \rangle$? → 1

[Video]

Four domains:

- Subcritical: $\langle k \rangle$, p
- Critical: $\langle k \rangle$, p
- Supercritical: $\langle k \rangle$, p
- Connected: $\langle k \rangle$, p

The Evolution of a Random Network



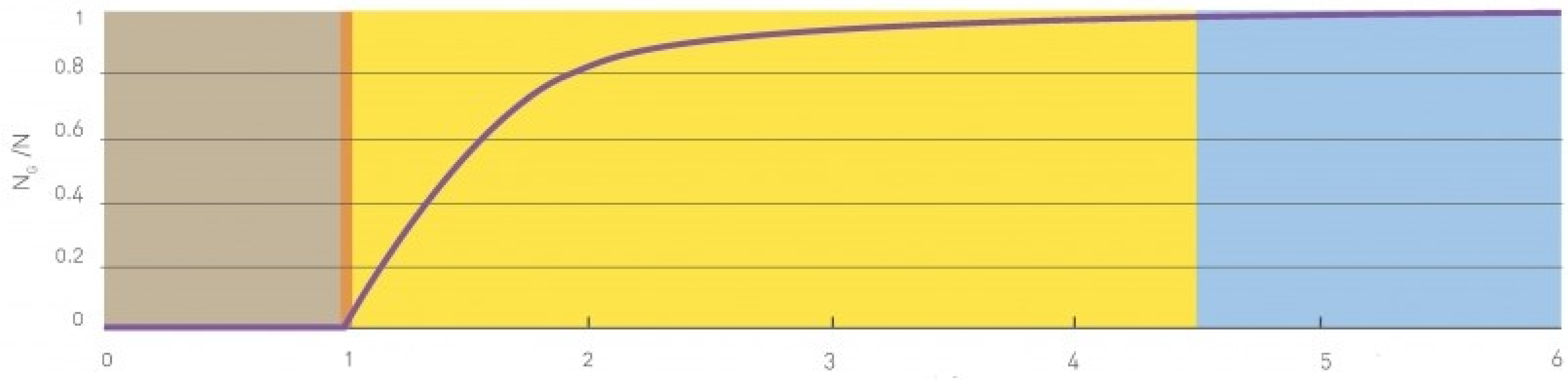
What is the critical value of $\langle k \rangle$? → 1

Video

Four domains:

- Subcritical: $\langle k \rangle > 0,\ p < \frac{1}{N}$
- Critical: $\langle k \rangle$ $,\ p$
- Supercritical: $\langle k \rangle$ $,\ p$
- Connected: $\langle k \rangle$ $,\ p$

The Evolution of a Random Network



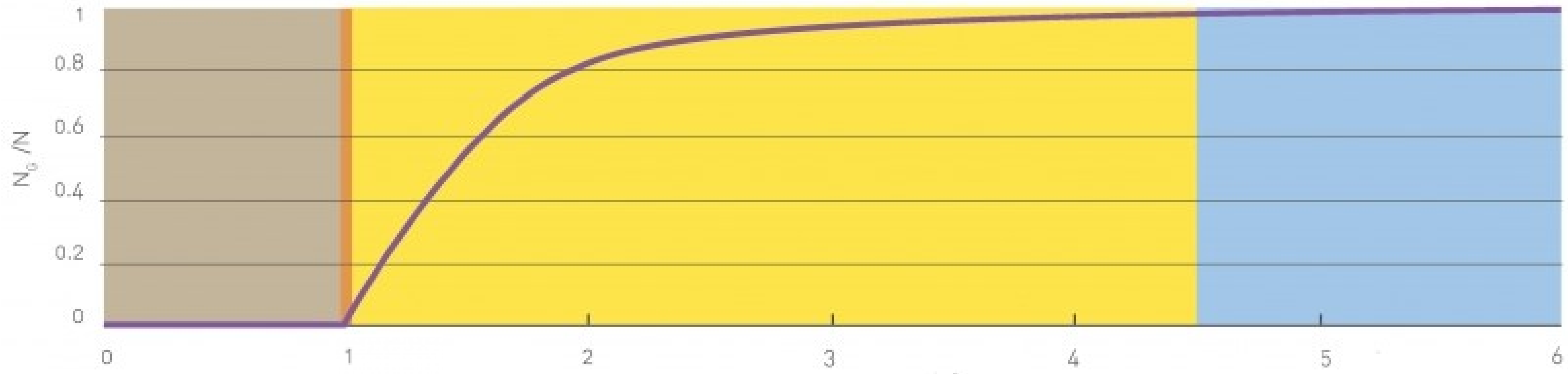
What is the critical value of $\langle k \rangle$? → 1

Video

Four domains:

- Subcritical: $\langle k \rangle > 0, p < \frac{1}{N}$
- Critical: $\langle k \rangle = 1, p = \frac{1}{N}$
- Supercritical: $\langle k \rangle$, p
- Connected: $\langle k \rangle$, p

The Evolution of a Random Network



What is the critical value of $\langle k \rangle$? → 1

Video

Four domains:

- Subcritical: $\langle k \rangle > 0,\ p < \frac{1}{N}$
- Critical: $\langle k \rangle = 1,\ p = \frac{1}{N}$
- Supercritical: $\langle k \rangle > 1,\ p > \frac{1}{N}$
- Connected: $\langle k \rangle > \ln(N),\ p > \frac{\ln(N)}{N}$

The Evolution of a Random Network

Subcritical domain:

- There is no giant component, or its relative size ($\frac{N_G}{N}$) is nearly 0.

Critical domain:

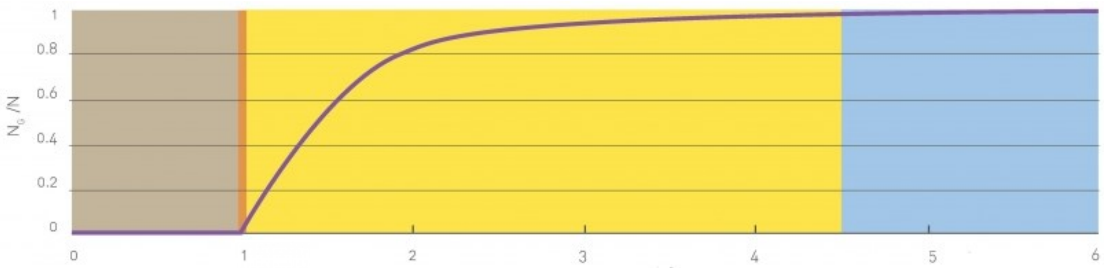
- N_G is 0 relatively to N.
- BUT!!, N_G is much larger, than $N_G \sim N^{\frac{2}{3}}$.
- In case of popularity ($7 * 10^9$) this means increase from $\sim 22{,}7$ to $\sim 3 * 10^6, \frac{N_G}{N} = 0.00043$.

Supercritical domain:

- Although there are separated components, the giant component includes most of the nodes.

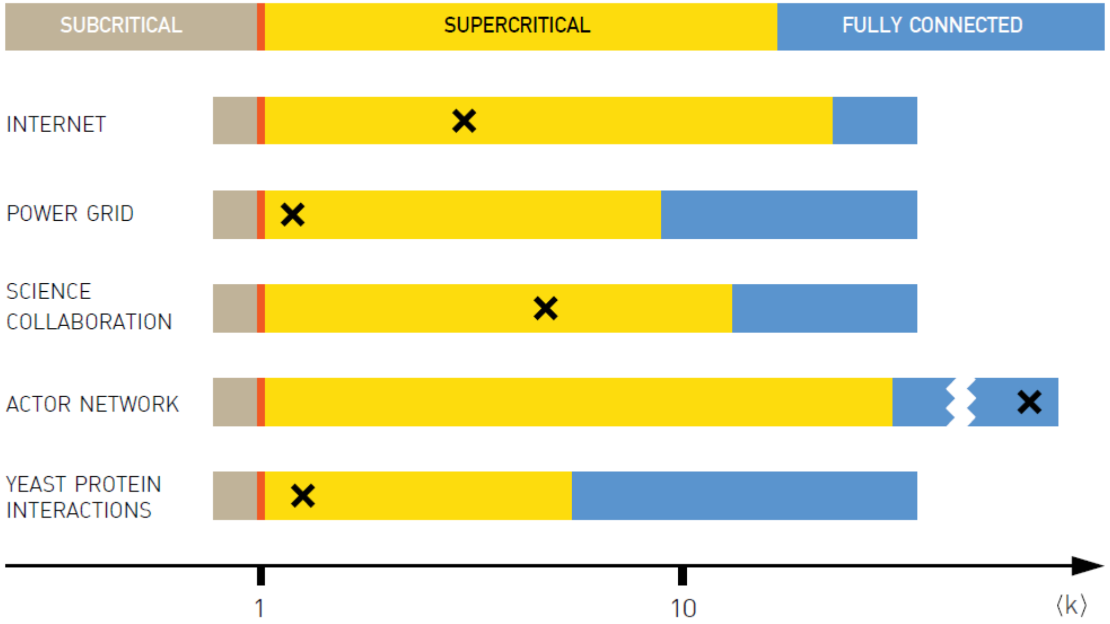
Connected domain:

- The giant component includes all of the nodes.
- The network is connected.



Real Networks are Supercritical

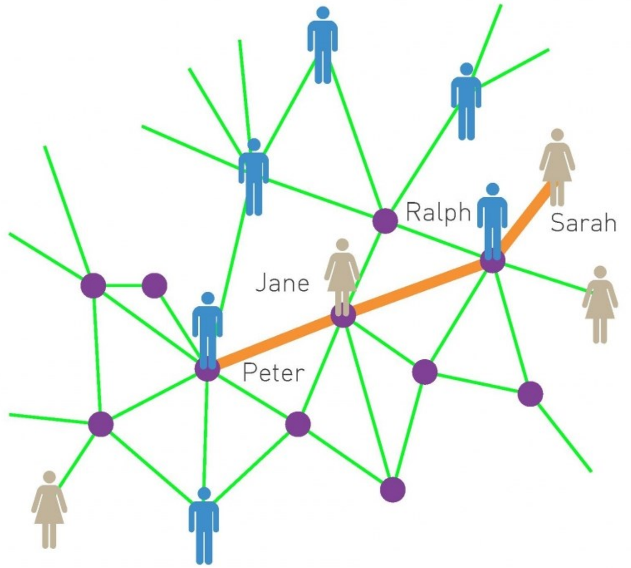
NETWORK	N	L	⟨k⟩	lnN
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	94,439	8.08	10.05
Actor Network	702,388	29,397,908	83.71	13.46
Protein Interactions	2,018	2,930	2.90	7.61



Small Worlds

Six degrees of separation

- In case of any two individuals on Earth, there is a path between them through at most six acquaintances.
- The information from Jane spreads rapidly.



An approach:

- $\langle k \rangle$ nodes at distance $d = 1$
- $\langle k \rangle^2$ nodes at distance $d = 2$
- …
- $\langle k \rangle^d$ nodes at distance d

E.g.: population

- $\langle k \rangle \cong 1000$
- 10^6 *people can be reached in two steps.*

Diameter $d_{\max}$

- $d_{\max} = \frac{\ln N}{\ln \langle k \rangle}$

Small World:

The diameter depends logarithmically on the system size.

Watts-Strogatz Model

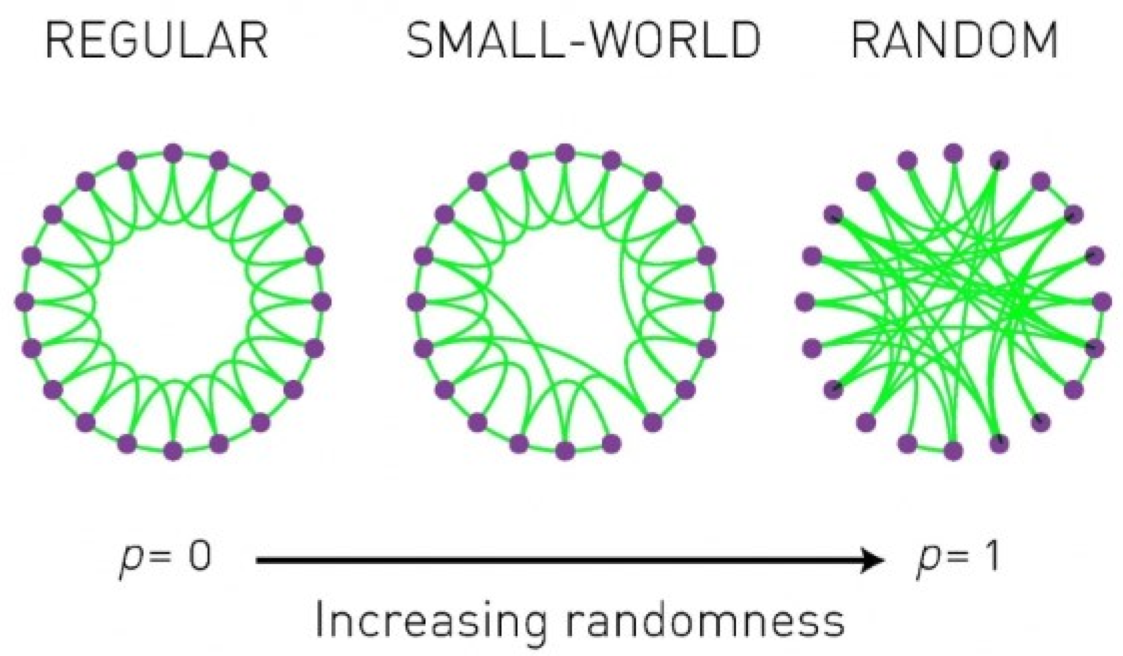
Watts-Strogatz model:

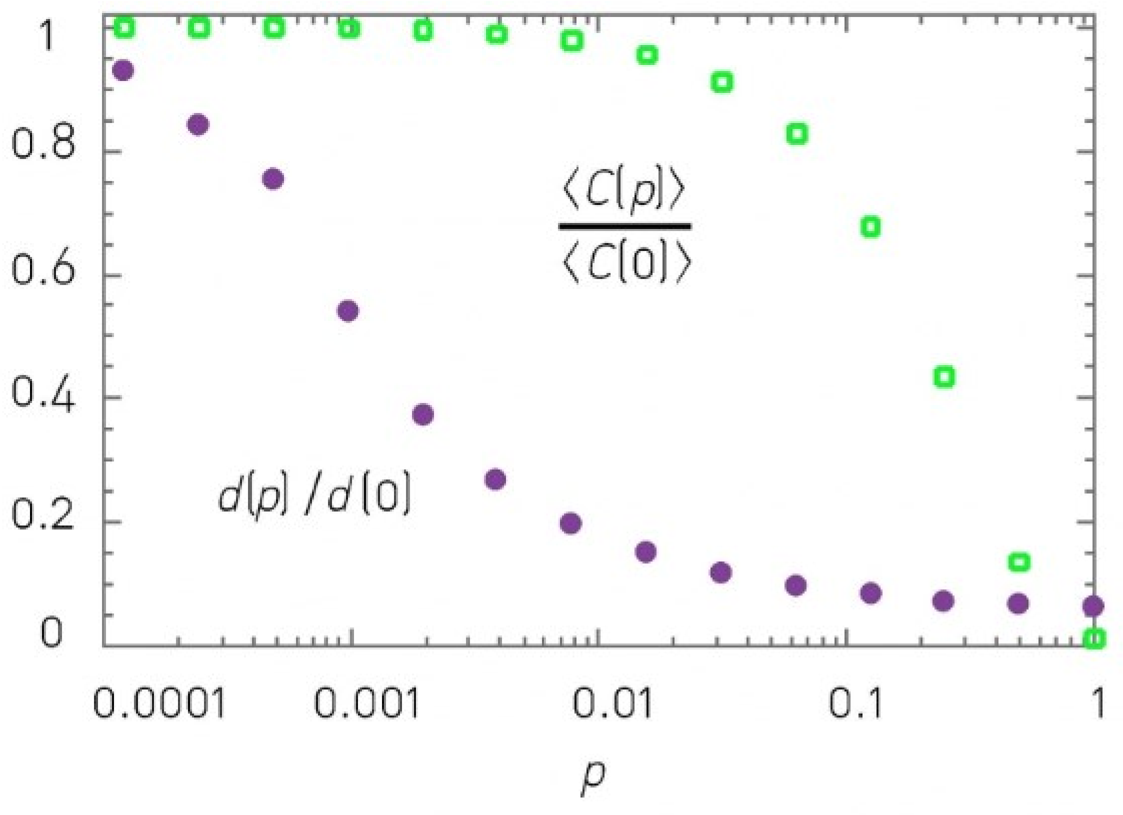
- Extension of the random network model.
- Motivated by:
  - Small World property
  - High clustering: The average clustering coefficient of real networks is much higher than expected for a random network.
- Intermediate status between regular lattice (high clustering, lack of small-world property) and random network (low clustering, but small-word property).

Algorithm:

1. Start from a ring of nodes, each node is connected to their immediate and next neighbors.
2. With probability *p* each link is rewired to a randomly chosen node.

Watts-Strogatz Model





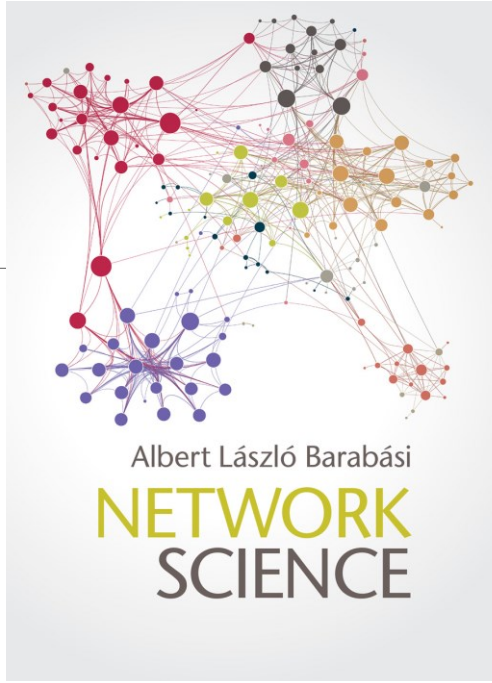
Network Analysis

04 – THE SCALE-FREE PROPERTY

Slides were created by: Agnes Vathy-Fogarassy

Network Science book (online)

Barabási, Albert-László. *Network Science*.
Cambridge University Press, 2016.



Introduction

The network of the nd.edu domain (University of Notre Dame): Video

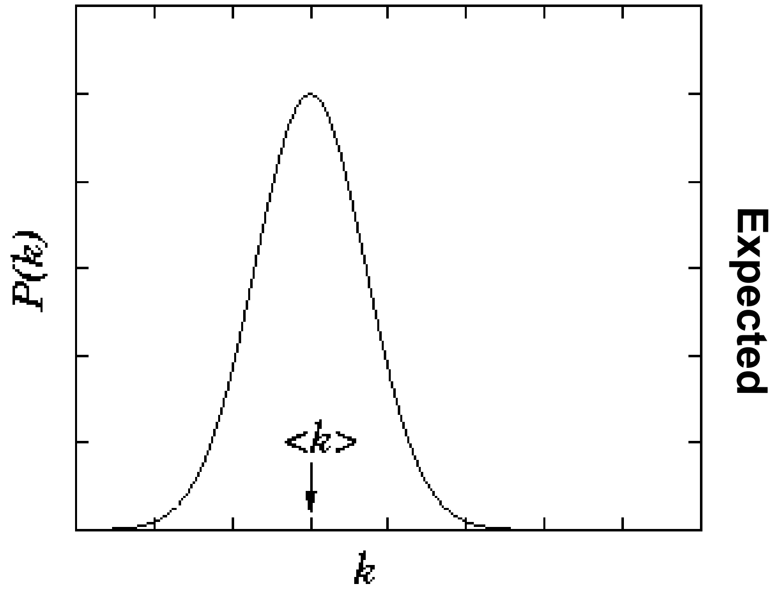
- 300,000 documents and
- 1.5 million links

Introduction

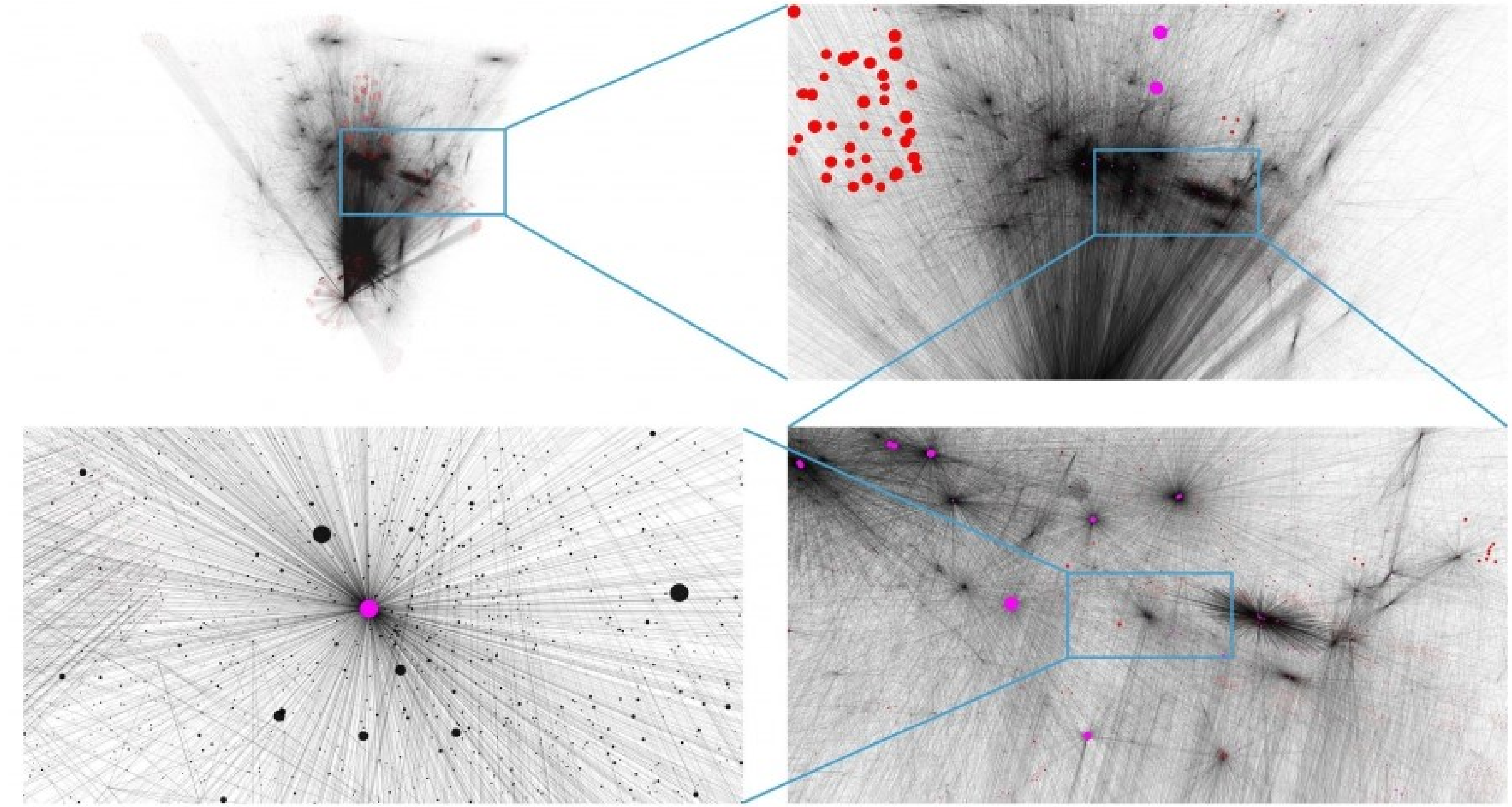
The network of the nd.edu domain (University of Notre Dame): Video

- 300,000 documents and
- 1.5 million links

With $N \approx 10^{12}$ document, WWW is the largest network humanity that has ever been built (human brain has $N \approx 10^{11}$ neurons)



Introduction



Expected

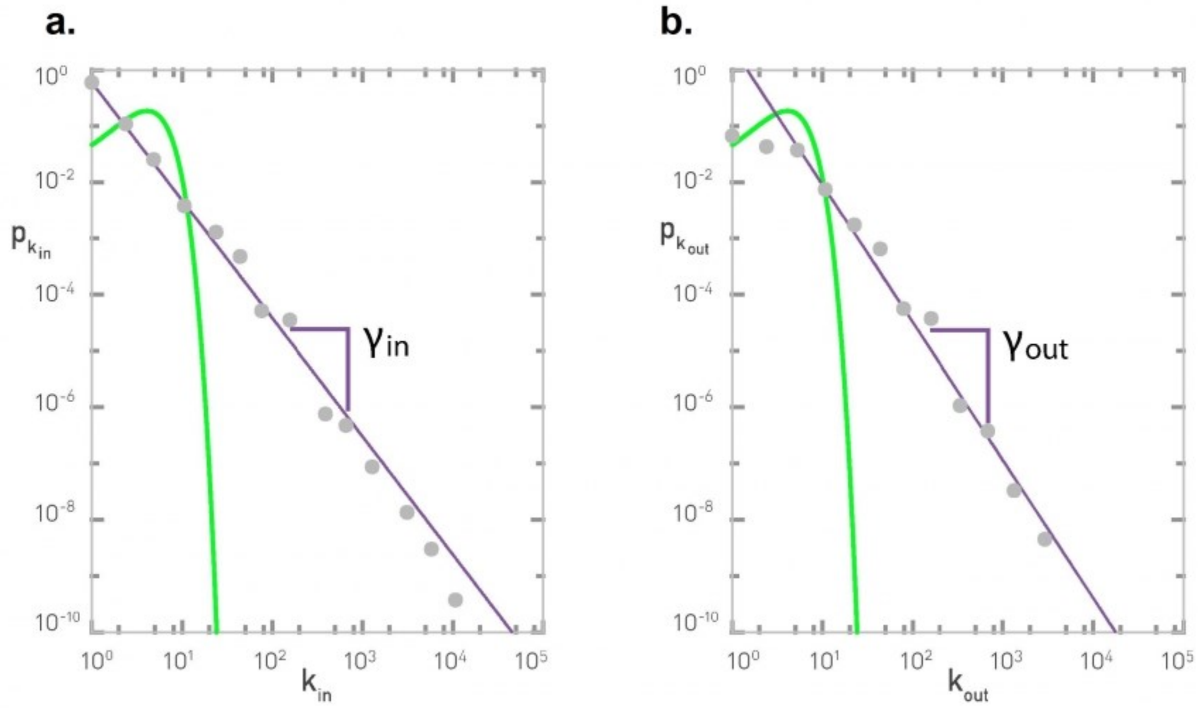
Power Laws and Scale-Free Networks

The real degree distribution of WWW

On a Log-Log scale the data form an almost straight line.

Degree follows Power Law, not Poisson distribution.

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



In Figure:

- $\gamma_{in} = 2.1$
- $\gamma_{out} = 2.45$
- $p_{k_{in}} \sim k^{-\gamma_{in}}$
- $p_{k_{out}} \sim k^{-\gamma_{out}}$

Power Laws and Scale-Free Networks

Definition:

- *A scale-free network is a network whose degree distribution follows a power law.*

Power Laws and Scale-Free Networks

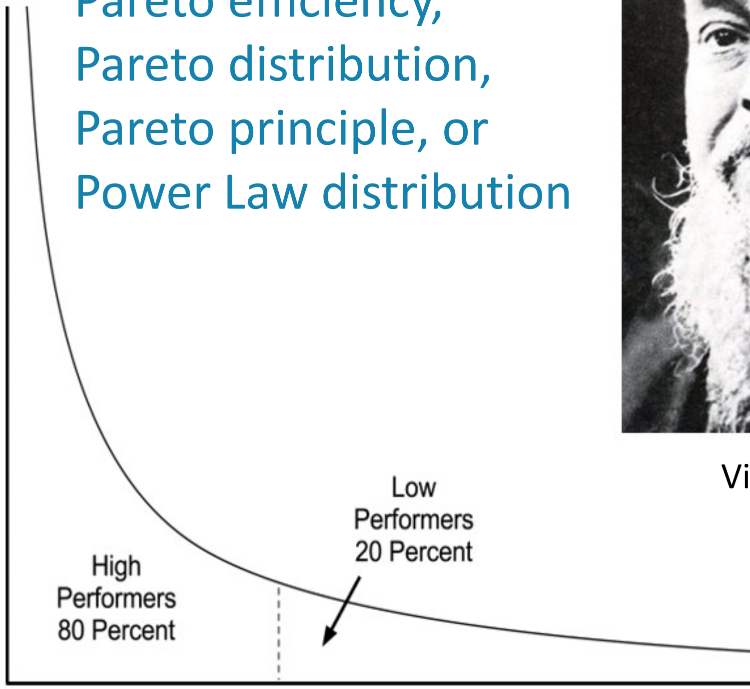
Definition:

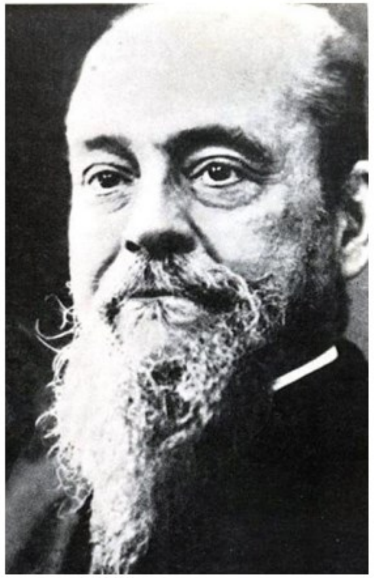
- *A scale-free network is a network whose degree distribution follows a power law.*

Discrete form:

- $p_k = Ck^{-\gamma}$

Pareto efficiency,
Pareto distribution,
Pareto principle, or
Power Law distribution





Vilfredo Federico
Damaso Pareto
(1848 – 1923)

Power Laws and Scale-Free Networks

Definition:

- *A scale-free network is a network whose degree distribution follows a power law.*

Discrete form:

- $p_k = Ck^{-\gamma}$

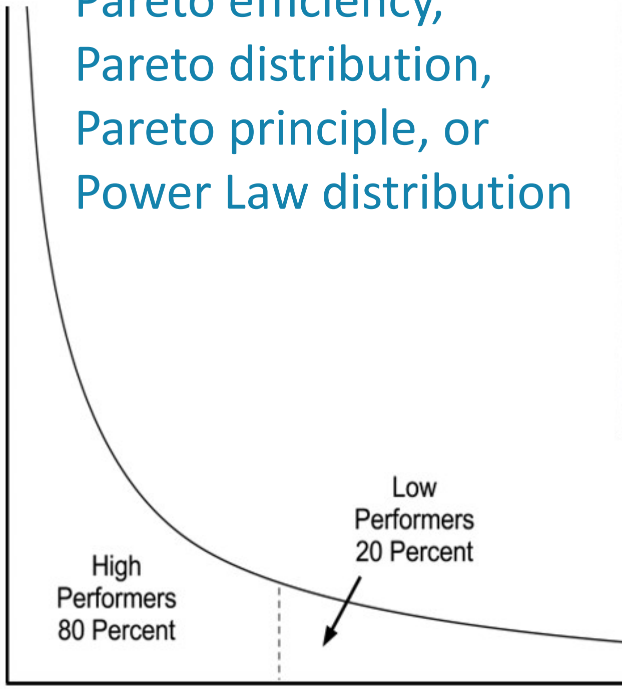
C is determined by the normalization condition:

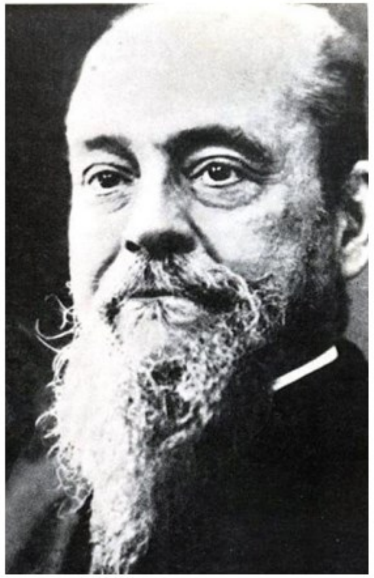
- $\sum_{k=1}^{\infty} p_k = 1$
- $C \sum_{k=1}^{\infty} k^{-\gamma} = 1 \rightarrow C = \frac{1}{\sum_{k=1}^{\infty} k^{-\gamma}} = \frac{1}{\xi(\gamma)}$

Thus,

- $p_k = \frac{k^{-\gamma}}{\xi(\gamma)}$
- BUT! It diverges at p_0, so we need to determine p_0 separately.

Pareto efficiency,
Pareto distribution,
Pareto principle, or
Power Law distribution



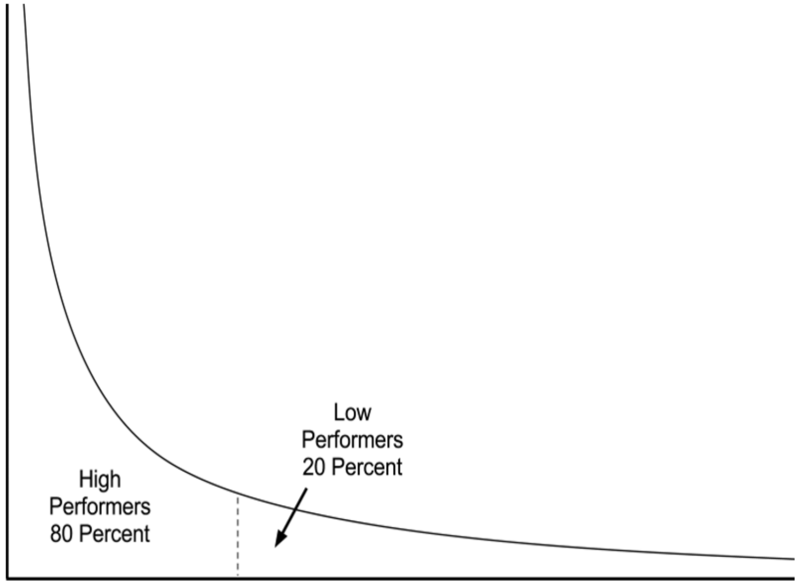


Vilfredo Federico
Damaso Pareto
(1848 – 1923)

Hubs

The main difference between Power Law and Poisson distribution:

- The tail.



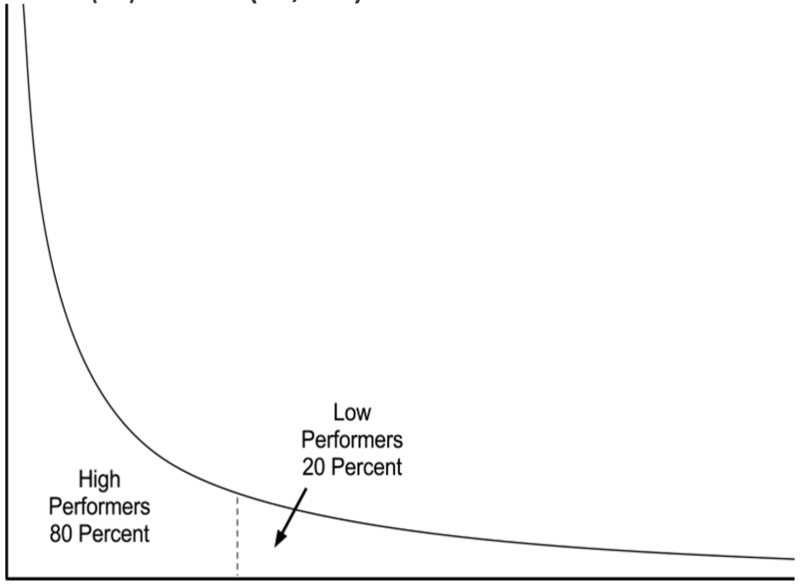
Hubs

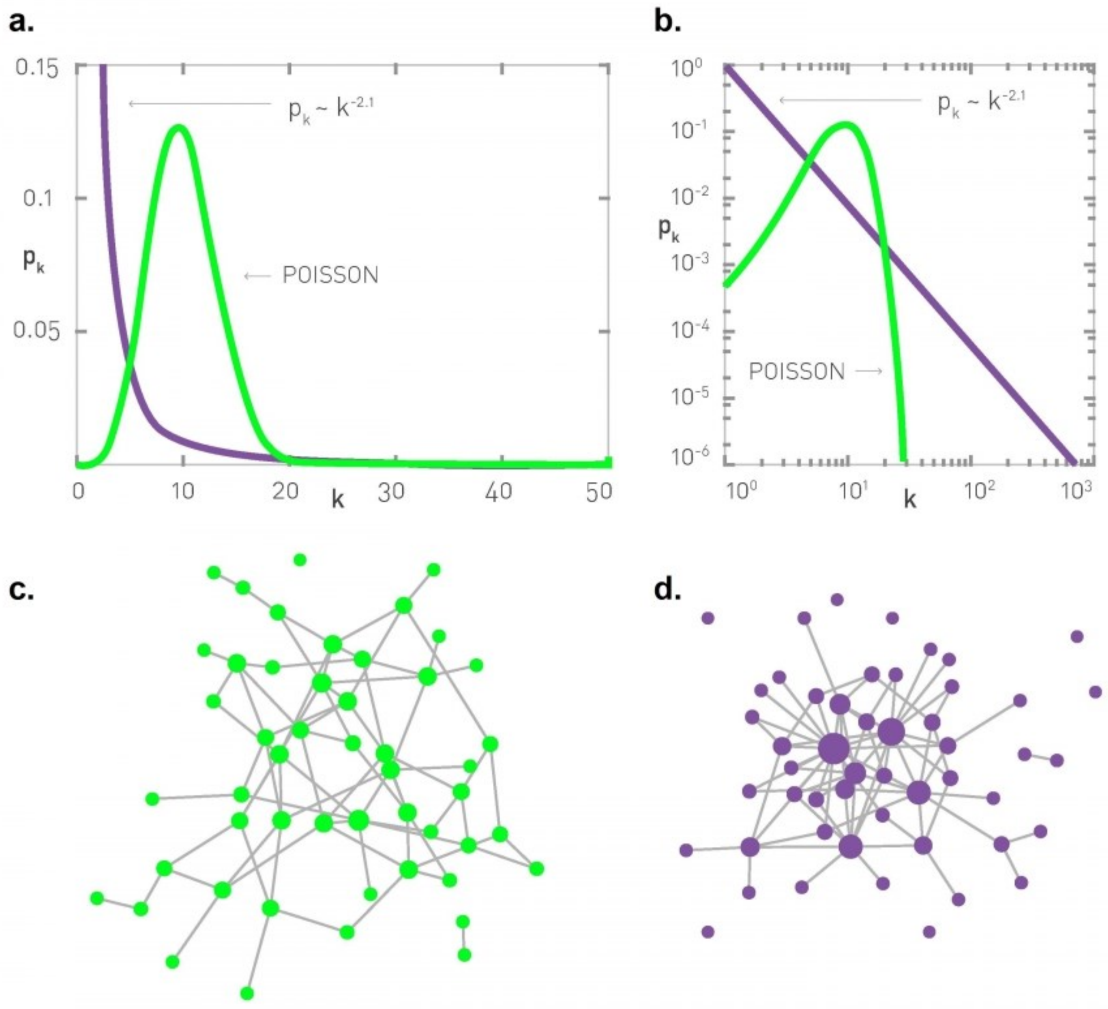
The main difference between Power Law and Poisson distribution:

- The tail.

Parameters:

- $\gamma = 2.1$
- $\langle k \rangle = 11$ (a., b.)
- $\langle k \rangle = 3$ (c., d.)





The Largest Hub

Network sizes:

- Web: $N \approx 10^{12}$
- Population: $N \approx 7 \times 10^9$
- Human gene network: $N \approx 2 \times 10^4$
- E.coli metabolic network: $N \approx 10^3$

How big is k_{max}?

The Largest Hub

Network sizes:

- Web: $N \approx 10^{12}$
- Population: $N \approx 7 \times 10^9$
- Human gene network: $N \approx 2 \times 10^4$
- E.coli metabolic network: $N \approx 10^3$

How big is k_{max}?

- Complete network:
- Random network:
- Scale-free network:

The Largest Hub

Network sizes:

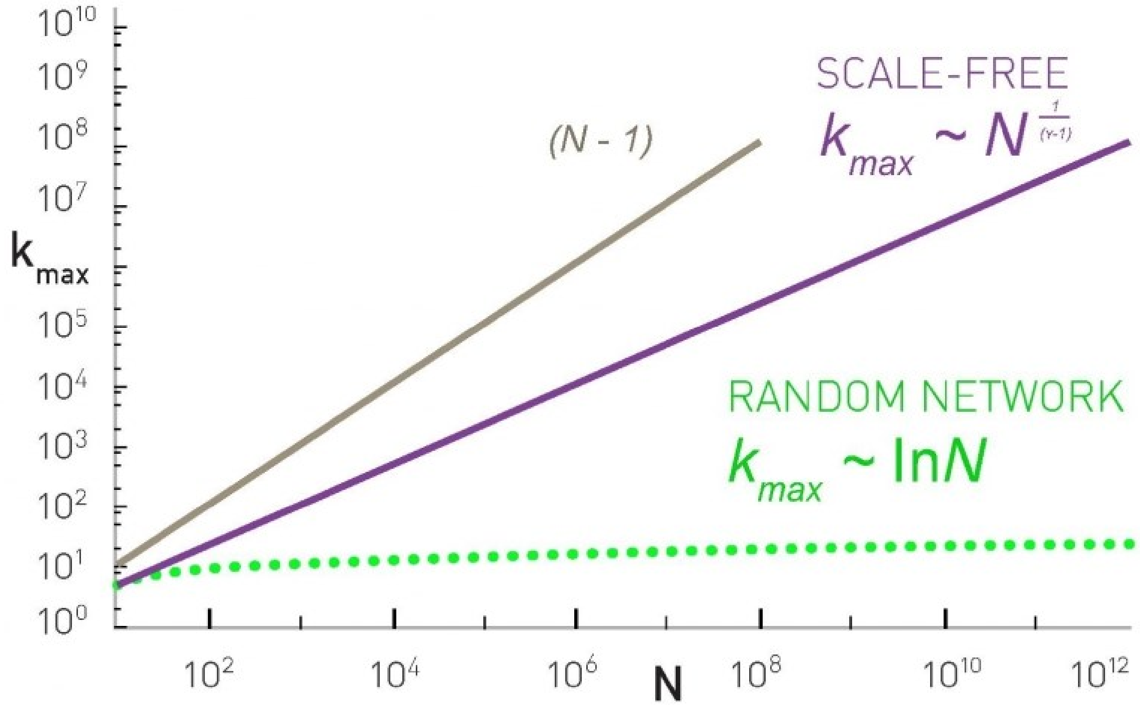
- Web: $N \approx 10^{12}$
- Population: $N \approx 7 \times 10^9$
- Human gene network: $N \approx 2 \times 10^4$
- E.coli metabolic network: $N \approx 10^3$

How big is k_{max}?

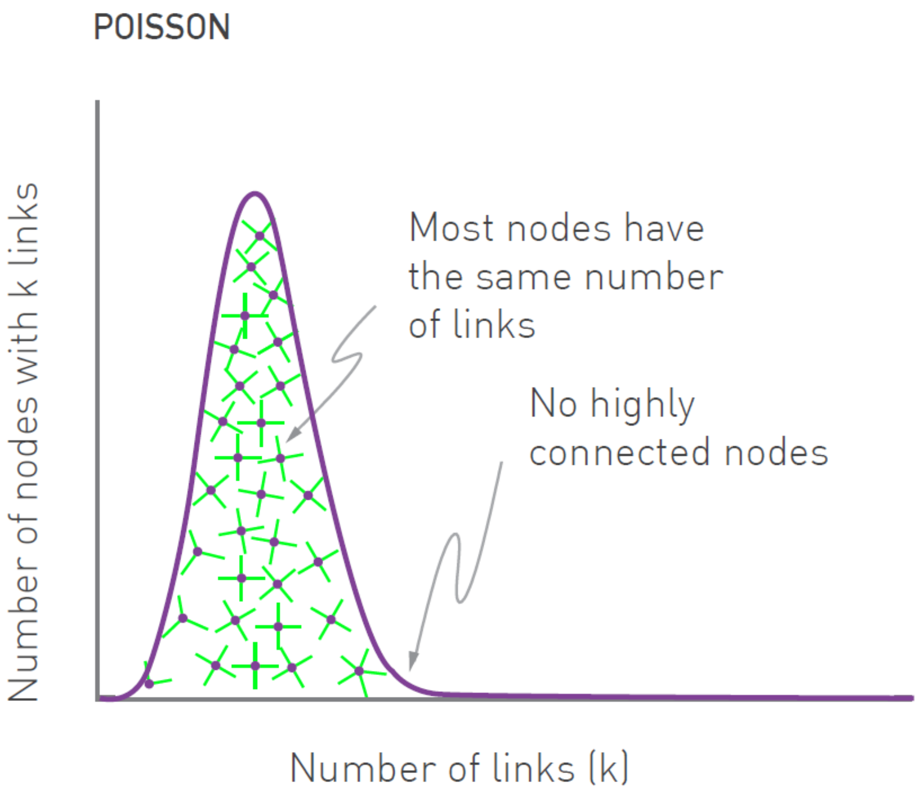
- Complete network: $k_{max} = N - 1$
- Random network: $k_{max} \sim \ln N$
- Scale-free network: $k_{max} \sim N^{\frac{1}{\gamma - 1}}$

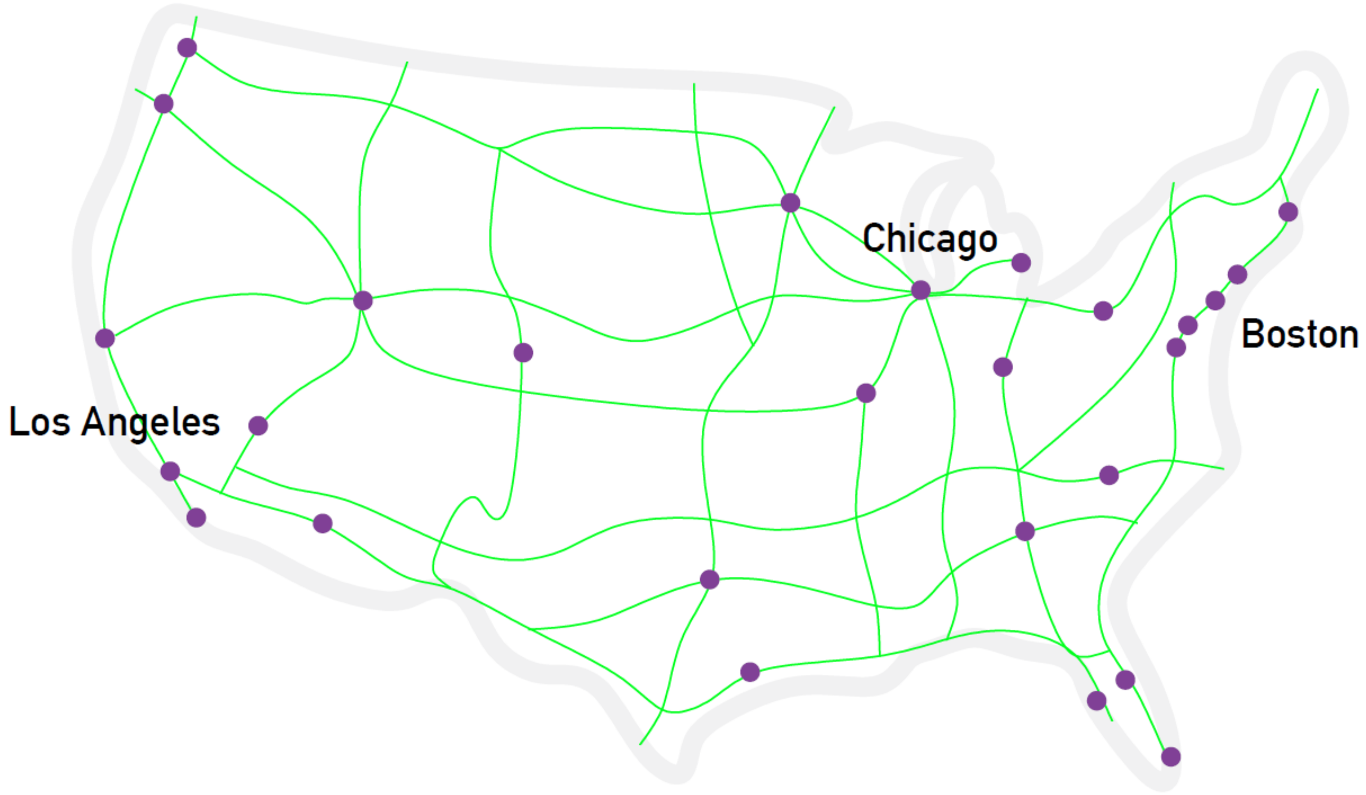
In figure:

- $\langle k \rangle = 3$
- $\gamma = 2.5$

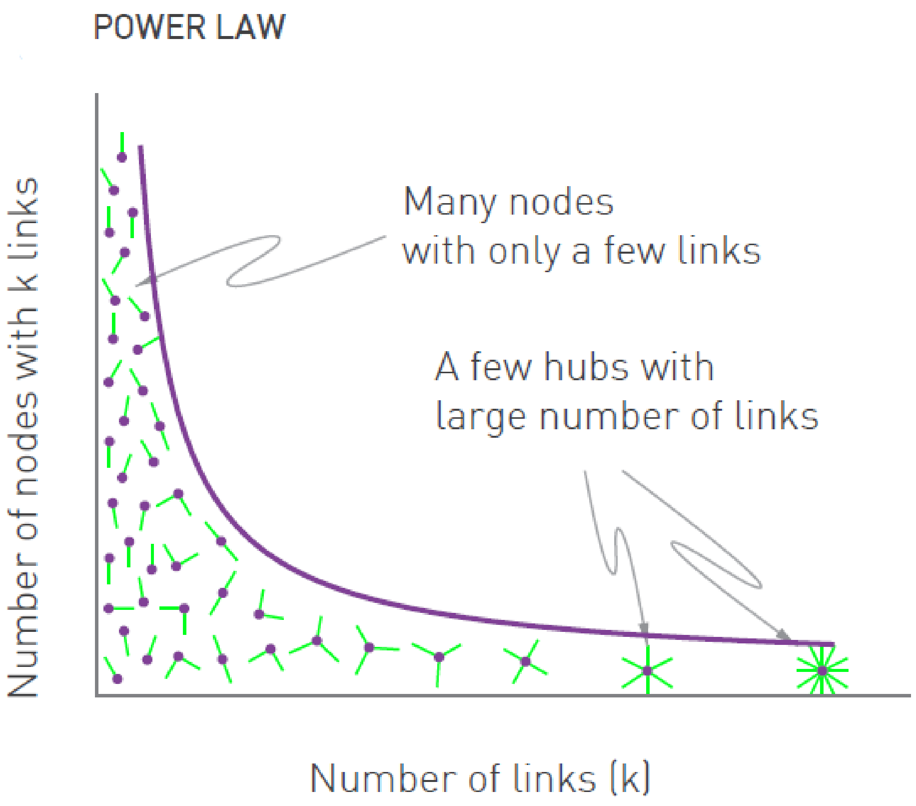


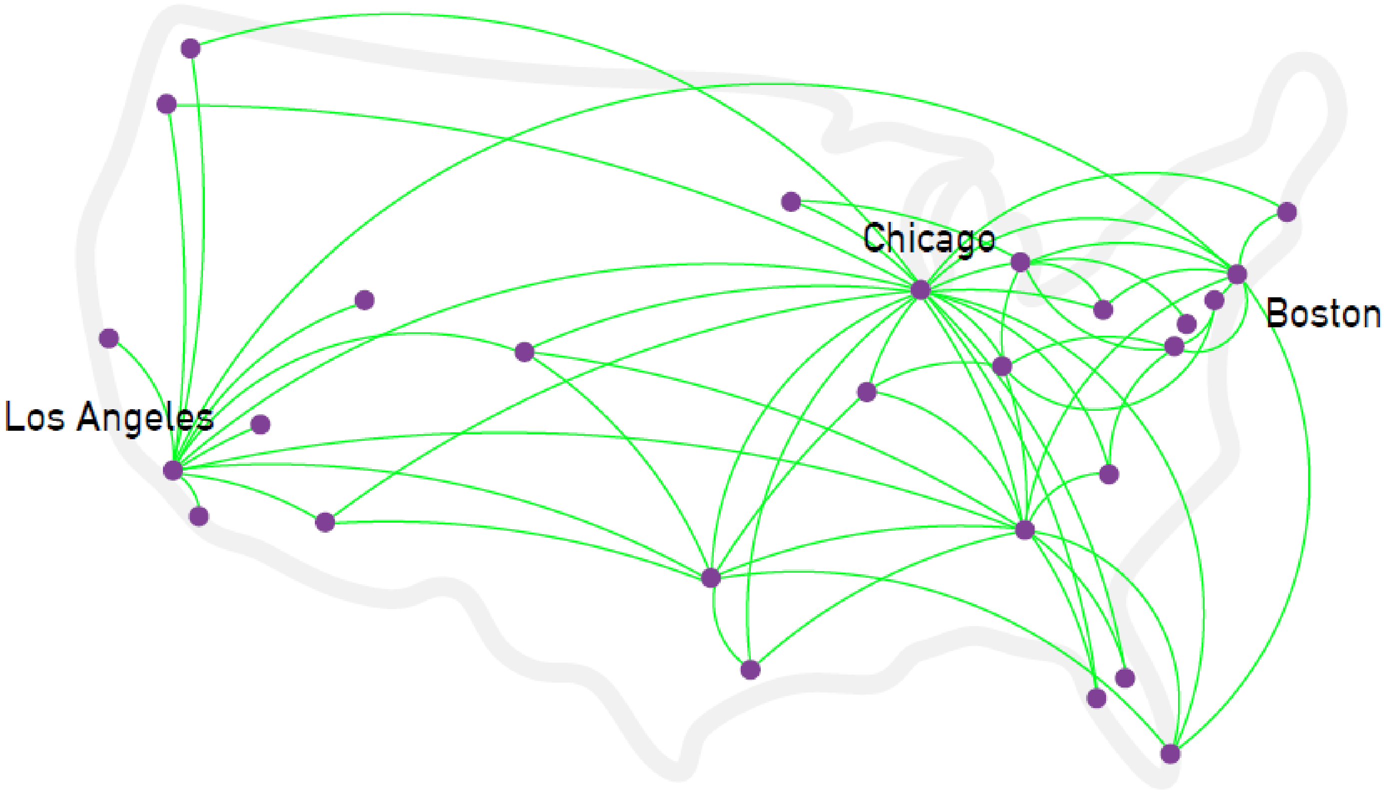
Example


POISSON
Most nodes have the same number of links
No highly connected nodes
Number of nodes with k links
Number of links (k)


Chicago
Boston
Los Angeles

Example


POWER LAW
Many nodes
with only a few links
A few hubs with
large number of links
Number of nodes with k links
Number of links (k)


Chicago
Boston
Los Angeles

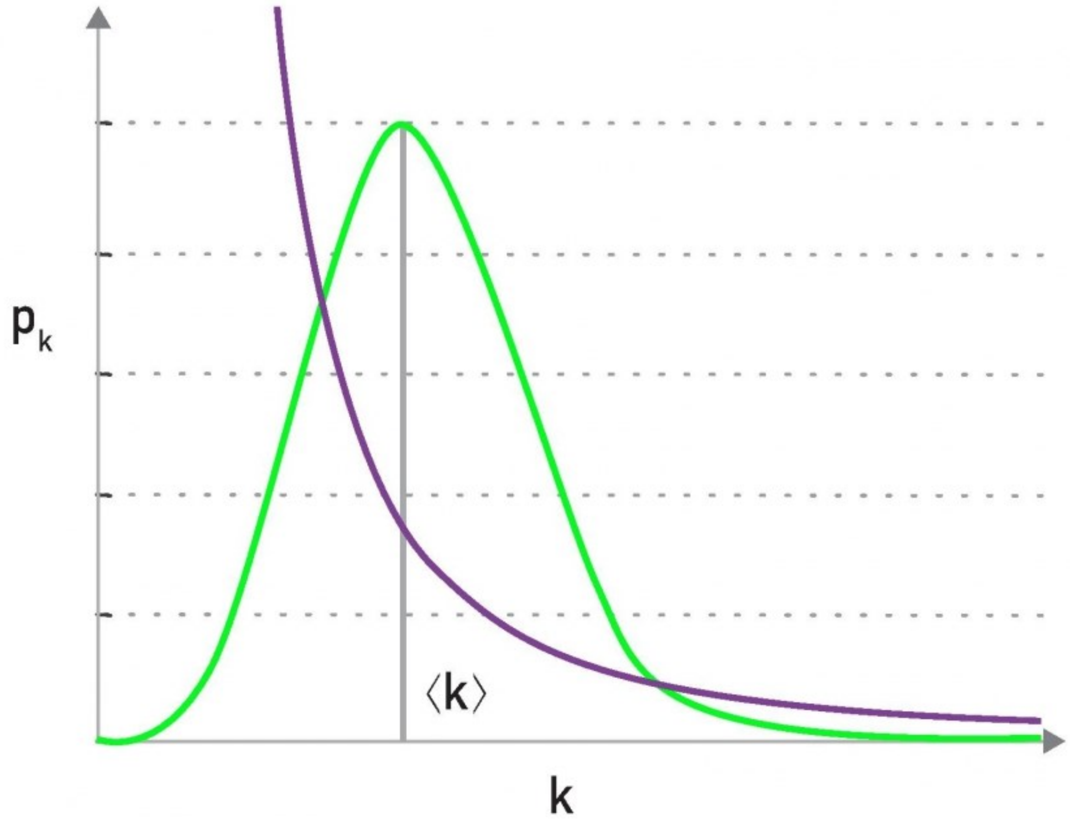
The Meaning of Scale-Free

Random Networks have a scale

- Due to Poisson distribution $\sigma_k = \langle k \rangle^{\frac{1}{2}}, \sigma < \langle k \rangle$
- Degrees of nodes are in the range $k = \langle k \rangle \pm \langle k \rangle^{\frac{1}{2}}$
- $\langle k \rangle$ serves a „scale" for random networks

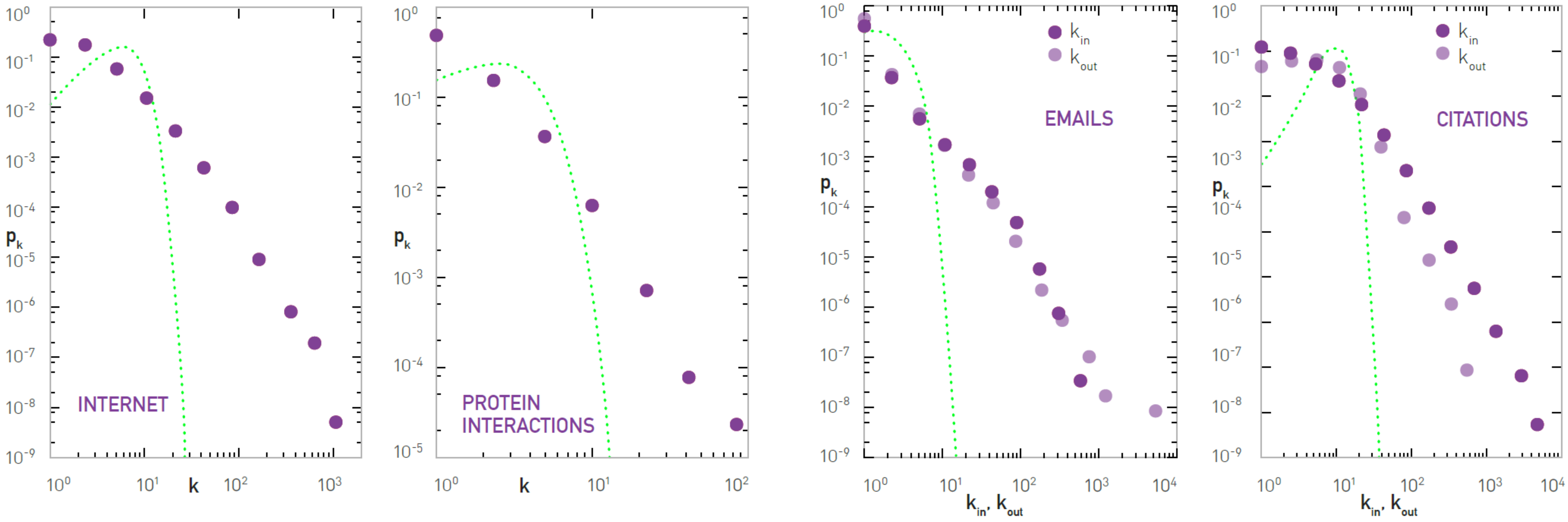
Scale-free Networks have no scale

- Network with a Power-law distribution with $\gamma < 3$
- Deviation from the average can be arbitrary large
- A randomly selected node can be:
  - tiny
  - huge



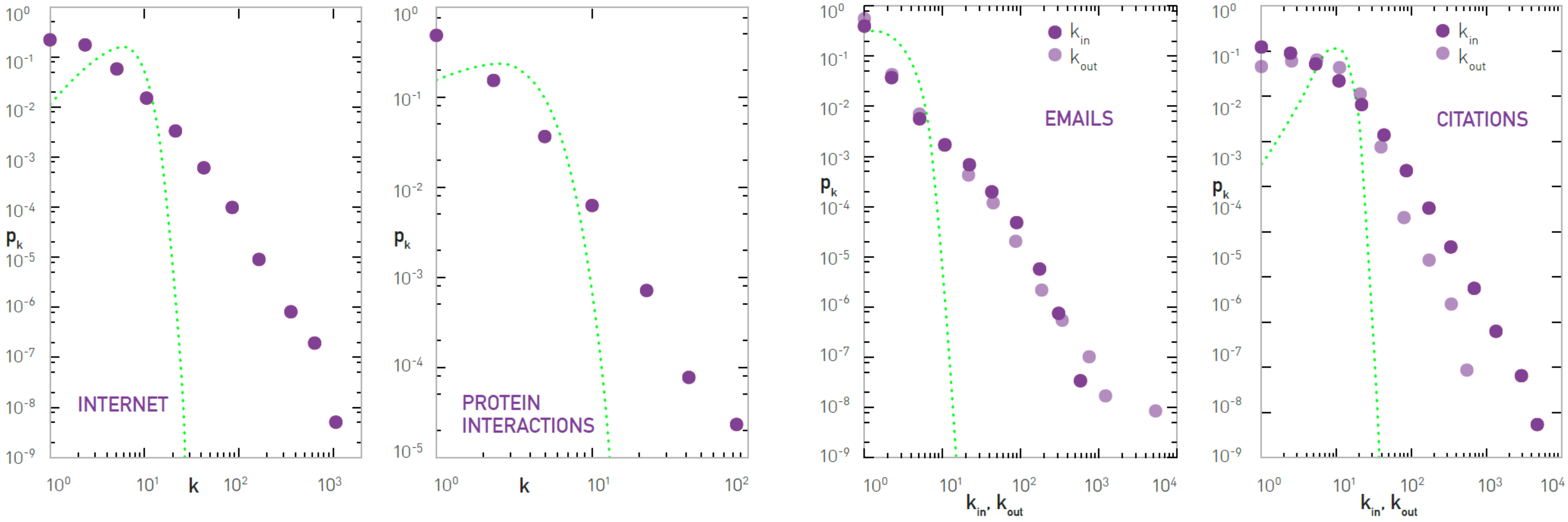
How can we determine γ?

Degree distribution of the real networks:



How can we determine γ?

Degree distribution of the real networks:



The degree exponent can be obtained by fitting a straight line to p_k on a log-log plot.

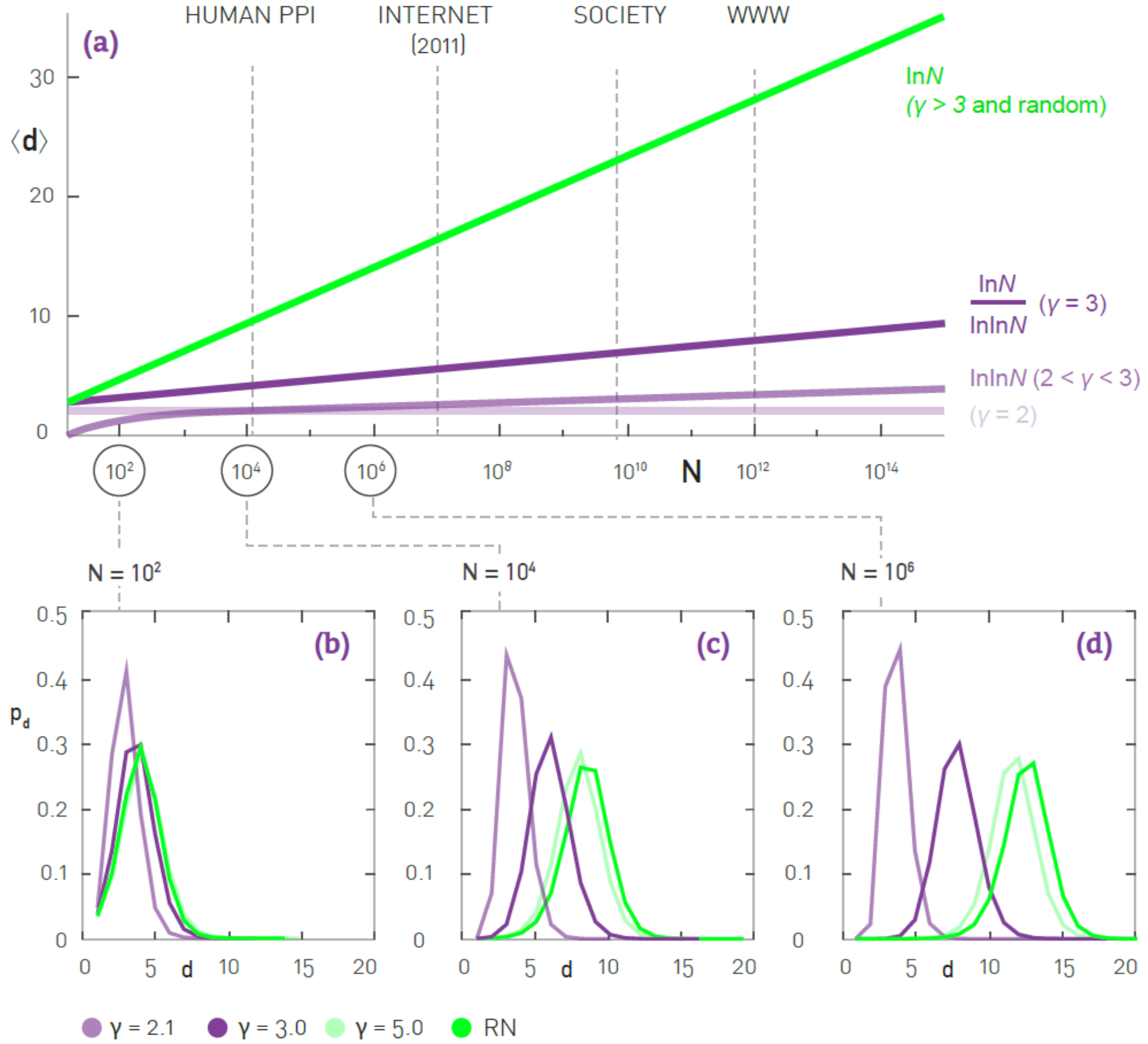
How can we determine γ?

Anomalous Regime $(\gamma = 2)$
- $k_{max} \approx N$
- $\langle d \rangle \sim const$

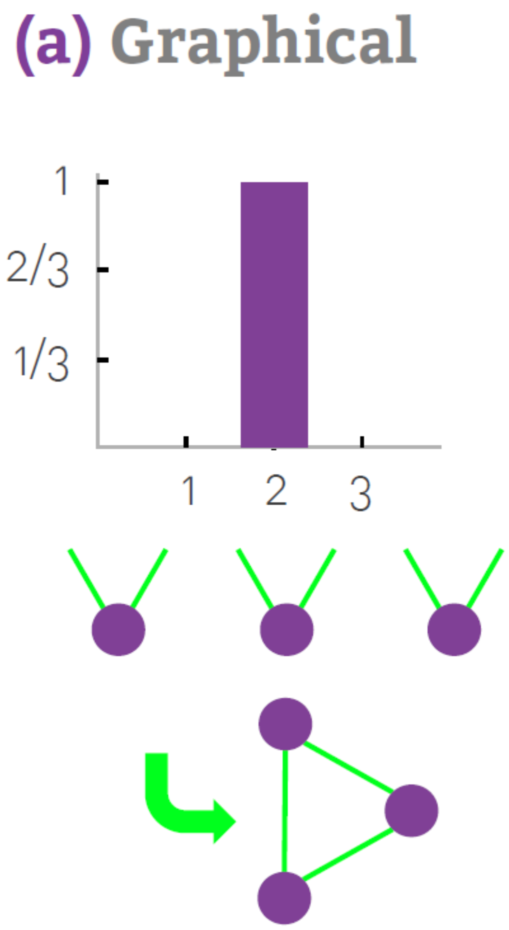
Ultra-Small World $(2 < \gamma < 3)$
- $\langle d \rangle \sim lnlnN$
- *Example*: Population: $N = 7 \times 10^9$
  - $lnN = 22.66$
  - $lnlnN = 3.12$

Critical Point $(\gamma = 3)$
- $\langle d \rangle \sim \frac{lnN}{lnlnN}$

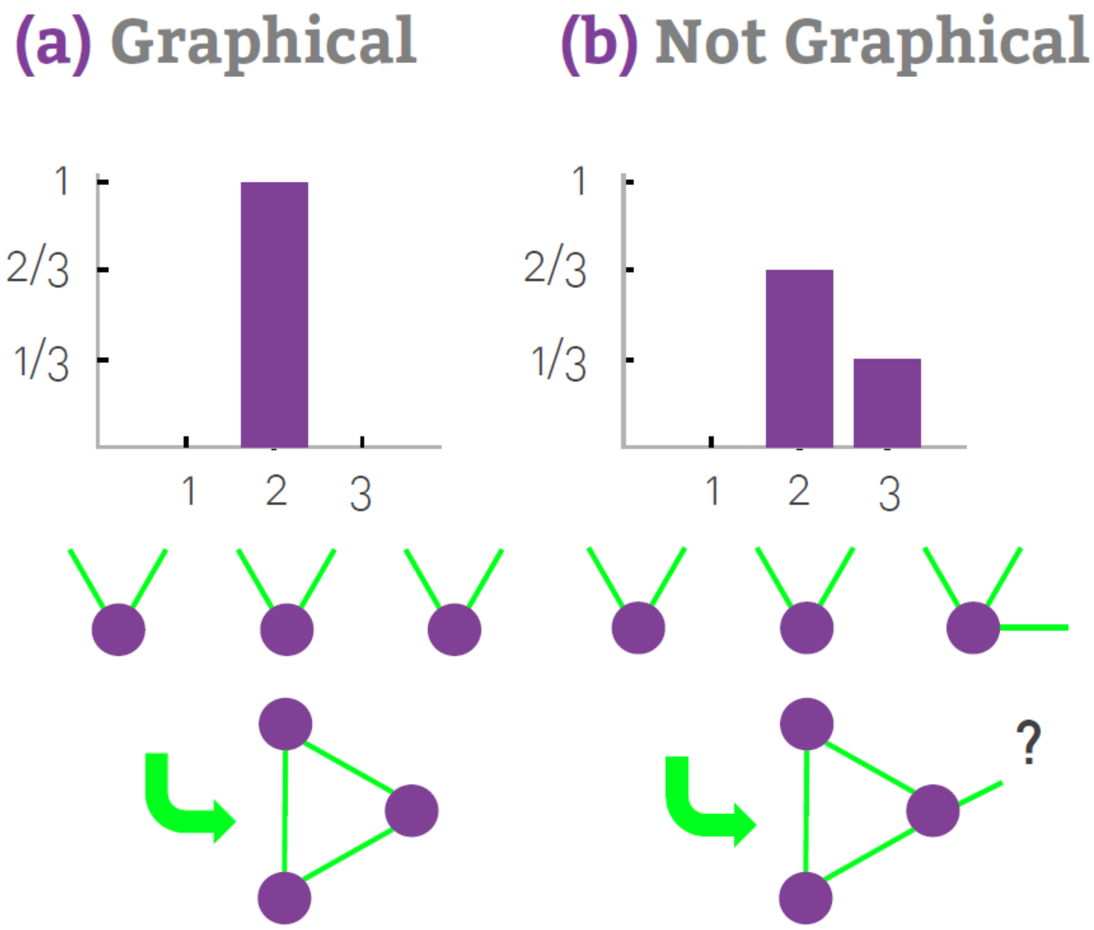
Small World $(\gamma > 3)$
- $\langle d \rangle \sim lnN$



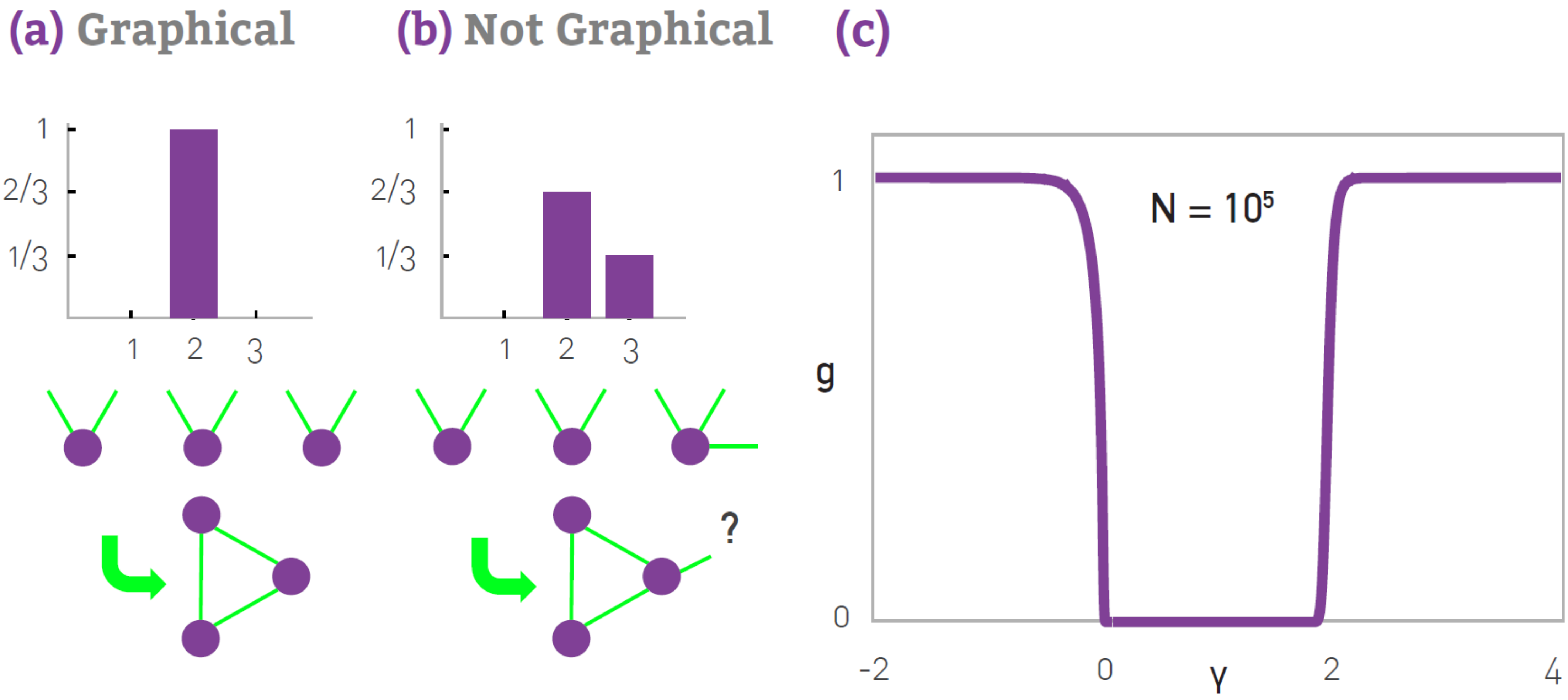
Why Scale-free networks with $\gamma < 2$ do not exist?



Why Scale-free networks with $\gamma < 2$ do not exist?



Why Scale-free networks with $\gamma < 2$ do not exist?



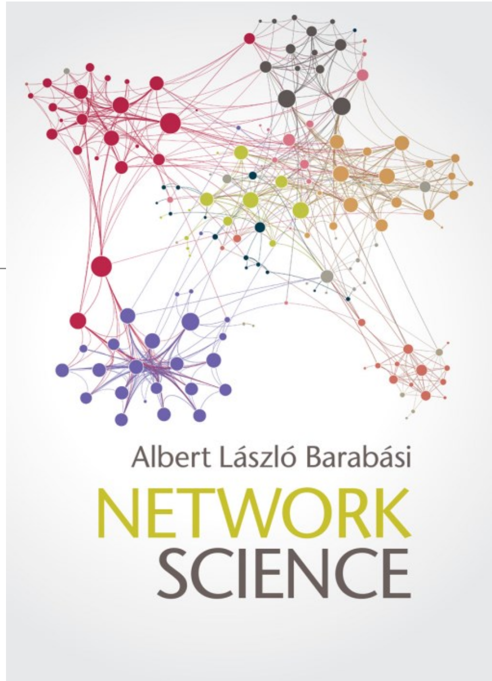
Network Analysis

05 – THE BARABÁSI-ALBERT MODEL

Slides were created by: Daniel Leitold

Network Science book (online)

Barabási, Albert-László. *Network Science*.
Cambridge University Press, 2016.



Introduction

Why do very different systems as the WWW and the cell both have scale-free architecture?

- The *nodes* of the cellular network are metabolites or proteins, while the nodes of the WWW are documents, representing information without a physical manifestation.
- The *links* within the cells are chemical reactions and binding interactions, while the links of the WWW are URLs, or small segments of computer codes.
- The *history* of these two systems could not be more different: The cellular network is shaped by 4 billion years of evolution, while the WWW is less than three decade old.
- The *purpose* of the metabolic network is to produce the chemical components the cell needs to stay alive, while the purpose of the WWW is information access and delivery.

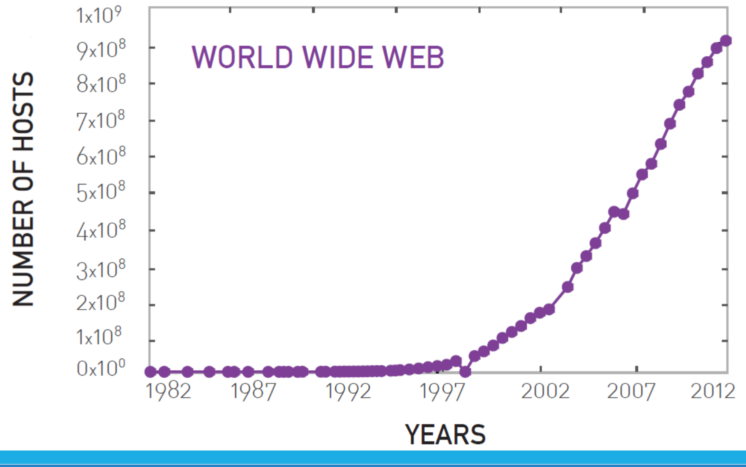
Why does the random network model of Erdős and Rényi fail to reproduce the hubs and the power laws observed in real networks?

We need to understand the mechanism responsible for the emergence of the scale-free property.

Growth and Preferential Attachment I

Why are hubs and power laws absent in random networks?

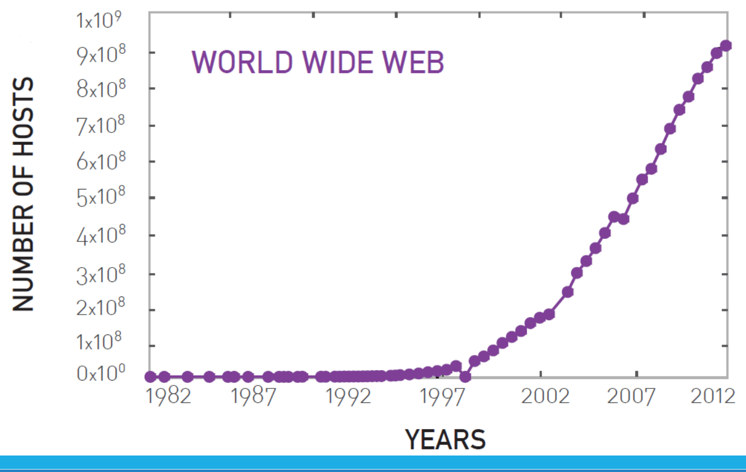
- In random network N is a fixed number.
- But! Networks expand through the addition of new nodes.
- Examples:
  - In 1991 the WWW had a single node, today the Web has over a trillion (10^{12}) documents.

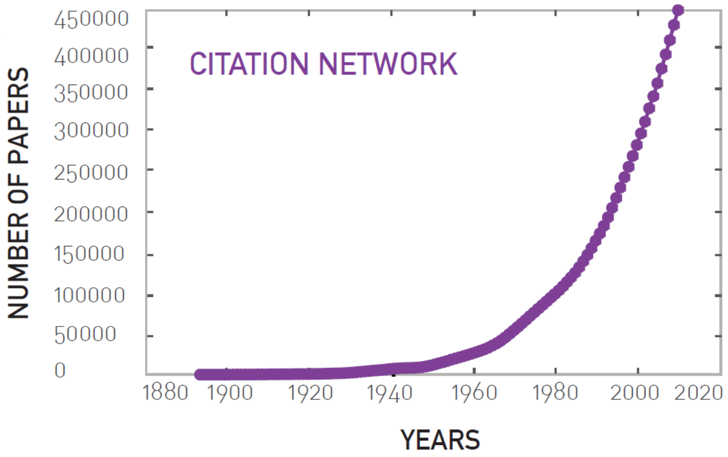


Growth and Preferential Attachment I

Why are hubs and power laws absent in random networks?

- In random network N is a fixed number.
- But! Networks expand through the addition of new nodes.
- Examples:
  - In 1991 the WWW had a single node, today the Web has over a trillion (10^{12}) documents.
  - The collaboration and the citation network continually expands through the publication of new research papers.

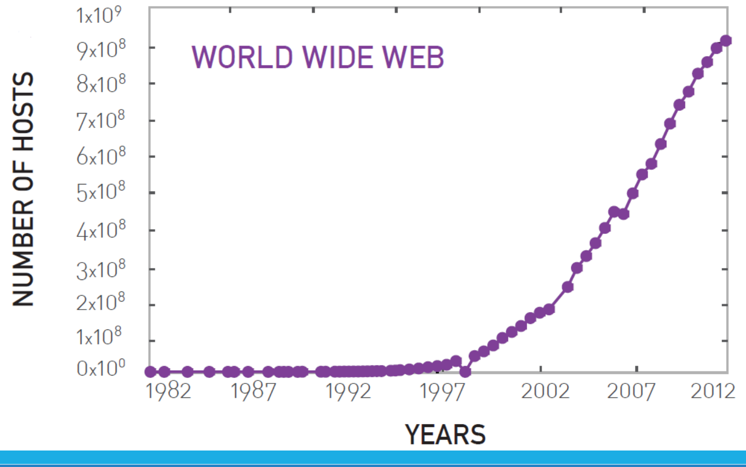


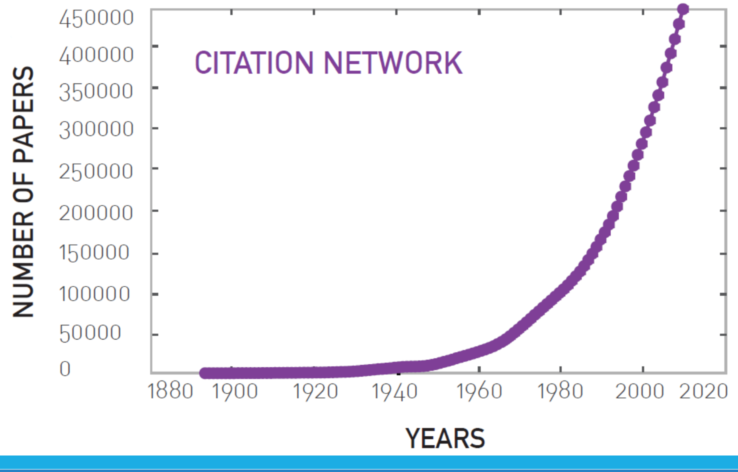


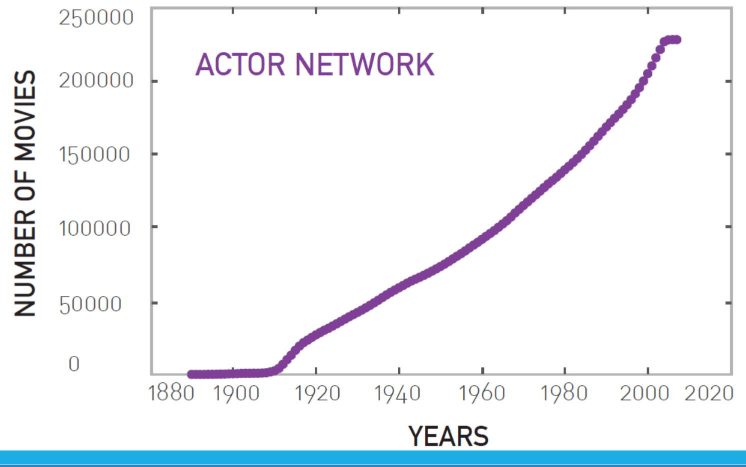
Growth and Preferential Attachment I

Why are hubs and power laws absent in random networks?

- In random network N is a fixed number.
- But! Networks expand through the addition of new nodes.
- Examples:
    - In 1991 the WWW had a single node, today the Web has over a trillion (10^{12}) documents.
    - The collaboration and the citation network continually expands through the publication of new research papers.
    - The actor network continues to expand through the release of new movies.
    - The number of genes has grown from a few to the over 20,000 genes that have appeared in a human cell over four billion years.



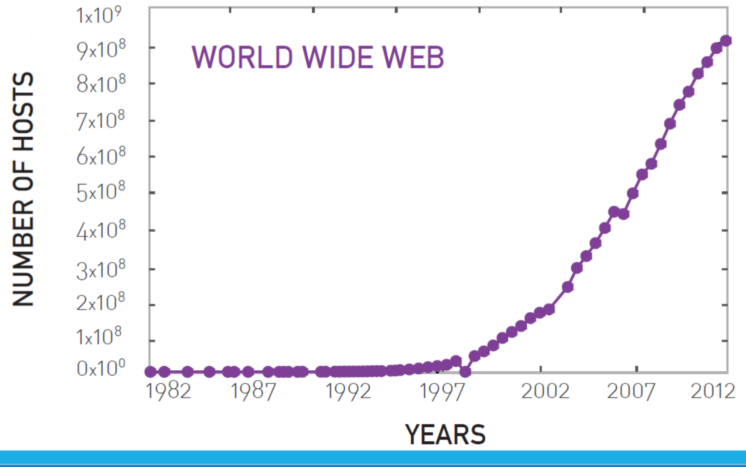


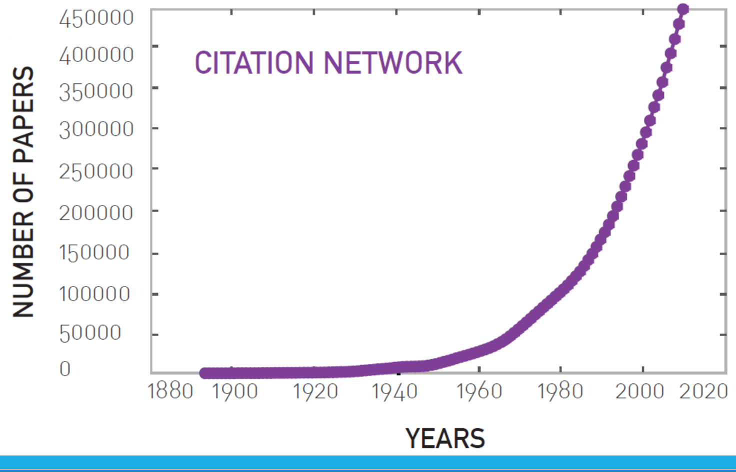


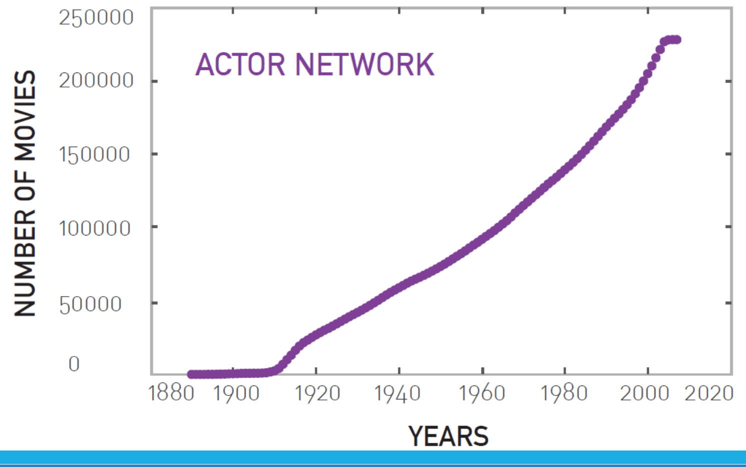
Growth and Preferential Attachment I

Why are hubs and power laws absent in random networks?

- In random network N is a fixed number.
- But! Networks expand through the addition of new nodes.
- Examples:
  - In 1991 the WWW had a single node, today the Web has over a trillion (10^{12}) documents.
  - The collaboration and the citation network continually expands through the publication of new research papers.
  - The actor network continues to expand through the release of new movies.
  - The number of genes has grown from a few to the over 20,000 genes that have appeared in a human cell over four billion years.
  - We need to use a dynamic model instead of a static one!







Growth and Preferential Attachment II

Why are hubs and power laws absent in random networks?

- The random network model selects the interaction partners randomly.
- But! In most of the real networks, new nodes prefer one with more connections.
- Examples:
  - We all know Google and Facebook, but we rarely encounter the billions of less-prominent nodes that populate the Web. We are more likely to link to a high-degree node than to a node with only few links.
  - The more cited is a paper, the more likely that we have heard about it. As we cite what we have read, our citations are biased towards the more cited publications, representing the high-degree nodes of the citation network.
  - The more movies an actor has played in, the more familiar is a casting director with his/her skills. Hence, the higher the degree of an actor in the actor network is, the higher are the chances that he/she will be considered for a new role.

In summary, the two differences:

- Growth
- Preferential attachment

The Barabási-Albert Model

Initializing:

- A network with m_0 nodes.
- Add links randomly to the network, until each node has at least one link.

Growth:

- Add a new node to the network,
- With $m \leq m_o$ new links such that,

Preferential Attachment:

- The probability to connect node i is: $\prod(k_i) = \frac{k_i}{\sum_j k_j}$

The Barabási-Albert Model

Initialising:

- A network with m_0 nodes.
- Add links randomly to the network, until each node has at least one link.

Growth:

- Add a new node to the network,
- With $m \leq m_o$ new links such that,

Preferential Attachment:

- The probability to connect node i is: $\prod(k_i) = \frac{k_i}{\sum_j k_j}$

Example:

- $m_0 = 2$

The Barabási-Albert Model

Initialising:

- A network with m_0 nodes.
- Add links randomly to the network, until each node has at least one link.

Growth:

- Add a new node to the network,
- With $m \leq m_o$ new links such that,

Preferential Attachment:

- The probability to connect node i is: $\prod(k_i) = \frac{k_i}{\sum_j k_j}$

Example:

- $m_0 = 2$

The Barabási-Albert Model

Initialising:

- A network with m_0 nodes.
- Add links randomly to the network, until each node has at least one link.

Growth:

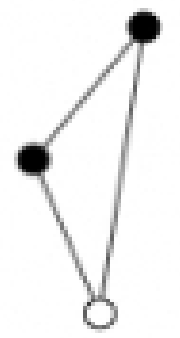
- Add a new node to the network,
- With $m \leq m_o$ new links such that,

Preferential Attachment:

- The probability to connect node i is: $\prod(k_i) = \frac{k_i}{\sum_j k_j}$

Example:

- $m_0 = 2$
- $m = 2$



The Barabási-Albert Model

Initialising:

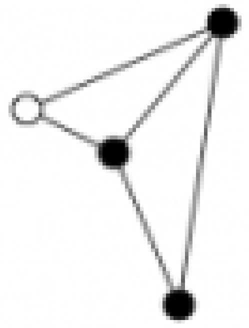
- A network with m_0 nodes.
- Add links randomly to the network, until each node has at least one link.

Growth:

- Add a new node to the network,
- With $m \leq m_o$ new links such that,

Preferential Attachment:

- The probability to connect node i is: $\prod(k_i) = \frac{k_i}{\sum_j k_j}$



Example:

- $m_0 = 2$
- $m = 2$

The Barabási-Albert Model

Initialising:

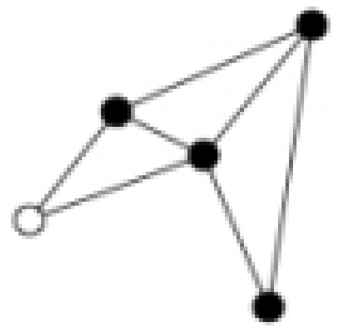
- A network with m_0 nodes.
- Add links randomly to the network, until each node has at least one link.

Growth:

- Add a new node to the network,
- With $m \leq m_o$ new links such that,

Preferential Attachment:

- The probability to connect node i is: $\prod(k_i) = \frac{k_i}{\sum_j k_j}$



Example:

- $m_0 = 2$
- $m = 2$

The Barabási-Albert Model

Initialising:

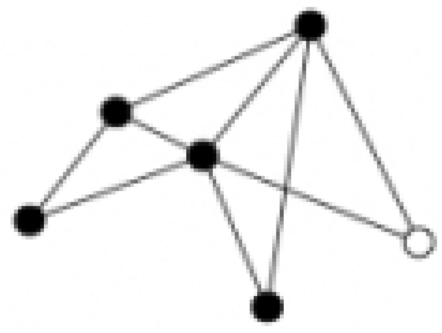
- A network with m_0 nodes.
- Add links randomly to the network, until each node has at least one link.

Growth:

- Add a new node to the network,
- With $m \leq m_o$ new links such that,

Preferential Attachment:

- The probability to connect node i is: $\prod(k_i) = \frac{k_i}{\sum_j k_j}$



Example:

- $m_0 = 2$
- $m = 2$

The Barabási-Albert Model

Initialising:

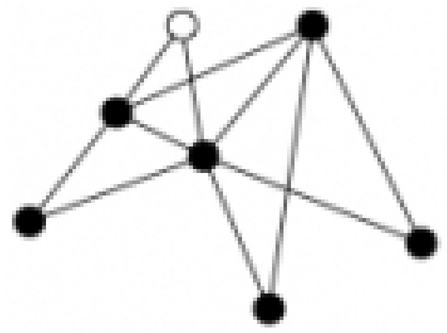
- A network with m_0 nodes.
- Add links randomly to the network, until each node has at least one link.

Growth:

- Add a new node to the network,
- With $m \leq m_o$ new links such that,

Preferential Attachment:

- The probability to connect node i is: $\prod(k_i) = \frac{k_i}{\sum_j k_j}$



Example:

- $m_0 = 2$
- $m = 2$
- Video

Network Analysis

06 - PRACTICE

Slides were created by: Agnes Vathy-Fogarassy

Cytoscape

Cytoscape webpage

Cytoscape 3.5.0 User Manual

NetworkAnalyzer Online Help

Welcome screen

With Empty network
- Create a network from scratch

From Network File

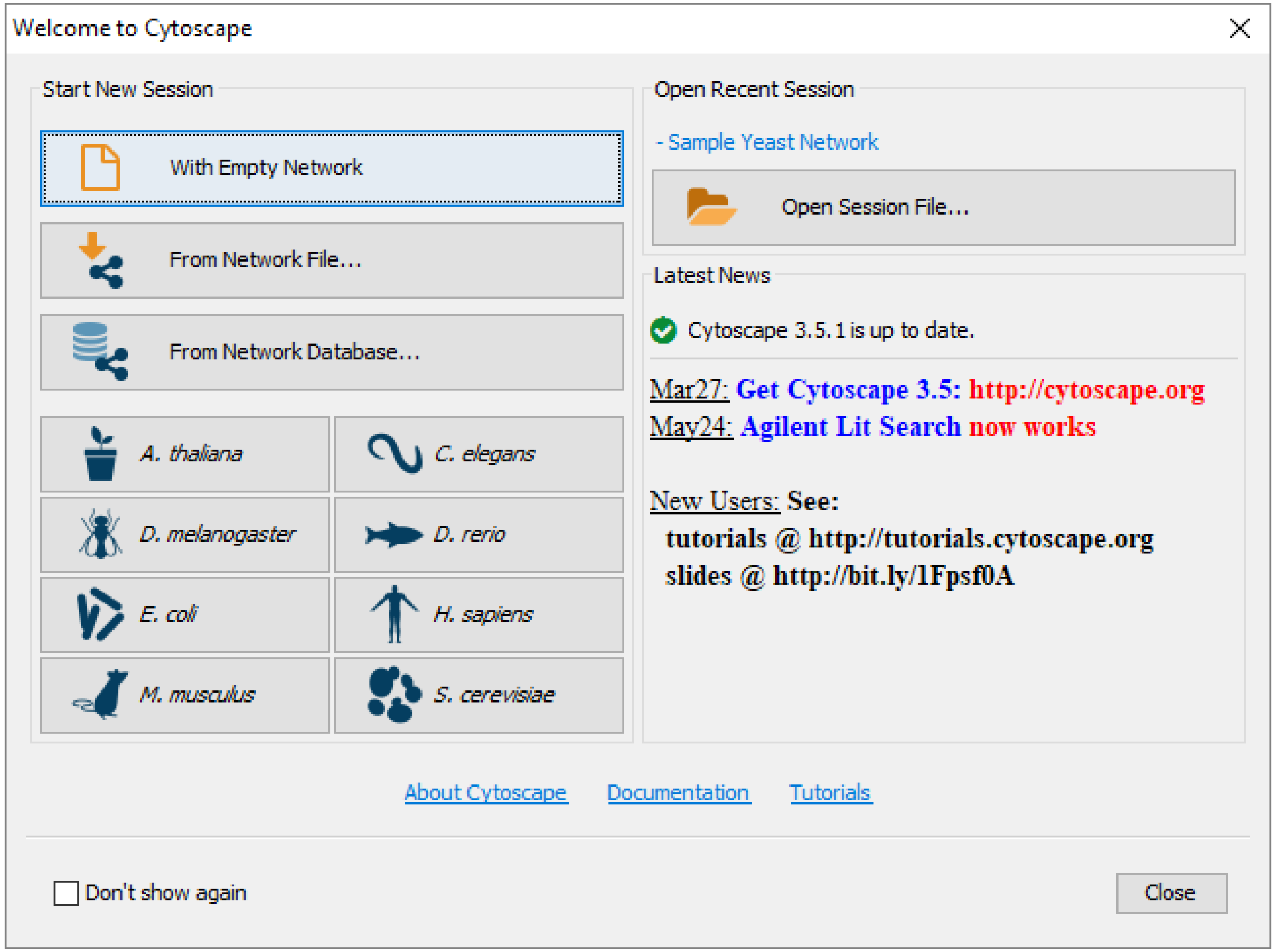
From Network Database
- Public network data
- Search for gene, e.g. “BRCA2”

Organism Networks

Open Recent Session

Open Session File
- List of opened sessions

Help -> Show Welcome Screen…



Basics

Load network

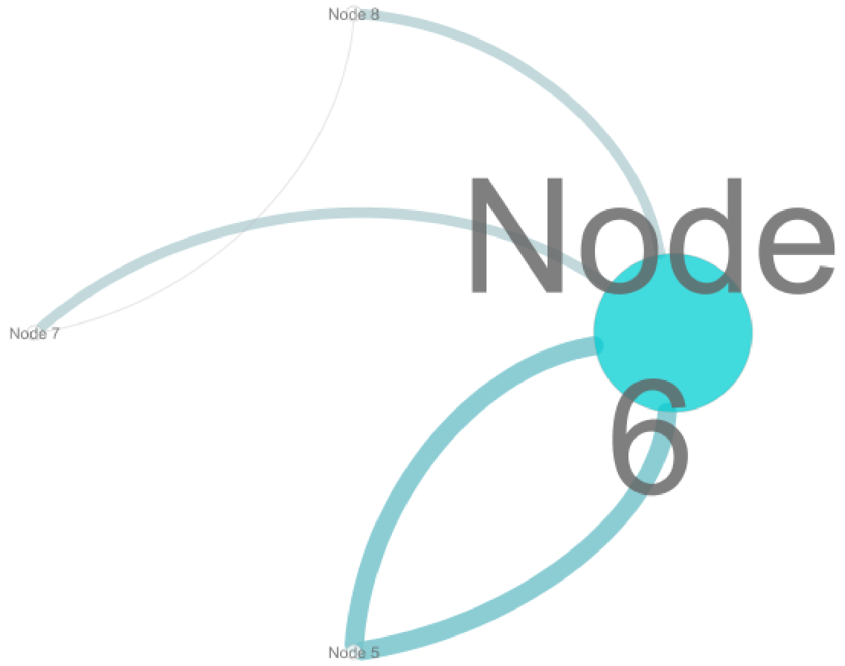
Create network

- Add Node
- Add Edge

Change Style

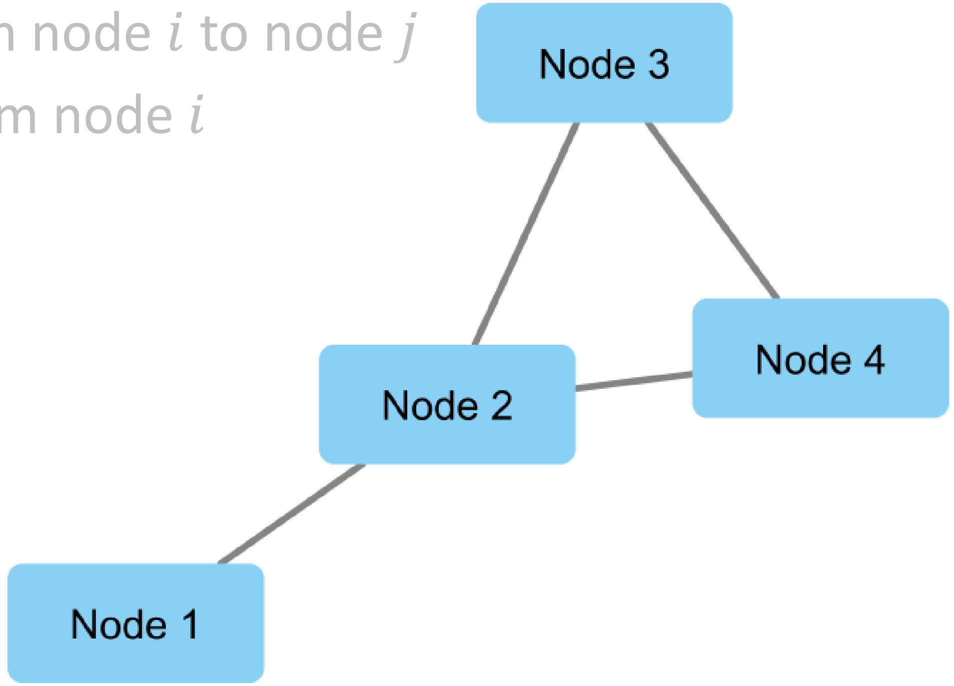
Change Layout

Select



NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i



Generated measures:

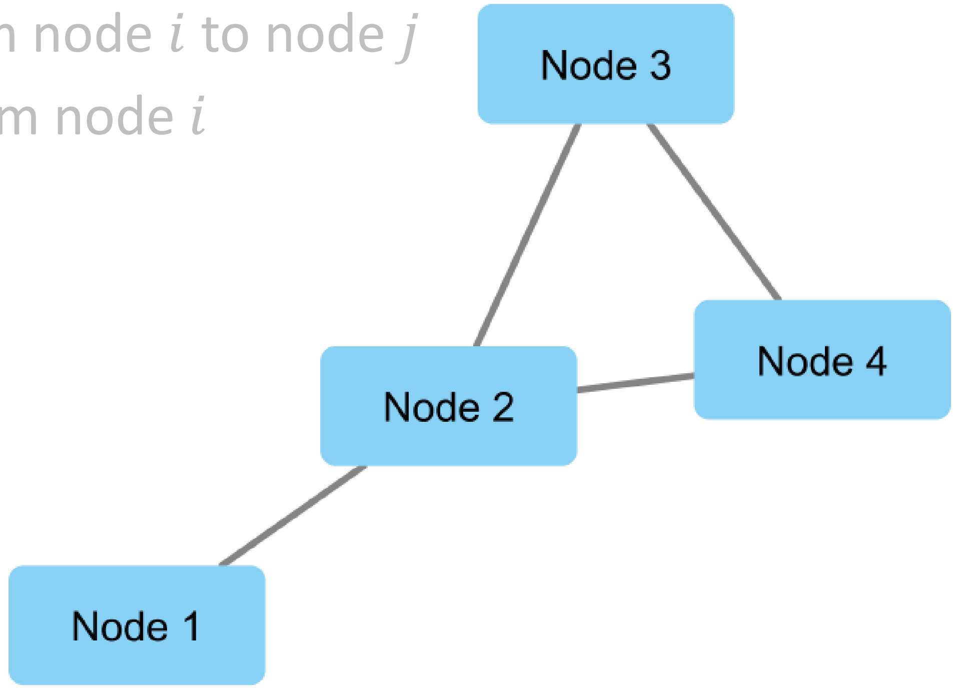
- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i

Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness



Mean of the length of geodesic paths

$\langle d \rangle(i) = \frac{\sum_{j \neq i} d_{ij}}{|d_i|}$ is the average shortest path of node i

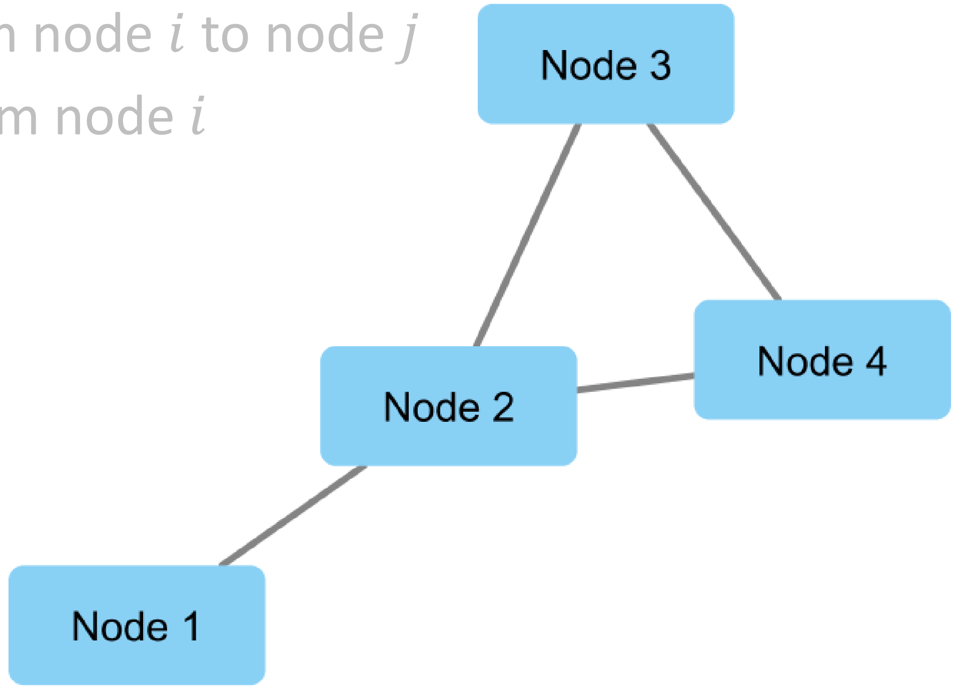
$$\langle d \rangle(1) = \frac{1+2+2}{3} = \frac{5}{3} = 1.6667$$

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i

Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness



Connectedness of neighbours

$C_i = \frac{2L_i}{k_i(k_i-1)}$ is the clustering coefficient of node i

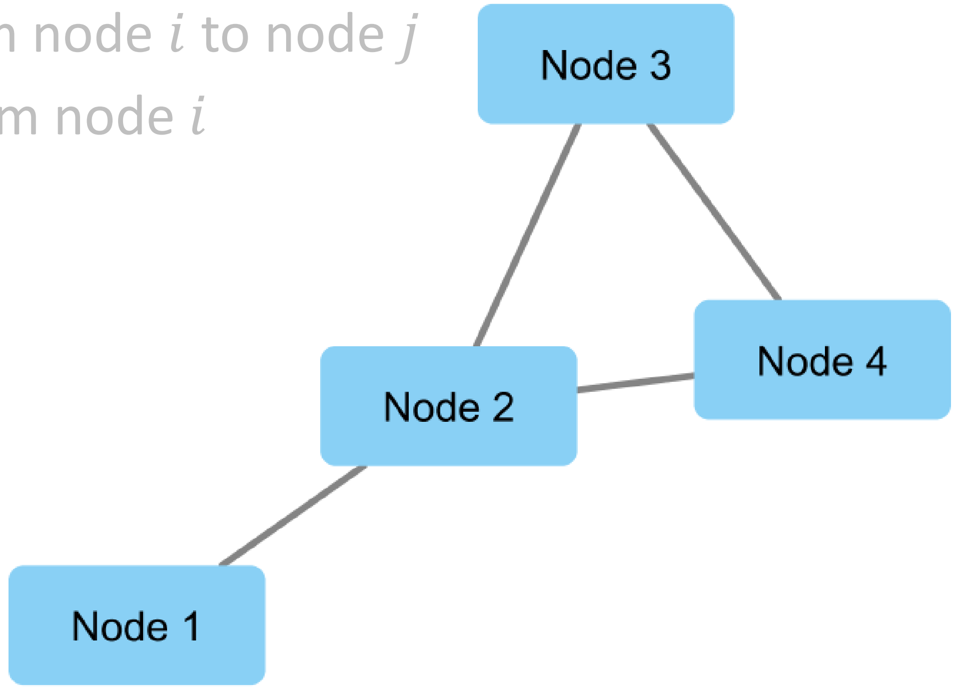
$C_2 = \frac{2}{6} = 0.3333$

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i

Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness



Reciprocal of average shortest path

$C_c(i) = \frac{1}{\langle d \rangle (i)}$ is the closeness centrality of node i

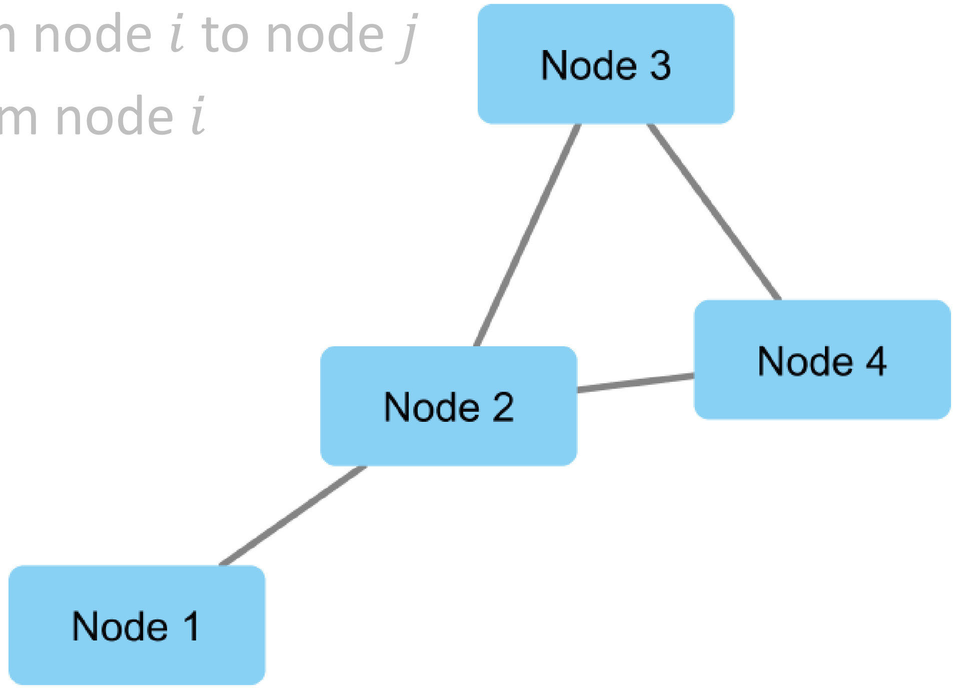
$C_c(1) = \frac{3}{5} = 0.6$

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i

Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness



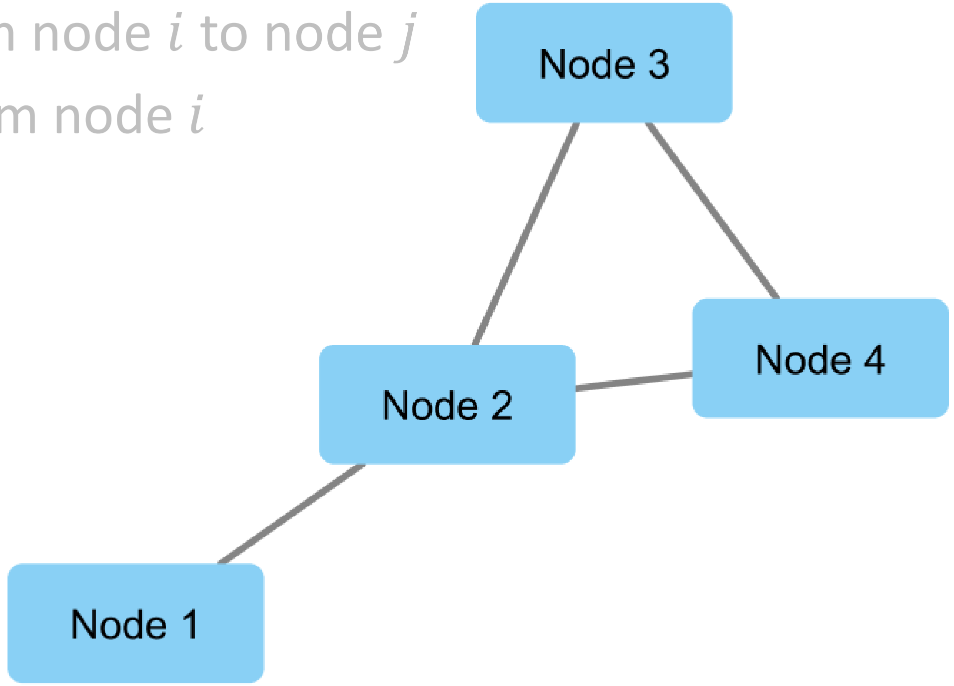
Maximum non-infinite of shortest path starts from node i

$E_c(i) = \max\left(d_{ij} \middle| i \neq j, d_{ij} \neq \infty\right)$ is the eccentricity of node i

$E_c(1) = \max(1,2,2) = 2$

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i



Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness

The number of the shortest paths going through node i

$St(i) = \sum_{s \neq t \neq i}(1|\sigma_{st}(i))$ is the stress of node i

$\sigma_{st}(i)$ the number of the shortest paths from node s to t that passes node i

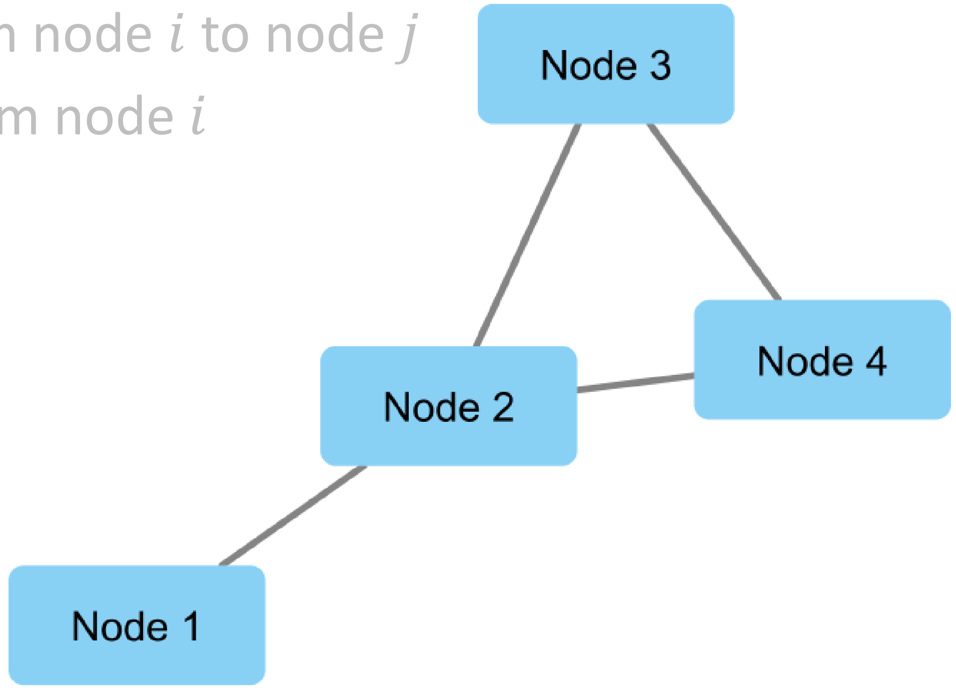
$St(2) = 4$

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i

Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness



The number of the connection of node i

k_i is the degree of node i

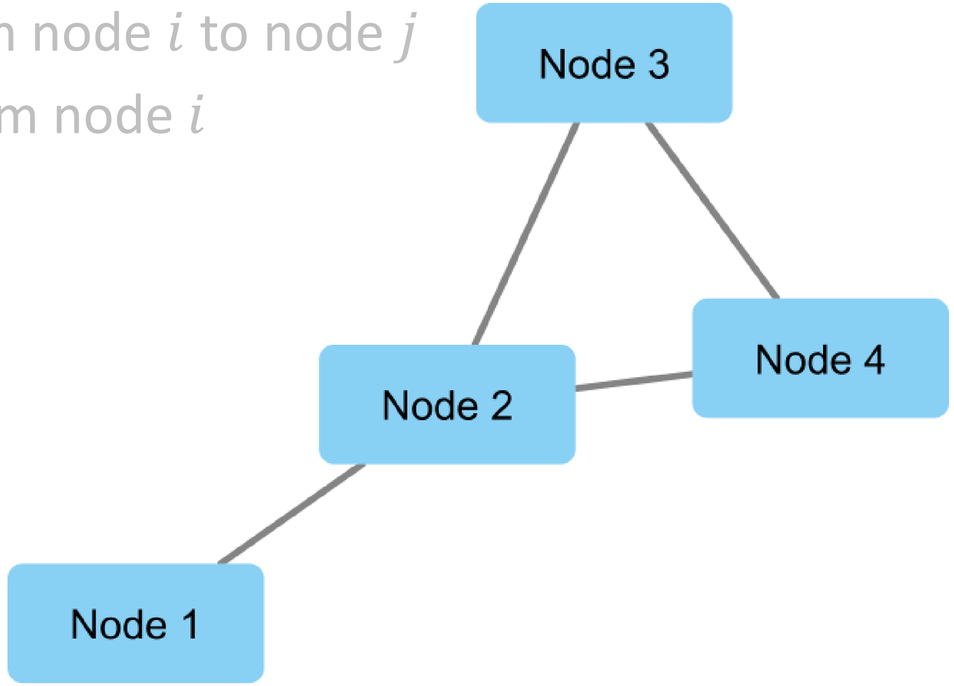
$k_2 = 3$

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i

Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness



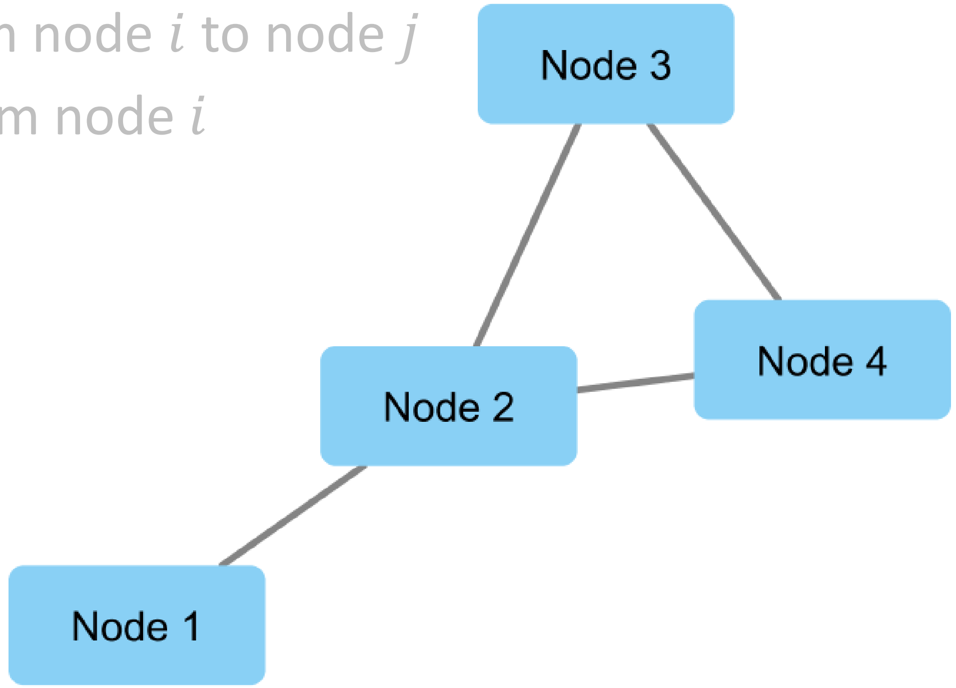
Proportion of appearance of node i in all of the shortest paths

$C_b(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$ is the betweenness centrality of node i

$C_b(2) = \frac{4}{6} = 0.6667$

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i



Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness

Average degree of neighbours of node i

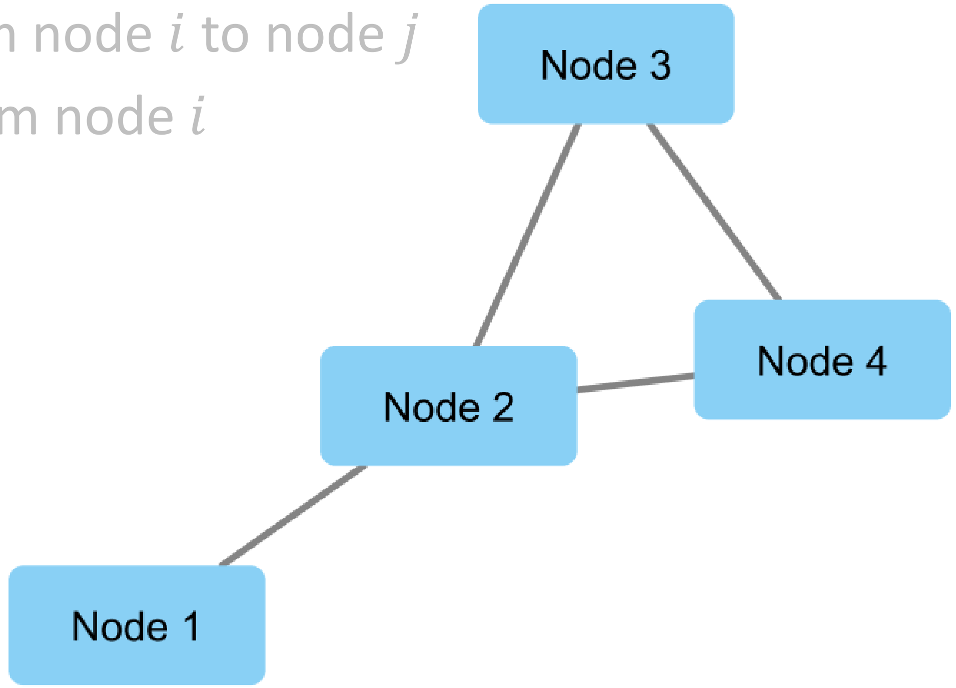
$C_n(i) = \frac{\sum_{j \in n_i} k_j}{k_i}$ is the neighbourhood connectivity of node i

n_i is the set of neighbours of node i

$C_n(1) = 3$

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i



Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness

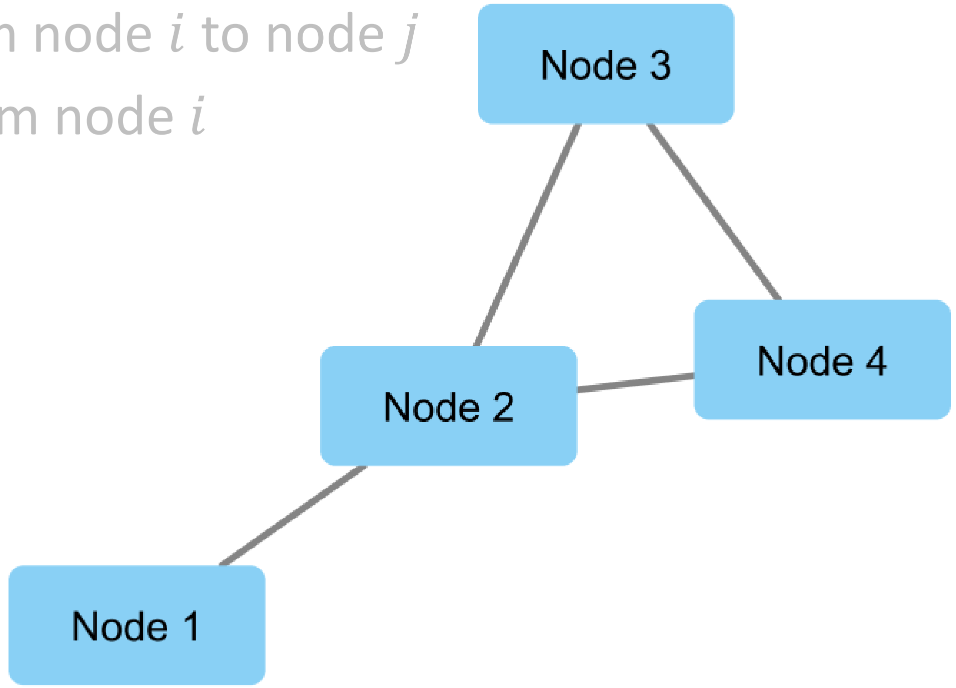
How ties of node i reach out into the network

$R(i) = \frac{(d_{max}+1)-\langle d\rangle(i)}{d_{max}}$ is the radiality of node i

$$R(1) = \frac{3-1.6667}{2} = 0.6667$$

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i



Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness

How node i shares interaction with other nodes

$T_i = \frac{avg(J(i,j))}{k_i}$ is the topological coefficient of node i

$J(i,j)$: number of common neighbours of node i and j, (+1, if i and j are neighbours)

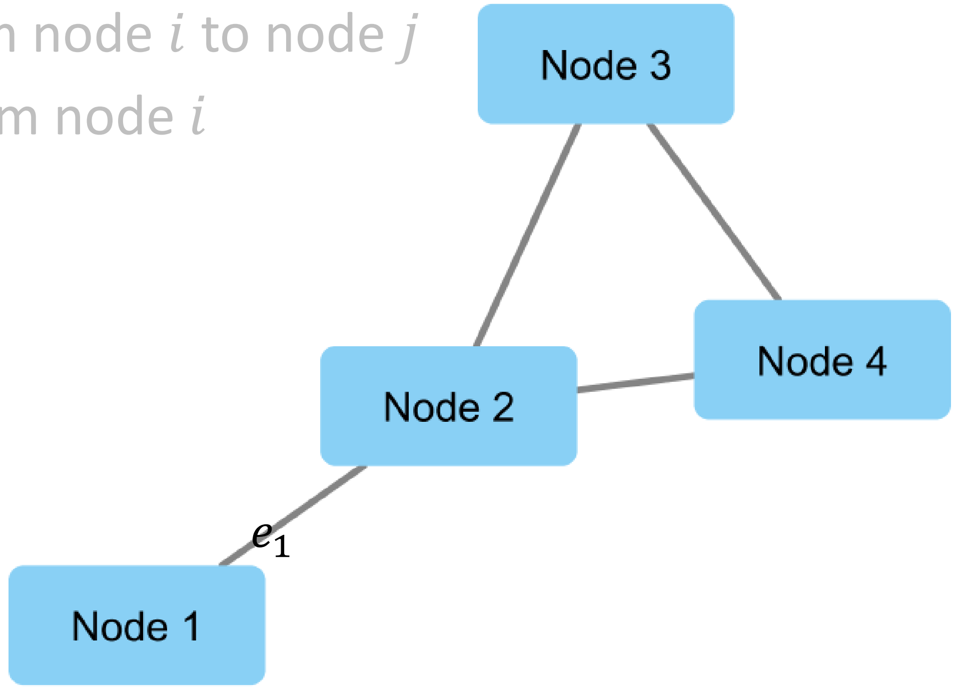
$T_1 = 0$ // k_i must be at least 2, $T_3 = 0.8333$

NetworkAnalyzer

- Let σ_{ij} the number of the shortest paths from node i to node j
- Let $|d_i|$ the number of the shortest paths from node i

Generated measures:

- Average shortest path
- Clustering Coefficient
- Closeness Centrality
- Eccentricity
- Stress
- Degree
- Betweenness Centrality
- Neighborhood Connectivity
- Radiality
- Topological Coefficient
- Edge Betweenness



The number of shortest paths going through edge $e = (i, j)$
$B_e(e) = \sum_{s,t}(1|\sigma_{st}(e))$ is the edge betweenness of edge e
$B_e(e_1) = 6$

Other applications

Gephi

- Random networks
- Basic network measures
- Website

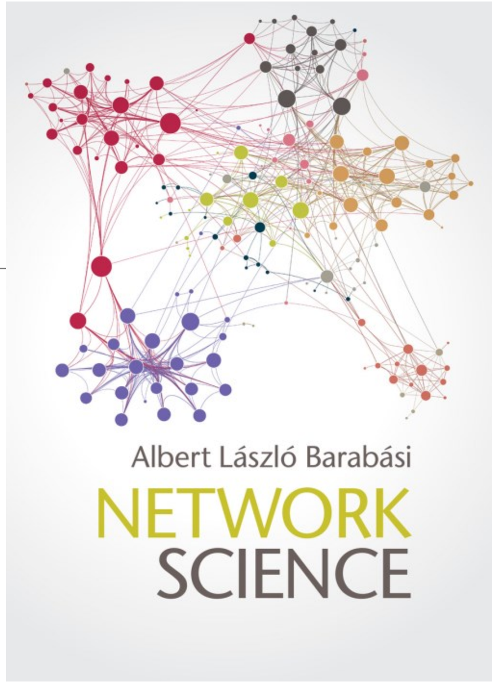
Netlogo

- Barabási-Albert model simulator
  - (Sample Models/Networks/Preferential Attachment)
- Small World simulator
  - (Sample Models/Networks/Small Worlds)
- Giant Component simulator
  - (Sample Models/Networks/Giant Component)
- Website

Network Analysis

07 – EVOLVING NETWORKS

Slides were created by: Daniel Leitold



[Network Science book (online)]

Barabási, Albert-László. *Network Science*.
Cambridge University Press, 2016.

Introduction

By the late 1990s, two search engines had been created with an early start:

- Alta Vista
- Inktomi

Six years after the birth of the WWW, Google was a latecomer to search, BUT:

- Became the leading search engine, and
- by 2000 had become the biggest hub of the Web.

Youngster Facebook:

- In 2011 it became the Web's biggest node.

https://blogs-images.forbes.com/adamhartung/files/2015/04/Facebook-v-Google.jpg?width=960

Introduction

The Web's competitive landscape highlights an important limitation of our modelling framework:

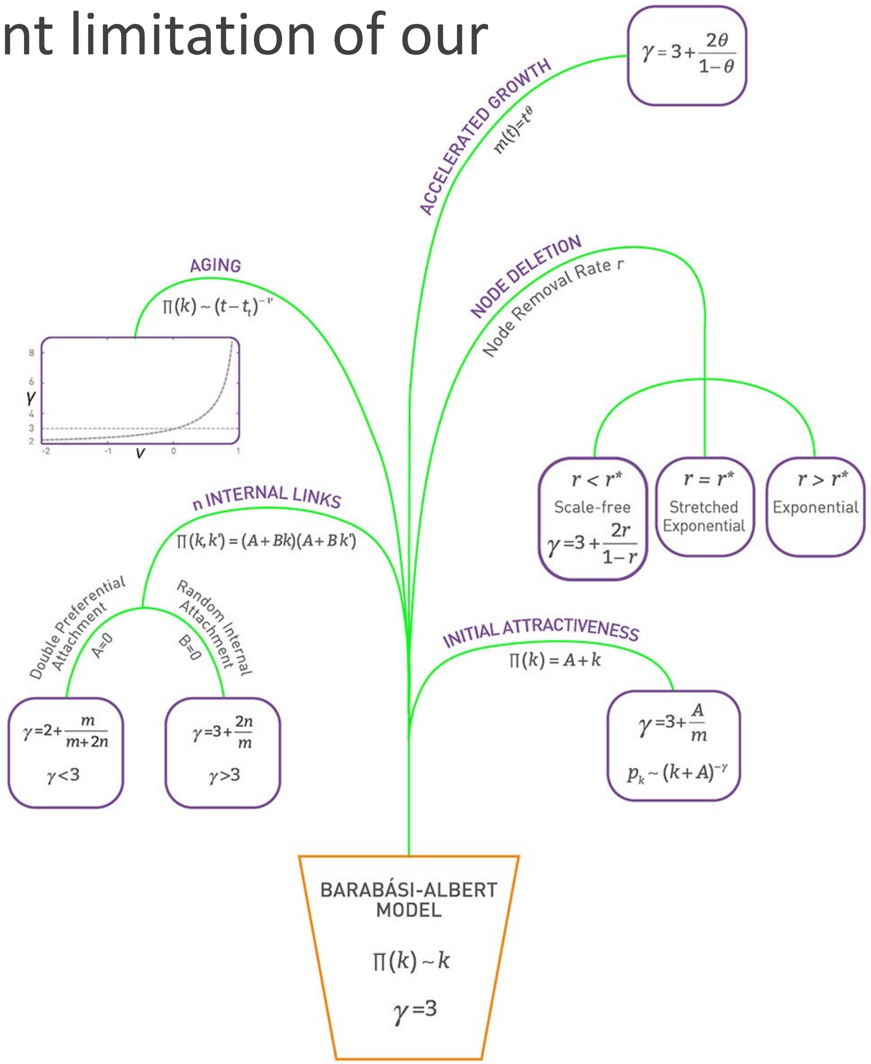
- None of the models is able to account for it.

The biggest node is

- Random in Erdős-Rényi model
- The oldest in Barabási-Albert model ($k(t) \sim t^{\frac{1}{2}}$)
  - first mover's advantage

We will explore

- Initial attractiveness
- n internal links
- Node deletion
- Aging of nodes
- Accelerated growth



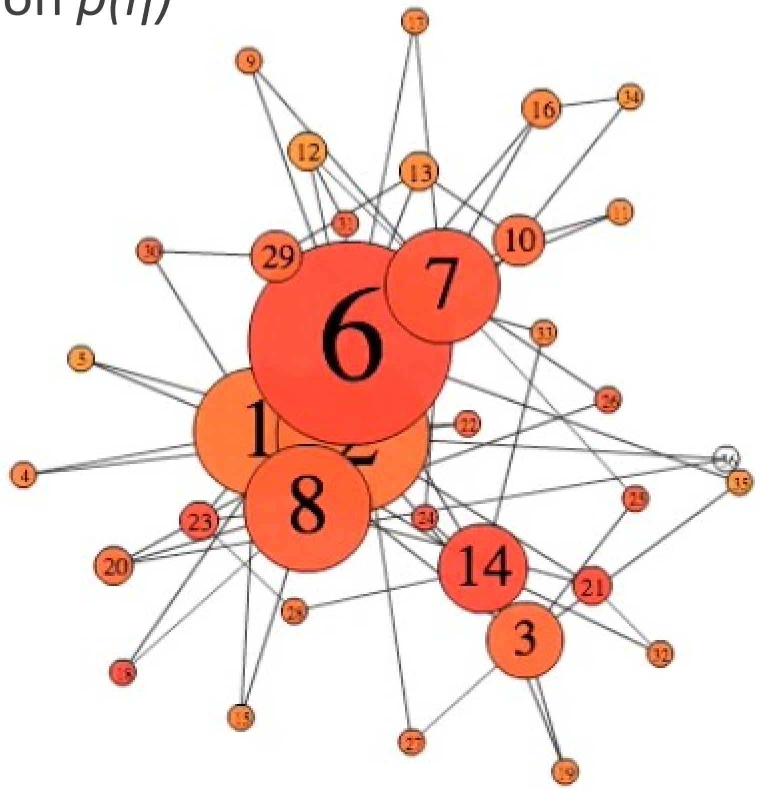
The Bianconi-Barabási Model

Intrinsic node property:

- Fitness (η): a random number chosen from a fitness distribution $\rho(\eta)$
- Video

The Bianconi-Barabási Model:

- Growth: a new node (j) has:
  - m new connections, and
  - η_j fitness
- Preferential Attachment: probability to connect to node i
  - Depends on degree (k_i) and fitness (η_i)
  - $\prod_i = \frac{\eta_i k_i}{\sum_j \eta_j k_j}$



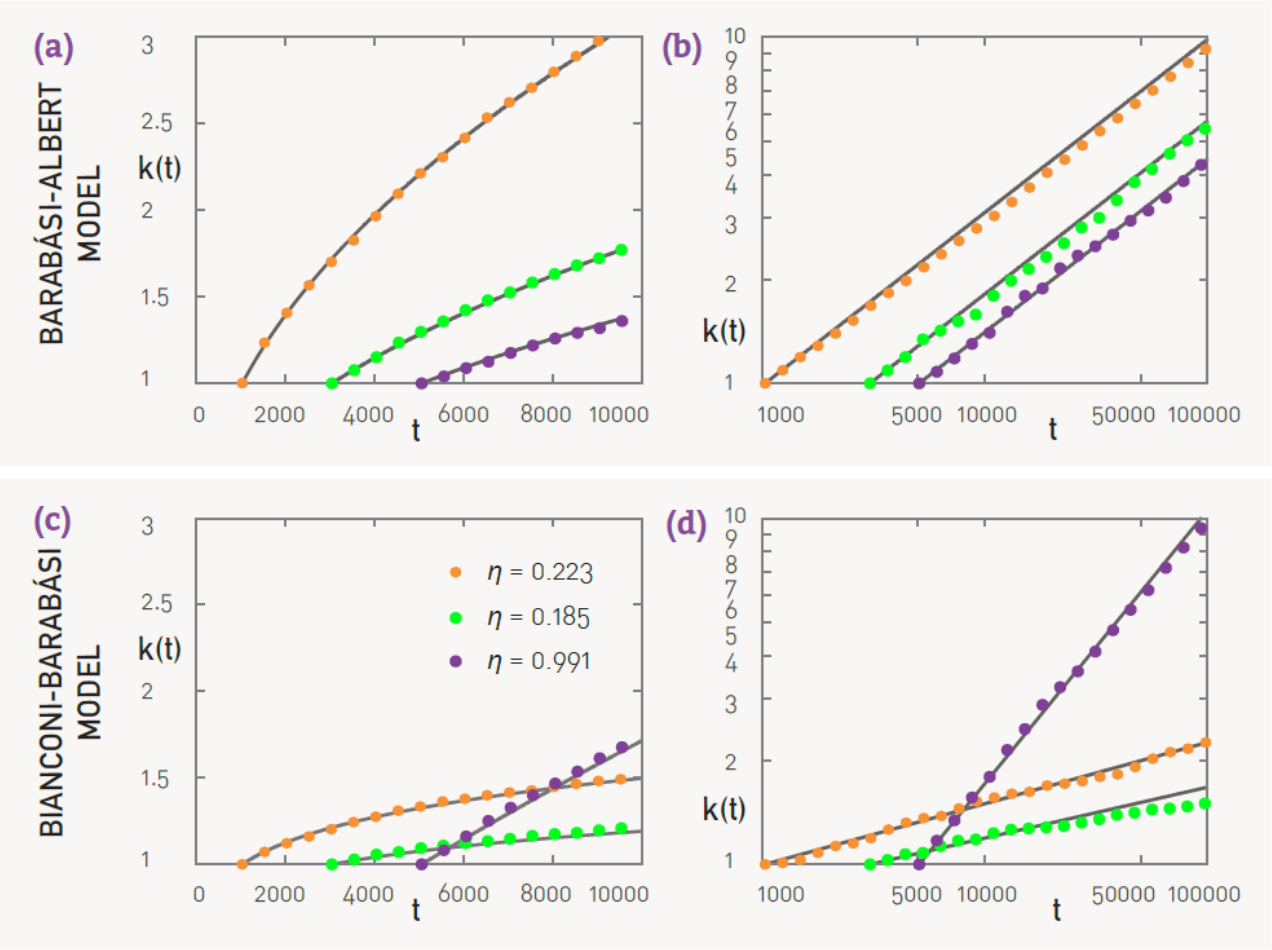
Degree Dynamics

We can predict each node's evolution

- $\frac{\partial k_i}{\partial t} = m \frac{\eta_i k_i}{\sum_j \eta_j k_j}$

The degree at time t

- $k(t, t_i, \eta_i) = m \left(\frac{t}{t_i}\right)^{\beta(\eta_i)}$



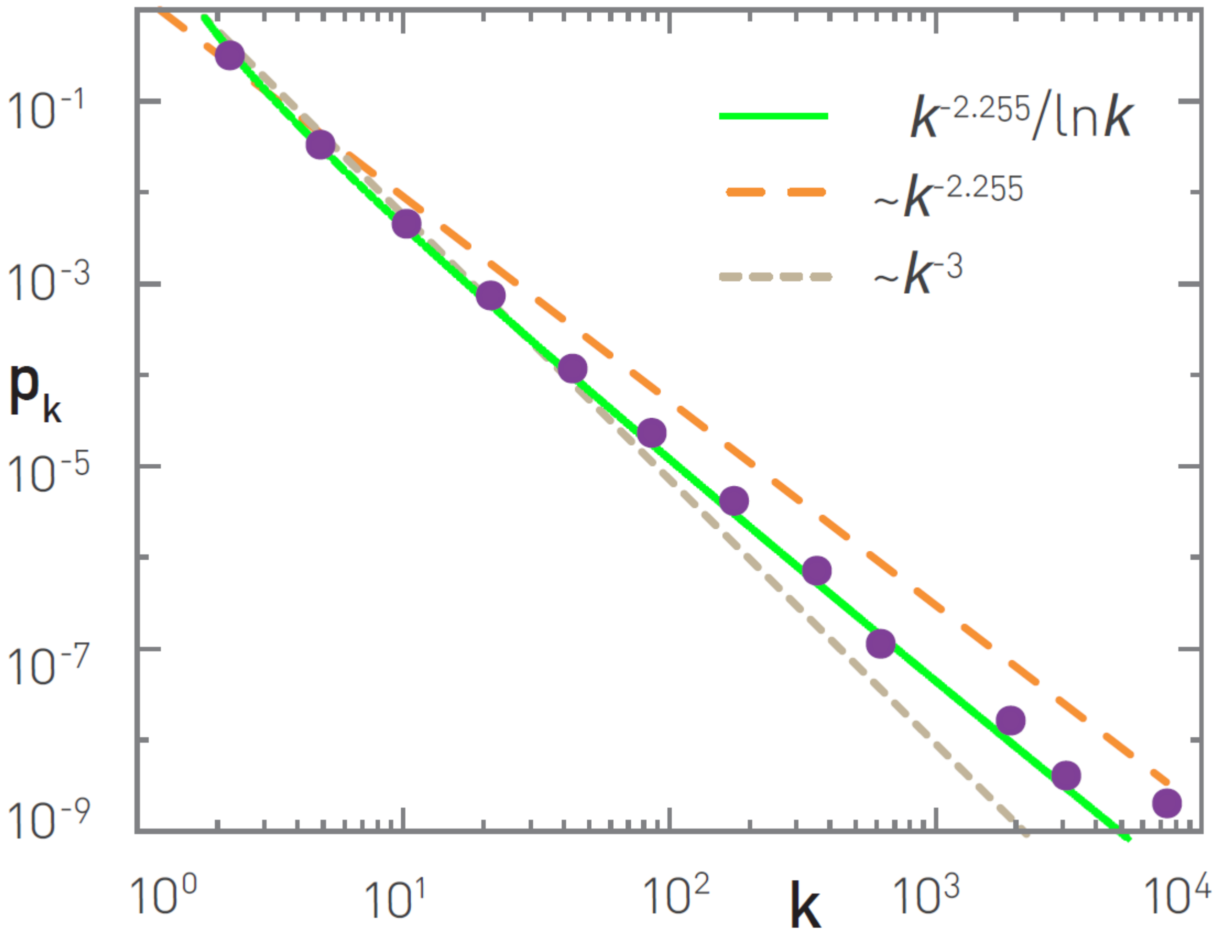
Degree Dynamics

We can predict each node's evolution

- $\frac{\partial k_i}{\partial t} = m \frac{\eta_i k_i}{\sum_j \eta_j k_j}$

The degree at time t

- $k(t, t_i, \eta_i) = m \left(\frac{t}{t_i}\right)^{\beta(\eta_i)}$



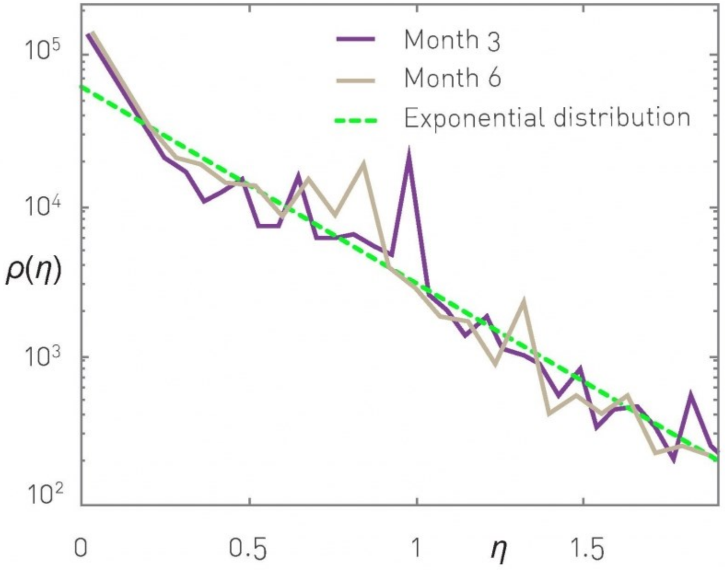
Degree distribution:

- Equal Fitnesses (BA model)
  - $p_k \sim k^{-3}$
- Uniform Fitness Distribution
  - p_k depends on $p(\eta)$

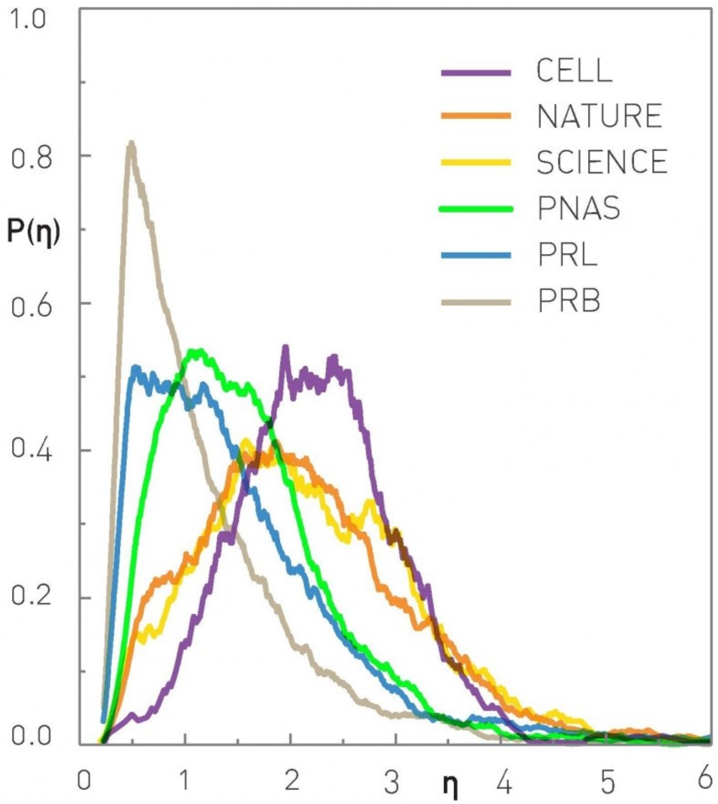
Measuring Fitness

Our ability to determine the fitness is prone to errors.

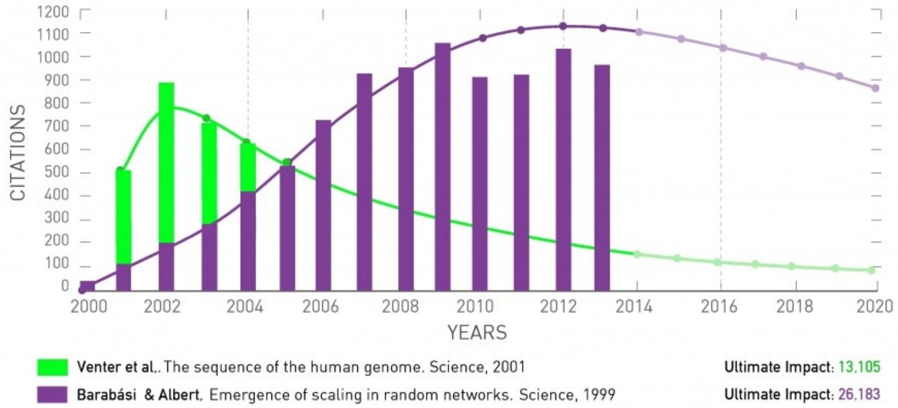
Fitness is determined:
- not by us
- BUT by nodes



The Fitness Distribution of the WWW



The Fitness Distribution of Research Papers



Predicting Ultimate Impact

Bose-Einstein Condensation

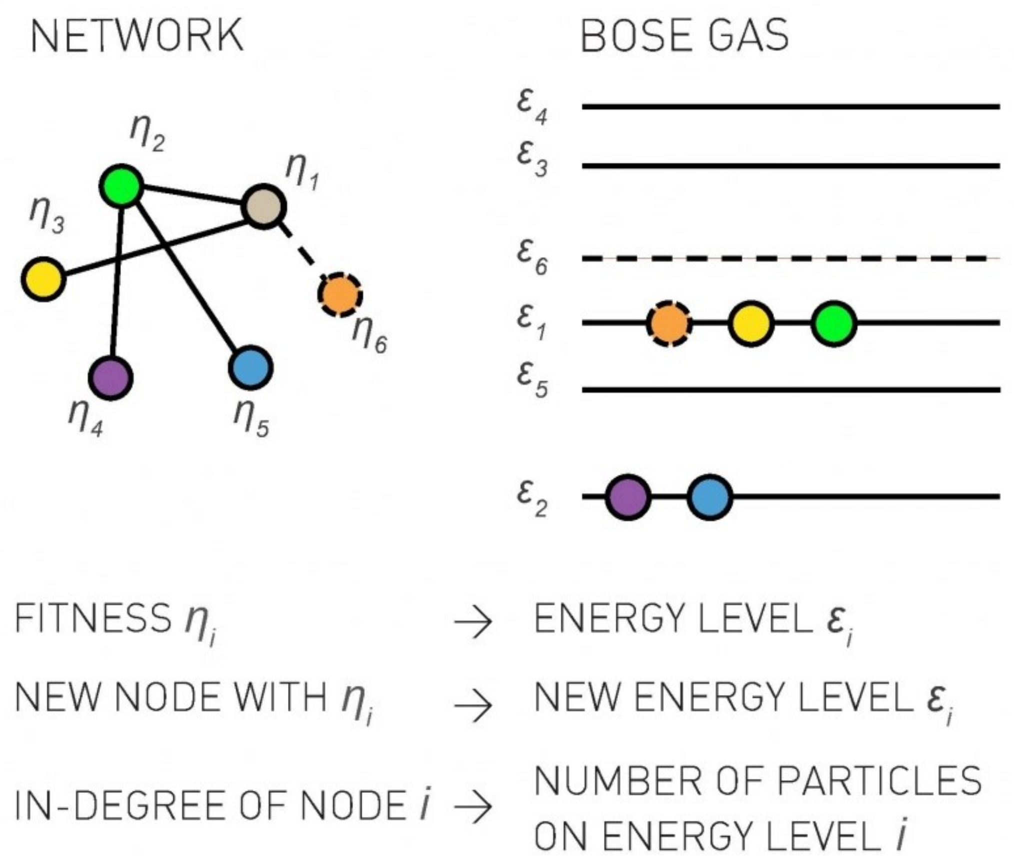
Some networks can undergo Bose-Einstein condensation.

- Fitness → Energy
- Links → Particles
- Nodes → Energy levels

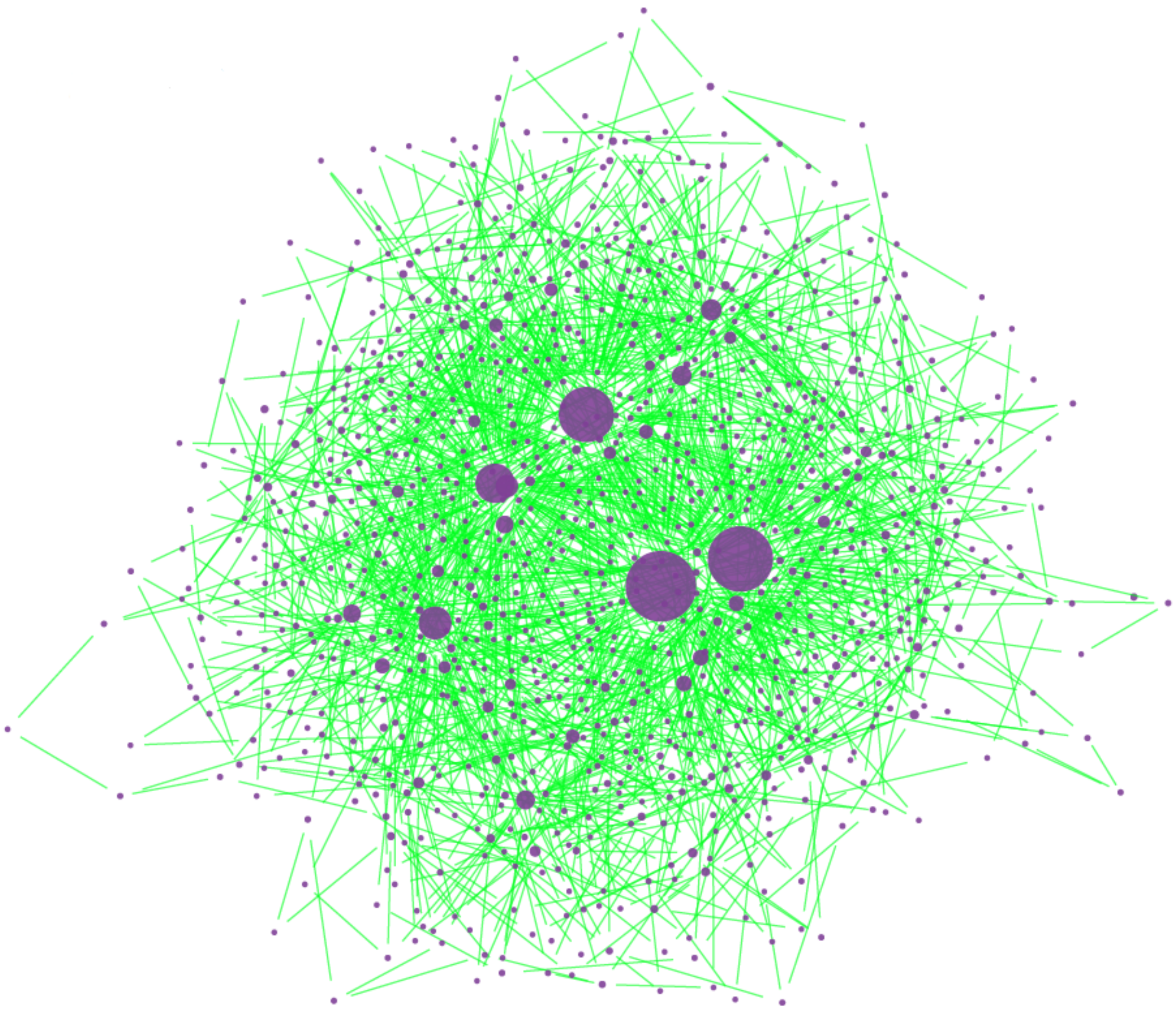
The links of the fitness model behave like subatomic particles in a quantum gas.

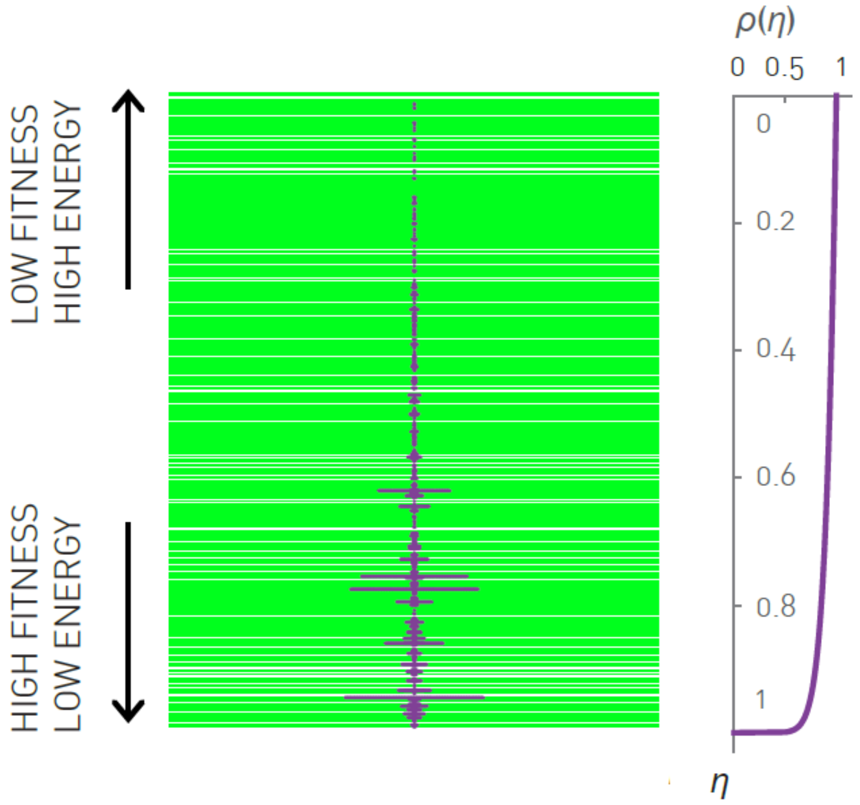
Based on fitness distributions

- Scale-free phase
  - Fit-gets-rich phenomenon
  - Degree distribution follows power-law
- Bose-Einstein condensation (Video)
  - Winner takes-all phenomenon
  - Hub and spoke topology

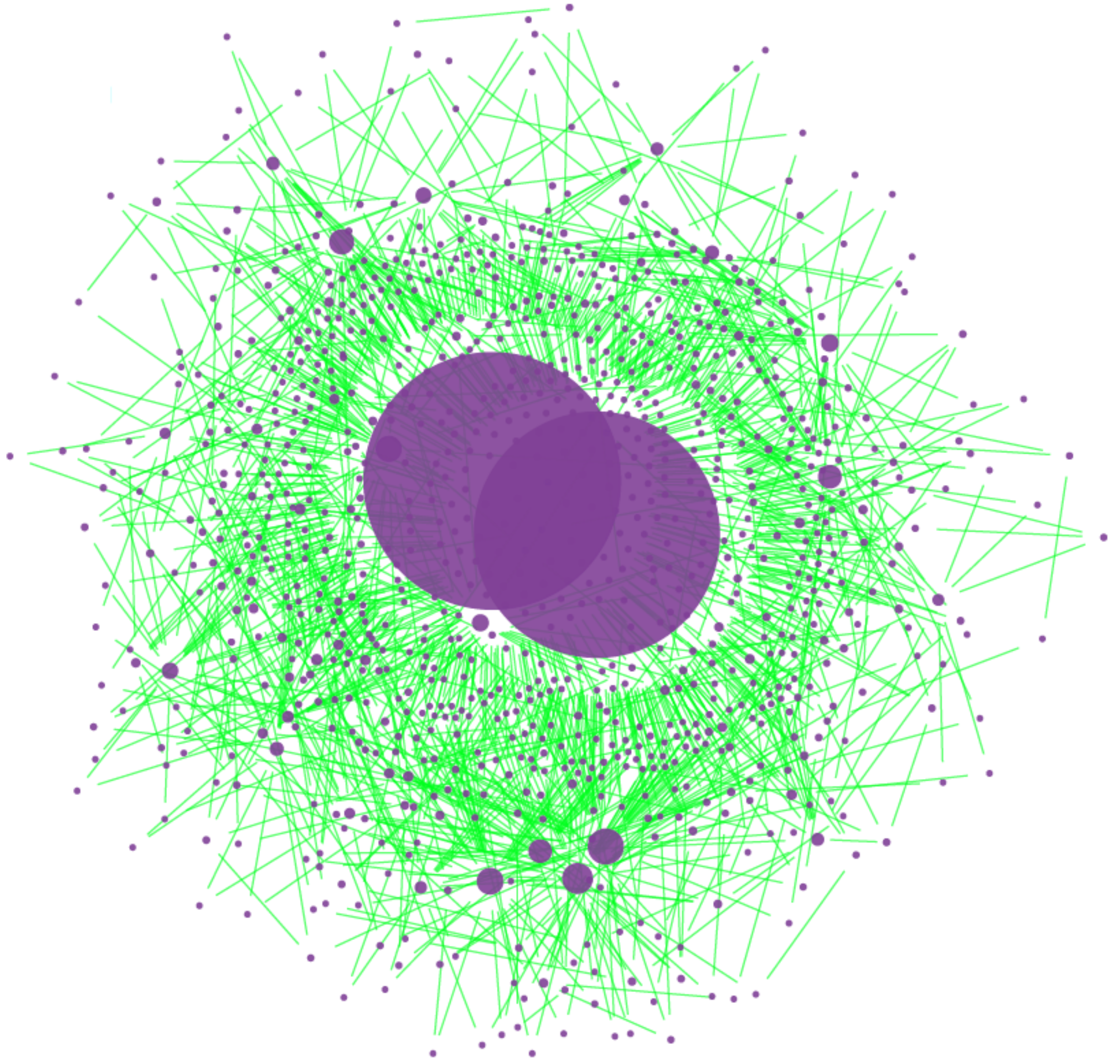


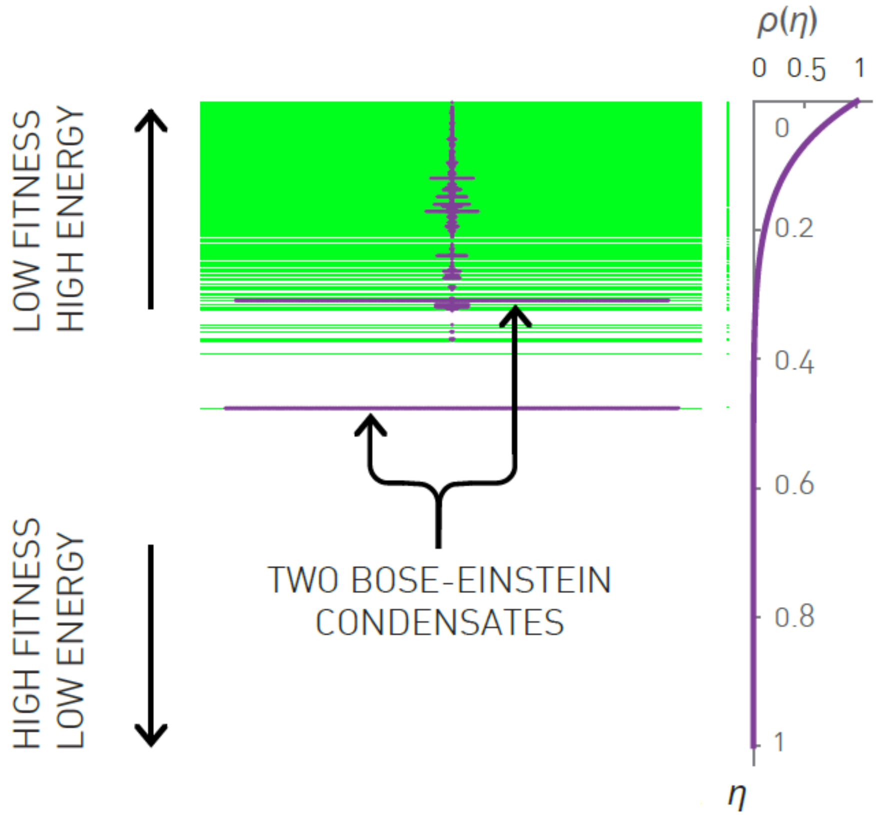
Bose-Einstein Condensation





Bose-Einstein Condensation





Evolving Networks

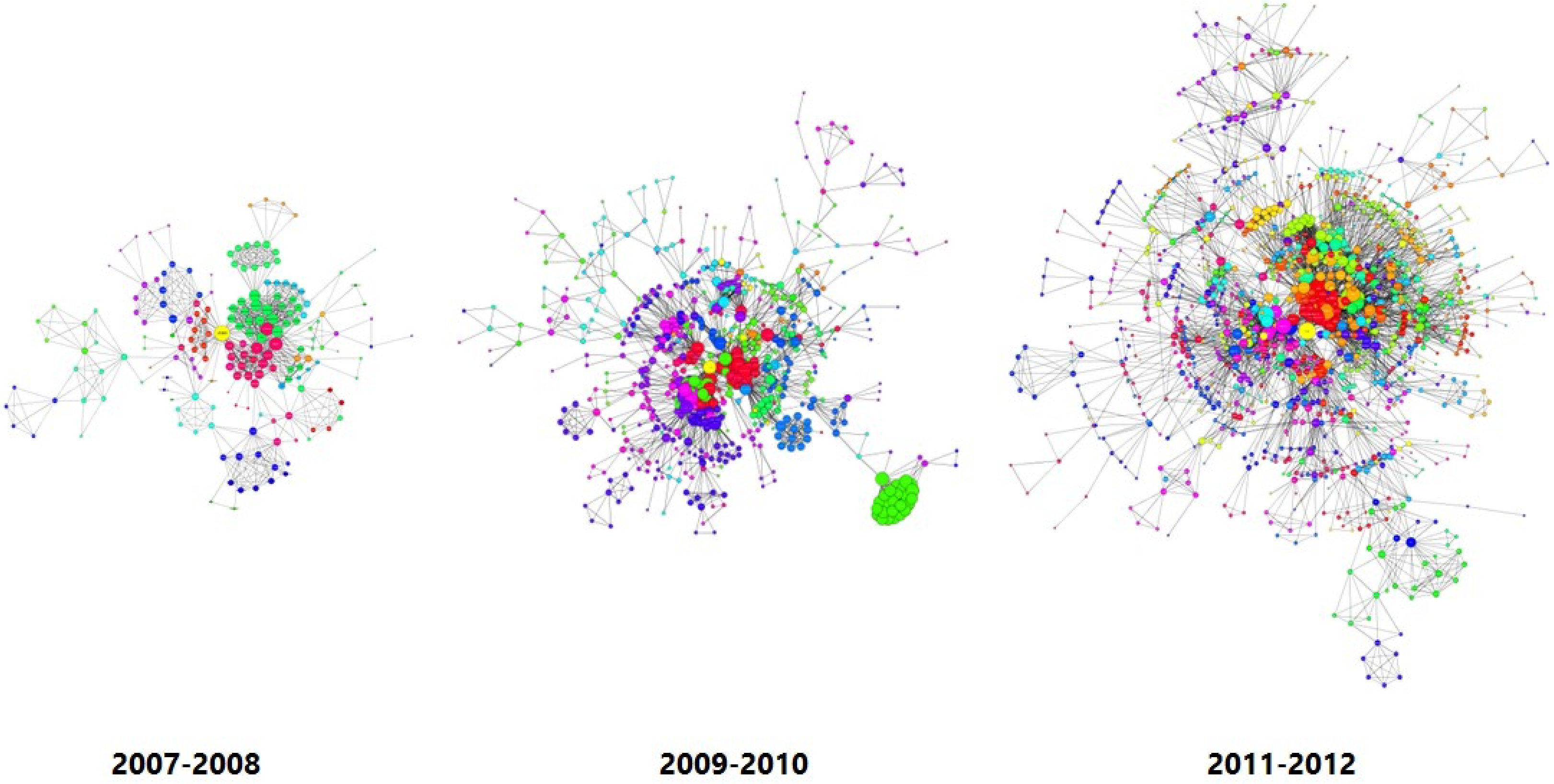
Initial Attractiveness

Internal Links

Node Deletion

Accelerated Growth

Aging



The largest components in Apple's inventor network over a 6-year period

https://www.kenedict.com/site_update/wp-content/uploads/2013/07/Apple_Evolution.png

Initial Attractiveness

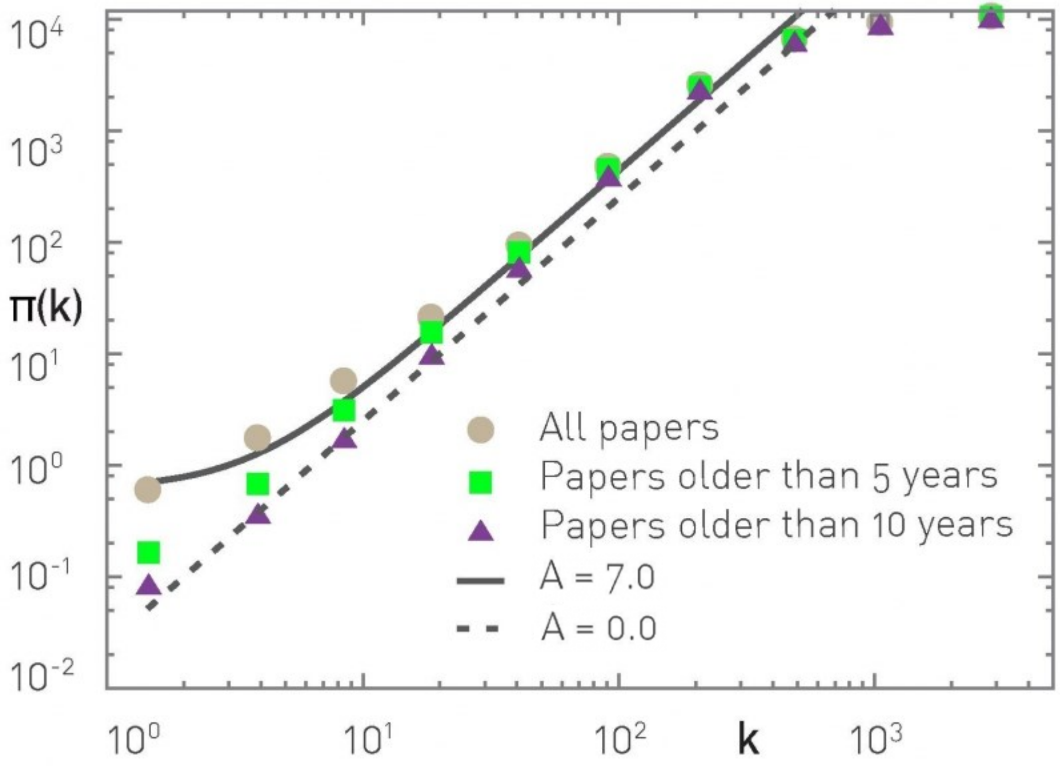
In the Barabási–Albert model an isolated node cannot acquire link.

BUT in reality:

- new research paper has $p > 0$ probability of being cited for the first time
- a person that moves to a new city quickly acquires acquaintances

Preferential attachment function :

- $\prod_k \sim A + k$
- Constant A is called initial attractiveness



The probability of a new paper to be cited for the first time ($A = 7$) is comparable to the citation probability of a paper with seven citations ($A = 0$).

Initial Attractiveness

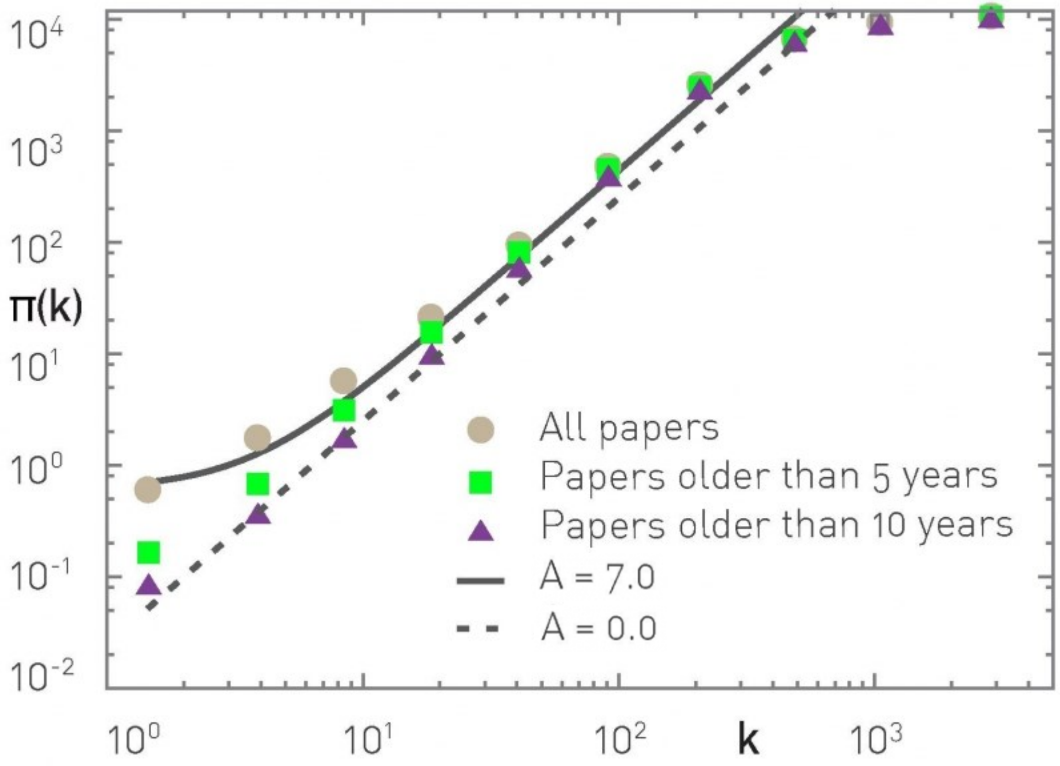
Effect to the Barabási–Albert model:

The Degree Exponent is increased:

- $\gamma = 3 + \frac{A}{m}$

Generates a Small-degree Saturation:

- $p_k = C(k + A)^{-\gamma}$
- pushes the small-k nodes toward higher degrees
- high degrees ($k \gg A$): the degree distribution follows the power law



The probability of a new paper to be cited for the first time ($A = 7$) is comparable to the citation probability of a paper with seven citations ($A = 0$).

Evolving Networks

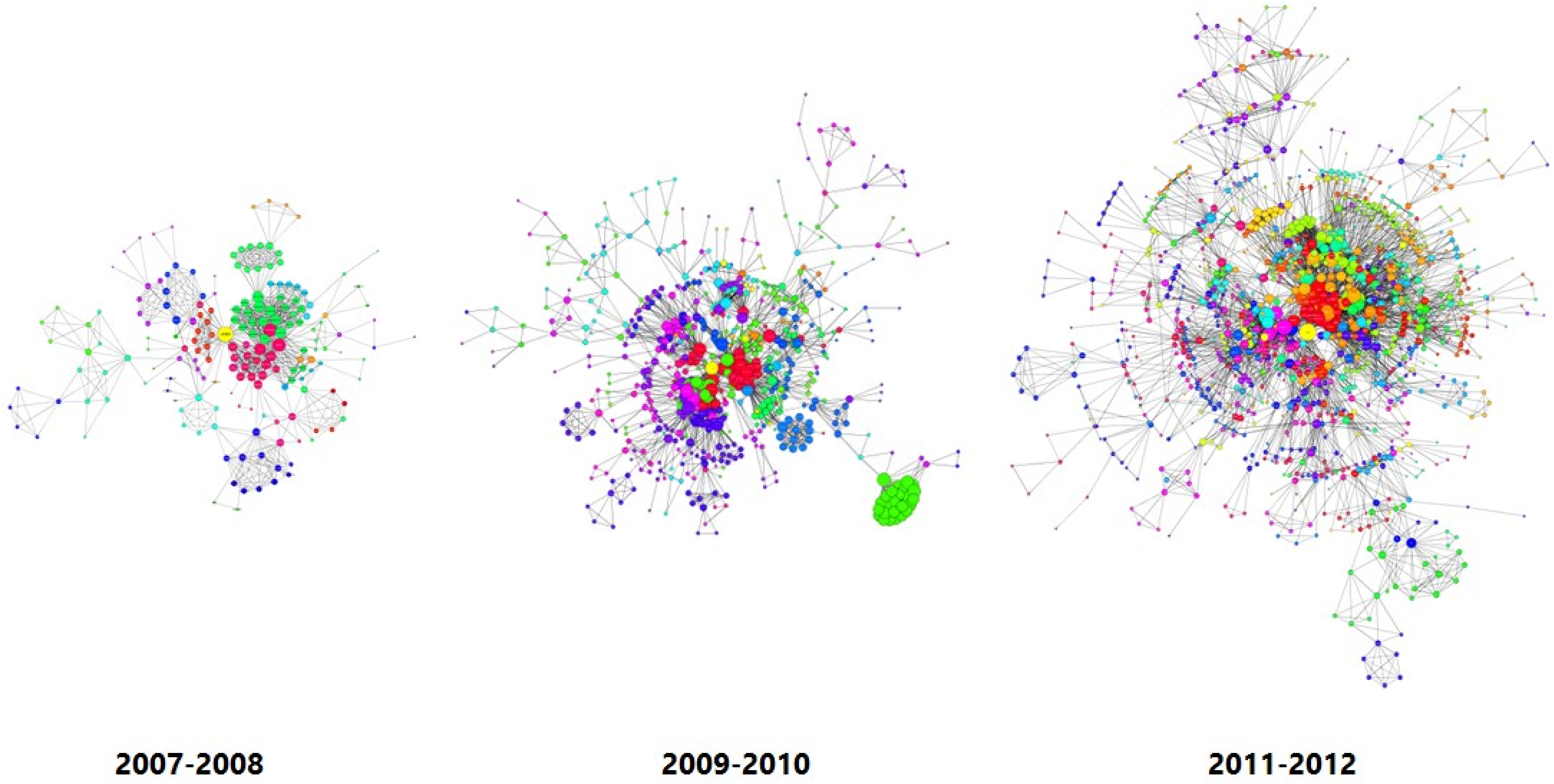
Initial Attractiveness

Internal Links

Node Deletion

Accelerated Growth

Aging



The largest components in Apple's inventor network over a 6-year period

https://www.kenedict.com/site_update/wp-content/uploads/2013/07/Apple_Evolution.png

Internal Links

In many networks new links do not only arrive with new nodes but are added between pre-existing nodes.

Preferential attachment function:

- $\prod(k, k') \sim (A + Bk)(A + Bk')$

Limiting cases of:

- Double Preferential Attachment (A=0)
  - $\gamma = 2 + \frac{m}{m+2n}$
  - Lowers the degree exponent from 3 to 2
- Random Attachment (B=0)
  - $\gamma = 3 + \frac{2n}{m}$
  - Degree exponent bigger than 3

Evolving Networks

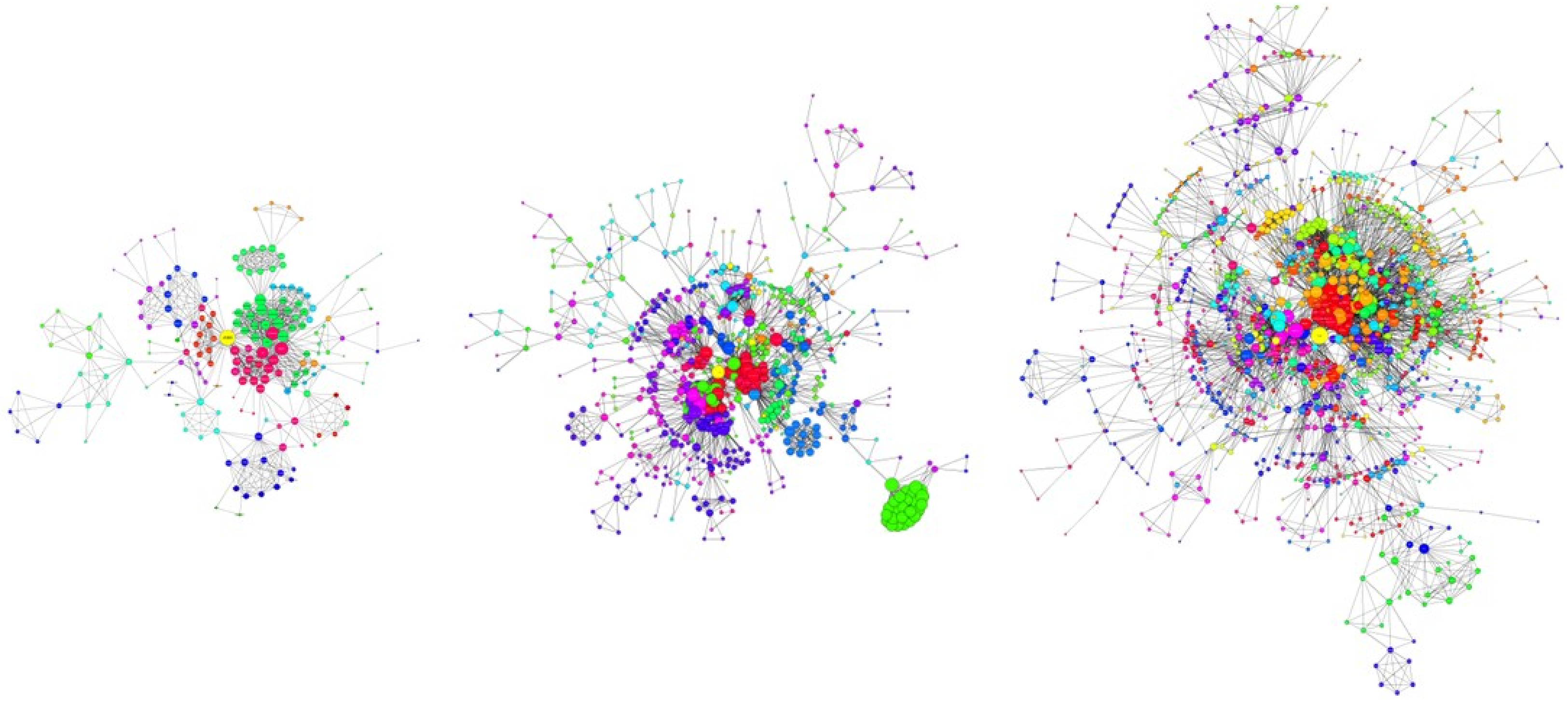
Initial Attractiveness

Internal Links

Node Deletion

Accelerated Growth

Aging



2007-2008 2009-2010 2011-2012

The largest components in Apple's inventor network over a 6-year period

https://www.kenedict.com/site_update/wp-content/uploads/2013/07/Apple_Evolution.png

Node Deletion

Real examples:

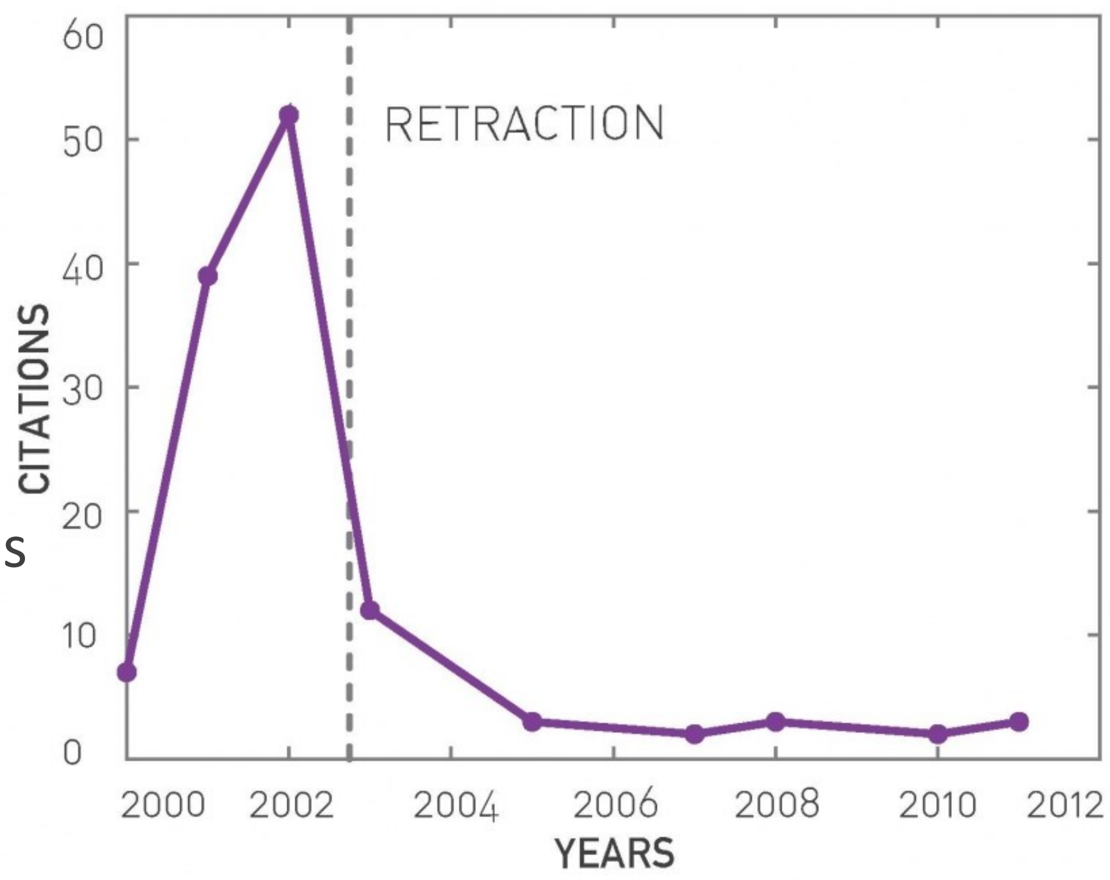
- employees leave the company
- web documents are removed

In Barabási–Albert model in each step:

- Add a node with m new links
- Remove a node with r rate

Based on r, there are three different phases

- Scale-free phase ($r < 1$)
  - $\gamma = 3 + \frac{2r}{1-r}$
- Exponential phase ($r = 1$)
  - $\gamma \to \infty$, N is constant, we loose scale-free property
- Declining Networks ($r > 1$)
  - Alzheimer's research focuses on the progressive loss of neurons



The Impossibility of Node Deletion
Retraction lead to a dramatic drop in citations, but the papers continue to be cited.

Evolving Networks

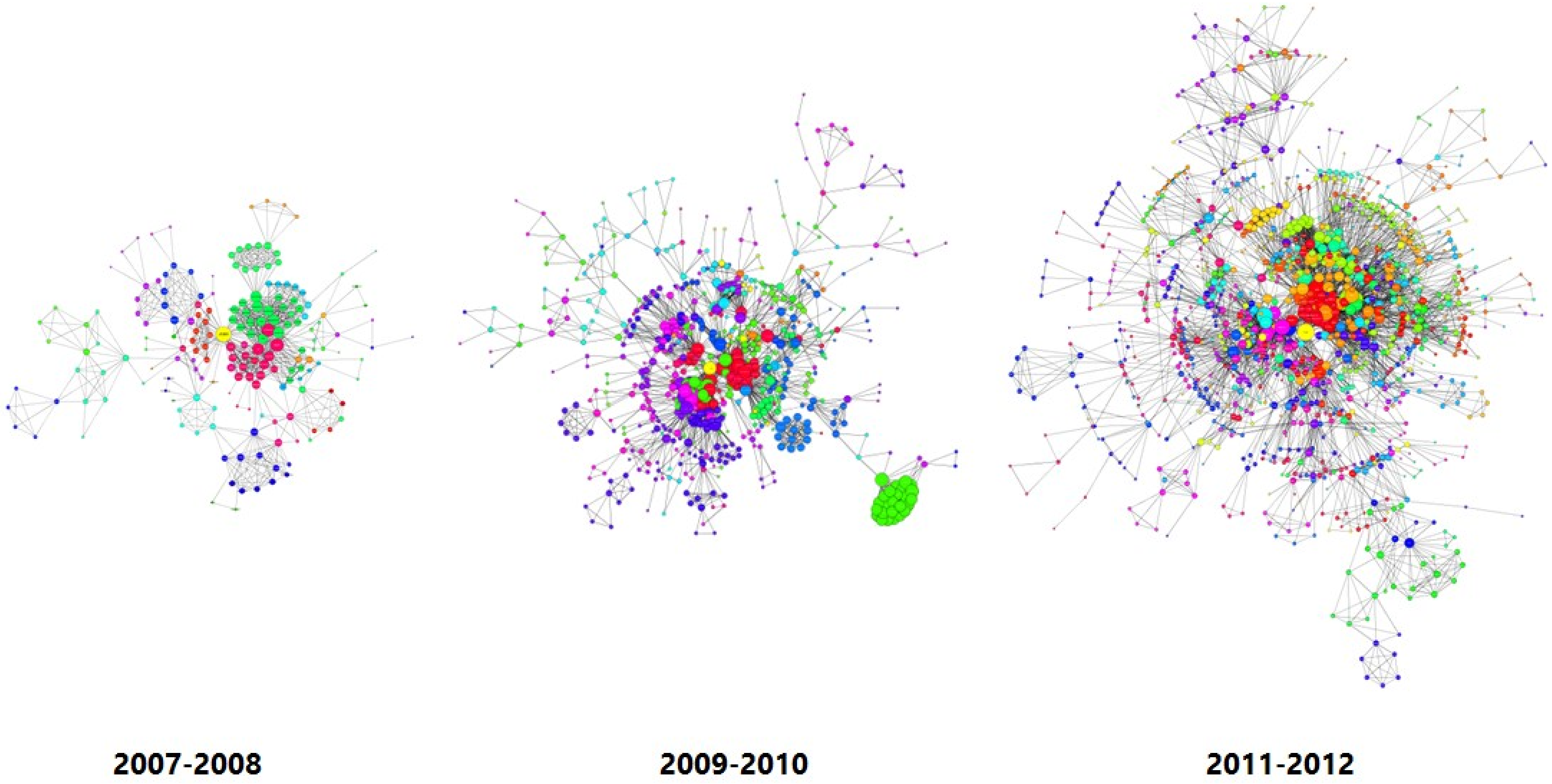
Initial Attractiveness

Internal Links

Node Deletion

Accelerated Growth

Aging



The largest components in Apple's inventor network over a 6-year period

https://www.kenedict.com/site_update/wp-content/uploads/2013/07/Apple_Evolution.png

Accelerated Growth

Real examples:

- Internet increased from $\langle k \rangle = 3.42$ in November 1997 to 3.96 by December 1998.
- WWW increased its average degree from 7.22 to 7.86 during a five month interval.
- In metabolic networks the average degree of the metabolites grows approximately linearly with the number of metabolites.

The number of links arriving with new nodes is as follows:

- $m(t) = m_0 t^{\theta}$
- If $\theta > 0$, the network follows accelerated growth.

Degree exponent

- $\gamma = 3 + \frac{2\theta}{1-\theta}$

For $\theta = 1$:

- The degree exponent diverges, leading to *hyper-accelerating growth*.
- In this case $\langle k \rangle$ grows linearly with time and the network looses its scale-free nature.

Evolving Networks

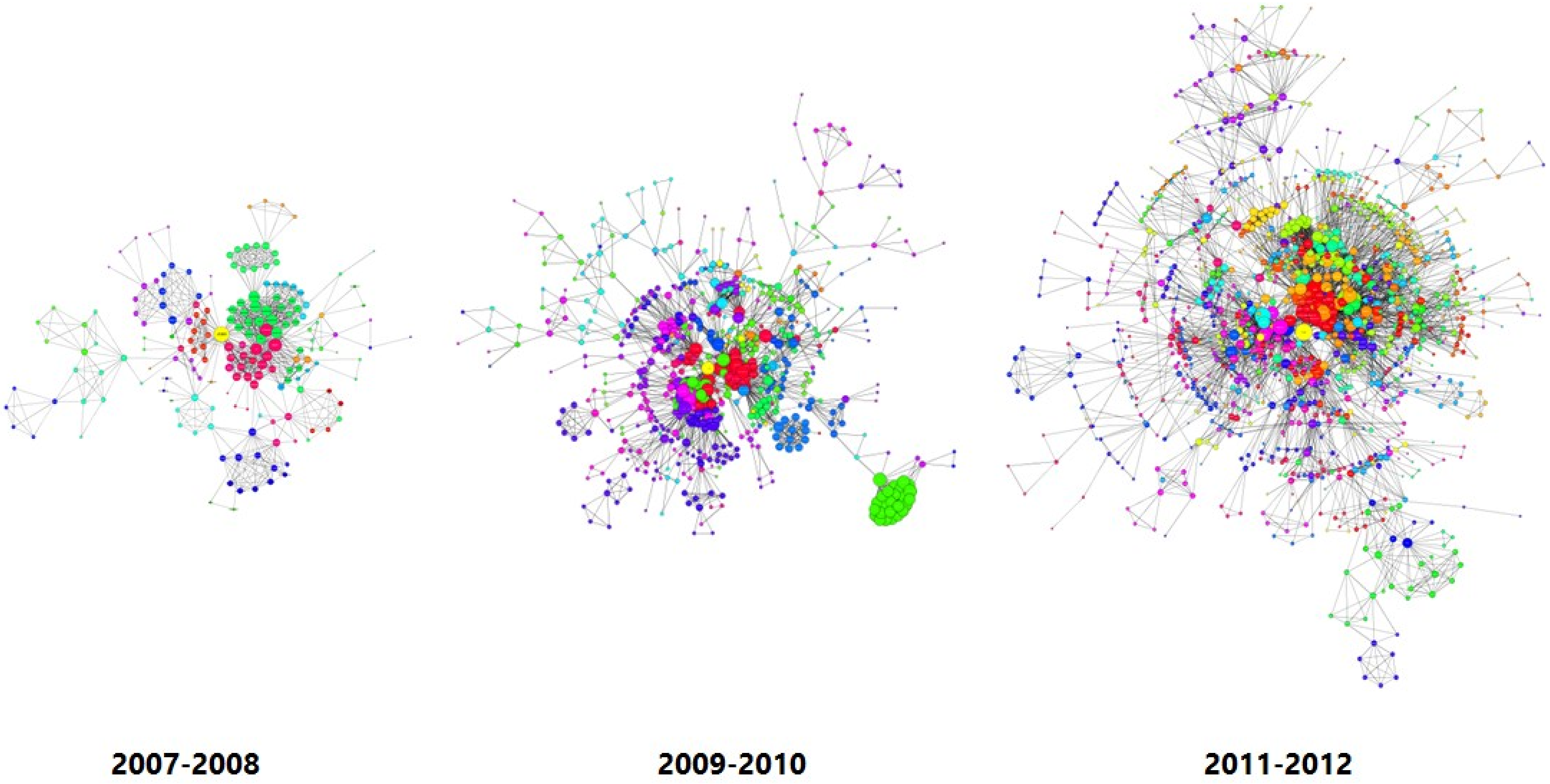
Initial Attractiveness

Internal Links

Node Deletion

Accelerated Growth

Aging



The largest components in Apple's inventor network over a 6-year period

https://www.kenedict.com/site_update/wp-content/uploads/2013/07/Apple_Evolution.png

Aging

Real examples:

- Actors have a finite professional life span.
- Scientists have a finite professional life span.

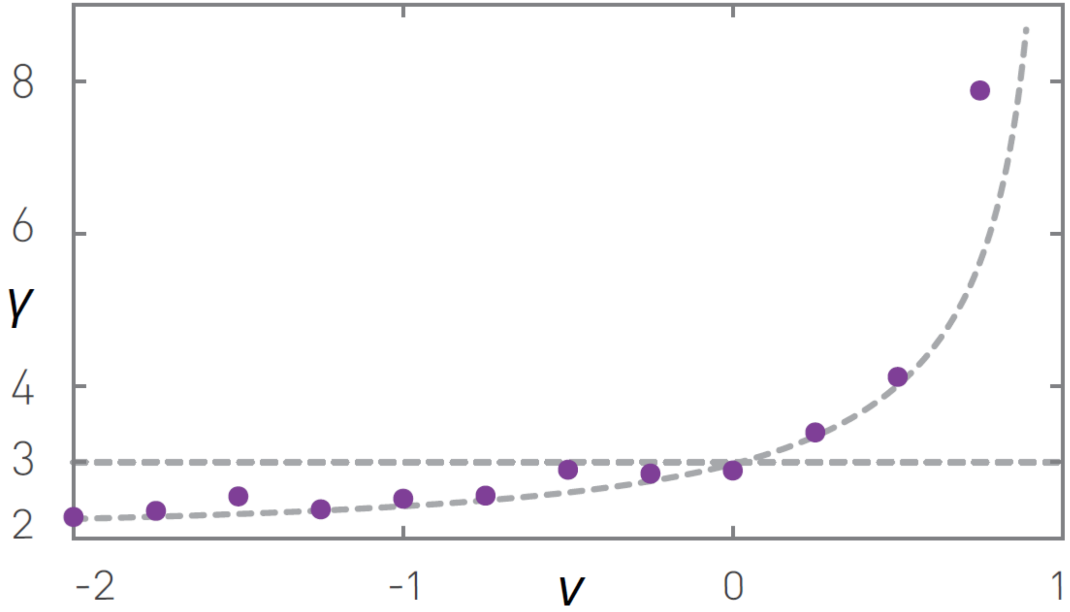
The probability that a new node connects to node i is:

- $\prod(k_i, t - t_i) \sim k(t - t_i)^{-\nu}$
  - t_i is the time node i was added to the network
  - t is the actual time
  - ν is a tuneable parameter

Aging

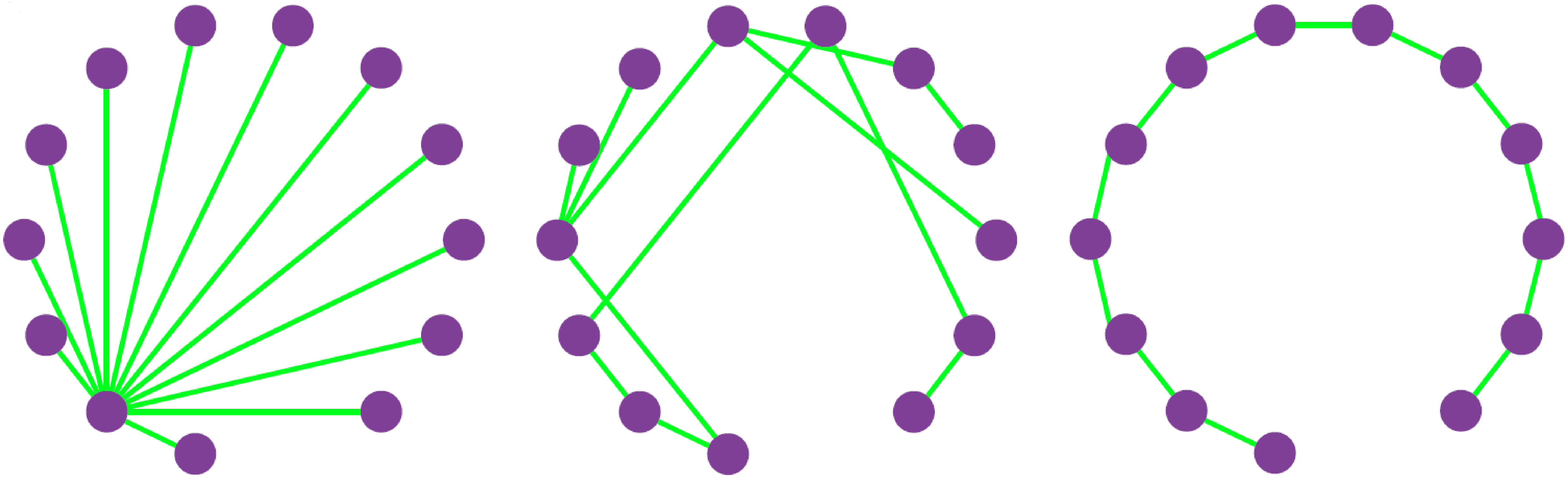
ν influences the network:

- Negative ν ($\nu < 0$)
  - enhances the role of the preferential attachment
  - In the extreme case, $\nu \rightarrow -\infty$ each new node connects to the oldest node, resulting in a hub and spoke topology.
- Positive ν
  - In the extreme case, $\nu \rightarrow \infty$ each node will connect to its immediate predecessor.
- $\nu > 1$
  - In this case, aging effect overcomes the role of preferential attachment.
  - Network looses its scale-free nature.



The Impact of Aging on Degree Exponent

Aging



$V = -10$ $V = 1$ $V = 10$

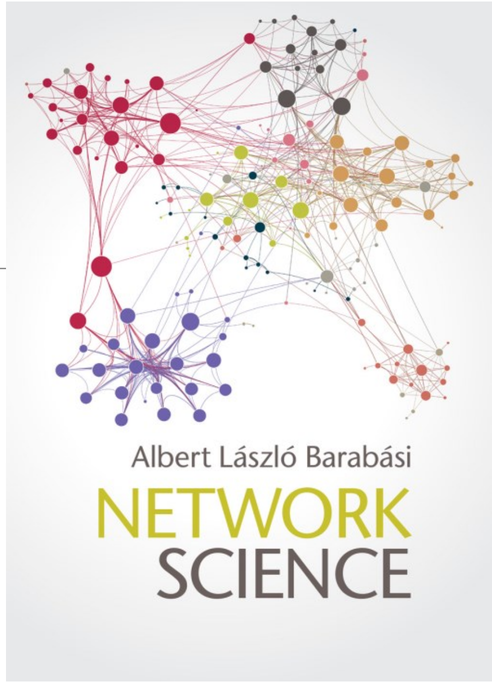
Network Analysis

08 – DEGREE CORRELATIONS

Slides were created by: Agnes Vathy-Fogarassy

Network Science book (online)

Barabási, Albert-László. *Network Science*.
Cambridge University Press, 2016.



Introduction

What is the common between the following celeb-pairs:

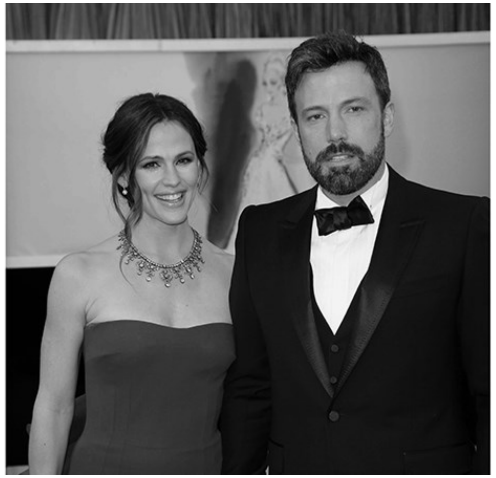
- Angelina Jolie and Brad Pitt
- Ben Affleck and Jennifer Garner
- Michael Douglas and Catherine Zeta-Jones
- Tom Cruise and Katie Holmes

They are married or were married.

- Who's Dated Who?

Why is it interesting?

- Number of eligible individuals: $\sim 10^8$
- List of celebrities: ~ 1000
- Probability they are married: $\sim 10^{-5}$



Introduction

If we do not care about the dating habits of celebrities, what this phenomenon tells us about the structure of the social network?

- Political leaders and CEOs: They know an exceptionally large number of individuals and they are known by even more. They are hubs.

Interesting property of the social networks:

- Hubs tend to have ties to other hubs.
- Is this true in other networks?

Introduction

If we do not care about the dating habits of celebrities, what this phenomenon tells us about the structure of the social network?

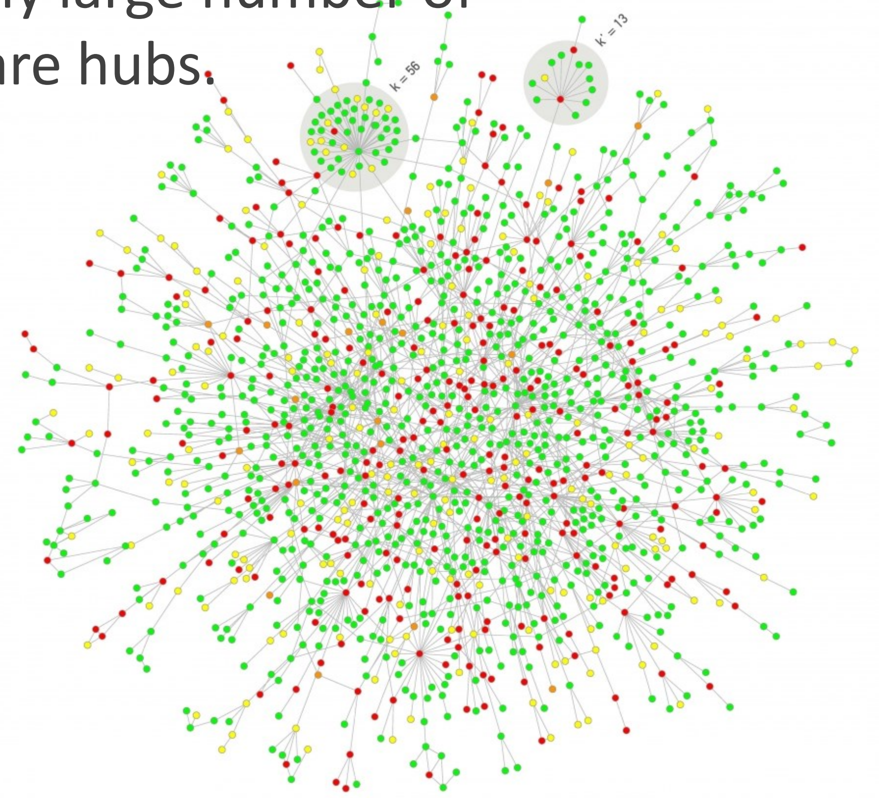
- Political leaders and CEOs: They know an exceptionally large number of individuals and they are known by even more. They are hubs.

Interesting property of the social networks:

- Hubs tend to have ties to other hubs.
- Is this true in other networks?

Counterexample: Protein-interaction network of yeast:

- $N = 1870, L = 2277$
- The two biggest hubs: $k = 56, k' = 13$
- Hubs link to many small-degree nodes.
- They generate hub-and-spoke patterns.



Introduction

Let's note the probability that the two hubs are connected to each other by:

- $p_{k,k'} = \frac{kk'}{2L}$
- In our case $p_{56,13} = 0.16$
- $(p_{1,2} = 0.0004)$

The number of links to nodes with small degree is surprising:

- $N_1 p_{1,56} \approx 12$ nodes
- So, we except that the node has 12 neighbours with $k = 1$.
- BUT: It has 46 neighbours with degree of one.

Summary:

- In case of social networks: hubs connect to hubs.
- In protein network: hubs avoid linking to hubs.
- We measure this phenomenon with degree correlations.

Assortativity and Disassortativity

The degree correlation matrix:

- e_{ij} - probability of the two ends of a randomly selected link has degrees i and j

What is the probability, that one end of a randomly selected link has degree k:

- $q_k = \frac{kp_k}{\langle k \rangle}$
- Connection to e_{ij}: $\sum_j e_{ij} = q_i$

Assortativity and Disassortativity

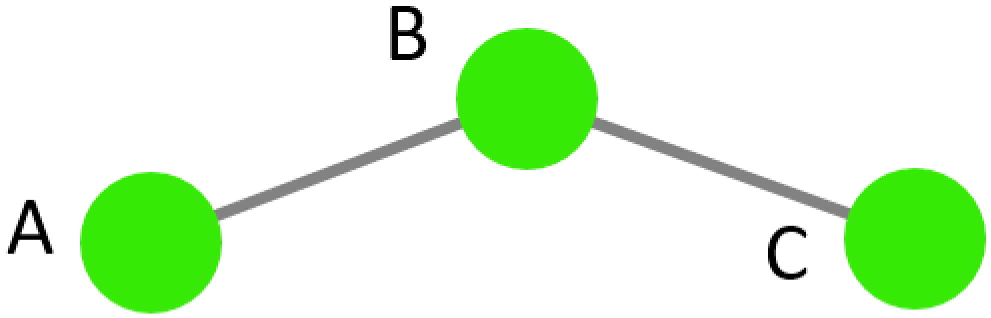
The degree correlation matrix:

- e_{ij} - probability of the two ends of a randomly selected link has degrees i and j

What is the probability, that one end of a randomly selected link has degree k:

- $q_k = \frac{kp_k}{\langle k \rangle}$
- Connection to e_{ij}: $\sum_j e_{ij} = q_i$

Let see an example for the following network:



Assortativity and Disassortativity

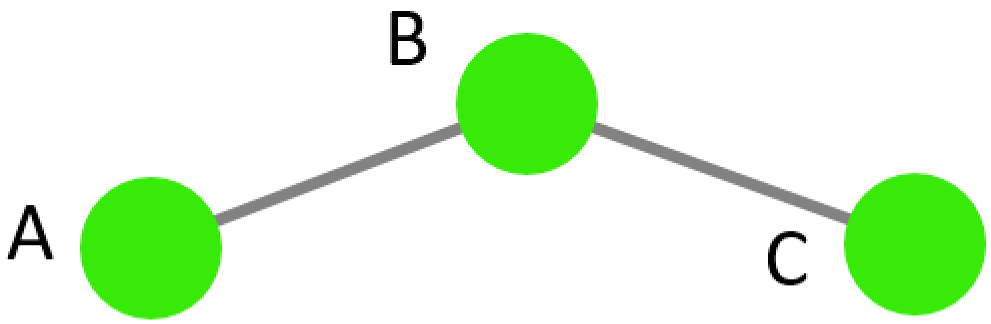
The degree correlation matrix:

- e_{ij} - probability of the two ends of a randomly selected link has degrees i and j

What is the probability, that one end of a randomly selected link has degree k:

- $q_k = \frac{kp_k}{\langle k \rangle}$
- Connection to e_{ij}: $\sum_j e_{ij} = q_i$

Let see an example for the following network:



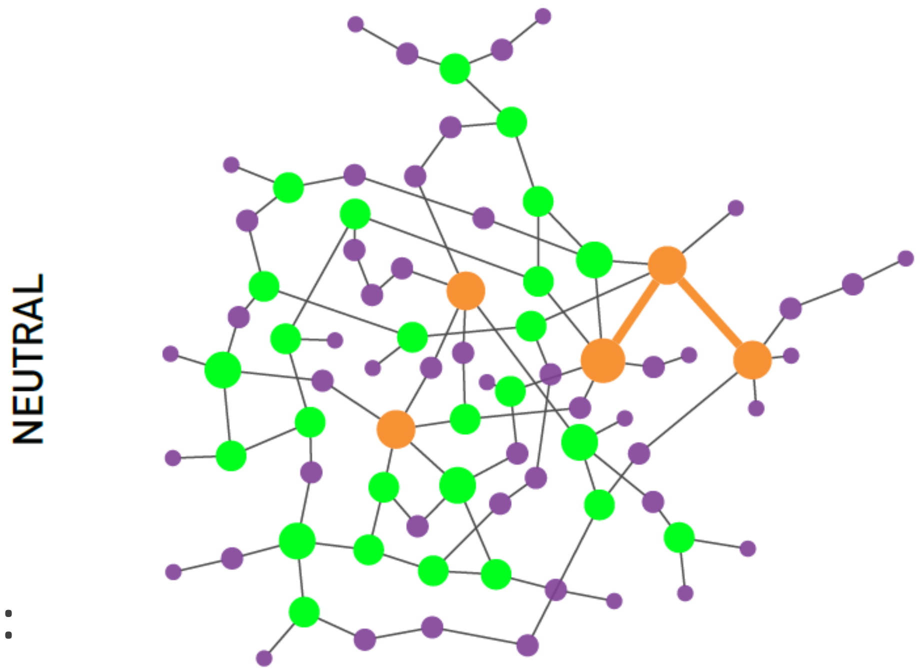
Three types of networks (based on degree correlation matrix):

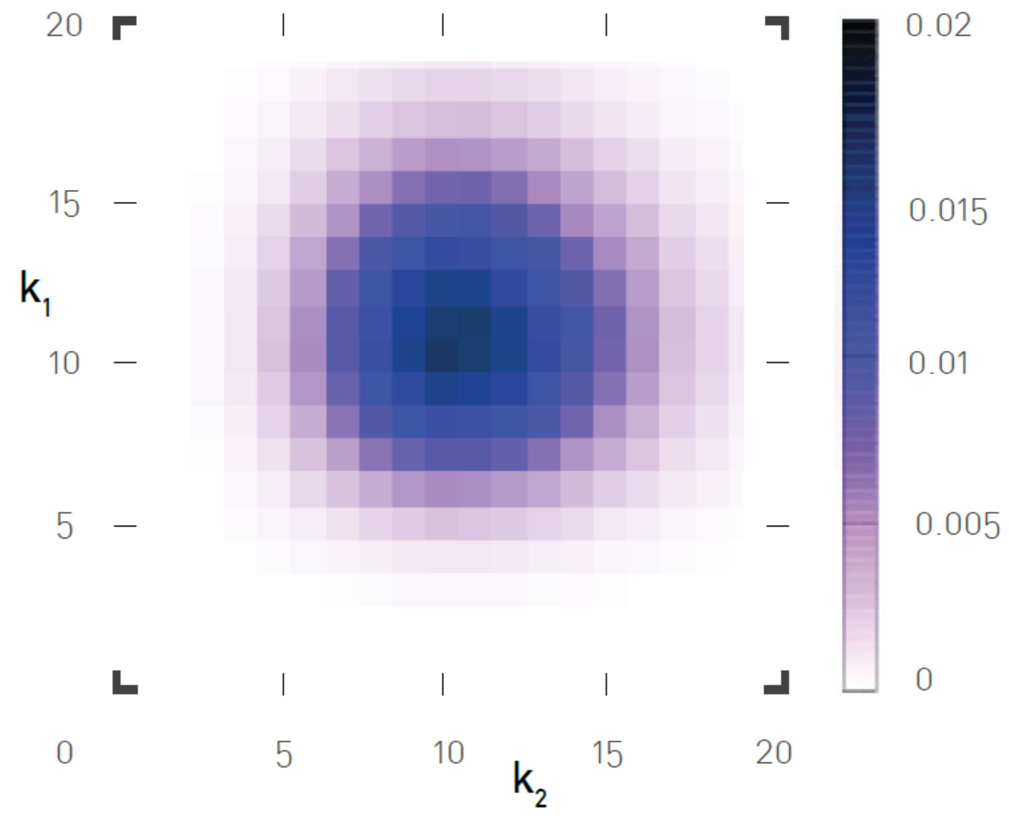
- Neutral
- Assortative
- Disassortative

Assortativity and Disassortativity

Neutral Network

- Connections are random





Colours:

5 biggest hubs

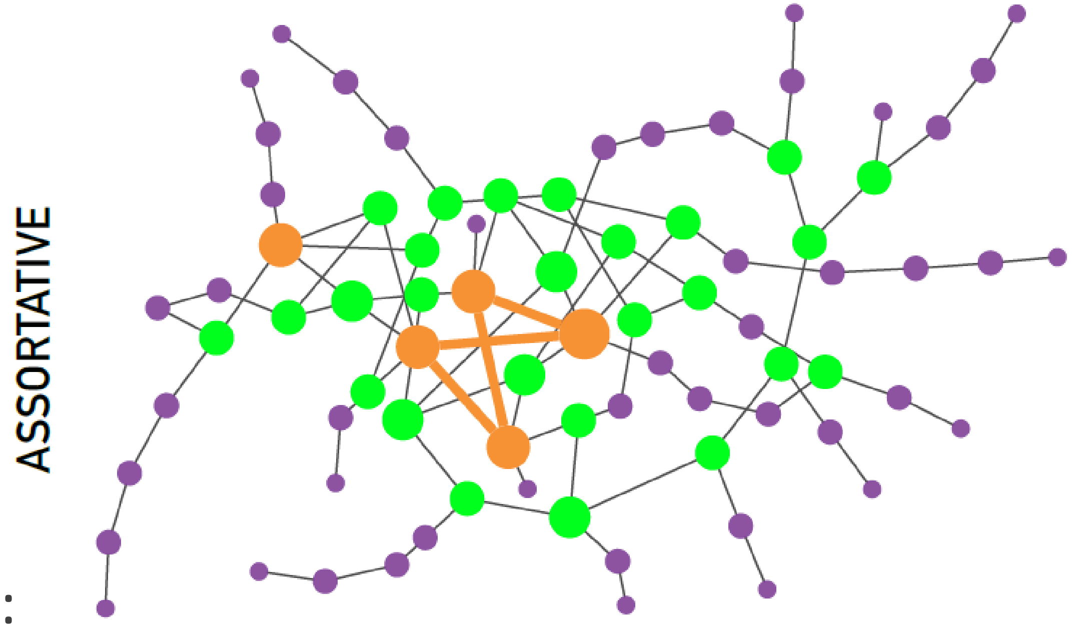
nodes with high degree

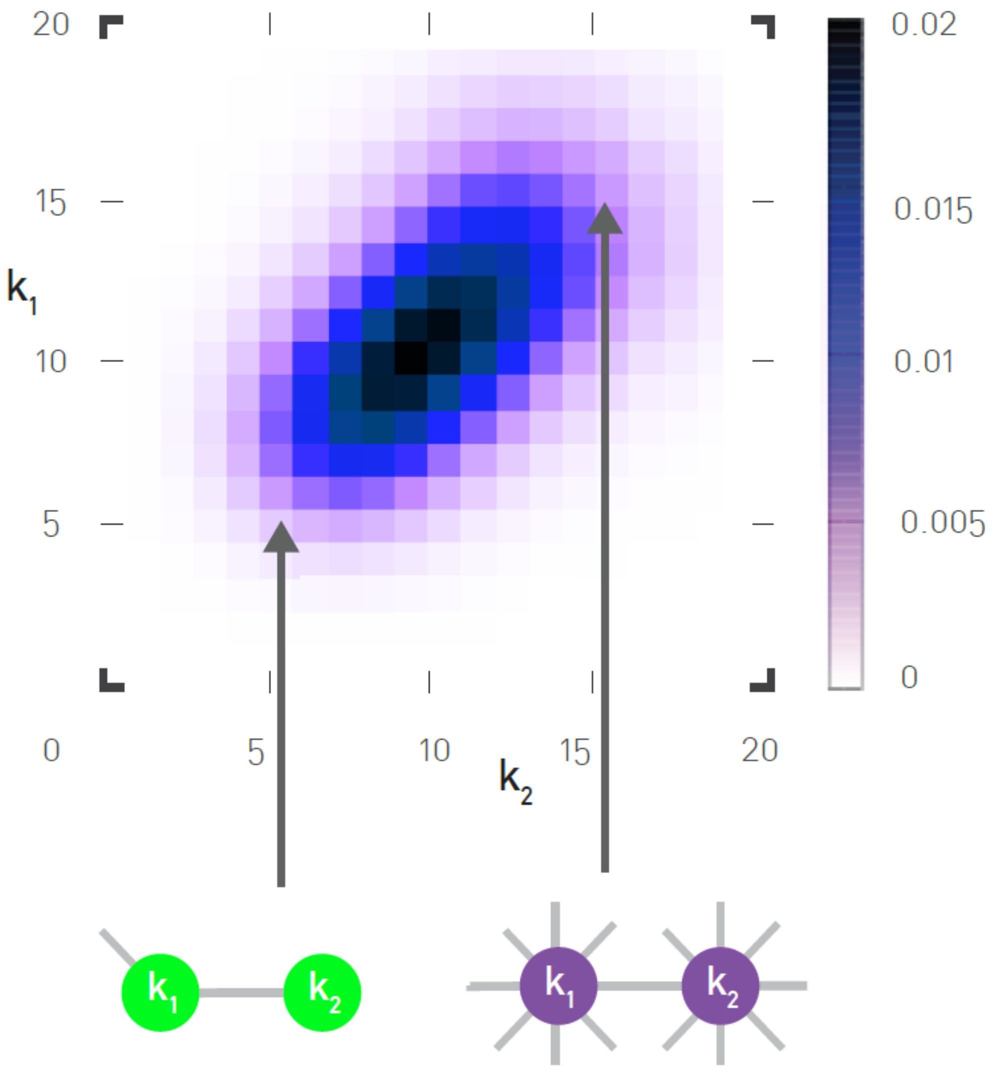
nodes with small degree

Assortativity and Disassortativity

Assortative Network

- Hubs connect to hubs





Colours:

5 biggest hubs

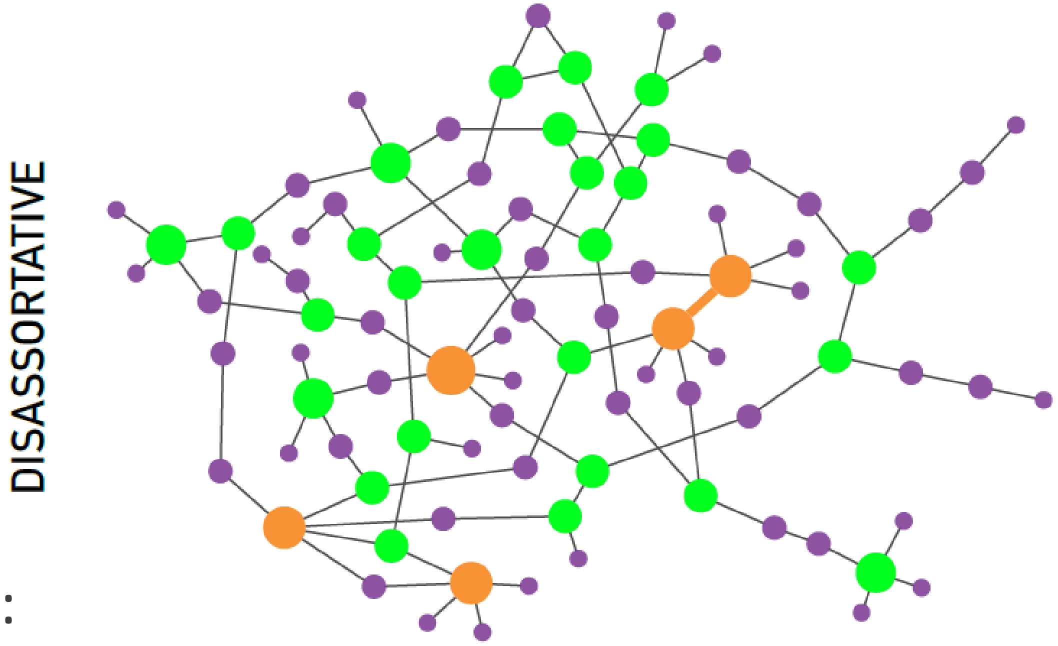
nodes with high degree

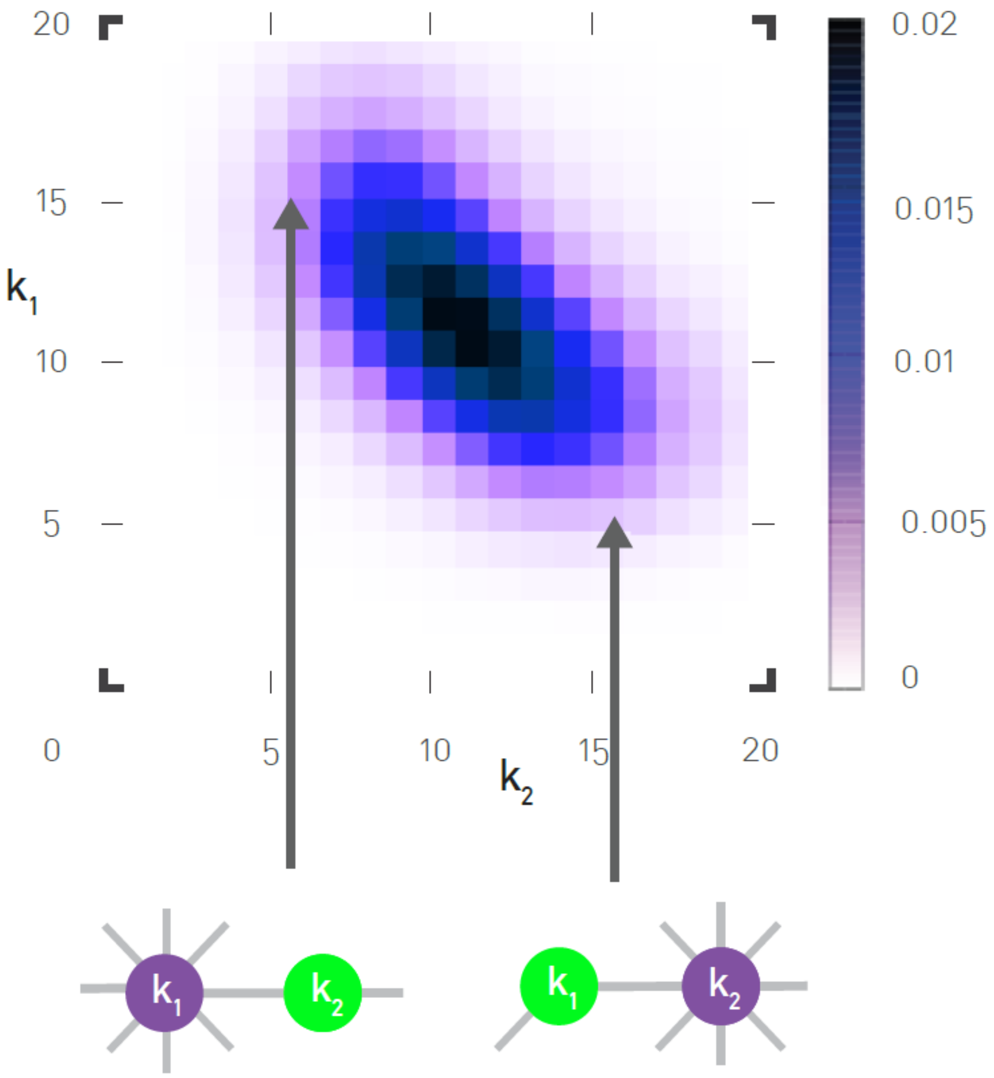
nodes with small degree

Assortativity and Disassortativity

Disassortative Network

- Hubs connect to nodes with small degree





Colours:

5 biggest hubs

nodes with high degree

nodes with small degree

Measuring Degree Correlations

Degree correlation function:

- $k_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} k_j$

  where $k_{nn}(k_i)$ is the k_{nn} value of node i (not of degree k_i).

For all nodes with degree k:

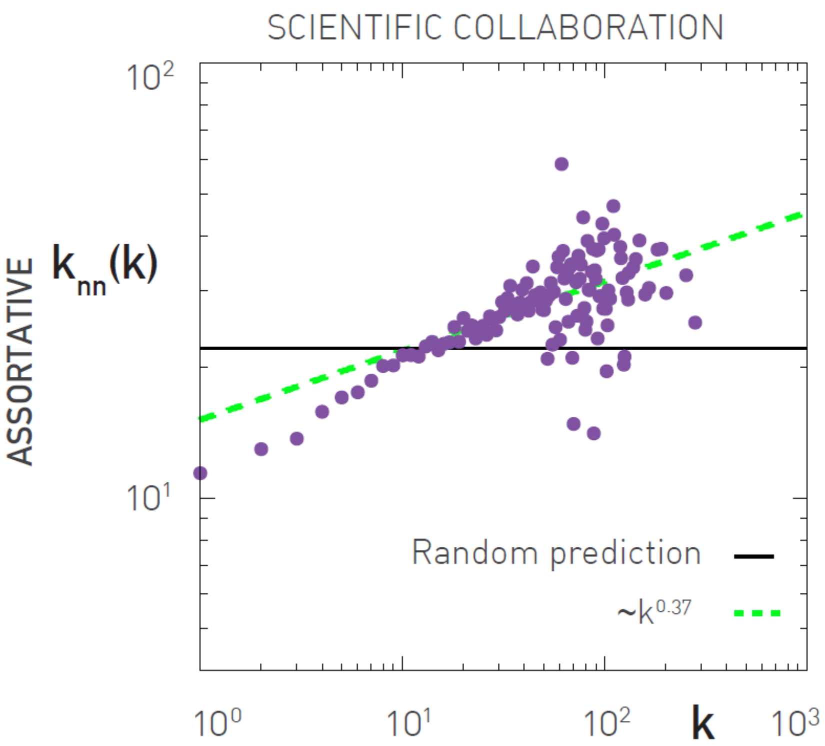
- $k_{nn}(k) = \sum_{k'} k' P(k'|k)$

  $P(k'|k)$ means the conditional probability that following a link of a k-degree node we reach a degree-k' node.

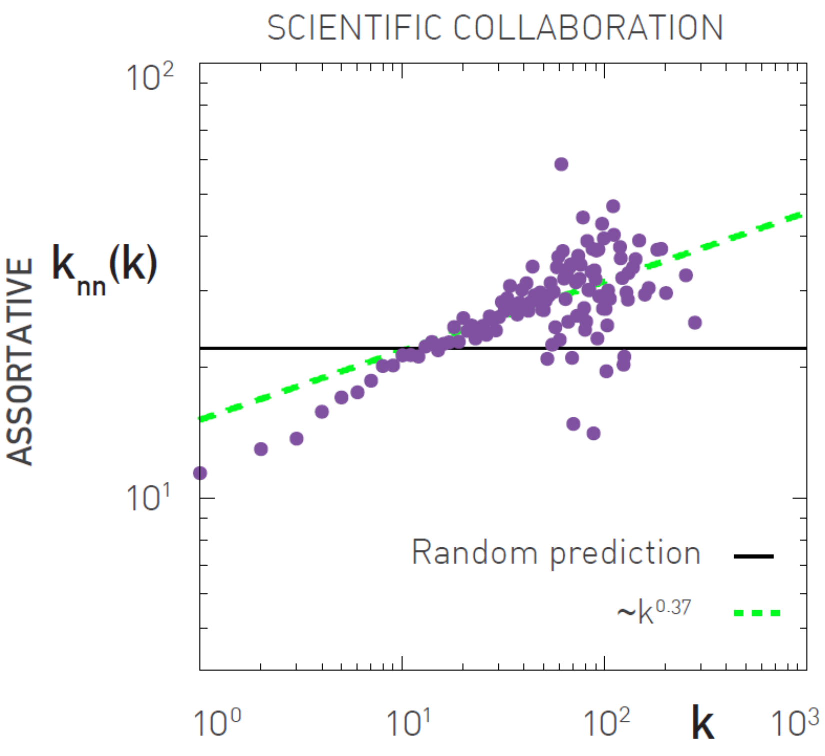
$k_{nn}(k)$ can be predicted by:

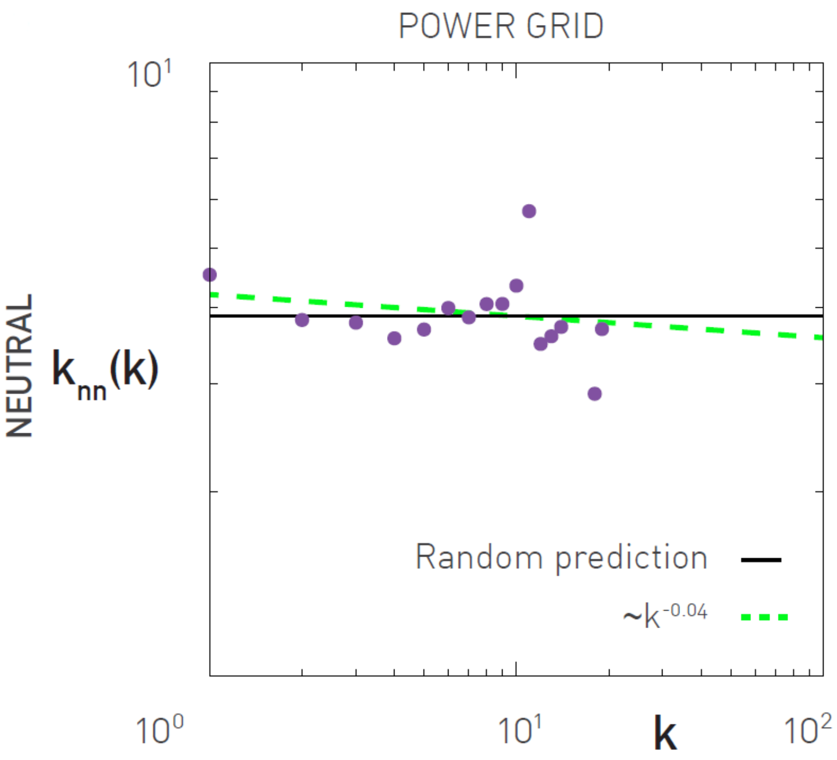
- $k_{nn}(k) = ak^{\mu}$, where μ is the correlation exponent:
  - Assortative: $\mu > 0$
  - Neutral: $\mu = 0$
  - Disassortative: $\mu < 0$

Measuring Degree Correlations

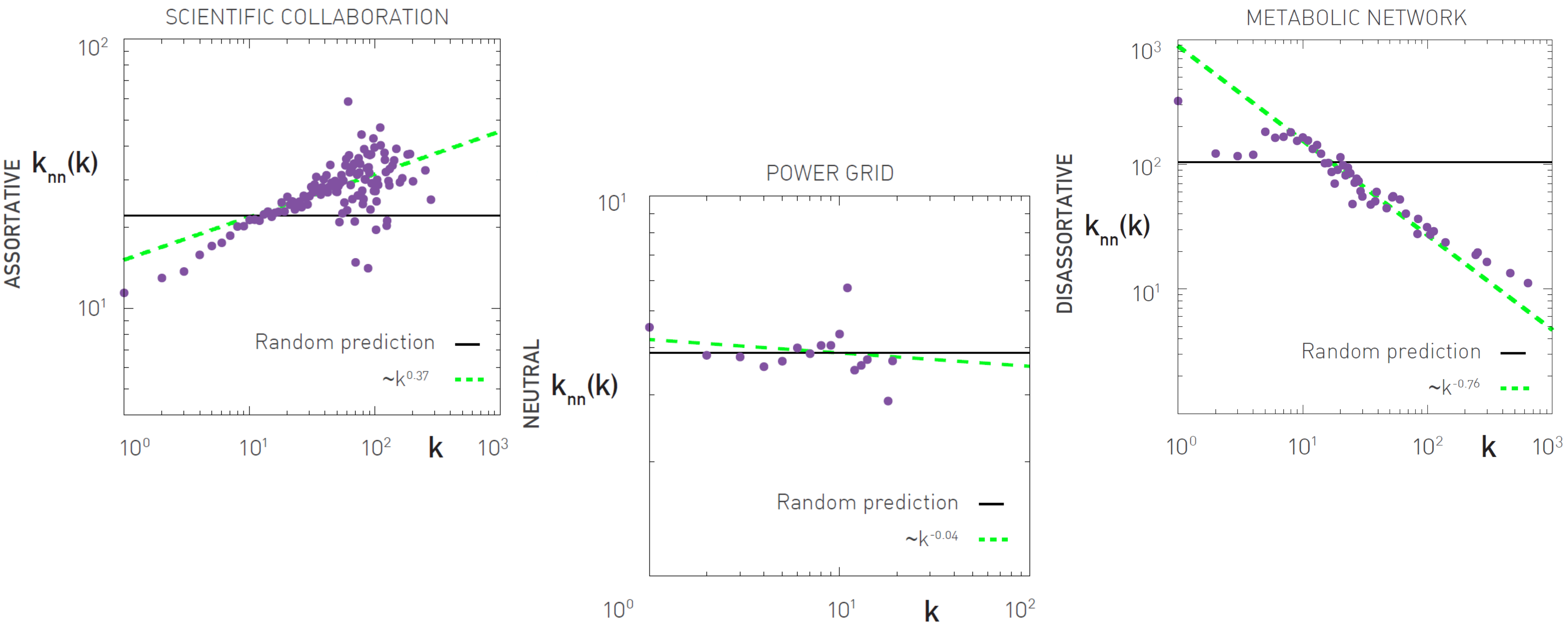


Measuring Degree Correlations





Measuring Degree Correlations



Degree Correlation Coefficient

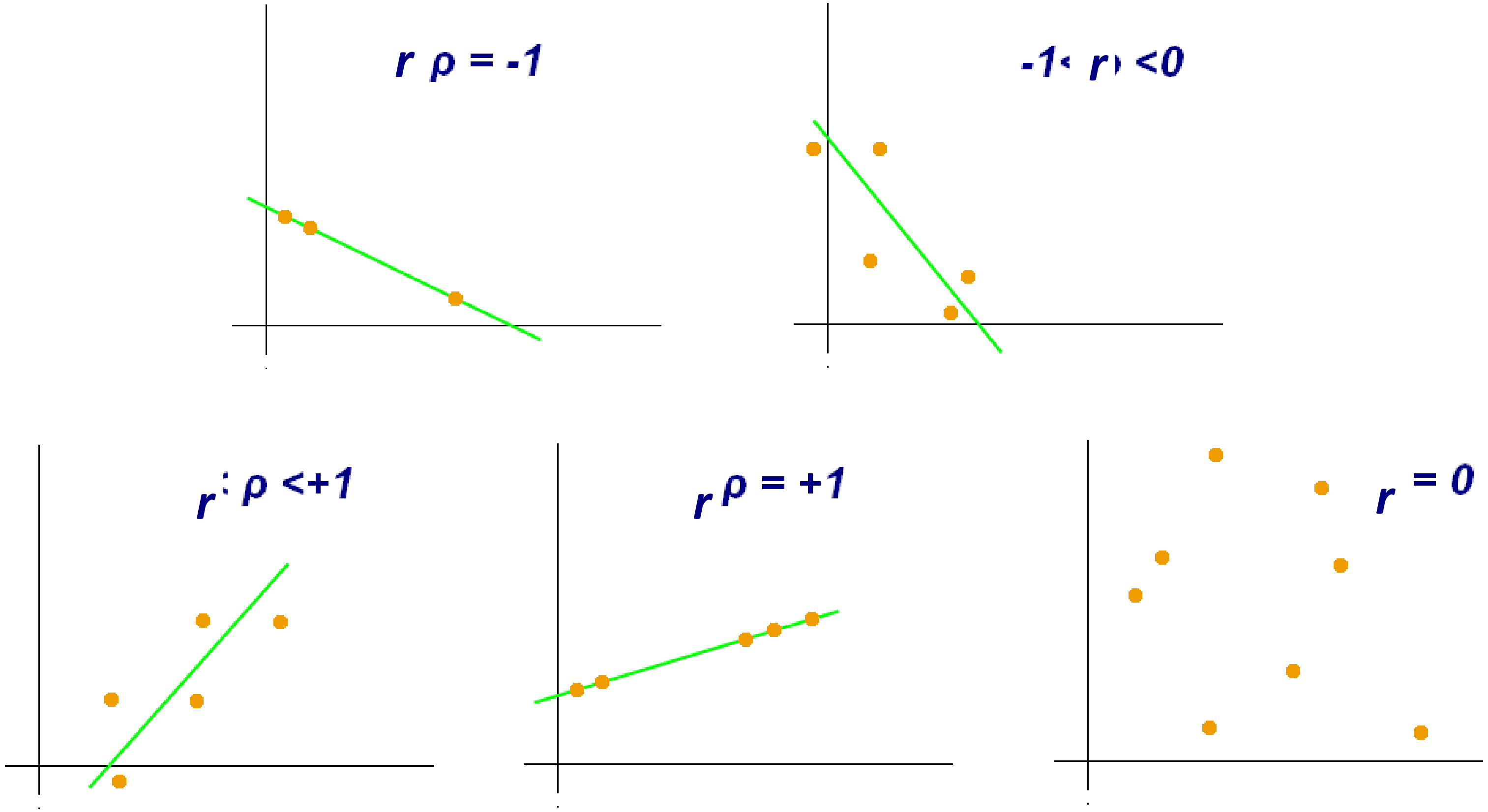
Degree Correlation Coefficient:

- Enables to characterise the network with a single number
- $r = \sum_{jk} \frac{jk(e_{jk} - q_j q_k)}{\sigma^2}$
- $\sigma^2 = \sum_k k^2 q_k - [\sum_k k q_k]^2$
- $-1 \leq r \leq 1$

The network is:

- disassortative, if $\mathrm{r} < 0$
- neutral, if $\mathrm{r} = 0$
- assortative, if $\mathrm{r} > 0$

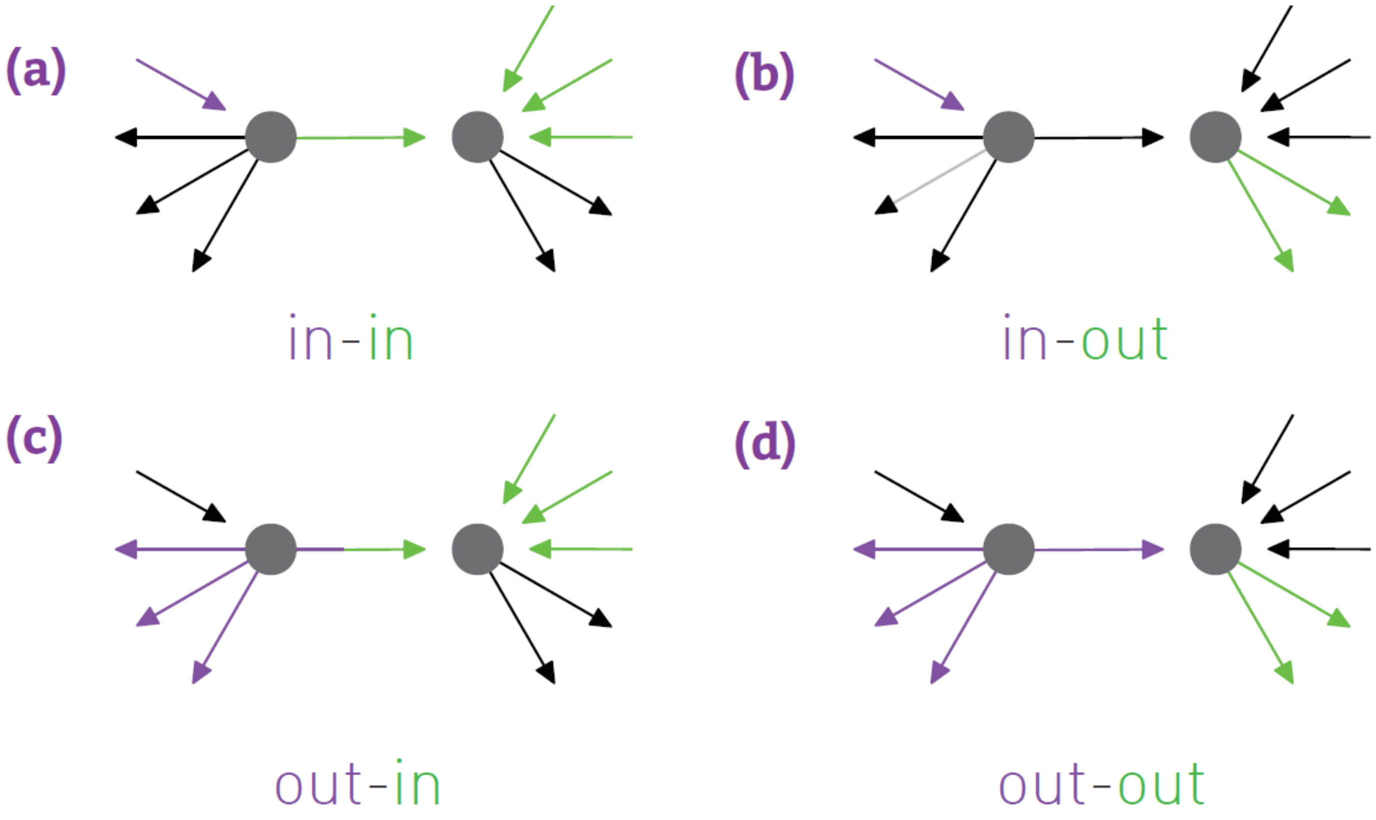
This coefficient is often called as Pearson coefficient.

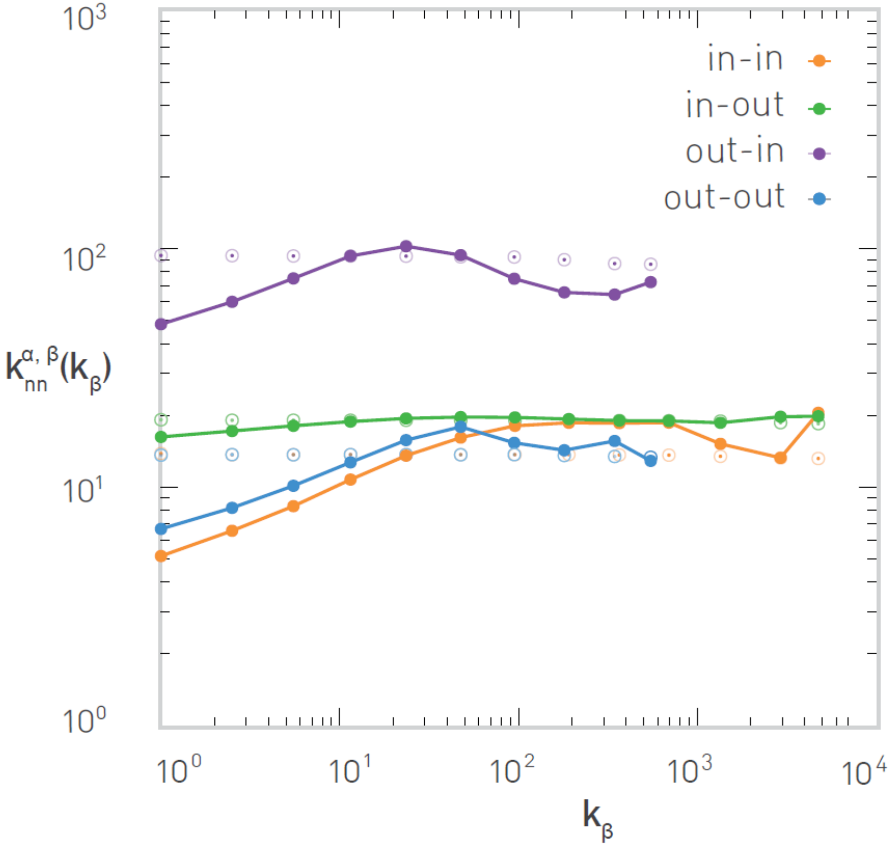


Correlations in Directed Networks

In directed networks:

- $k_{nn}^{\alpha,\beta}(k)$ is defined, where α and β refer to the in and out indices.





Xalvi-Brunet & Sokolov algorithm

Generates networks with desired degree correlations.

Step 1: Choose at random two links. Label the four nodes of these two links with a, b, c, and d such that their degrees are ordered as: $k_a \geq k_b \geq k_c \geq k_d$.

Step 2: Break the selected links and rewire them to form new pairs.

Step 2A: To achieve an *assortative* network:

- Pairing the two highest degree nodes (a with *b*) and the two lowest degree nodes (*c* with *d*).

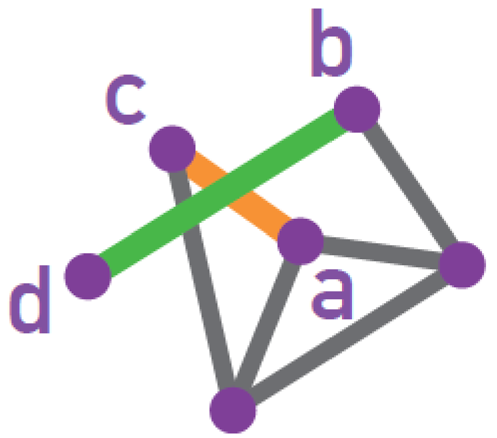
Step 2B: To achieve *disassortative* network:

- Pairing the highest and the lowest degree nodes (a with *d* and *b* with *c*).

Xalvi-Brunet & Sokolov algorithm

STEP 1 LINK SELECTION

STEP 2 REWIRE



$k_a \geq k_b \geq k_c \geq k_d$

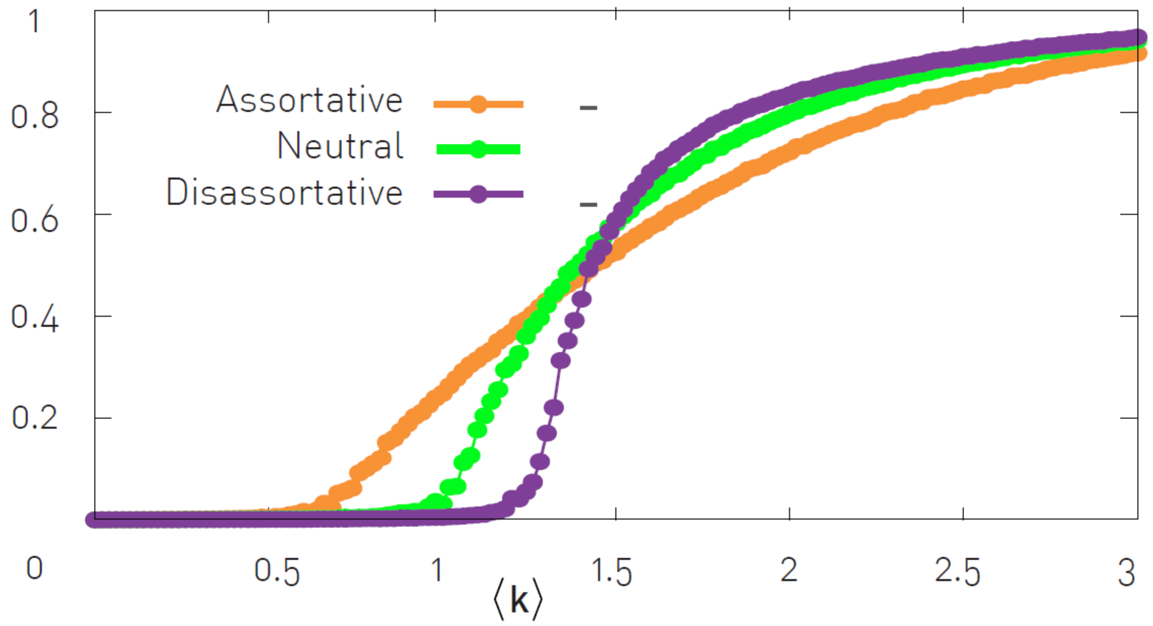
ASSORTATIVE

DISASSORTATIVE

The Impact of Degree Correlations

Giant component:

- Assortative network:
  - Phase transition point is smaller ($\langle k \rangle < 1$)
- Neutral network:
  - Erdős-Rényi network, $\langle k \rangle = 1$
- Disassortative network:
  - The phase transition point is delayed ($\langle k \rangle > 1$)



Why is it important? The giant component influences:

- Spread of disease
- Robustness of the network
  - Assortative networks are more robust
  - Disassortative networks are less robust

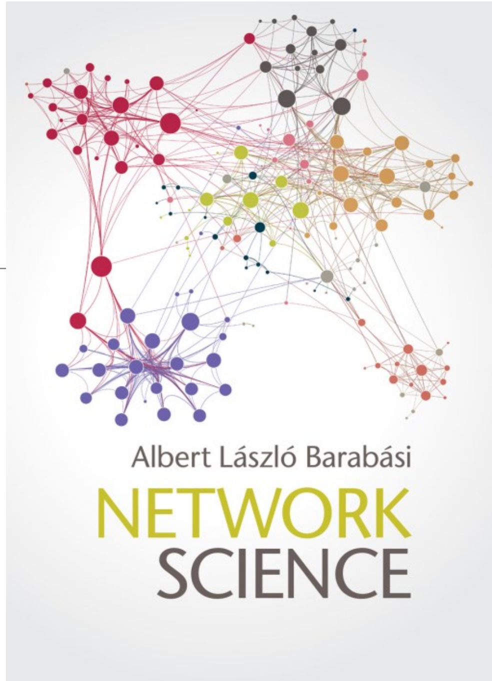
Network Analysis

09 – NETWORK ROBUSTNESS

Slides were created by: Daniel Leitold

[Network Science book (online)]

Barabási, Albert-László. *Network Science*.
Cambridge University Press, 2016.



Introduction

Errors and failures can corrupt all human designs:

- Failure of a component in your car's engine may force you to call for a tow truck.
- Wiring error in your computer chip can make your computer useless.

In natural and social systems:

- While there are countless protein misfolding errors and missed reactions in our cells, we rarely notice their consequences.
- Large organizations can function despite numerous absent employees.

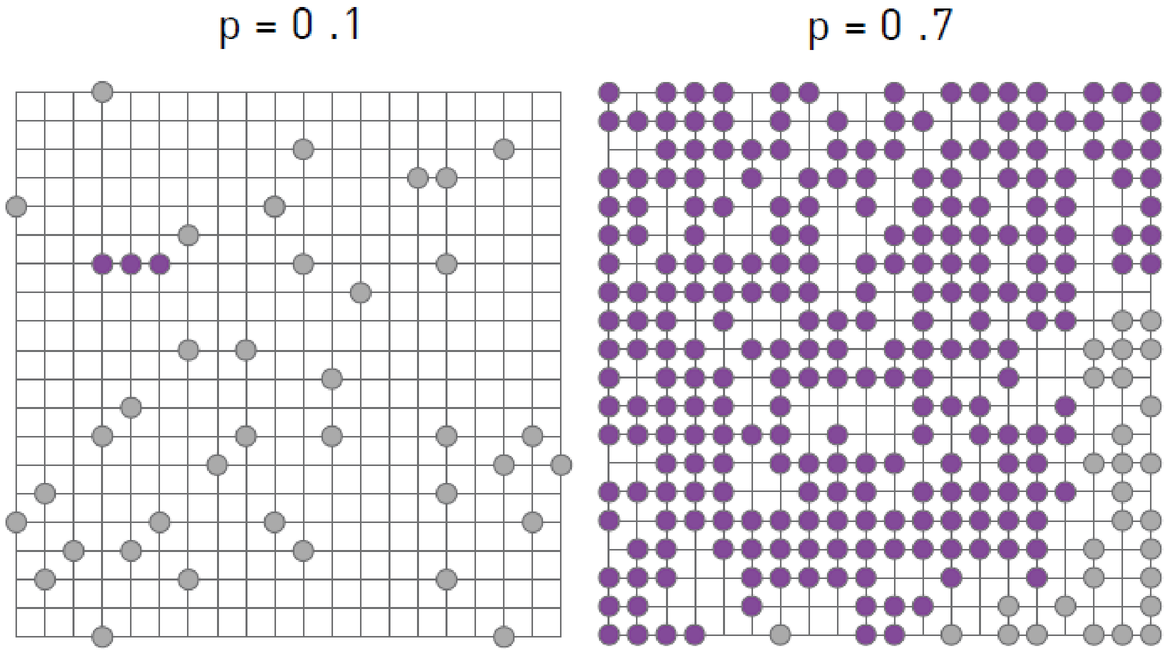
"Robust" comes from the latin Quercus Robur, meaning oak, the symbol of strength and longevity in the ancient world.

Percolation Theory

Percolation theory is a highly developed subfield of statistical physics and mathematics.

A typical problem addressed by the illustration:

- showing a square lattice
- we place pebbles with probability p at each intersection
- neighbouring pebbles are considered connected, forming clusters



Questions:

- What is the expected size of the largest cluster?
- What is the average cluster size?

Percolation Theory

A key prediction of percolation theory is that the cluster size does not change gradually with p.

- For a wide range of p the lattice is populated with numerous tiny clusters.
- If p approaches a critical value p_c, these small clusters grow and coalesce, leading to a large cluster.

Percolation Theory

A key prediction of percolation theory is that the cluster size does not change gradually with p.

- For a wide range of p the lattice is populated with numerous tiny clusters.
- If p approaches a critical value p_c, these small clusters grow and coalesce, leading to a large cluster.

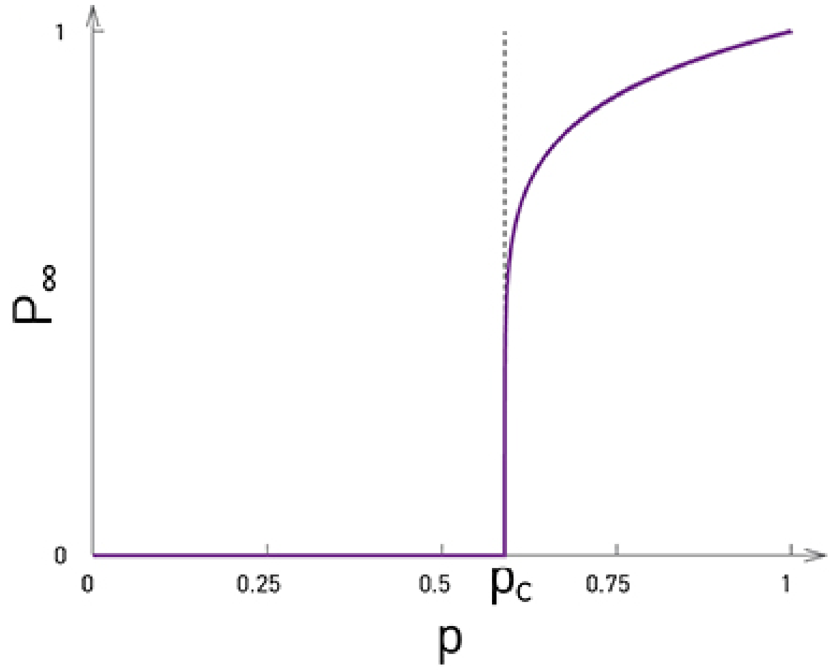
Three main quantities:

- $\langle S \rangle$: average size of all finite clusters.
- P_∞: order parameter, probability that a randomly chosen pebble belongs to the largest cluster.
- ξ: mean distance between two pebbles that belong to the same cluster.

Percolation Theory

A key prediction of percolation theory is that the cluster size does not change gradually with p.

- For a wide range of p the lattice is populated with numerous tiny clusters.
- If p approaches a critical value p_c, these small clusters grow and coalesce, leading to a large cluster.



Three main quantities:

- $\langle S \rangle$: average size of all finite clusters.
- P_∞: order parameter, probability that a randomly chosen pebble belongs to the largest cluster.
- ξ: mean distance between two pebbles that belong to the same cluster.

Inverse Percolation Transition and Robustness

Let us view a square lattice as a network whose nodes are the intersections.

Then, remove a fraction f of nodes randomly

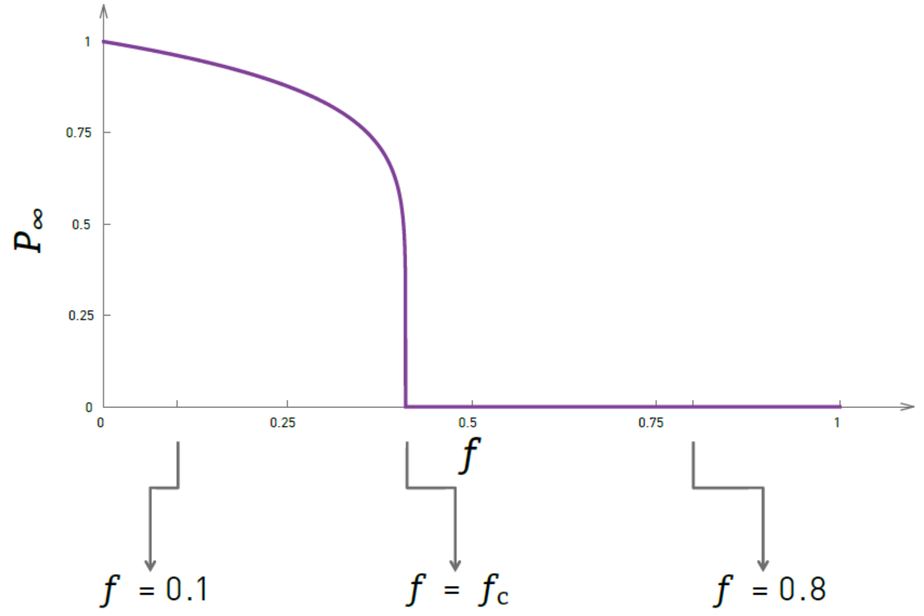
- If f is small, the damage is little.
- Increasing f can isolate chunks of nodes.
- For large f the giant component breaks into tiny disconnected components.

This fragmentation process is not gradual

- It is characterized by a critical threshold f_c

Summary

- Breakdown of a random network under random node removal is not a gradual process.



Robustness of Scale-free Networks

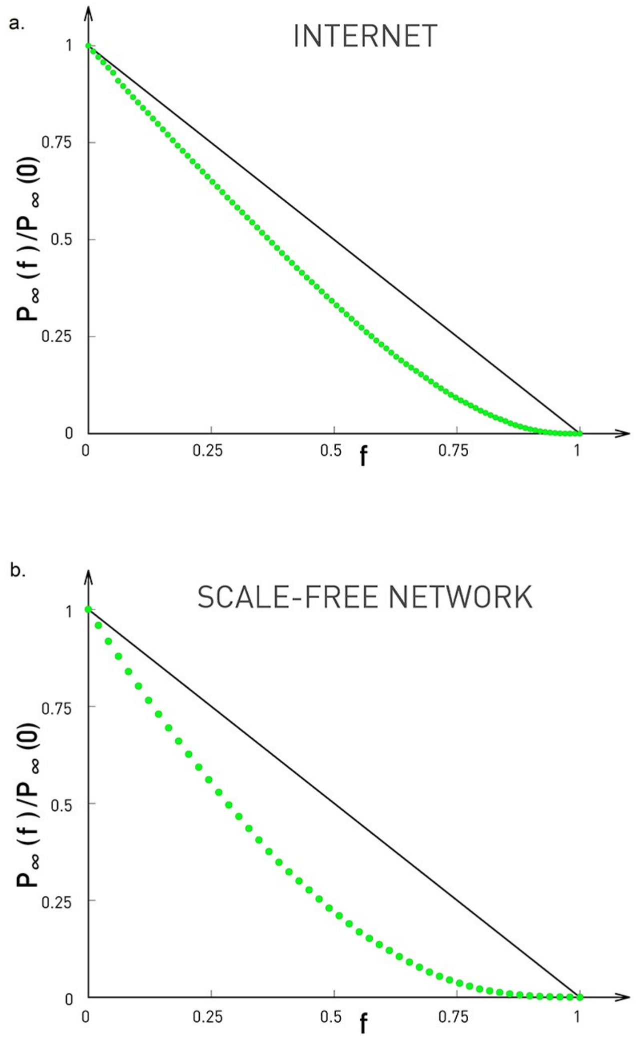
Percolation theory focuses mainly on regular lattices.

But:

- Internet refuses to break apart even in case of dramatic number of node failures.
- For a scale-free network with degree exponent $\gamma = 2.5$, identical pattern can be observed.

In case of random node removal the giant component fails to collapse at some finite f_c

- Giant component vanishes if f is close to 1.
- Video



Molloy-Reed Criterion

We need to determine f_c to the scale free network.

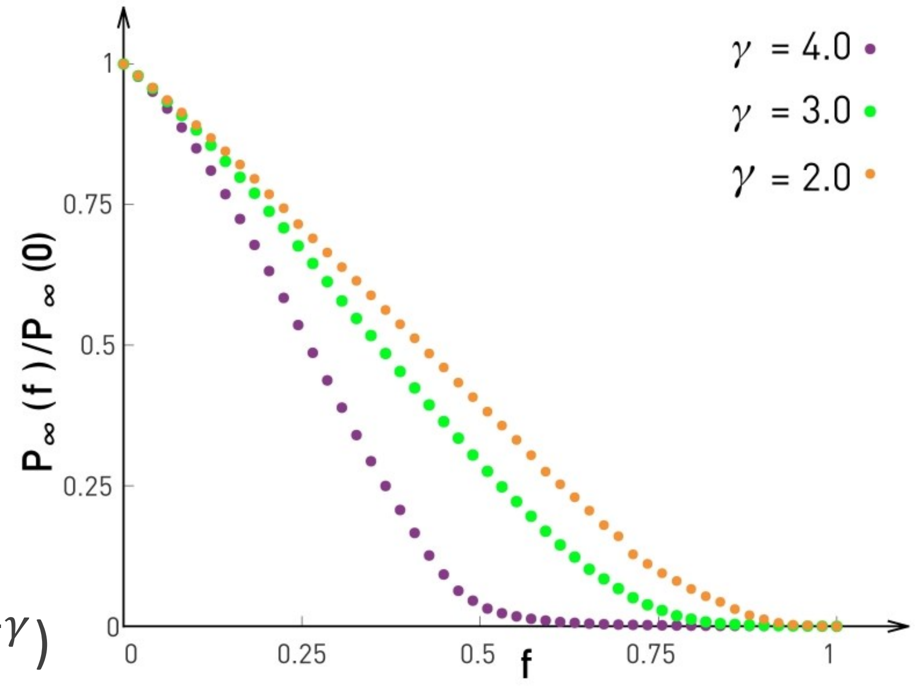
We need to determine if there is a giant component in the network.

Molloy-Reed criterion

- There is a giant component in the network if
- $\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} > 2$
- If $\kappa < 2$, then there is no giant component in the network.

Critical threshold:

- $f_c = 1 - \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1}$
- In a scale-free network, $\langle k \rangle$ depends on γ ($p_k = k^{-\gamma}$)



Robustness of Finite Networks

Scale-free networks are more robust than random networks

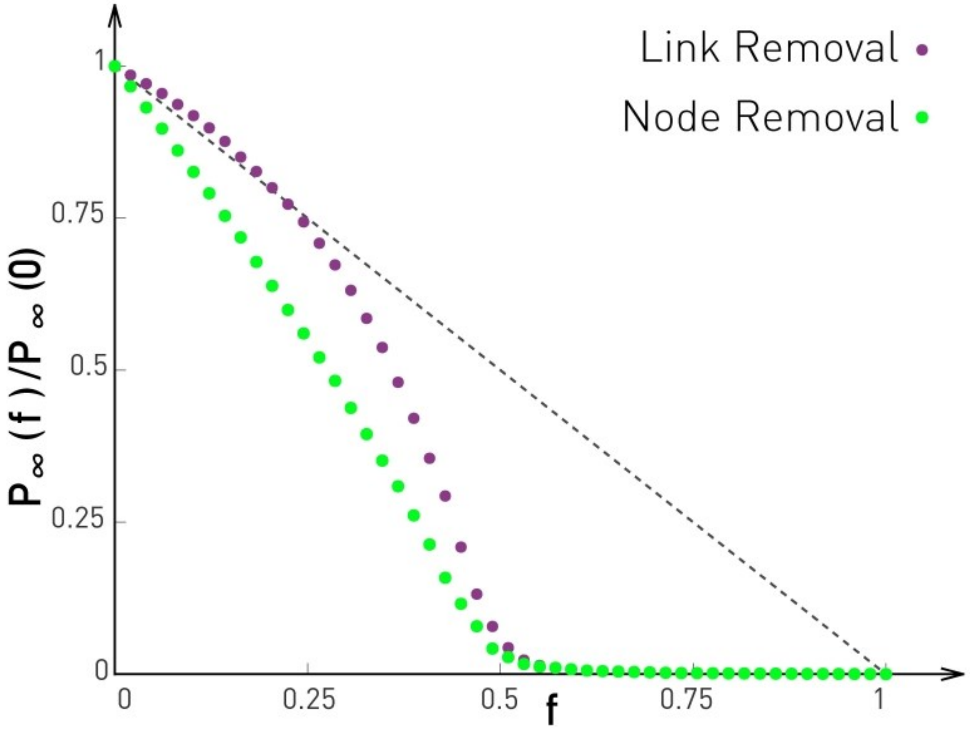
- $f_c > f_c^{ER}$

In case of Internet

- $f_c = 0.972$
- $N = 192\,244$
- $97\% \rightarrow 186\,861$ routers should fail simultaneously.

The enhanced robustness is valid for

- Nodes
- Links



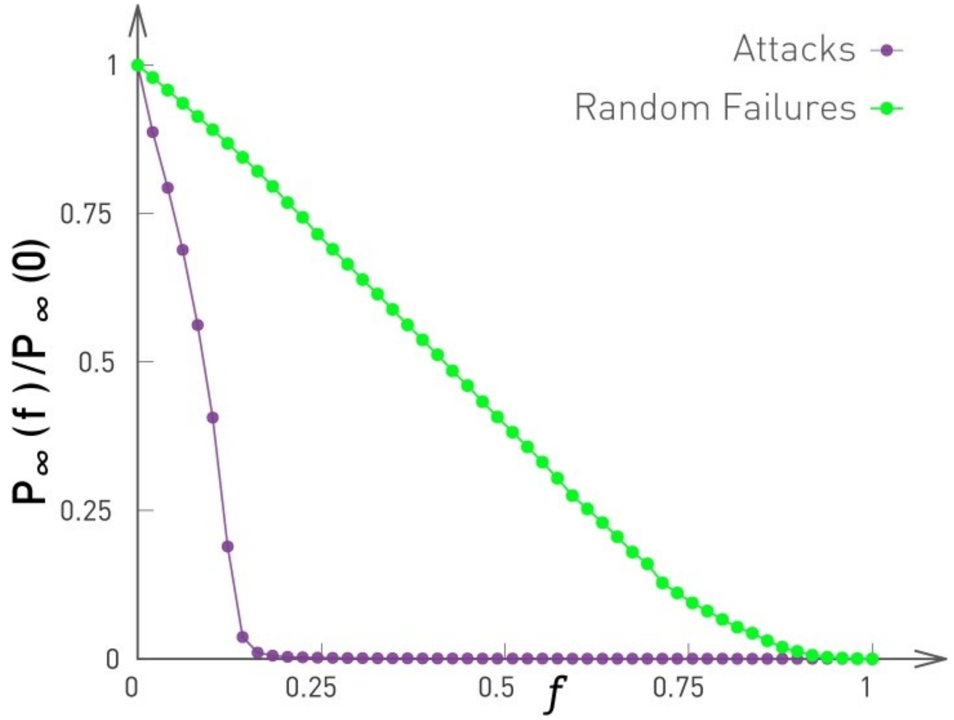
Attack Tolerance

Hubs play important role in holding the network together.

- What if we remove the hubs?
- The likelihood that nodes would break in this descending order by degree under normal conditions is essentially zero.

An attack

- Assumes a detailed knowledge of the network topology
- An ability to target the hubs
- And a desire to deliberately cripple the network



[Video]

Critical Threshold Under Attack

An attack on a scale-free network has two consequences:

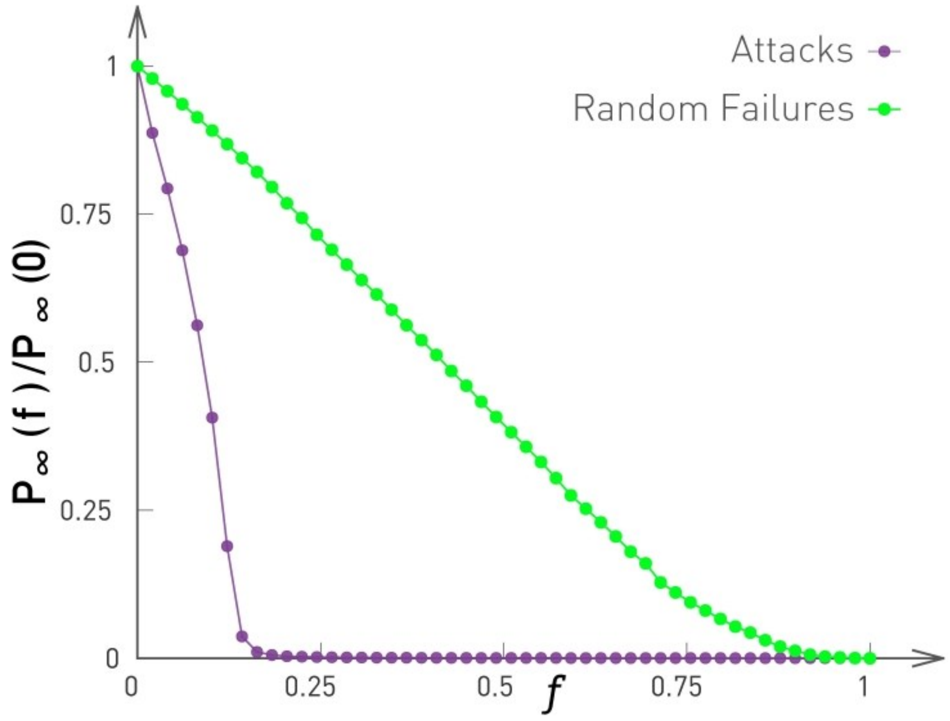
- The critical threshold f_c is smaller than $f_c = 1$, indicating that under attacks a scale-free network can be fragmented by the removal of a finite fraction of its hubs.
- The observed f_c is remarkably low, indicating that we need to remove only a tiny fraction of the hubs to cripple the network.

The results of the removed hubs:

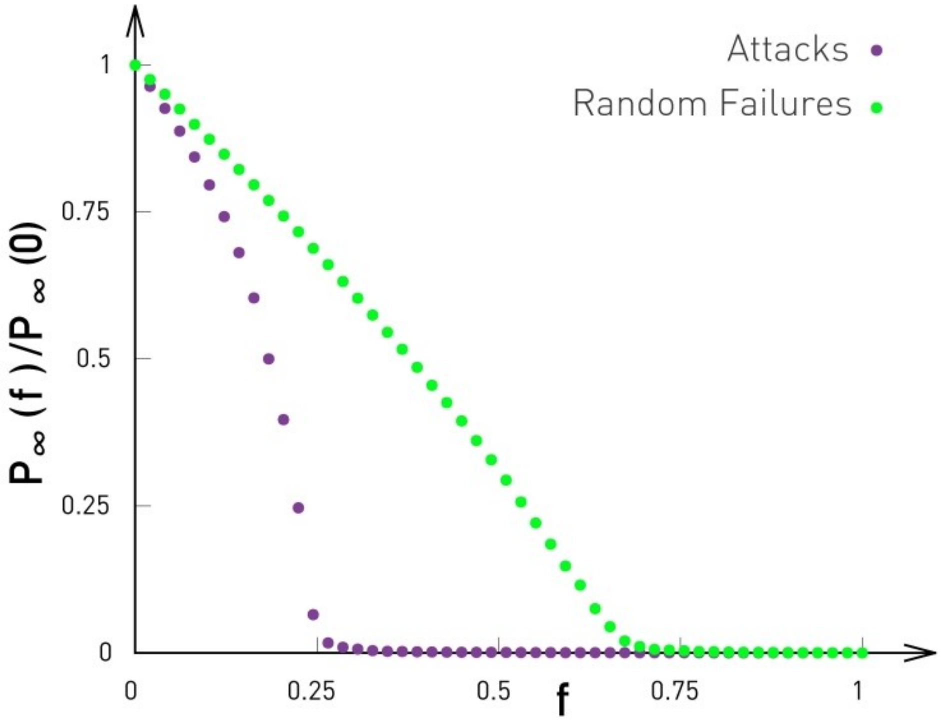
- It changes the maximum degree of the network from $k_{max}$to $k'_{\max}$ as all nodes with degree larger than $k'_{\max}$ have been removed.
- The degree distribution of the network changes from p_k to $p'_{k'}$, as nodes connected to the removed hubs will loose links, altering the degrees of the remaining nodes.

By combining these two changes we can map the attack problem into the robustness problem discussed in the previous section.

Attacks and Failures in Random and Scale-free Networks



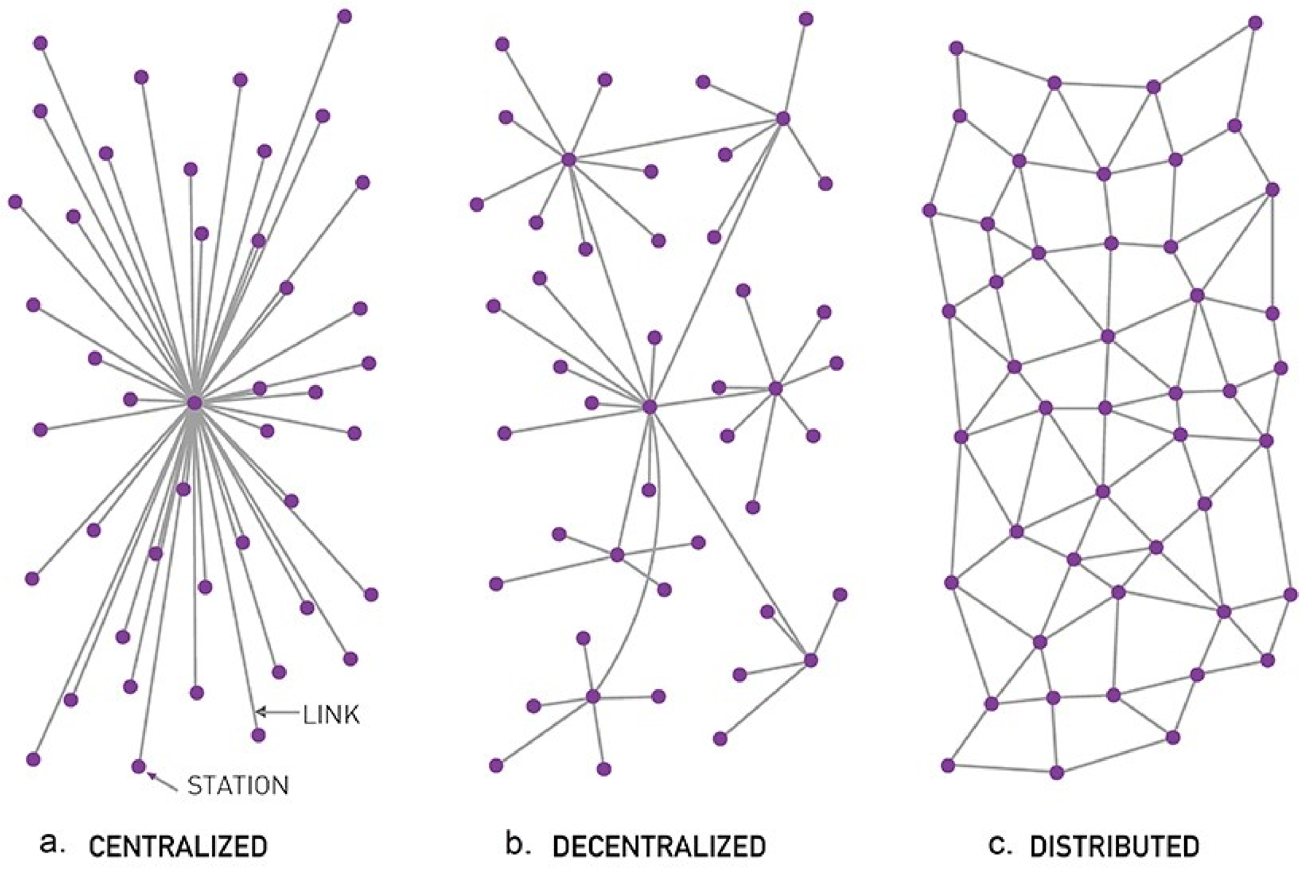
Scale-free network



Random network

Possible configurations of communication networks

Envisioned by Paul Baran in 1959. (Paul Baran was assigned to develop a communication system that can survive a Soviet nuclear attack.)



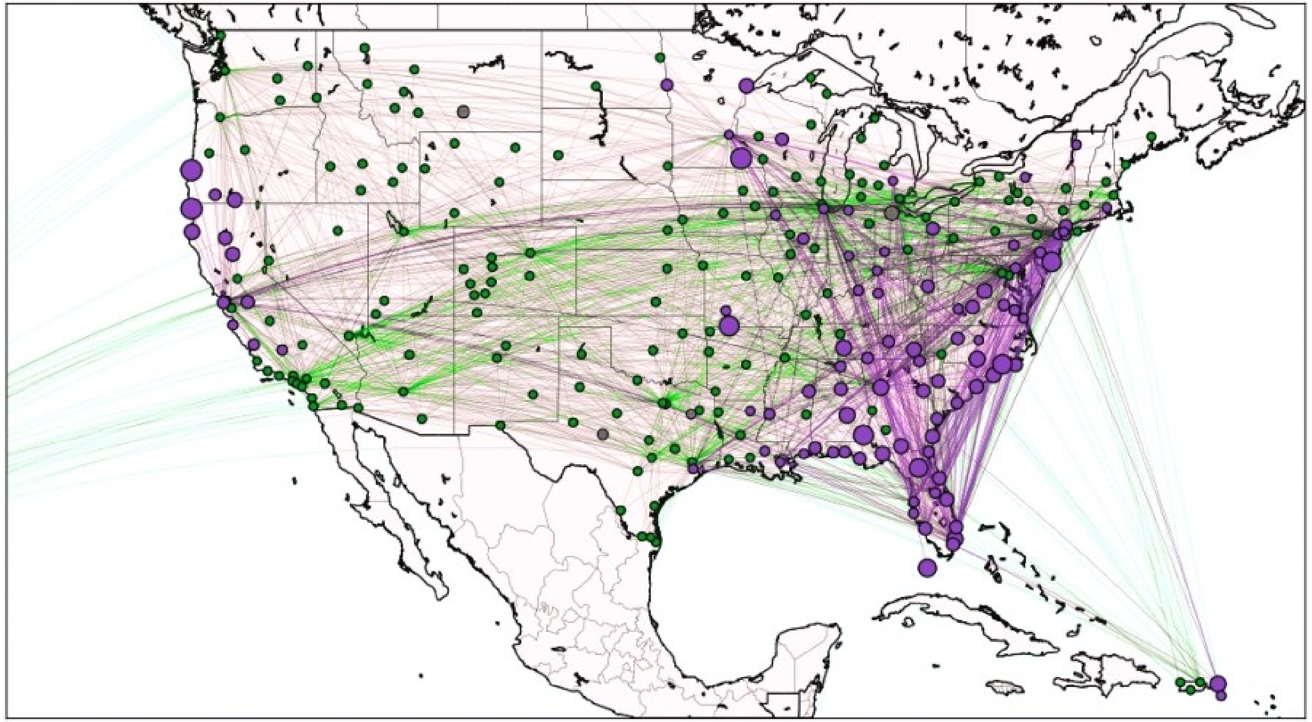
Cascading Failures

So far we have assumed that each node failure is a random event, hence the nodes of a network fail independently from each other.

In reality, in a network the activity of each node depends on the activity of its neighbouring nodes.

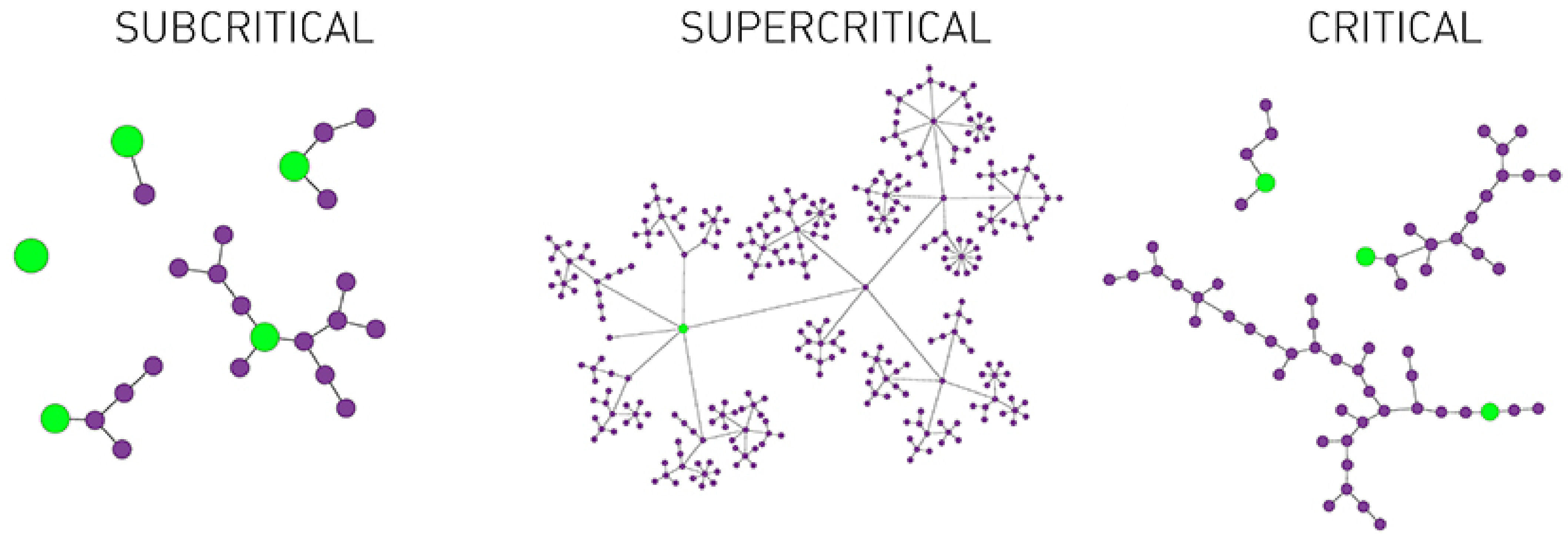
Real examples:

- Blackouts (Power Grid)
- Denial of Service Attacks (Internet)
- Financial Crises
- Flight delays
  - Have an economic impact of over $40 billion per year
    - congested airports
    - normal traffic



Three phase of cascading networks

Based on average degree, three phases can be determined. (Section 8.6)



$\langle k \rangle < 1$ $\langle k \rangle > 1$ $\langle k \rangle = 1$

Building Robustness

Can we maximize the robustness of a network to both random failures and targeted attacks without changing the cost?

Cost to build and maintain a network is:

- Proportional to average degree $\langle k \rangle$

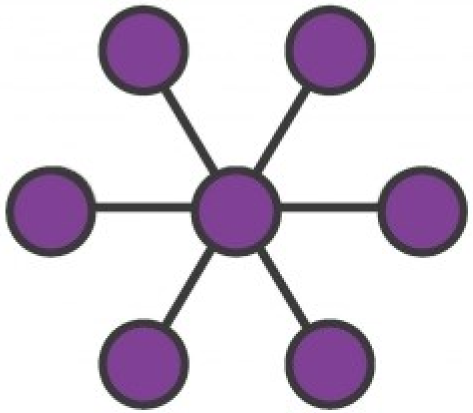
In order to enhance network robustness:

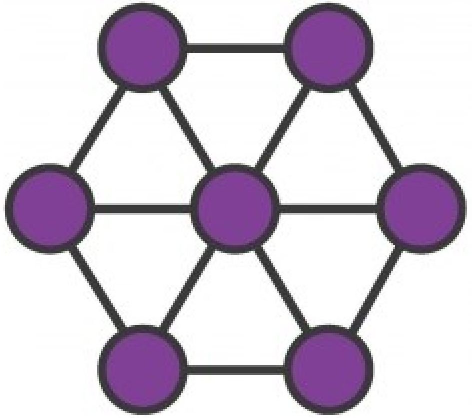
- We must increase f_c
- But f_c depends on $\langle k \rangle$ and $\langle k^2 \rangle$
- Thus, we need to maximize $\langle k^2 \rangle$, if we wish to keep the cost $\langle k \rangle$ fixed.

To maximize $\langle k^2 \rangle$

- Two type of nodes:
  - With k_{min}
  - With k_{max} ($k_{\max} = AN^{\frac{2}{3}}, A = \frac{\left(2\langle k\rangle^2(\langle k\rangle-1)^2\right)^{\frac{1}{3}}}{2\langle k\rangle-1}$)

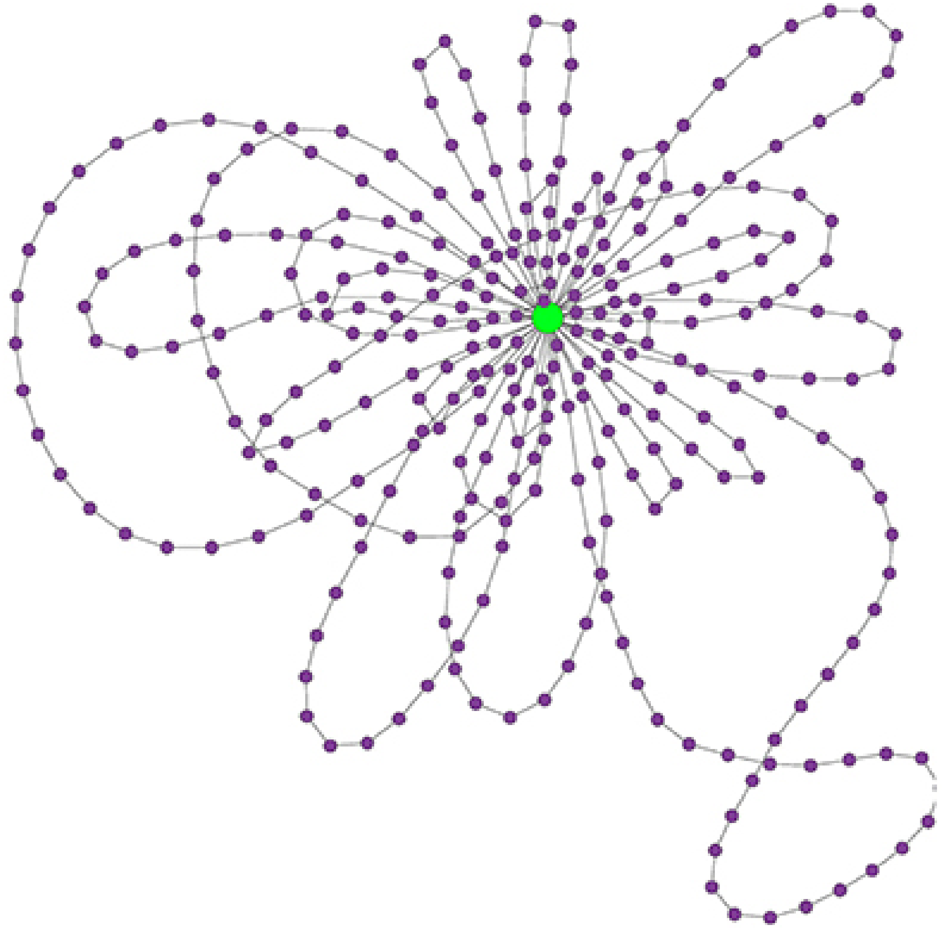
Optimal solution: one node with k_{max}, others with k_{min} (if $\mathrm{k_{min}} > 1$)

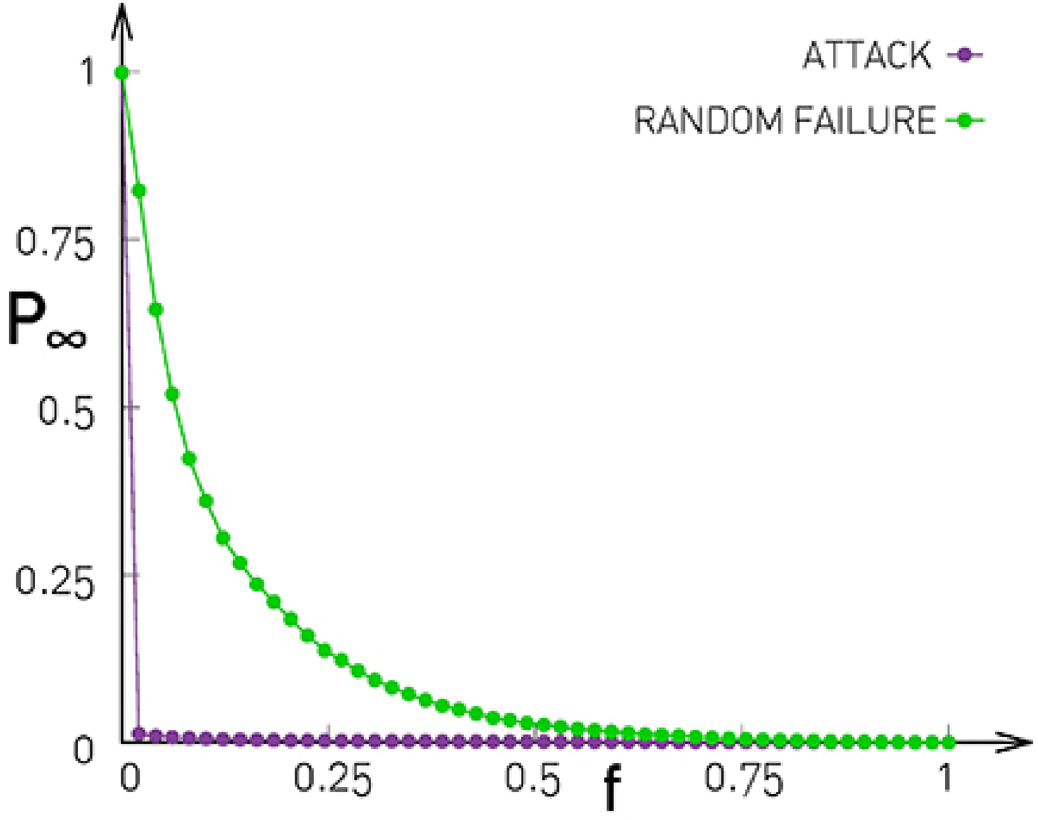




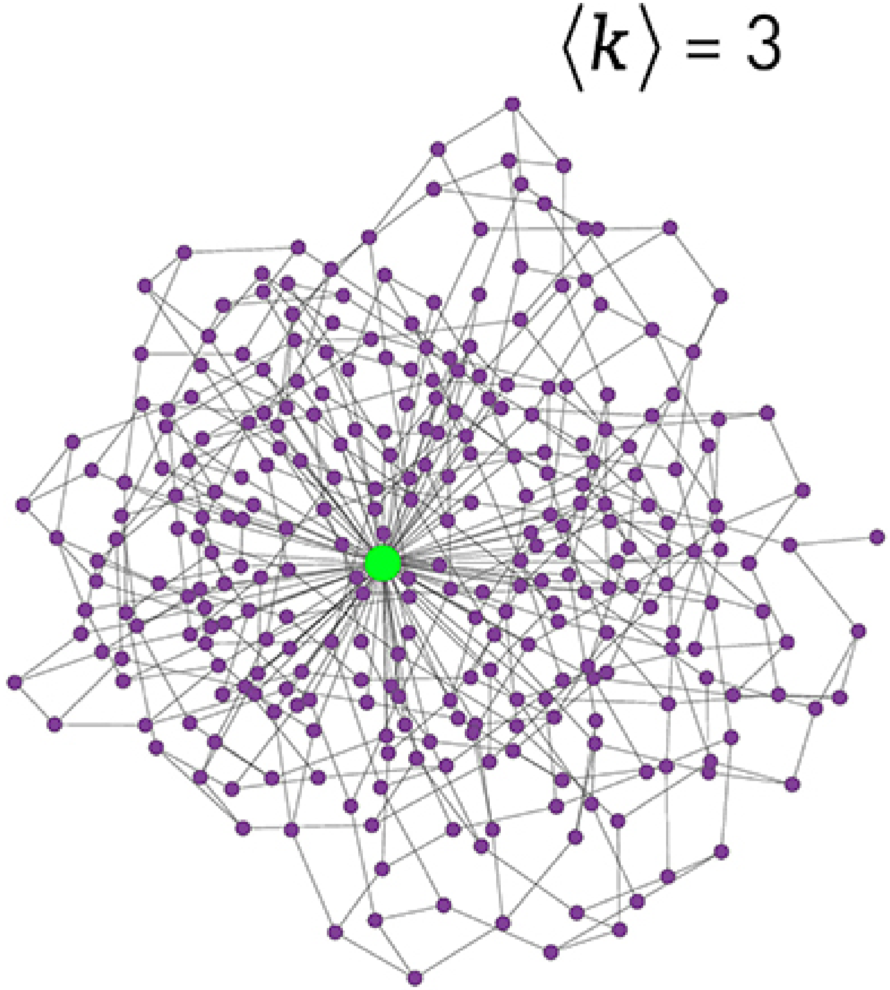
Examples

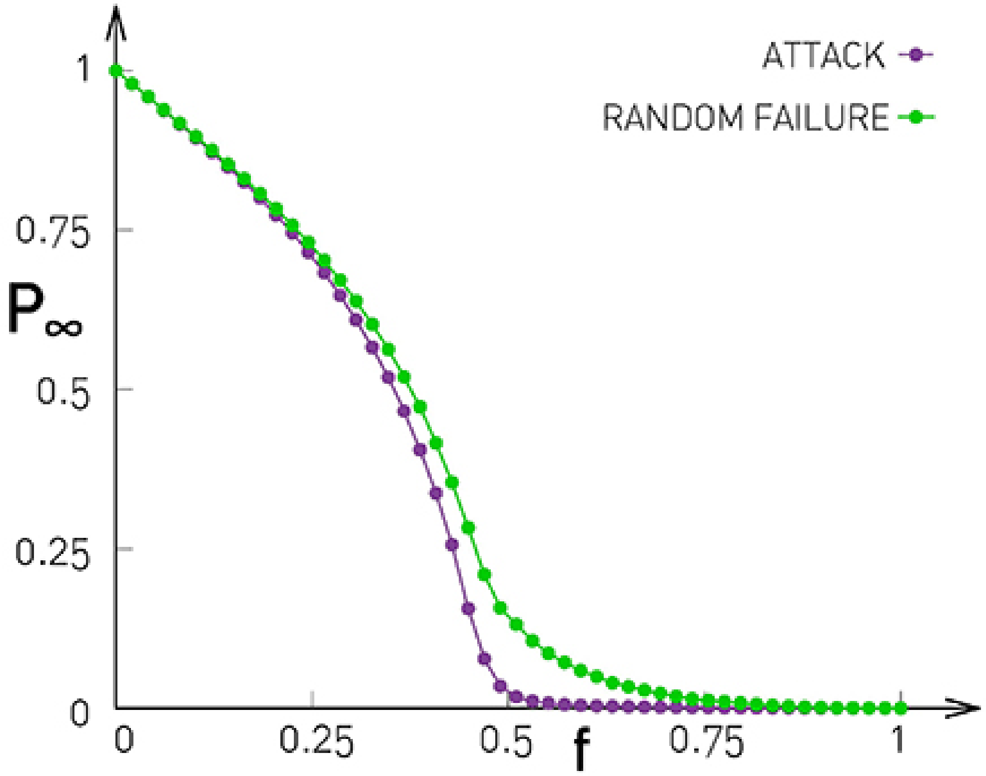
$\langle k \rangle = 2$



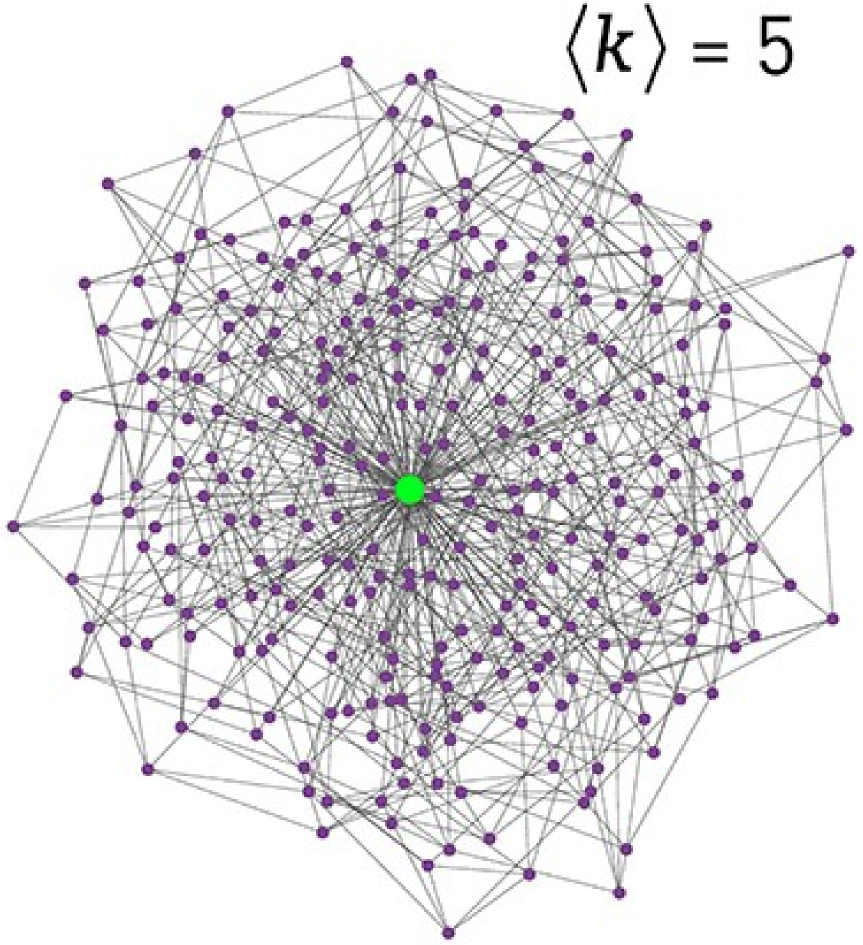


Examples

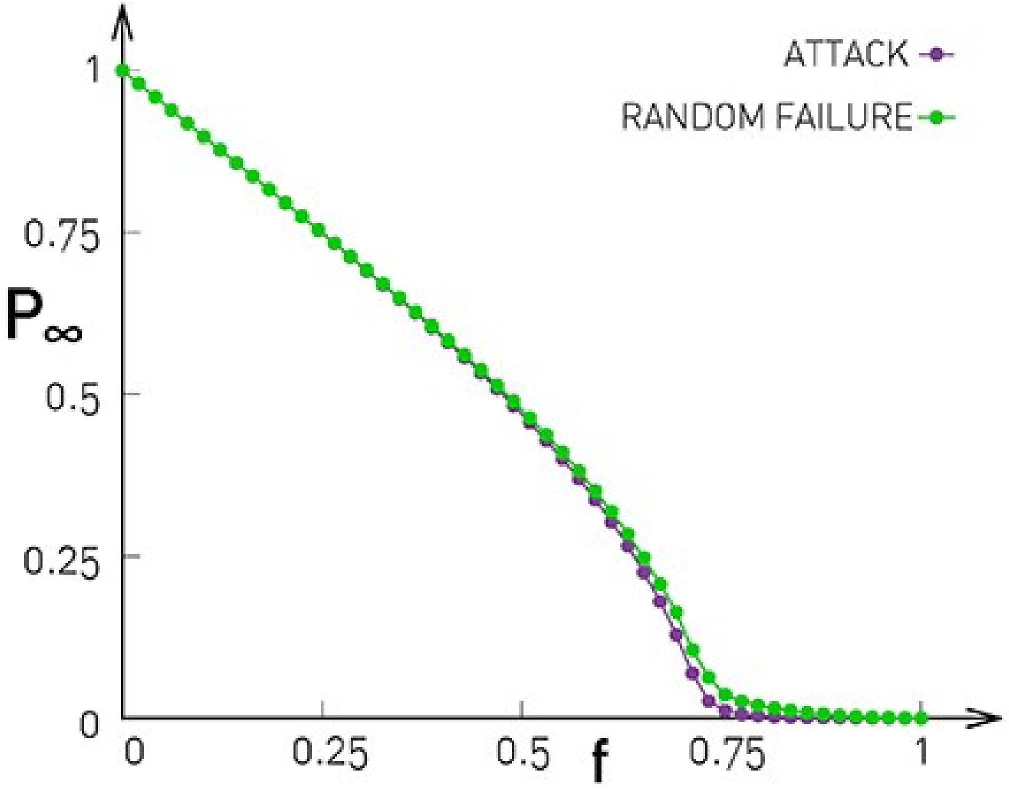

$\langle k \rangle = 3$


ATTACK
RANDOM FAILURE
P_∞
1
0.75
0.5
0.25
0
0
0.25
0.5
f
0.75
1

Examples

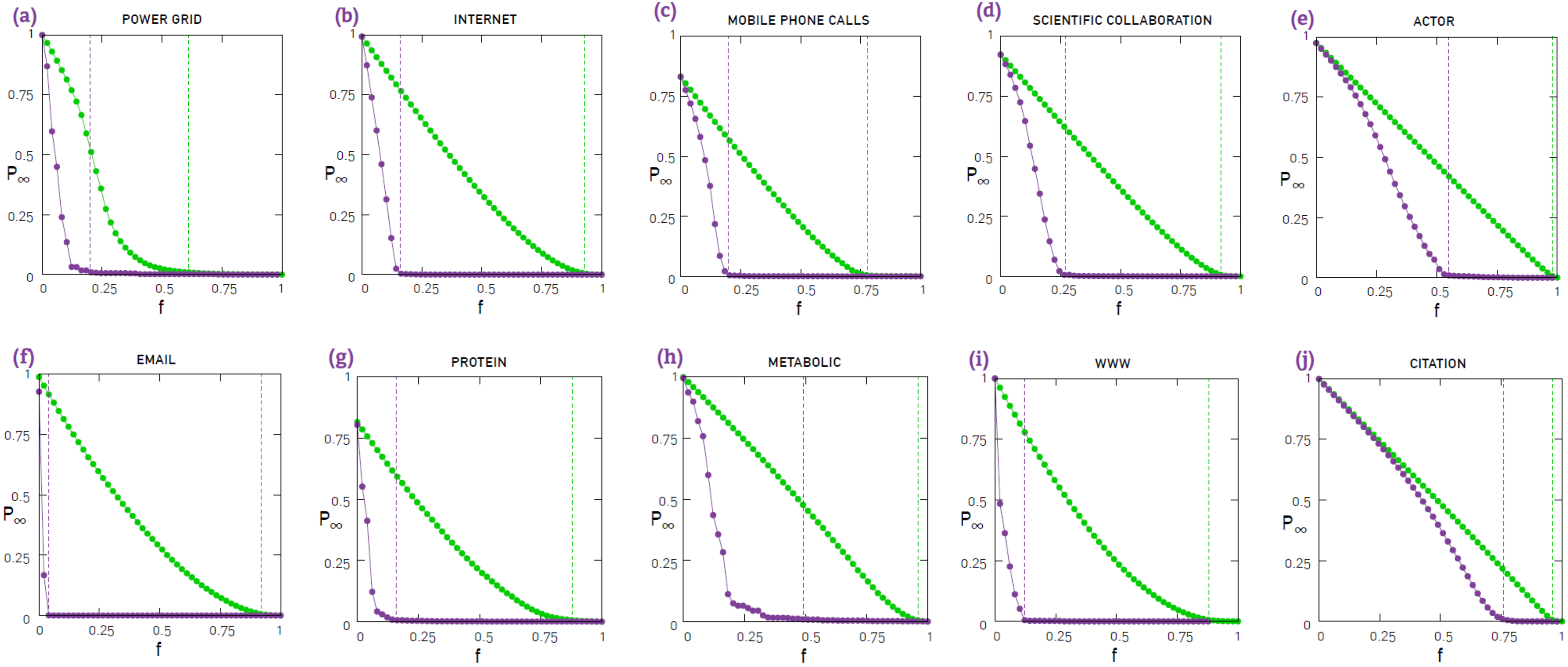


$\langle k \rangle = 5$



ATTACK
RANDOM FAILURE
P_∞
f
0
0.25
0.5
0.75
1

Real networks robustness



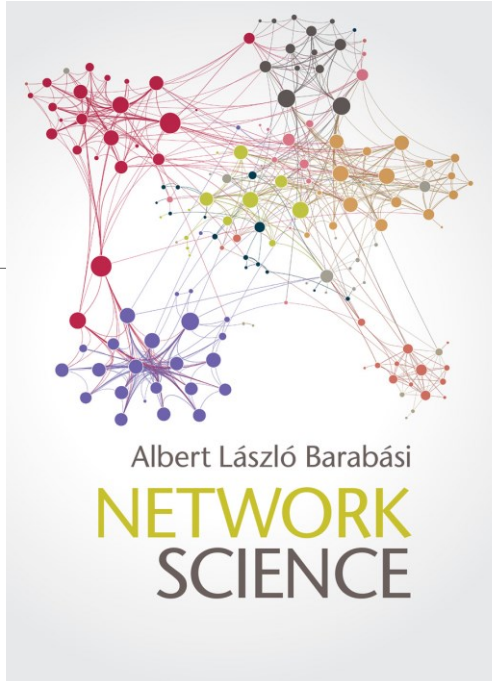
Network Analysis

10 – COMMUNITIES

Slides were created by: Agnes Vathy-Fogarassy

Network Science book (online)

Barabási, Albert-László. *Network Science*. Cambridge University Press, 2016.



Introduction

Belgium is a bicultural society:

- 59% of its citizens are Flemish, speaking Dutch.
- 40% are Walloons who speak French.

Introduction

Belgium is a bicultural society:

- 59% of its citizens are Flemish, speaking Dutch.
- 40% are Walloons who speak French.

Multiethnic countries break up all over the world.

How has this country fostered the peaceful coexistence of these two ethnic groups since 1830?

Introduction

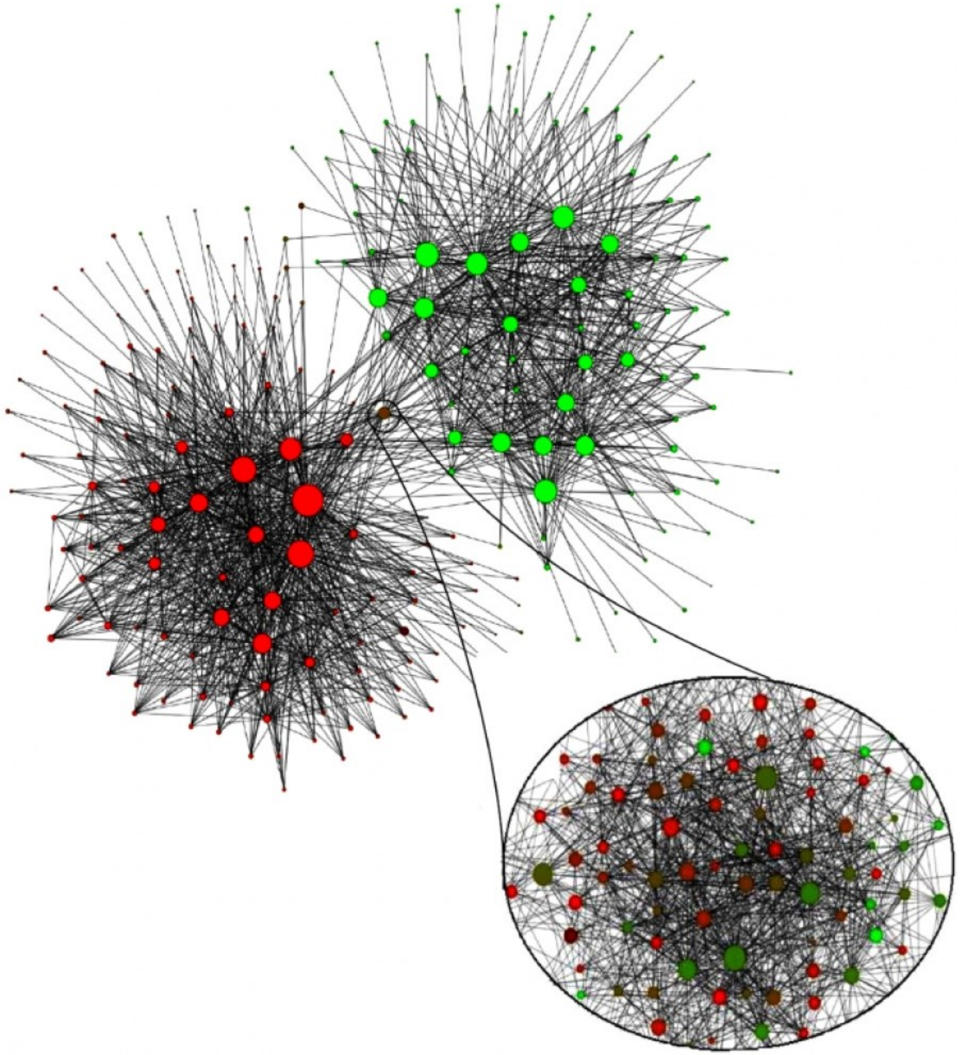
Belgium is a bicultural society:

- 59% of its citizens are Flemish, speaking Dutch.
- 40% are Walloons who speak French.

Multiethnic countries break up all over the world.

How has this country fostered the peaceful coexistence of these two ethnic groups since 1830?

The community structure was identified by mobile call network.



Introduction

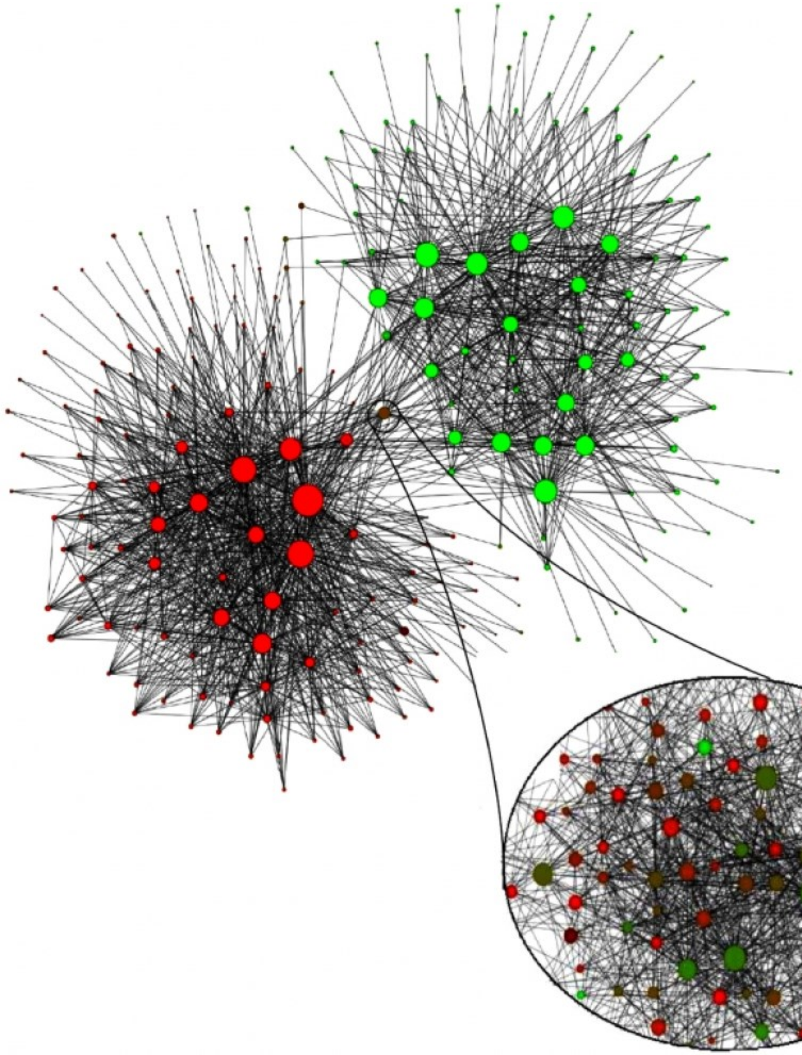
Belgium is a bicultural society:
- 59% of its citizens are Flemish, speaking Dutch.
- 40% are Walloons who speak French.

Multiethnic countries break up all over the world.

How has this country fostered the peaceful coexistence of these two ethnic groups since 1830?

The community structure was identified by mobile call network.

Community: group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities.



Introduction

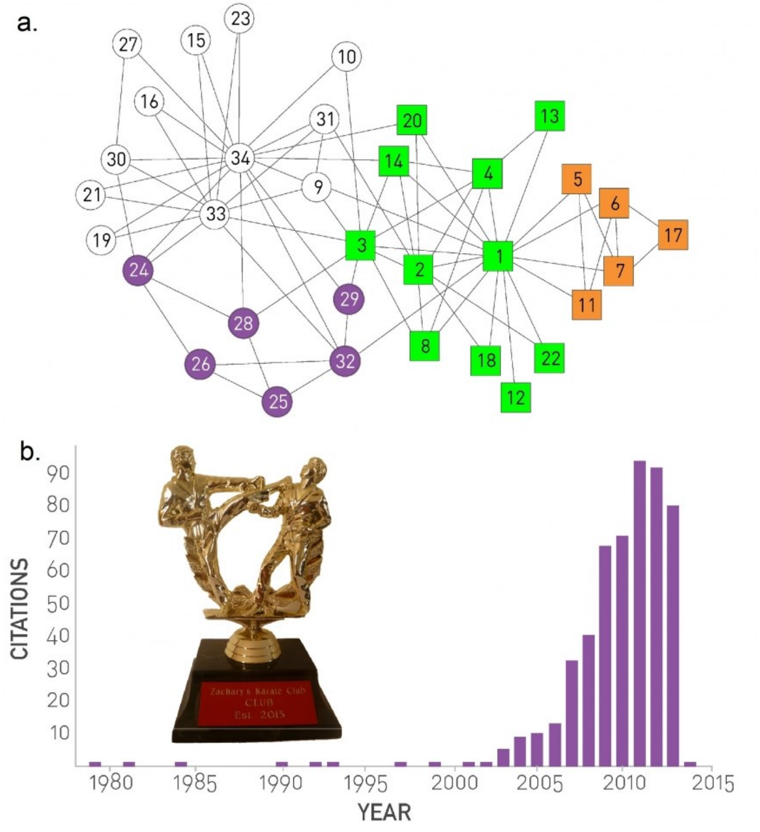
Two areas where communities play a particularly important role

- Social Network:
  - Employees of a company

Introduction

Two areas where communities play a particularly important role

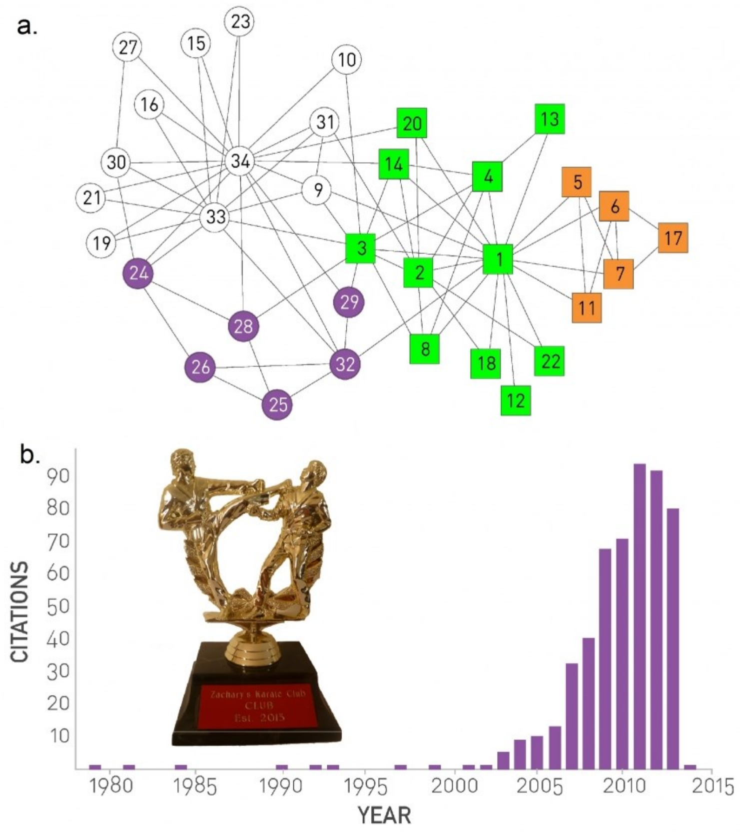
- Social Network:
  - Employees of a company
  - Zachary’s Karate Club
    - 34 members
    - Who regularly interacted outside the club.
    - Conflict between the club’s president and the instructor split the club into two.



Introduction

Two areas where communities play a particularly important role:

- Social Network:
  - Employees of a company
  - Zachary's Karate Club
    - 34 member
    - Who regularly interacted outside the club.
    - Conflict between the club's president and the instructor split the club into two.
- Biological Network:
  - For a long time biology has been focusing on single genes.
  - Disease module hypothesis:
    - Each disease can be linked to a well-defined neighbourhood (or environment) of the cellular network.



H1: Fundamental Hypothesis

- A network's community structure is uniquely encoded in its wiring diagram (A_{ij}).

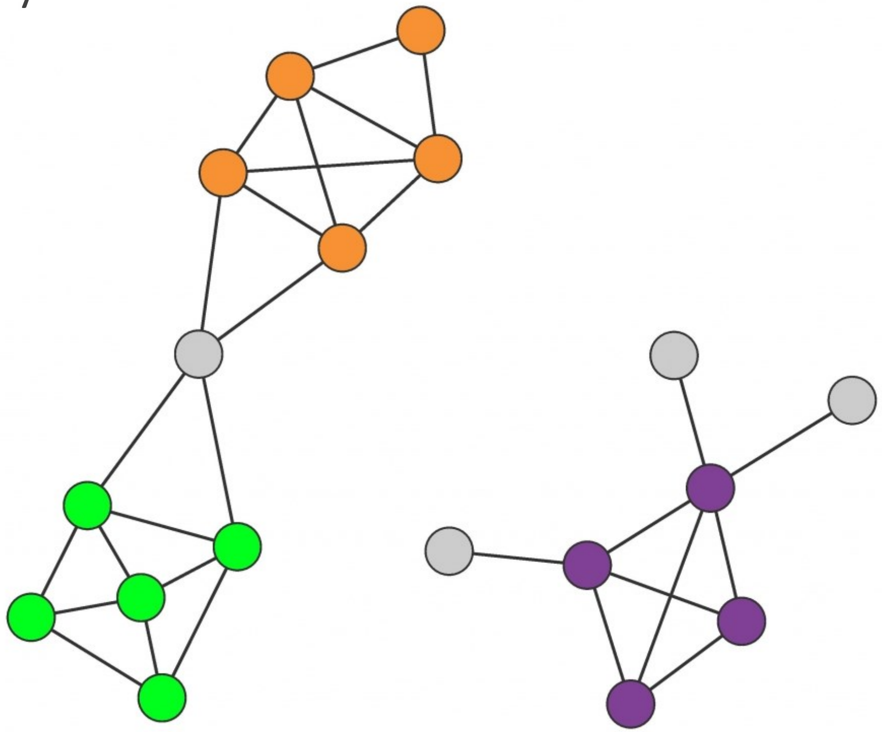
Basics of Communities

H2: Connectedness and Density Hypothesis

- A community is a locally densely connected subgraph in a network.
  - Connected – each node reach all the others.
  - Dense – a node connects to its community with higher probability.

Maximum Cliques

- Clique: complete subgraph
  - Community is a group of nodes where all know each other. (First approach in 1994)
  - Triangles are common, but bigger cliques are rare.
  - With the strict requirement potential groups are excluded.



Basics of Communities

Strong and Weak Communities

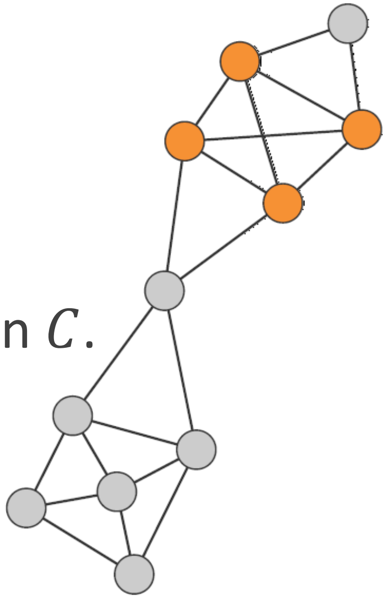
- Let C be a connected subnetwork with N_C node
- Let k_i^{int} be the number of links between node i and nodes in C.
- Let k_i^{ext} be the number of links between node i and nodes not in C.
  - If $k_i^{ext} = 0$, then C is a good community for node i.
  - If $k_i^{int} = 0$, node i should be assigned to a different community.

Basics of Communities

Strong and Weak Communities

- Let C be a connected subnetwork with N_C node
- Let k_i^{int} be the number of links between node i and nodes in C.
- Let k_i^{ext} be the number of links between node i and nodes not in C.
  - If $k_i^{ext} = 0$, then C is a good community for node i.
  - If $k_i^{int} = 0$, node i should be assigned to a different community.
- Strong community:
  - If the internal degree exceeds external degree in case of each node.
  - $k_i^{int}(C) > k_i^{ext}(C), \forall i \in C$
- Weak community:
  - If the total internal degree exceeds the total external degree.
  - $\sum_{i \in C} k_i^{int}(C) > \sum_{i \in C} k_i^{ext}(C)$

Basics of Communities



Strong and Weak Communities

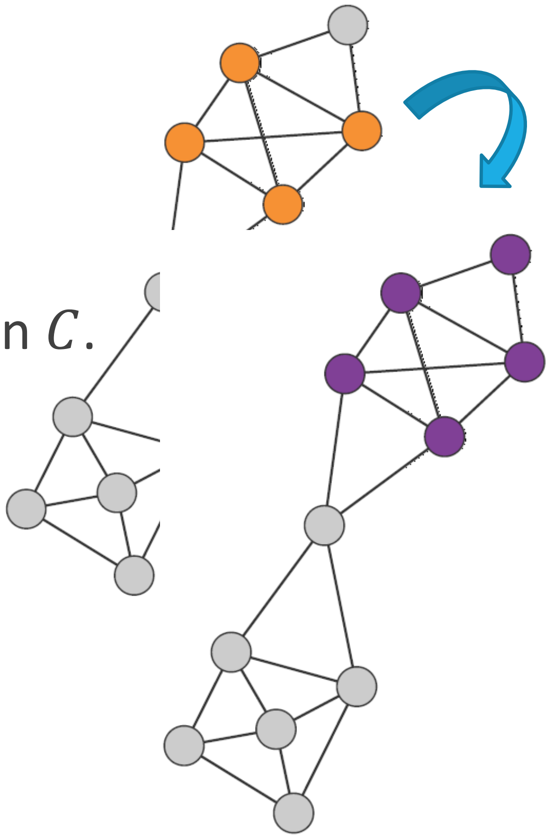
- Let C be a connected subnetwork with N_C node
- Let k_i^{int} be the number of links between node i and nodes in C.
- Let k_i^{ext} be the number of links between node i and nodes not in C.
  - If $k_i^{ext} = 0$, then C is a good community for node i.
  - If $k_i^{int} = 0$, node i should be assigned to a different community.
- Strong community
  - If the internal degree exceeds external degree in case of each node.
  - $k_i^{int}(C) > k_i^{ext}(C), \forall i \in C$
- Weak community
  - If the total internal degree exceeds the total external degree.
  - $\sum_{i \in C} k_i^{int}(C) > \sum_{i \in C} k_i^{ext}(C)$

Clique

Basics of Communities

Strong and Weak Communities

- Let C be a connected subnetwork with N_C node
- Let k_i^{int} be the number of links between node i and nodes in C.
- Let k_i^{ext} be the number of links between node i and nodes not in C.
  - If $k_i^{ext} = 0$, then C is a good community for node i.
  - If $k_i^{int} = 0$, node i should be assigned to a different community.
- Strong community
  - If the internal degree exceeds external degree in case of each node.
  - $k_i^{int}(C) > k_i^{ext}(C), \forall i \in C$
- Weak community
  - If the total internal degree exceeds the total external degree.
  - $\sum_{i \in C} k_i^{int}(C) > \sum_{i \in C} k_i^{ext}(C)$



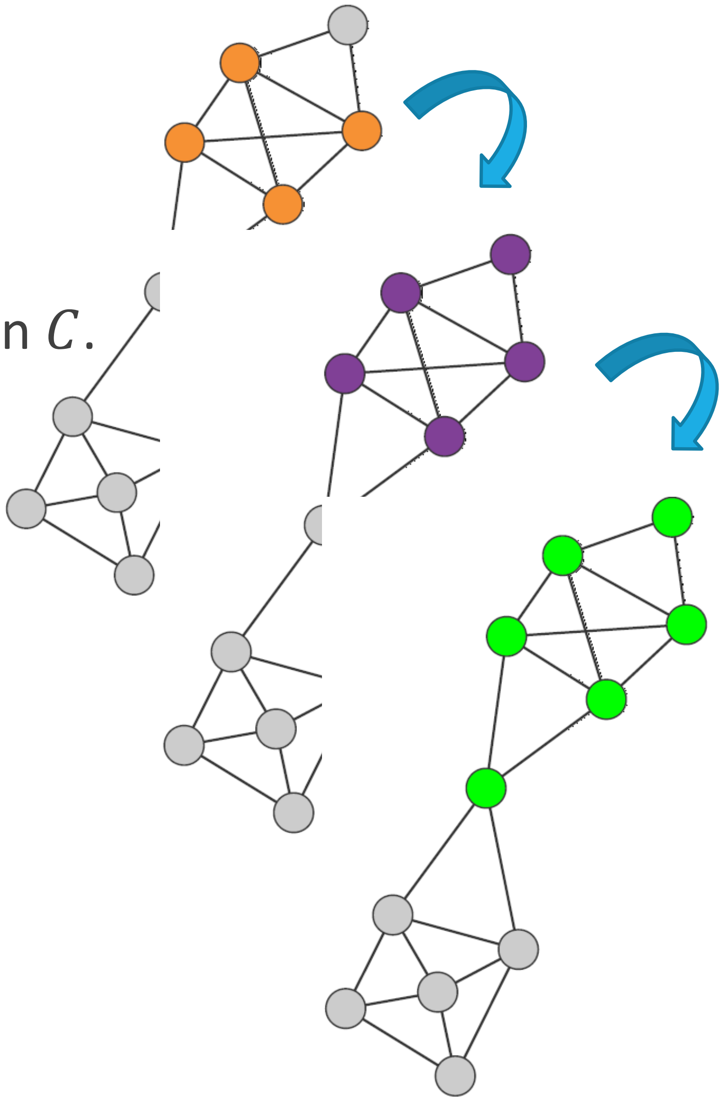
Clique ⊆ Strong community

Basics of Communities

Strong and Weak Communities

- Let C be a connected subnetwork with N_C node
- Let k_i^{int} be the number of links between node i and nodes in C.
- Let k_i^{ext} be the number of links between node i and nodes not in C.
  - If $k_i^{ext} = 0$, then C is a good community for node i.
  - If $k_i^{int} = 0$, node i should be assigned to a different community.
- Strong community
  - If the internal degree exceeds external degree in case of each node.
  - $k_i^{int}(C) > k_i^{ext}(C), \forall i \in C$
- Weak community
  - If the total internal degree exceeds the total external degree.
  - $\sum_{i \in C} k_i^{int}(C) > \sum_{i \in C} k_i^{ext}(C)$

Clique ⊆ Strong community ⊆ Weak community

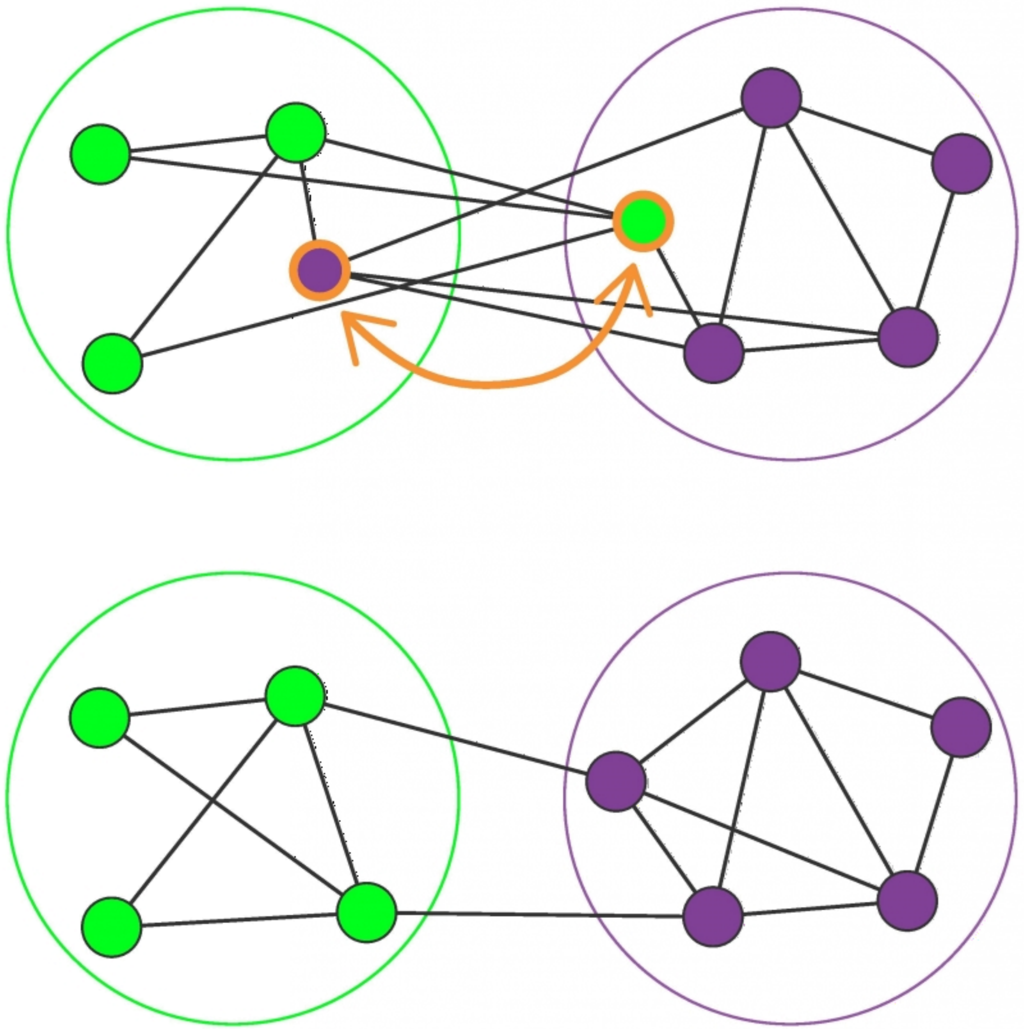


Basics of Communities

Number of communities:

- Simplest solution: graph bisection.
  - Minimize the cut size.
  - *E.g. 1:*
    - $N = 10$
    - $N_1 = N_2 = 5$
    - Check 252 bisection → **suppose** that it takes 1 millisecond (10^{-3} second).
  - *E.g. 2:*
    - $N = 100$
    - $N_1 = N_2 = 50$
    - $\sim 10^{29}$ bisection → **then** it takes 10^{16} years on the same computer.

What if we do not know the size and number of the community?



Hierarchical Clustering

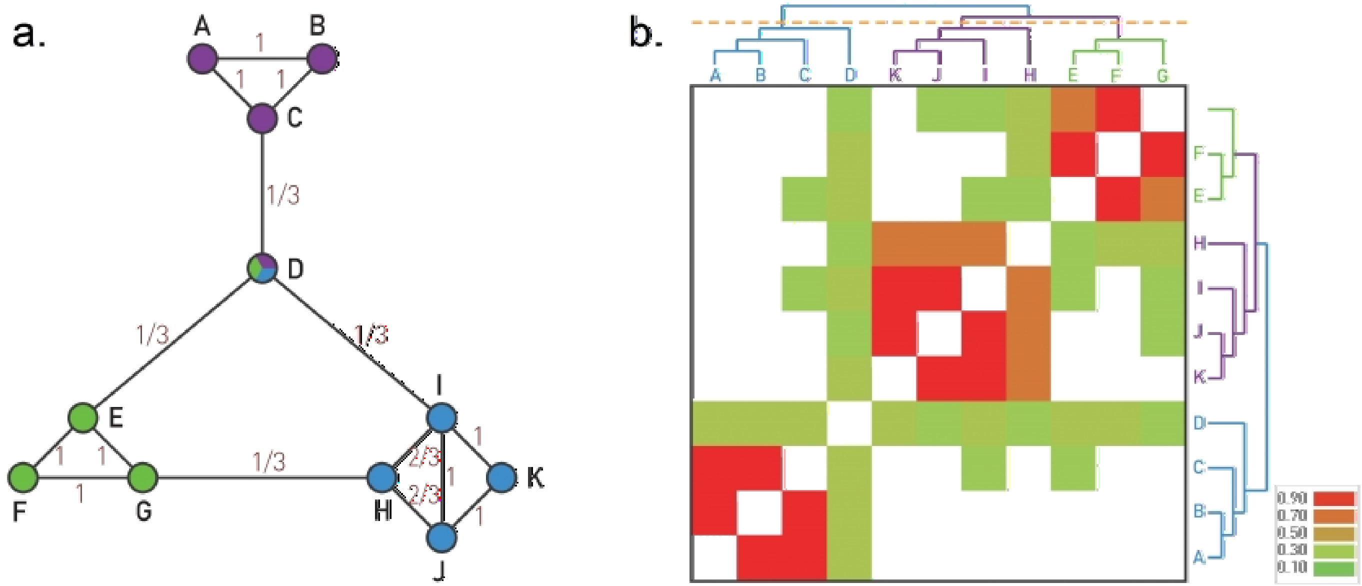
Two different procedures

- Agglomerative algorithms
  - Merge nodes into the same community.
  - Ravasz algorithm
- Divisive algorithms
  - Isolate communities by removing links.
  - Girvan-Newman algorithm

Hierarchical Clustering

Two different procedures

- Agglomerative algorithms
  - Merge nodes into the same communit
  - Ravasz algorithm
- Divisive algorithms
  - Isolate communities by removing links
  - Girvan-Newman algorithm

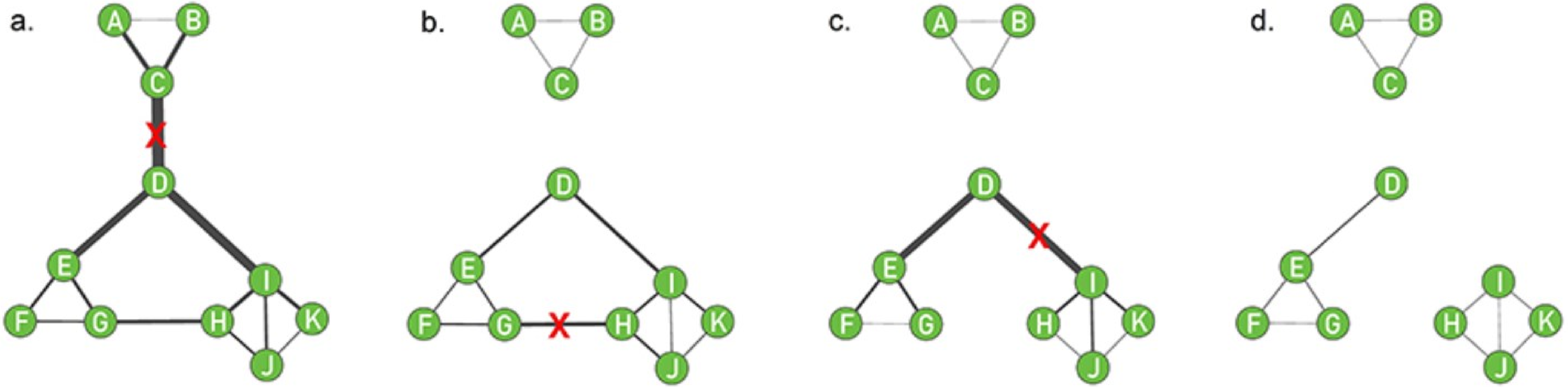


Ravasz algorithm:

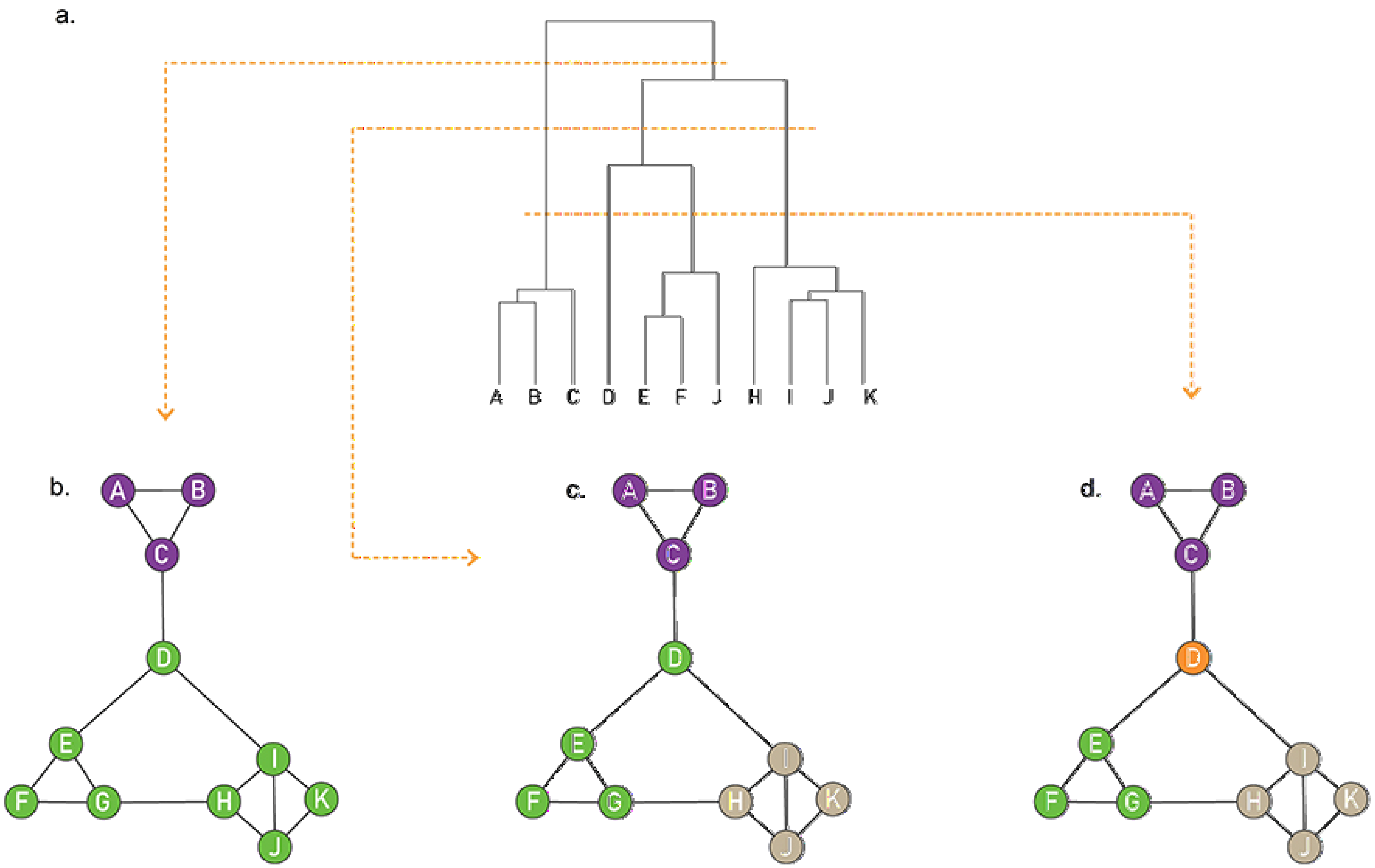
1. Assign each node to a community of its own and evaluate x_{ij} similarity for all node pairs. x_{ij} is calculated by neighbours, degrees and number of links (Section 9.3)
2. Find the community pair or the node pair with the highest similarity and merge them into a single community.
3. Calculate the similarity between the new community and all other communities.
4. Repeat Steps 2 and 3 until all nodes form a single community.

Girvan-Newman algorithm

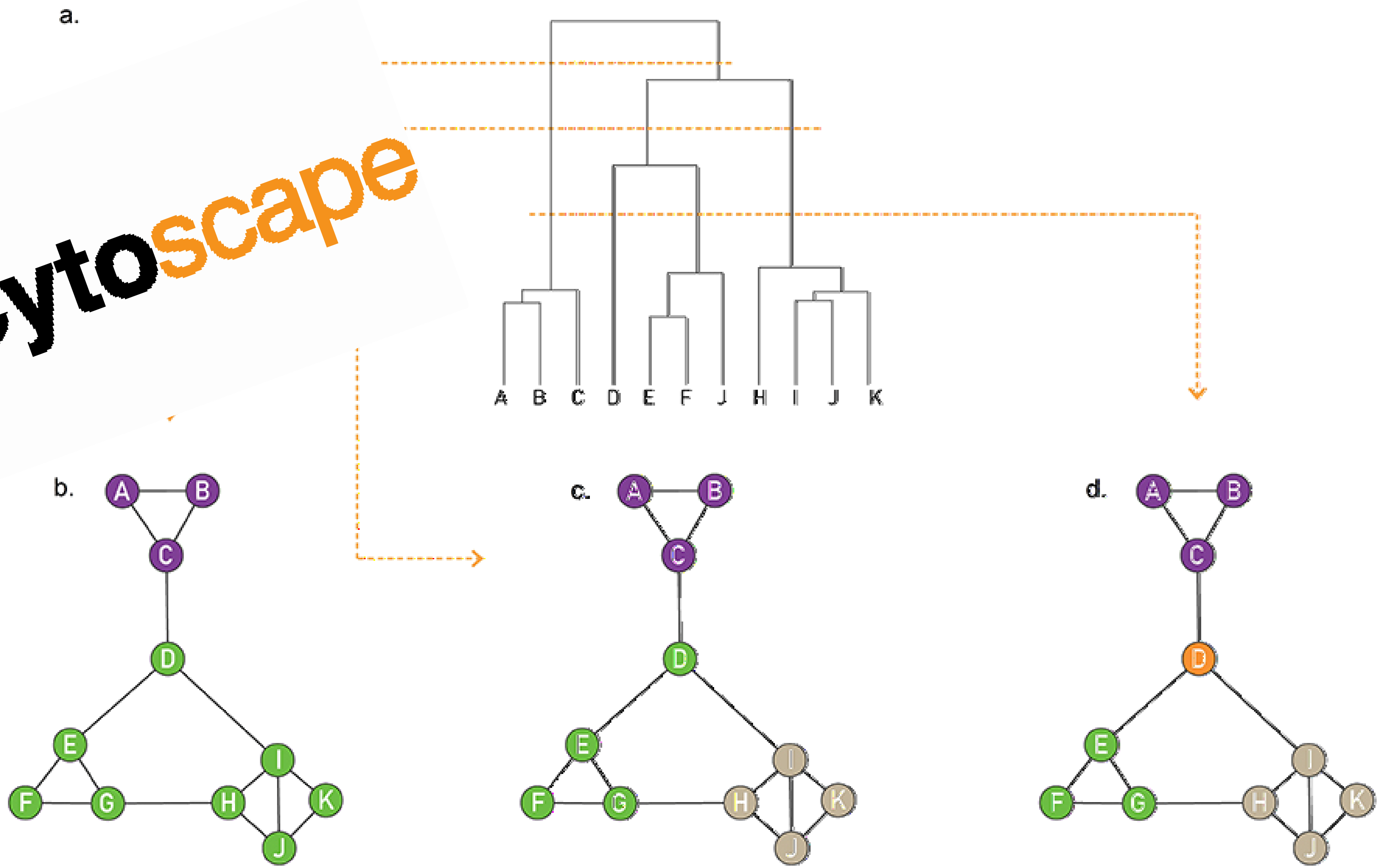
1. Compute the edge betweenness centrality x_{ij} of each link.
2. Remove one of the links with the largest centrality.
3. Recalculate the centrality of each link for the altered network.
4. Repeat steps 2 and 3 until all links are removed.



Hierarchical Clustering



Hierarchical Clustering



Modularity

H3: Random Hypothesis

- Randomly wired networks lack an inherent community structure.

Modularity

H3: Random Hypothesis

- Randomly wired networks lack an inherent community structure.

Modularity

- Allows us to decide if a community partition is better than some other ones.
- $M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$
  - n_c: number of communities
  - L_c: number of links in community C_c
  - k_c: sum of degrees of nodes in community C_c

Modularity

H3: Random Hypothesis

- Randomly wired networks lack an inherent community structure.

Modularity

- Allows us to decide if a community partition is better than some other ones.
- $M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$
  - n_c: number of communities
  - L_c: number of links in community C_c
  - k_c: sum of degrees of nodes in community C_c
- Higher modularity implies better partition.
- If the whole network is a single community, then $M = 0$
- If each node form a separate community, then M is negative

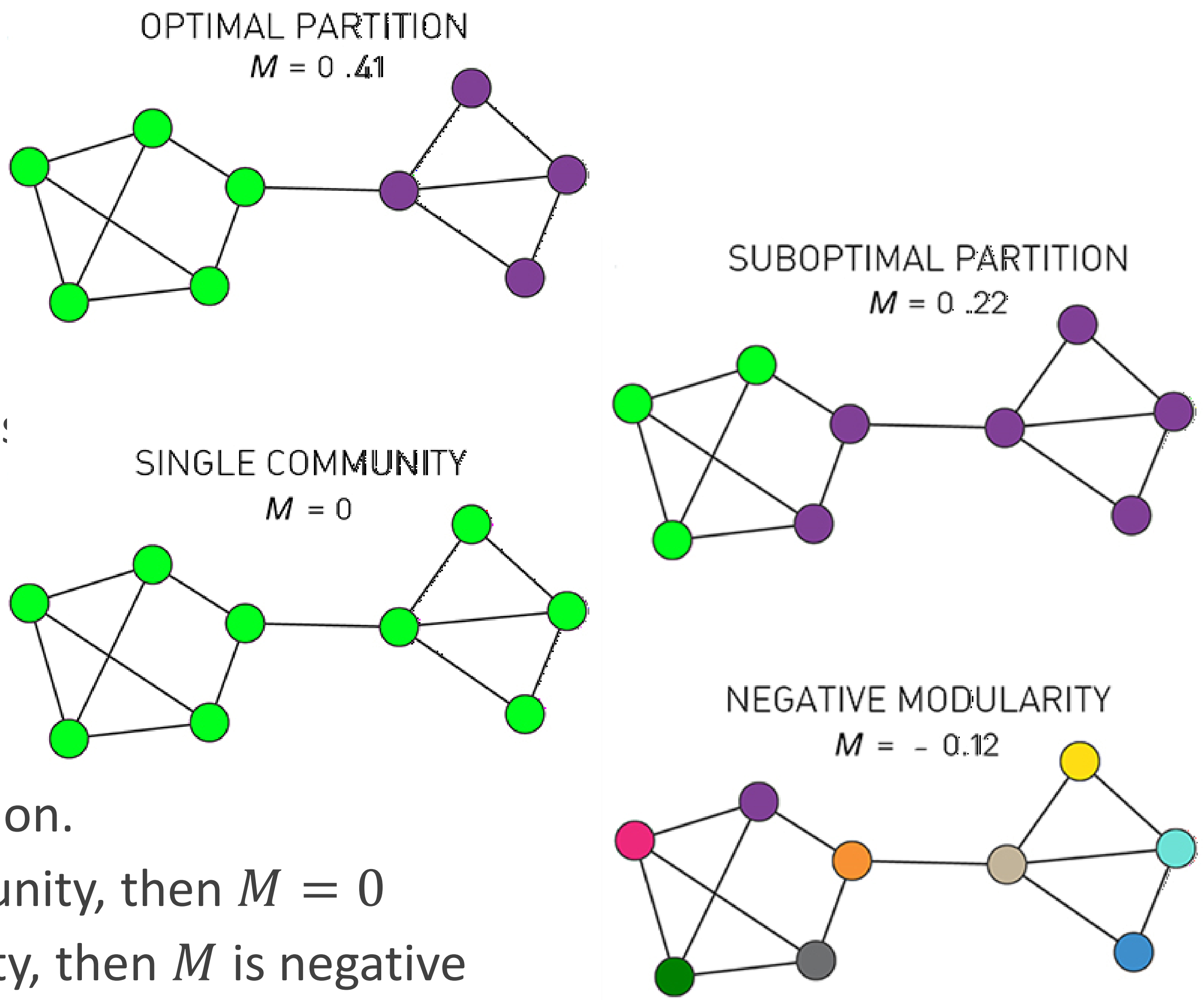
Modularity

H3: Random Hypothesis

- Randomly wired networks lack an inherent community structure.

Modularity

- Allows us to decide if a community partition is better than some other ones
- $M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$
  - n_c: number of communities
  - L_c: number of links in community C_c
  - k_c: sum of degrees of nodes in community C_c
- Higher modularity implies better partition.
- If the whole network is a single community, then $M = 0$
- If each node form a separate community, then M is negative



The Greedy Algorithm

H4: Maximal Modularity Hypothesis

- For a given network the partition with maximum modularity corresponds to the optimal community structure.

The Greedy Algorithm

H4: Maximal Modularity Hypothesis

- For a given network the partition with maximum modularity corresponds to the optimal community structure.

Greedy Algorithm to produce maximal M

1. Assign each node to a community, starting with N communities of single nodes.
2. Inspect each community pair connected by at least one link and compute the modularity difference ΔM obtained if we merge them. Identify the community pair for which ΔM is the largest and merge them. Note that modularity is always calculated for the full network.
3. Repeat Step 2 until all nodes merge into a single community, recording M for each step.
4. Select the partition for which M is maximal.

The Greedy Algorithm

H4: Maximal Modularity Hypothesis

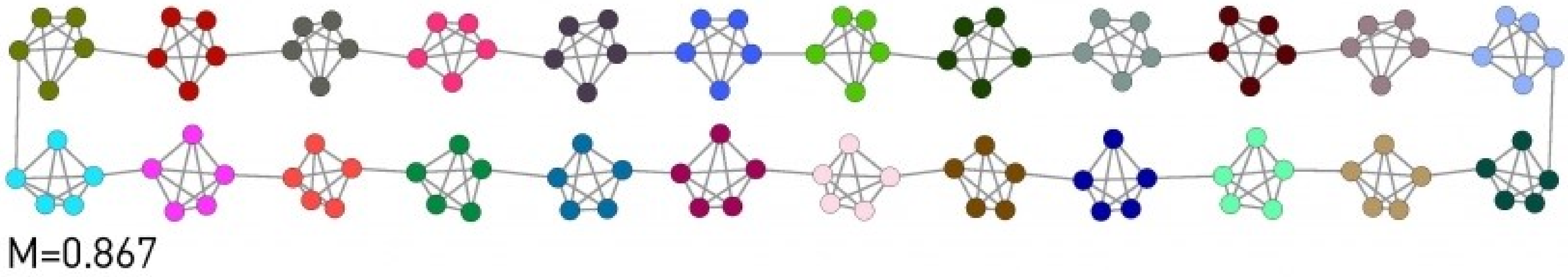
- For a given network the partition with maximum modularity corresponds to the optimal community structure.

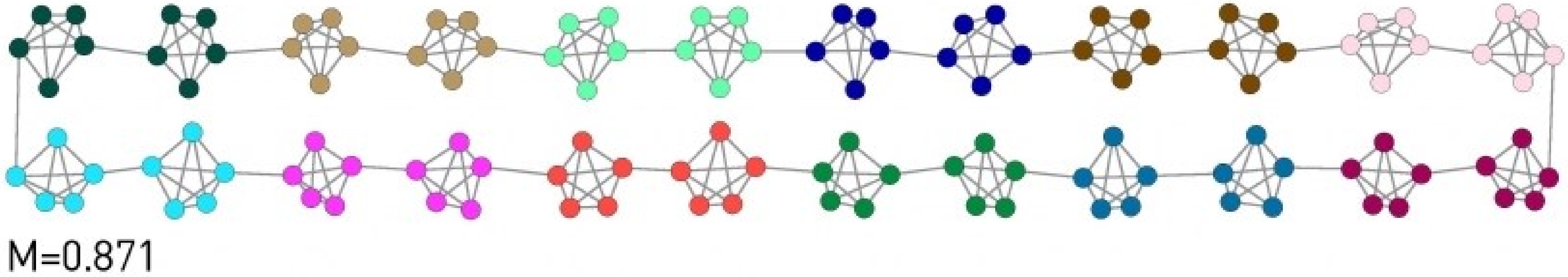
Greedy Algorithm to produce maximal M

1. Assign each node to a community, starting with N communities of single nodes.
2. Inspect each community pair connected by at least one link and compute the modularity difference ΔM obtained if we merge them. Identify the community pair for which ΔM is the largest and merge them. Note that modularity is always calculated for the full network.
3. Repeat Step 2 until all nodes merge into a single community, recording M for each step.
4. Select the partition for which M is maximal.

Disadvantage: increase of M results in merged small communities ($k < \sqrt{2L}$)

The Greedy Algorithm





Overlapping Communities

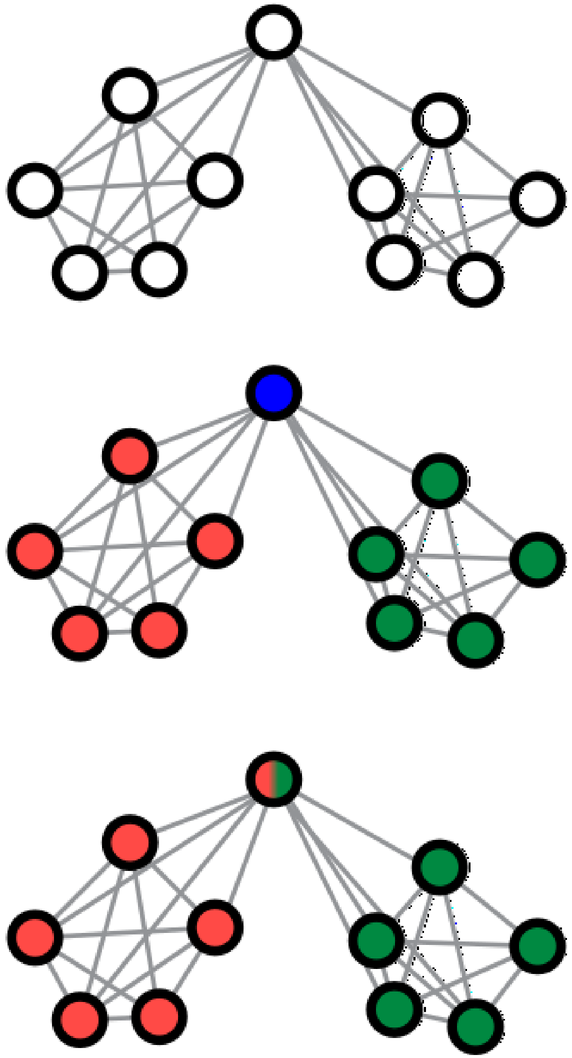
Real example:

- A teacher holds two courses, and knows most of the students.
- The students from the two courses do not know each other.
- How are the communities evolved in this case?

Until now, we have strictly distinguished the communities.

Two algorithms that enable overlapping communities:

- Clique Percolation
- Link Clustering



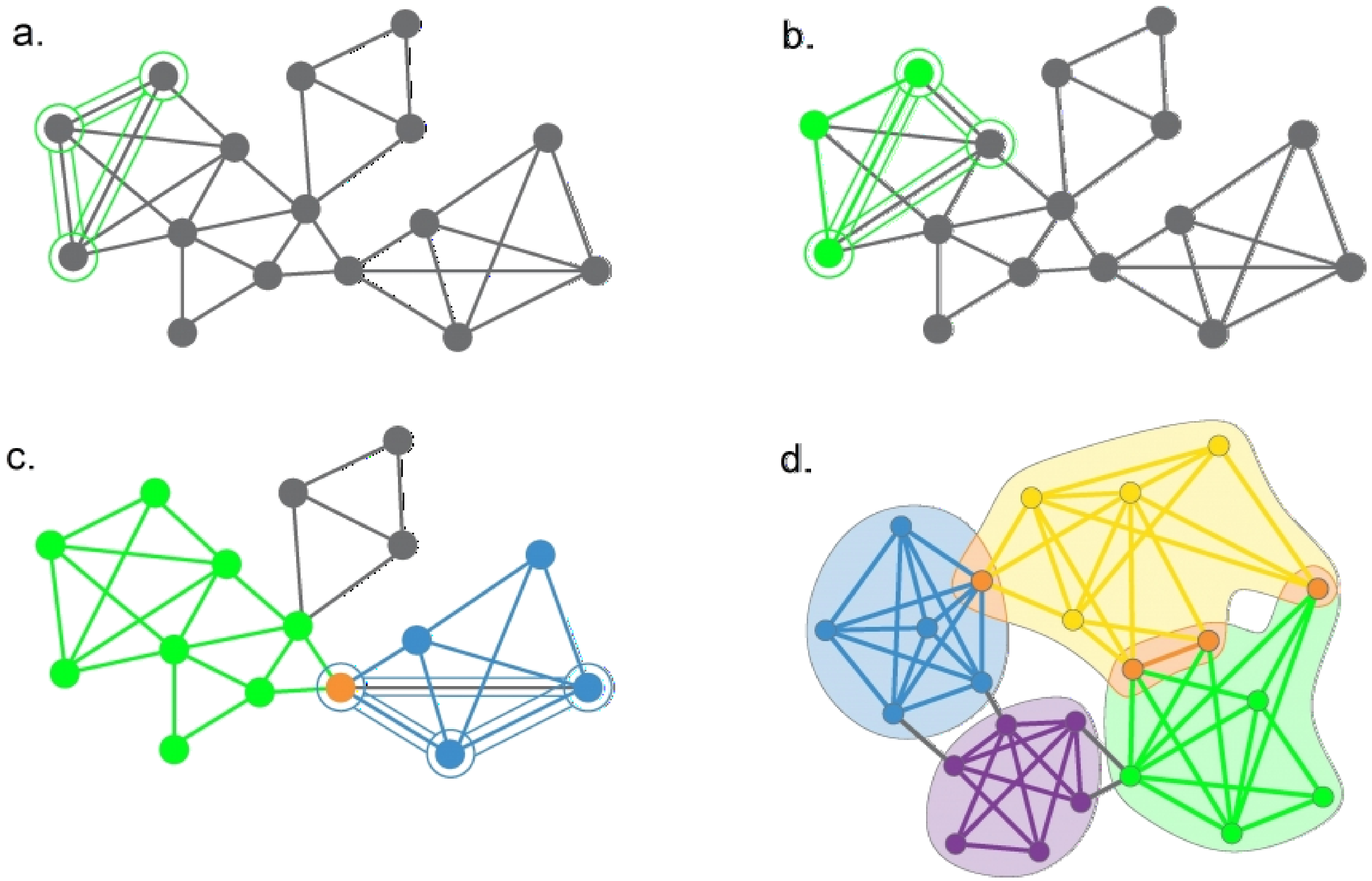
Clique Percolation

Often called *Cfinder*

Two k-cliques are considered adjacent if they share $k - 1$ nodes.

A k-clique community is the largest connected subgraph obtained by the union of all adjacent k-cliques.

If two k-cliques are not adjacent with each other, then they are belong to different communities.

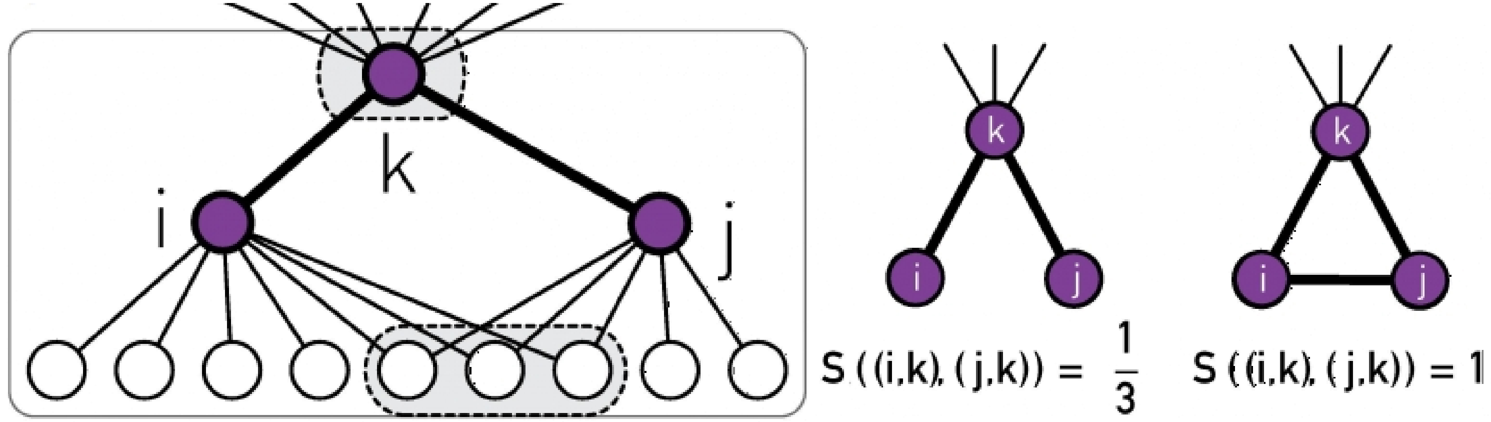


Link Clustering

Links can provide the communities.

Step 1: Define Link Similarity

- $S\big((i,k),(j,k)\big) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}$
- $n_+(i)$: set of neighbours of node i including node i itself



Link Clustering

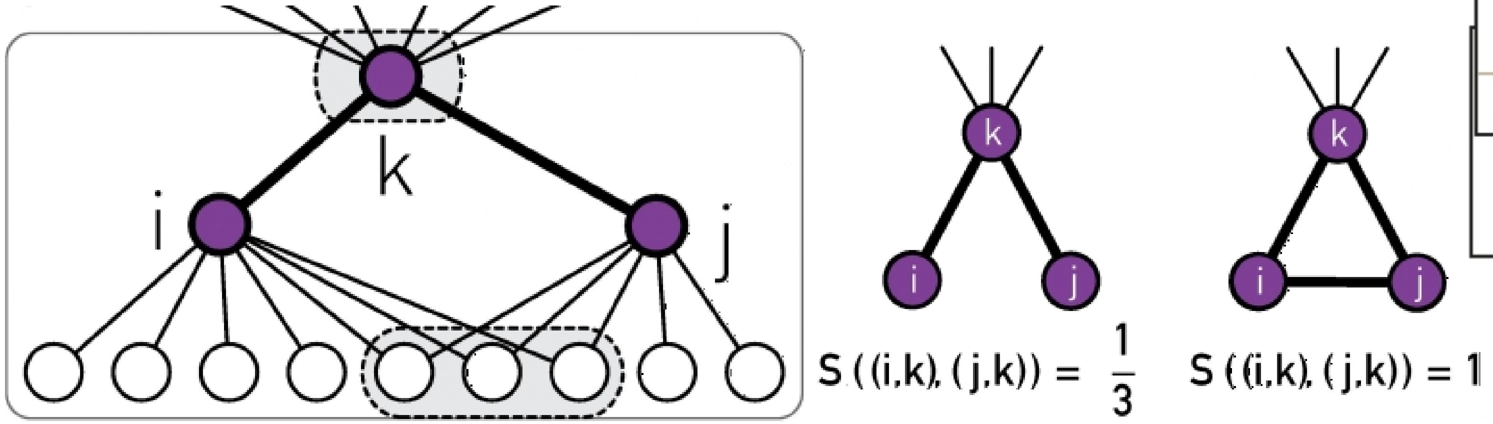
Links can provide the communities.

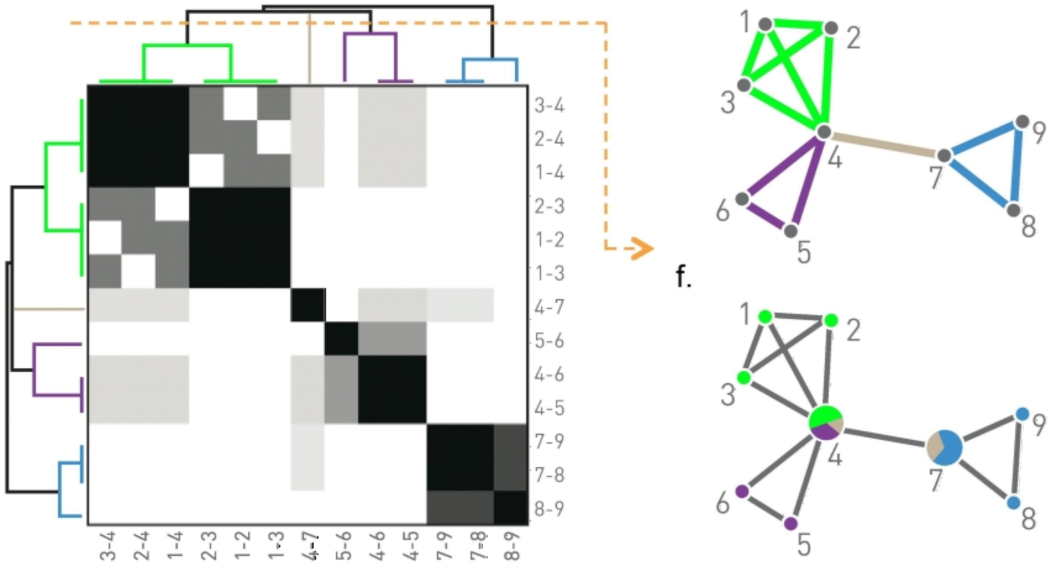
Step 1: Define Link Similarity

- $S\big((i,k),(j,k)\big) = \dfrac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}$
- $n_+(i)$: set of neighbours of node i including node i itself

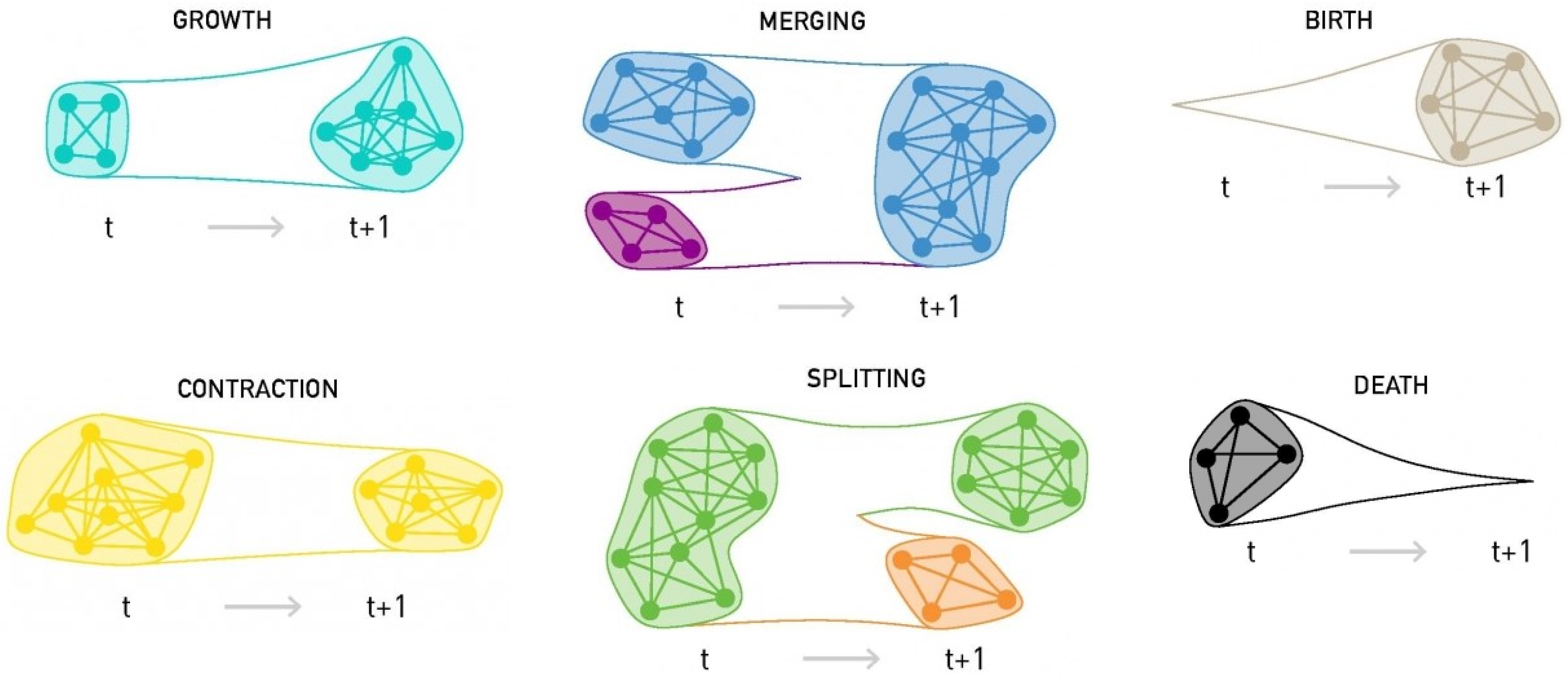
Step 2: Apply Hierarchical Clustering

- Iteratively merging communities with the largest similarity link pairs.





Community Evolution



Summary

Do we really have communities?

- How do we know that there are indeed communities in a particular network?

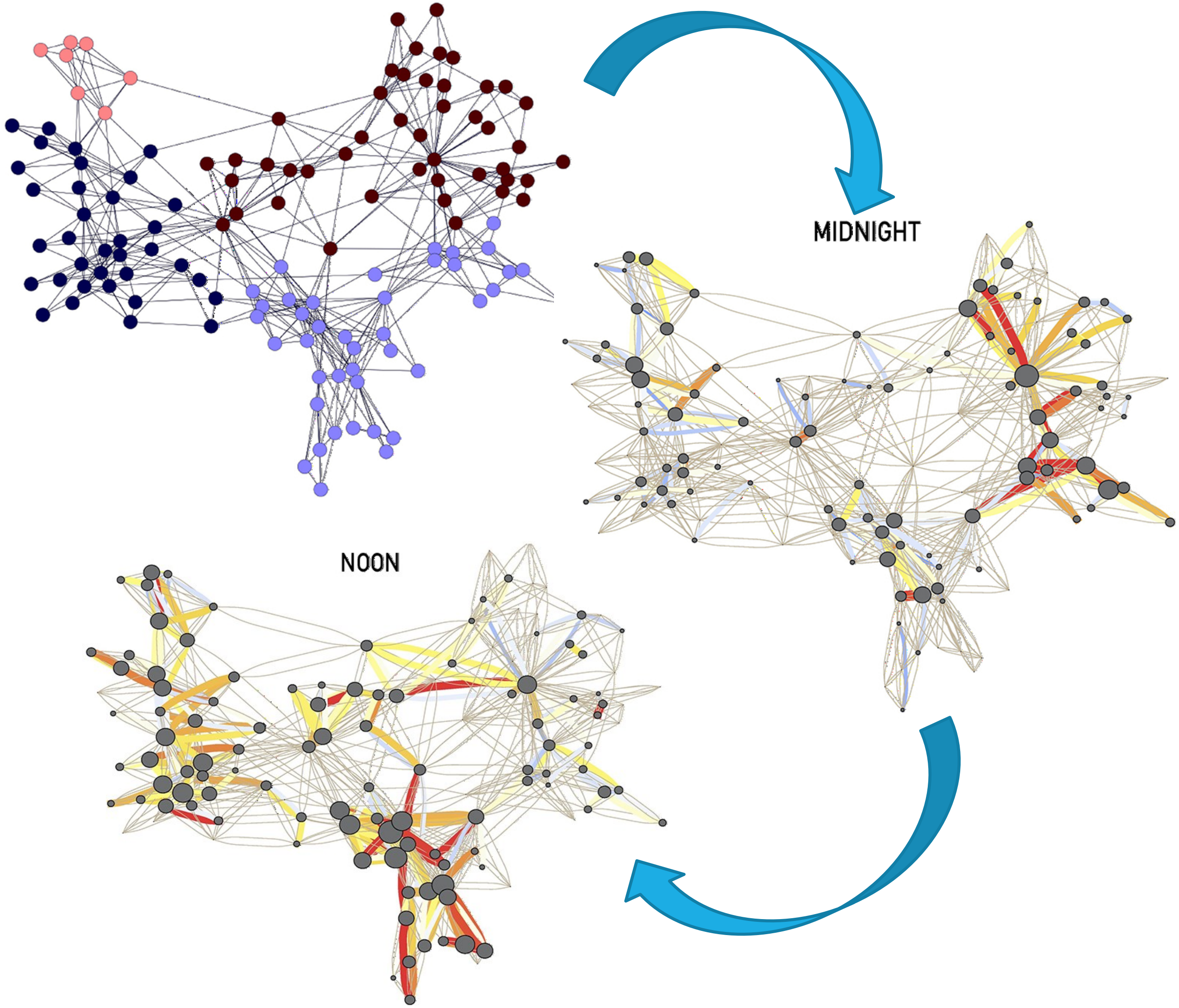
Hypotheses or theorems?

Do all the nodes need to belong to com

Dense vs. sparse communities.

Do communities matter?

- Image: neighbourhood of the mobile call network.



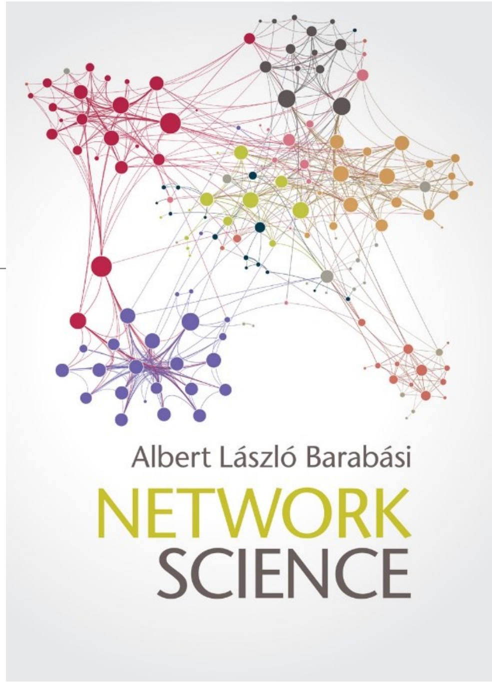
Network Analysis

11 – SPREADING PHENOMENA

Slides were created by: Daniel Leitold

Network Science book (online)

Barabási, Albert-László. *Network Science*.
Cambridge University Press, 2016.



Introduction

In February 21, 2003, a physician from Guangdong Province in southern China checked in the Metropole Hotel in Hong Kong.

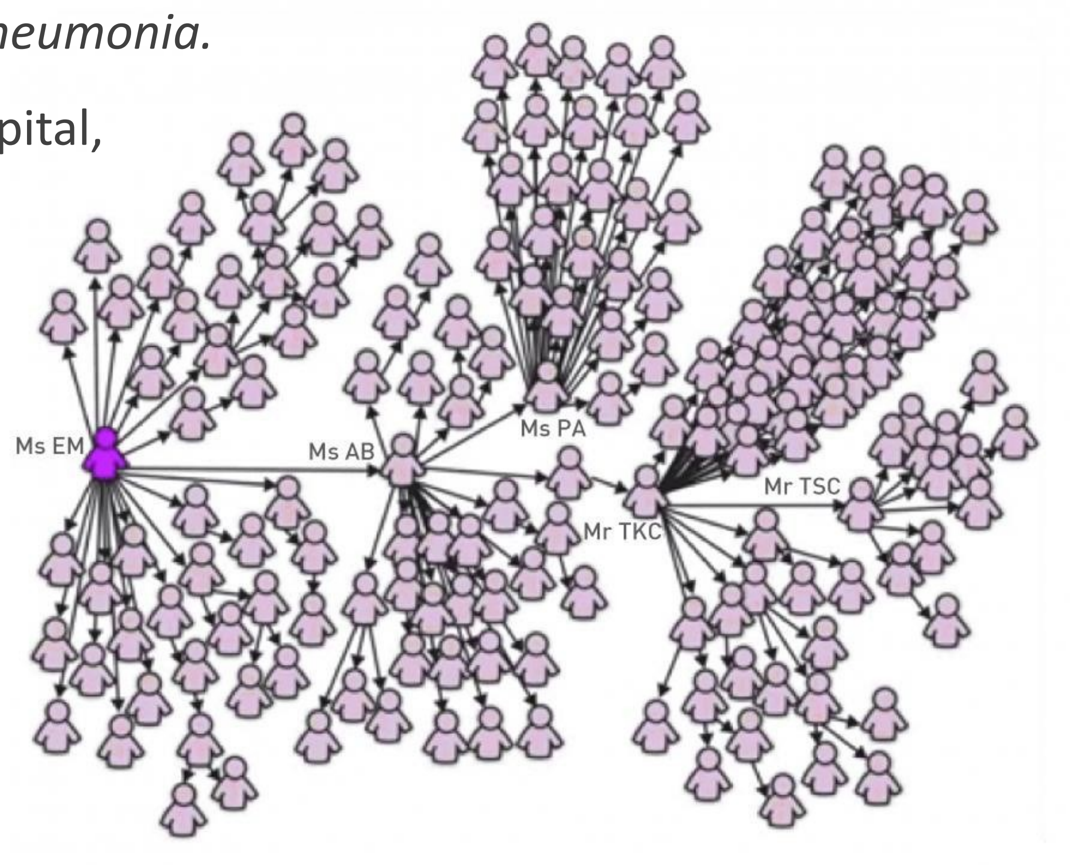
He previously treated people with a diagnosis: *atypical pneumonia.*

Next day, after leaving the hotel, he went to the local hospital, this time as a patient. He died there several days later of atypical pneumonia.

That night sixteen other guests of the Metropole Hotel and one visitor also contracted the disease: *Severe Acute Respiratory Syndrome*, or *SARS*.

These guests carried the SARS virus with them to Hanoi, Singapore, and Toronto

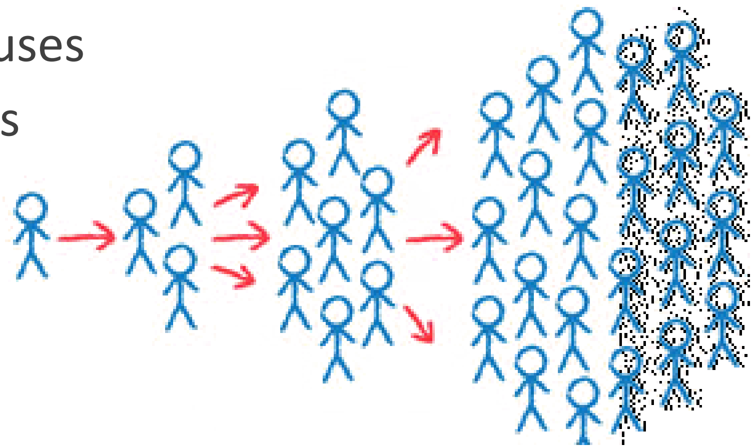
Epidemiologists later traced close to half of the 8,100 documented cases of SARS back to the Metropole Hotel.



Introduction

In this chapter: spreading processes

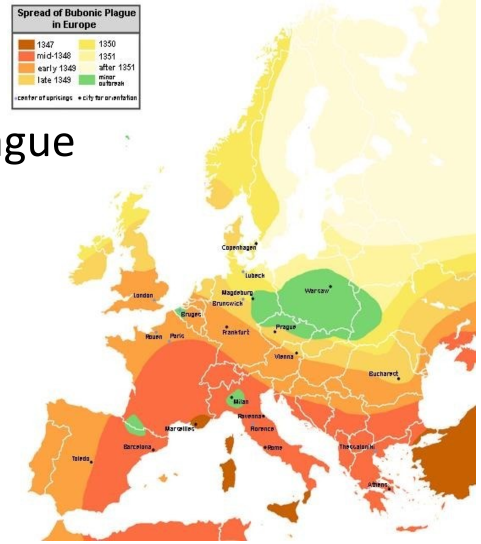
- **Biological**
  - *Airborne diseases*: Influenza, SARS, tuberculosis
  - *Contagious diseases* and parasites: Ebola, HIV, malaria
  - *Cancer-causing viruses*: HPV, EBV
- **Digital**
  - Computer viruses
  - Mobile viruses
  - Worms



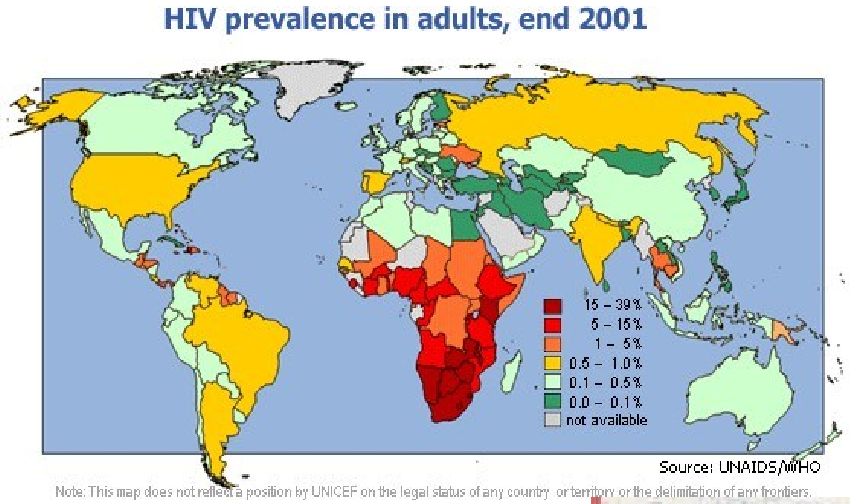
- **Social**
  - Innovations
  - Knowledge
  - Business practices
  - Products
  - Behaviour
  - Rumours
  - Memes

Biological

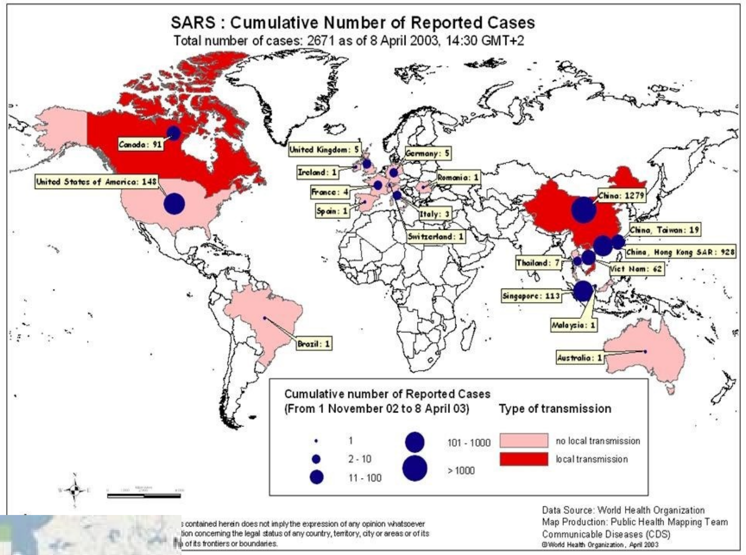
The Great Plague



HIV



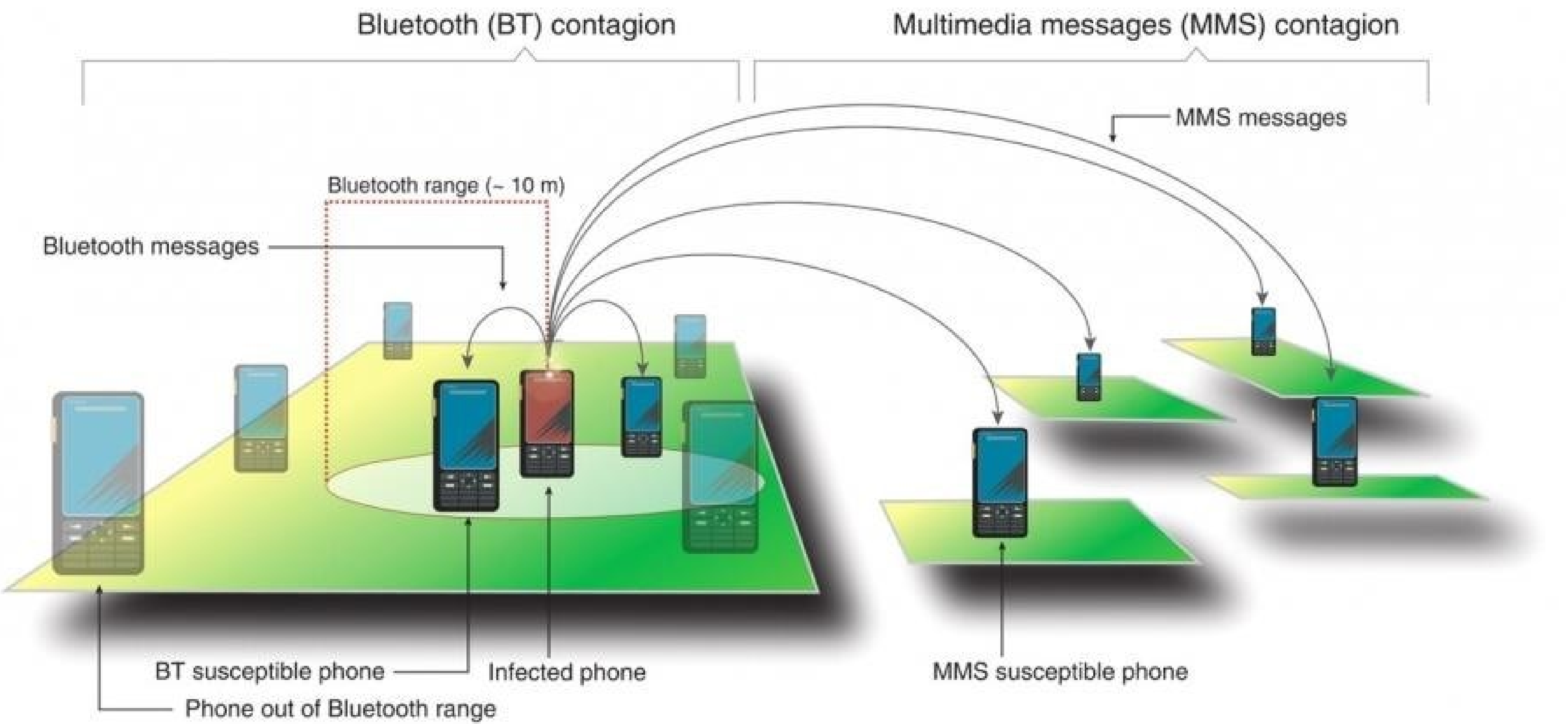
SARS



1918 Spanish flu

H1N1 flu

Digital



Bluetooth (BT) contagion
Multimedia messages (MMS) contagion
MMS messages
Bluetooth range (~ 10 m)
Bluetooth messages
BT susceptible phone
Infected phone
Phone out of Bluetooth range
MMS susceptible phone

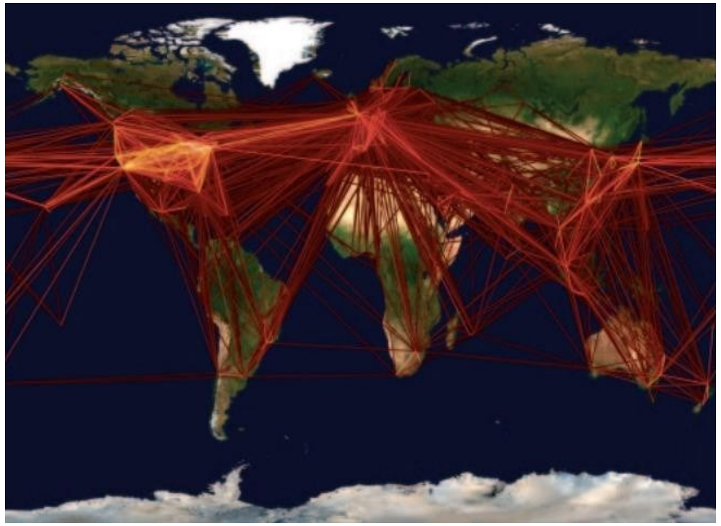
Social

PHENOMENA	AGENT	NETWORK
Venereal Disease	Pathogens	Sexual Network
Rumor Spreading	Information, Memes	Communication Network
Diffusion of Innovations	Ideas, Knowledge	Communication Network
Computer Viruses	Malwares, Digital viruses	Internet
Mobile Phone Virus	Mobile Viruses	Social Network/Proximity Network
Bedbugs	Parasitic Insects	Hotel - Traveler Network
Malaria	Plasmodium	Mosquito - Human network

Epidemic spreading – Why does it matter now?

High population density

High mobility



perfect conditions for epidemic spreading

Epidemic Modelling

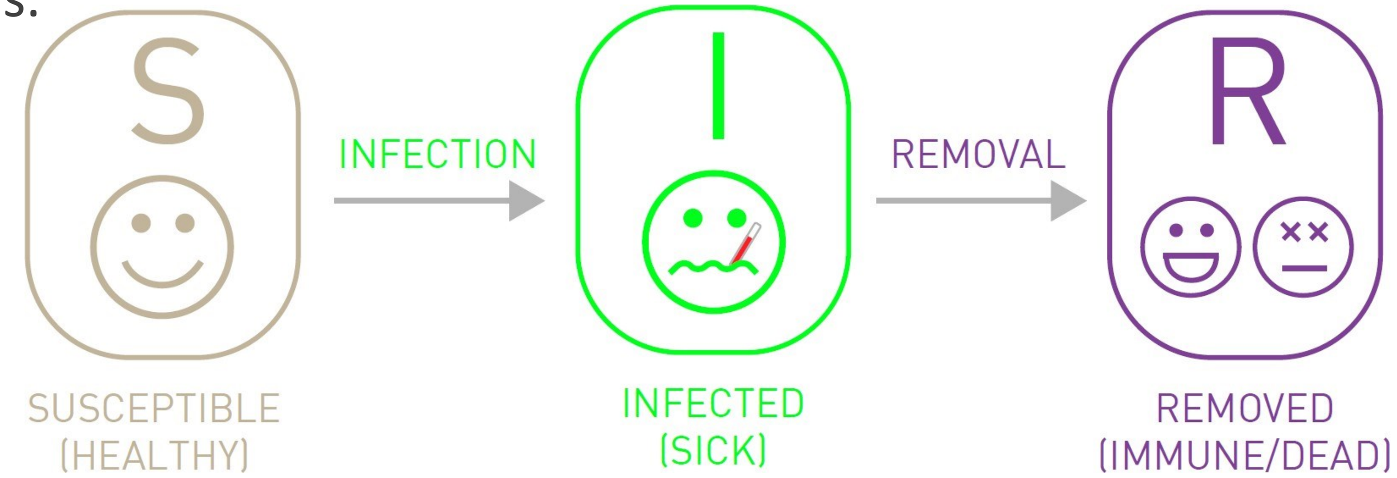
Epidemiology relies on two fundamental hypotheses:

Epidemic models classify each individual based on the stage of the disease affecting them.

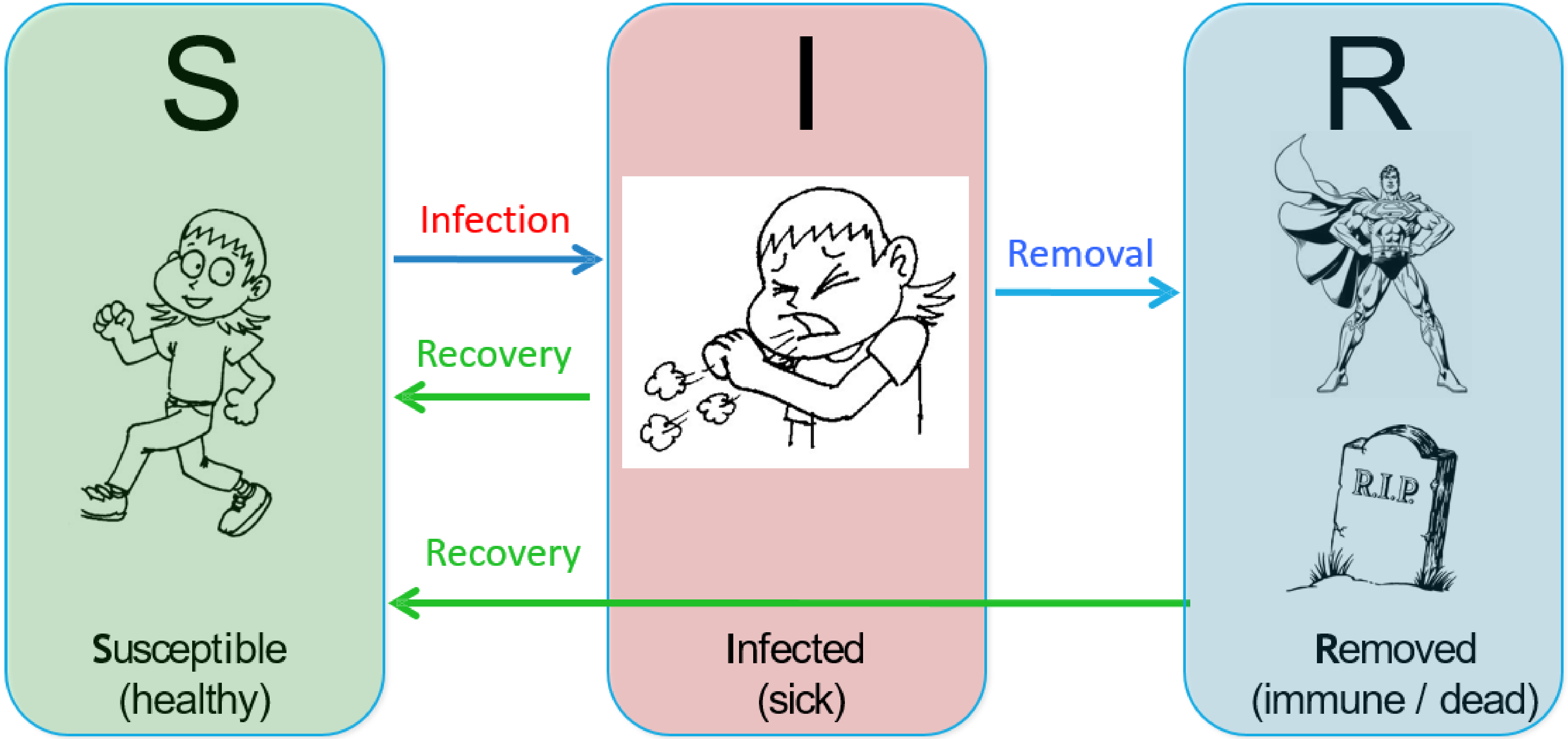
- **Susceptible** (**S**): Healthy individuals who have not yet contacted the pathogen.
- **Infectious** (**I**): Contagious individuals who have contacted the pathogen and hence can infect others.
- **Recovered** (**R**): Individuals who have been infected before, but have recovered from the disease, hence are not infectious.

Homogenous Mixing

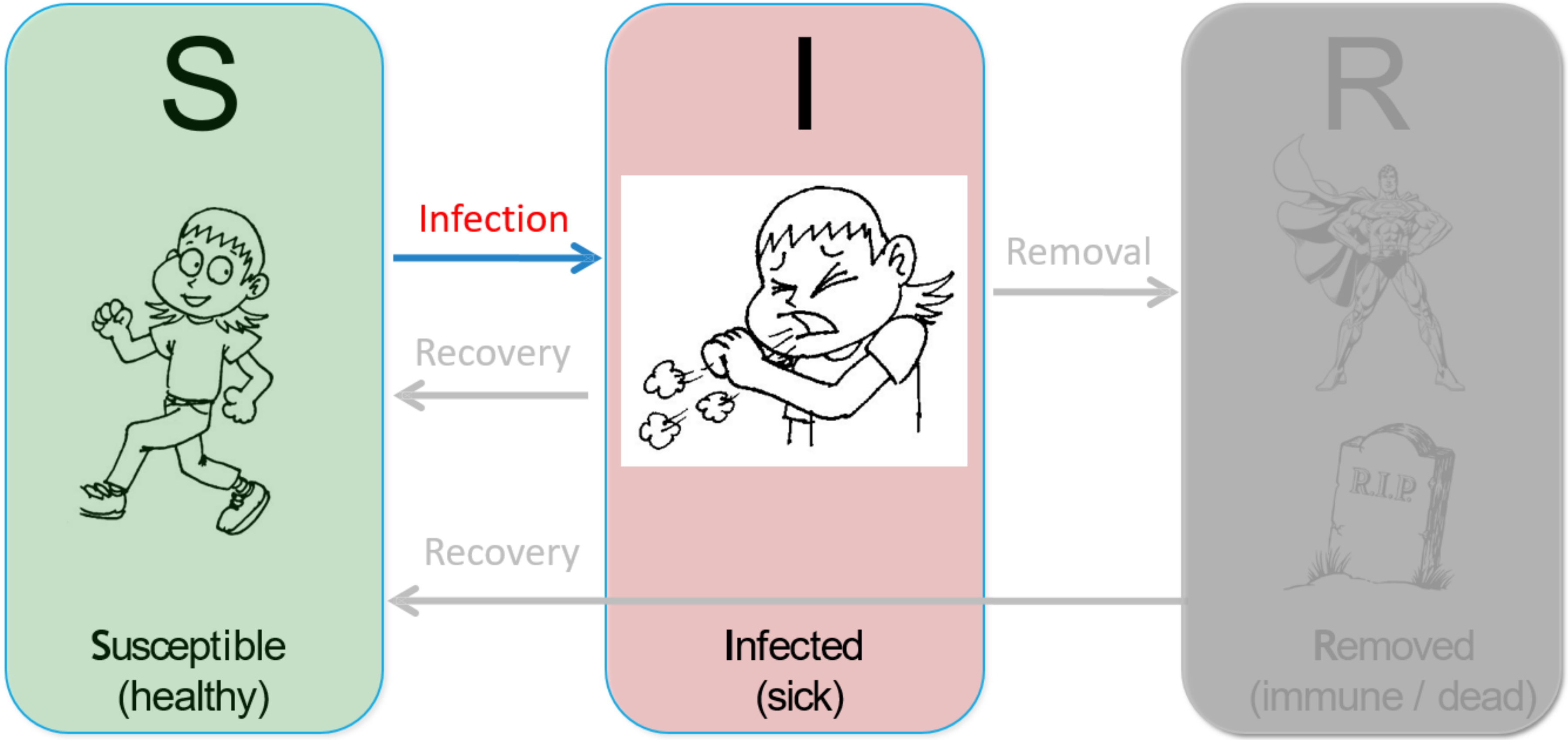
- Each individual has the same chance of coming into contact with an infected individual.



Classical Epidemic Models – Basic States



SI model


S
I
R
Infection
Recovery
Removal
Recovery
R.I.P
Susceptible
(healthy)
Infected
(sick)
Removed
(immune / dead)

SI model

N entity

$S(t)$ – number of healthy entity at time t

$I(t)$ – number of infected entity at time t

β – likelihood that the disease will be transmitted from an infected to a susceptible individual in a unit time

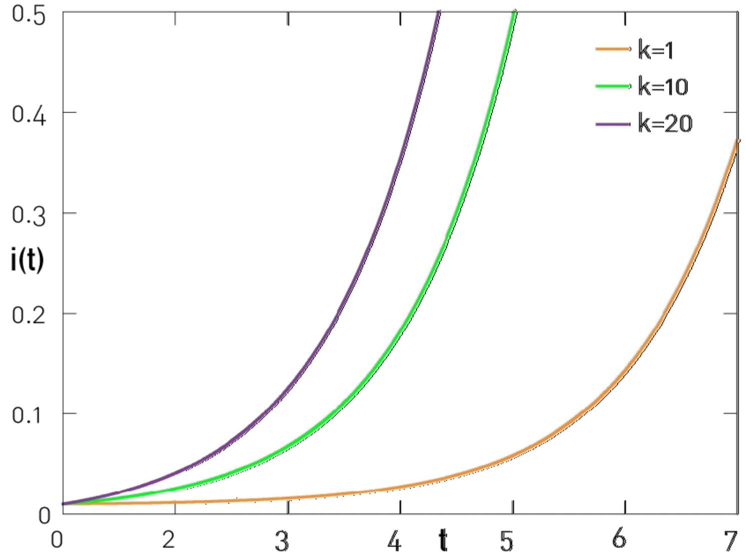
$S(0) = N - 1$

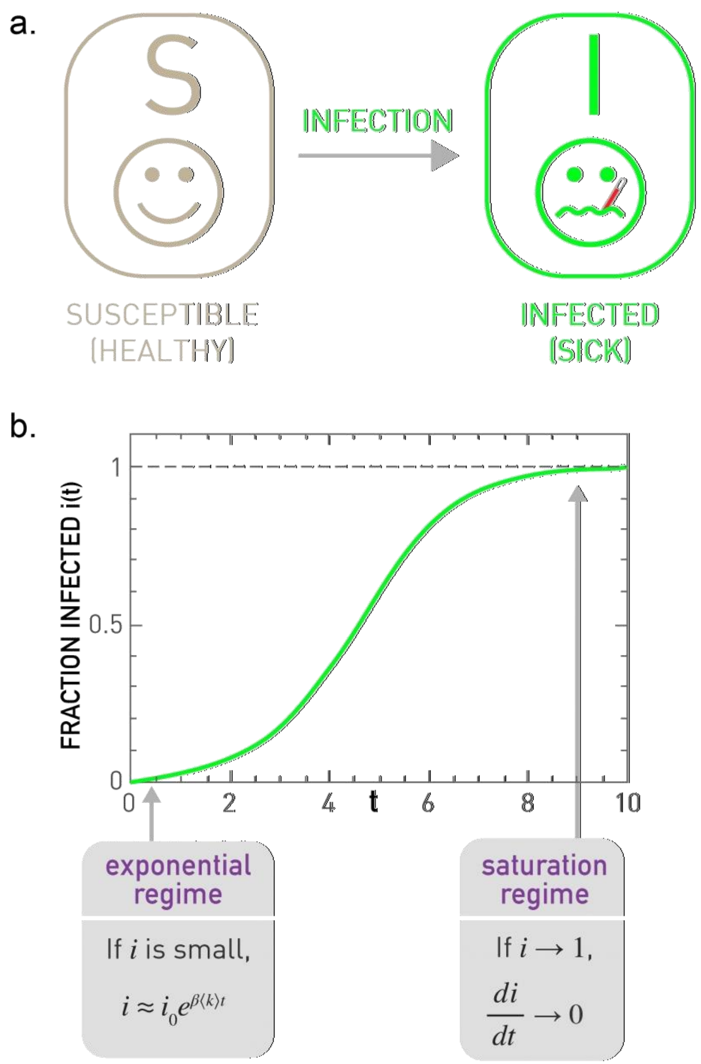
$I(0) = 1$

Dynamics:

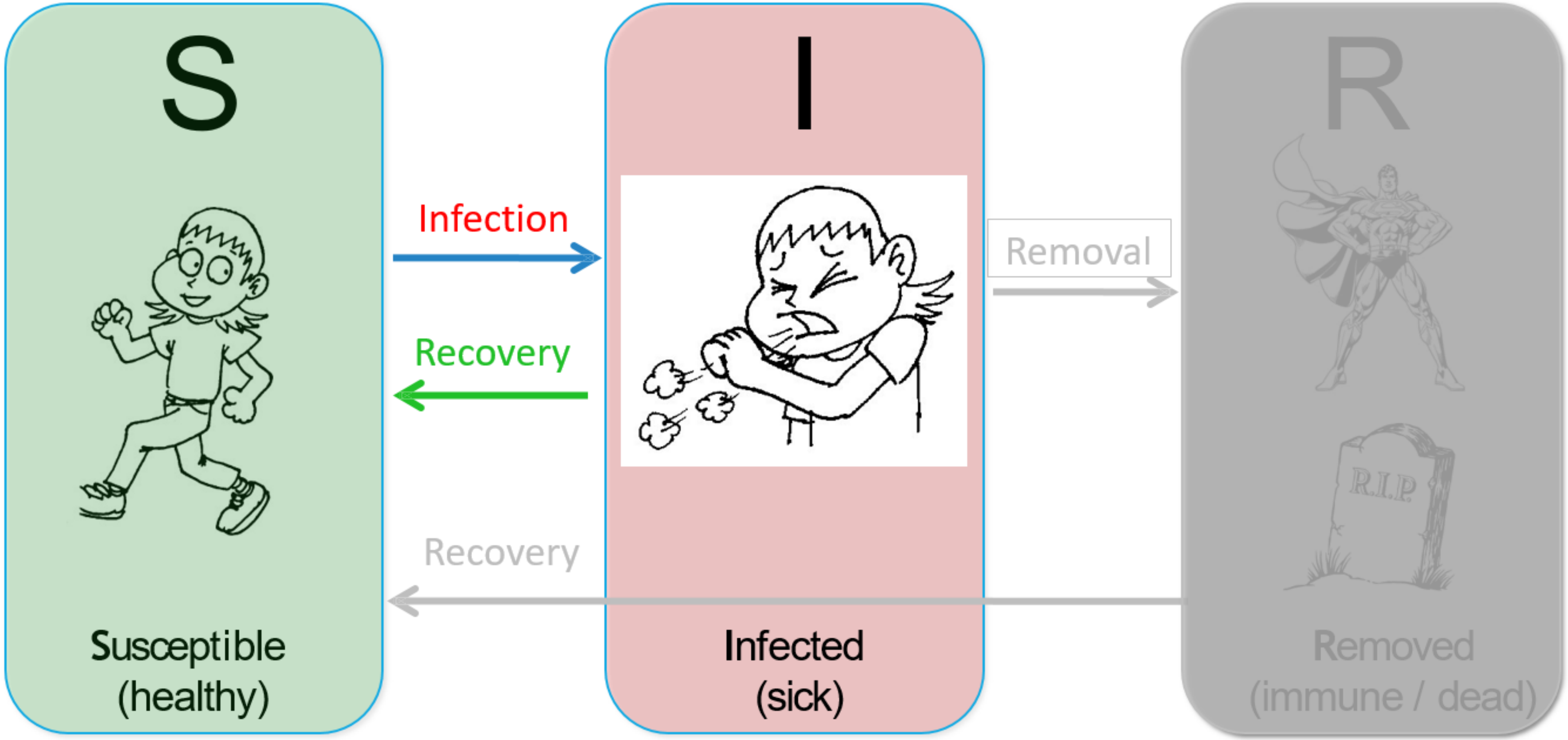
$$\frac{dI(t)}{dt} = \beta\langle k\rangle \frac{S(t)I(t)}{N}$$

E.g.: Toxoplasmosis





SIS model


S
I
R
Infection
Recovery
Removal
Recovery
R.I.P
Susceptible
(healthy)
Infected
(sick)
Removed
(immune / dead)

SIS model

Difference to *SI* model

- μ – recovery rate

In *SI* model, each entity gets infected

In case of *SIS* model:

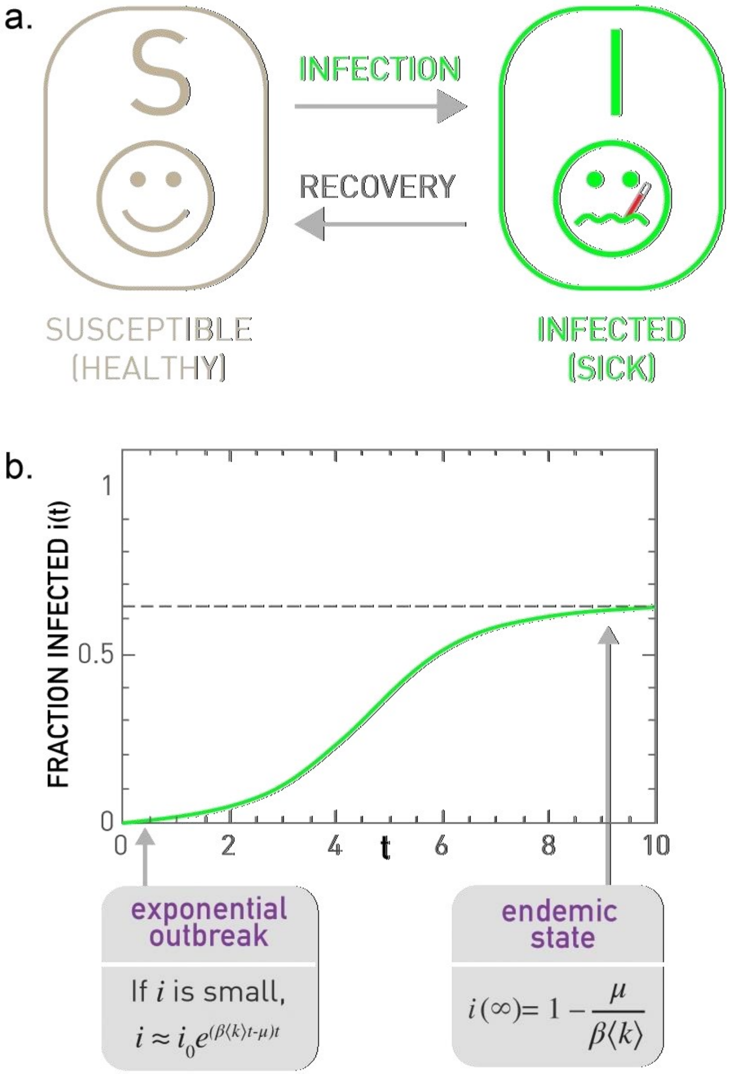
- μ – recovery rate: $I(\infty) = 1 - \frac{\mu}{\beta\langle k\rangle}$

In the *SIS* model the epidemic has two possible outcomes:

- Endemic State (In Hungarian: népbetegség)
  - $\mu < \beta\langle k\rangle$
- Disease-free State
  - $\mu > \beta\langle k\rangle$

Dynamics: $\frac{di}{dt} = \beta\langle k\rangle i(1-i) - \mu i$

E.g.: Common cold



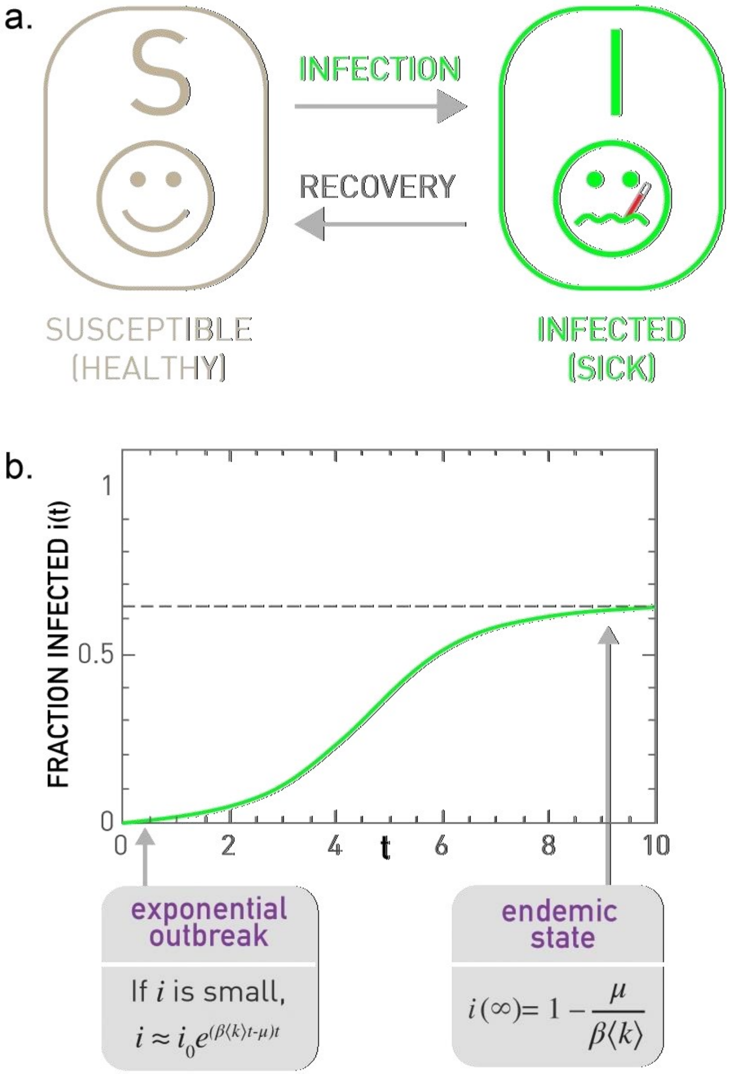
R_0 basic reproductive number

$R_0 = \frac{\beta\langle k \rangle}{\mu}$

R_0: number of susceptible that will be infected by an infected individual while he/she is infected

The reproductive number predicts the long-term fate of an epidemic

- $R_0 > 1$ the epidemic is in the endemic state
- $R_0 < 1$ the epidemic dies out

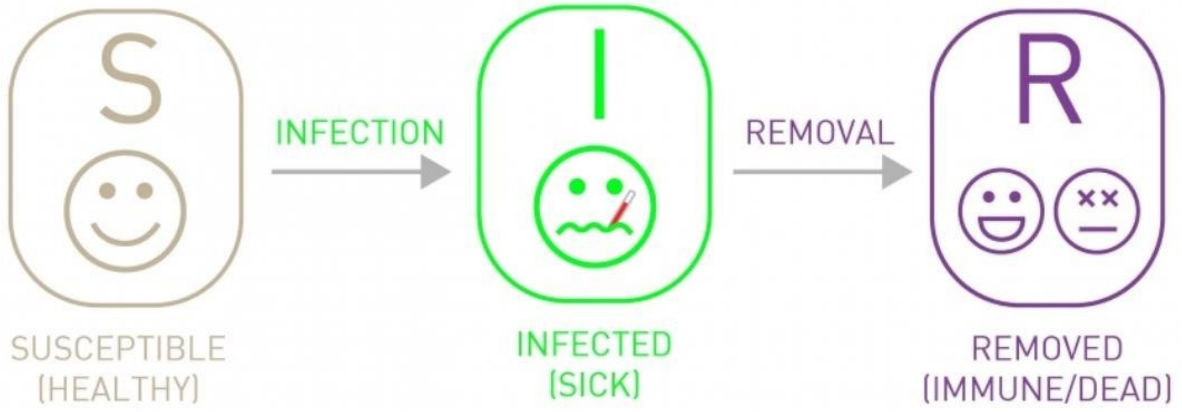


SIR model

In the *SIR* model recovered individuals enter a recovered state.

E.g.

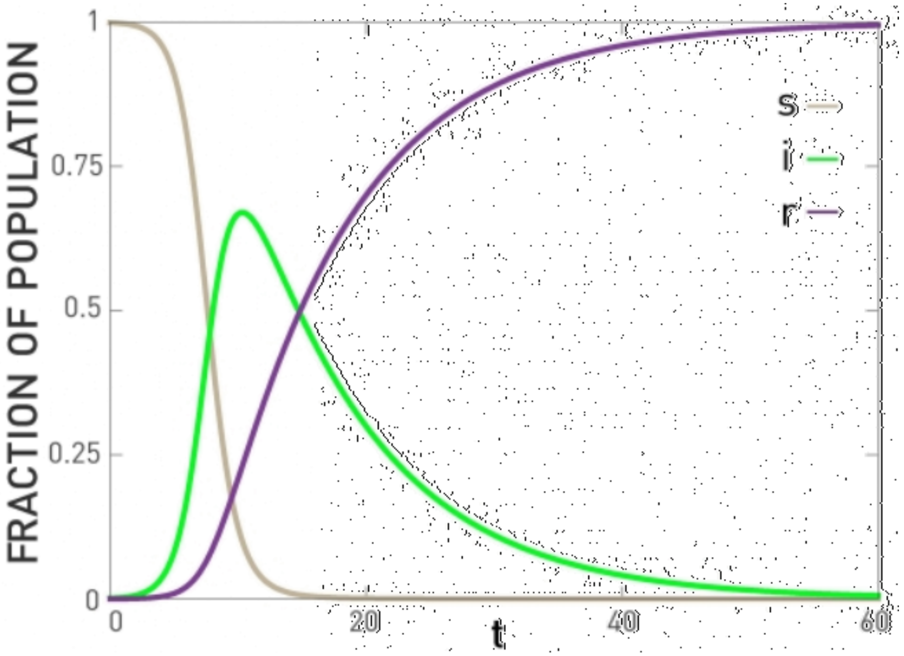
- Flu
- SARS
- Plague



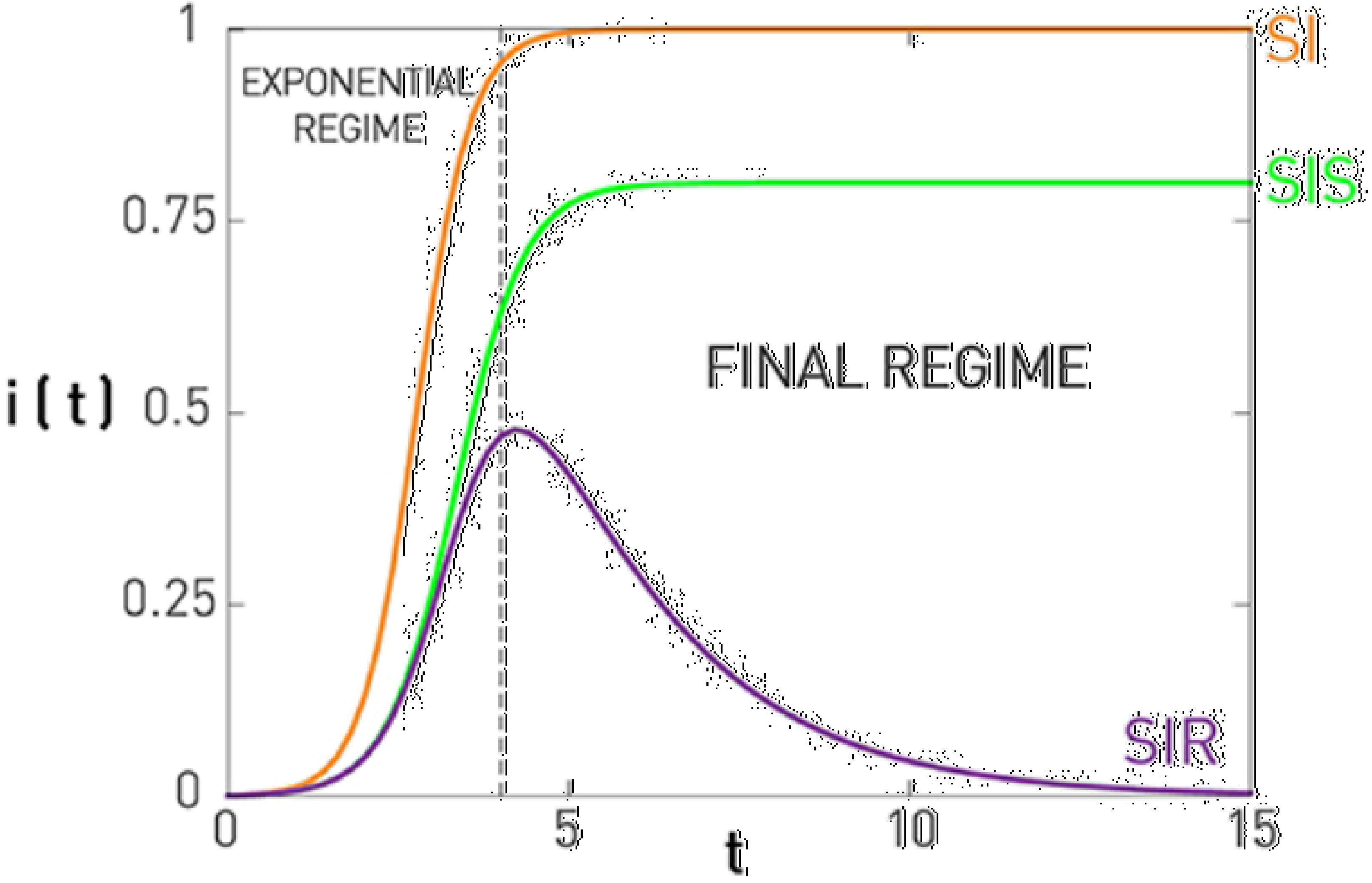
The reproductive number predicts the long-term fate of an epidemic:

- $R_0 < 1$ the pathogen persists in the population
- $R_0 > 1$ the pathogen dies out naturally

Dynamics: No closed solution



SI – SIS – SIR model comparison



EFOP-3.4.3-16-2016-00009

A felsőfokú oktatás minőségének és hozzáférhetőségének együttes javítása a Pannon Egyetemen

THANK YOU FOR YOUR KIND ATTENTION!

SZÉCHENYI 2020

MAGYARORSZÁG KORMÁNYA

Európai Unió
Európai Szociális Alap

BEFEKTETÉS A JÖVŐBE