



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Social Networks Analysis



Instructor

- Prof. Sabrina Gaito
- sabrina.gaito@unimi.it
- Via Celoria 18
- Zoom meeting by appointment via e-mail



Objectives

The learning objective of the course is provide students with the main concepts and methods of social network analysis.

Students will learn to manage data about network structure and to analyze, model and visualize such data to get valuable insights.

At the end of the course students will be able to design and carry out large-scale social network analysis studies.



Short Description

This course is an introduction to the concepts and methods of social network analysis.

It provides the main theories, models and methods in social network mining, as well as algorithms to handle large-scale networks efficiently.

By completing the course the students will be able to understand the basic concepts of social networks, to manage the fundamental concepts in analysing the large-scale data that are derived from social networks, to perform mining on large social networks and to visualize and get conclusions from the results.



Program

- Basic notions for networks from graph theory
- Networks models: Random model, scale-free networks, small-world networks
- Connected components
- Node centrality
- Link strength and reciprocity
- Transitivity, Triadic closure and Clustering coefficient
- Ego-networks
- Node similarity
- Node assortativity
- Dense subgraphs and Community detection
- Information diffusion

- Network visualization and basic analyses with Gephi
- Social network analysis with Python: the NetworkX library



Schedule

It will be held on zoom and the link will be posted before each lesson.

Please check any news on Ariel.

The recorded lesson will be made available after the lesson.

Course timetable:

Monday: 14.45 - 16.30

Thursday: 13.00 - 14.30



Final Examination

Oral exam: questions about definitions, methods, algorithms, concepts and calculations on the topics covered in the course, as well as discussions on real-data case studies.

A small project on the visualization and analysis of a social network from publicly available datasets (Gephi and/or Python)

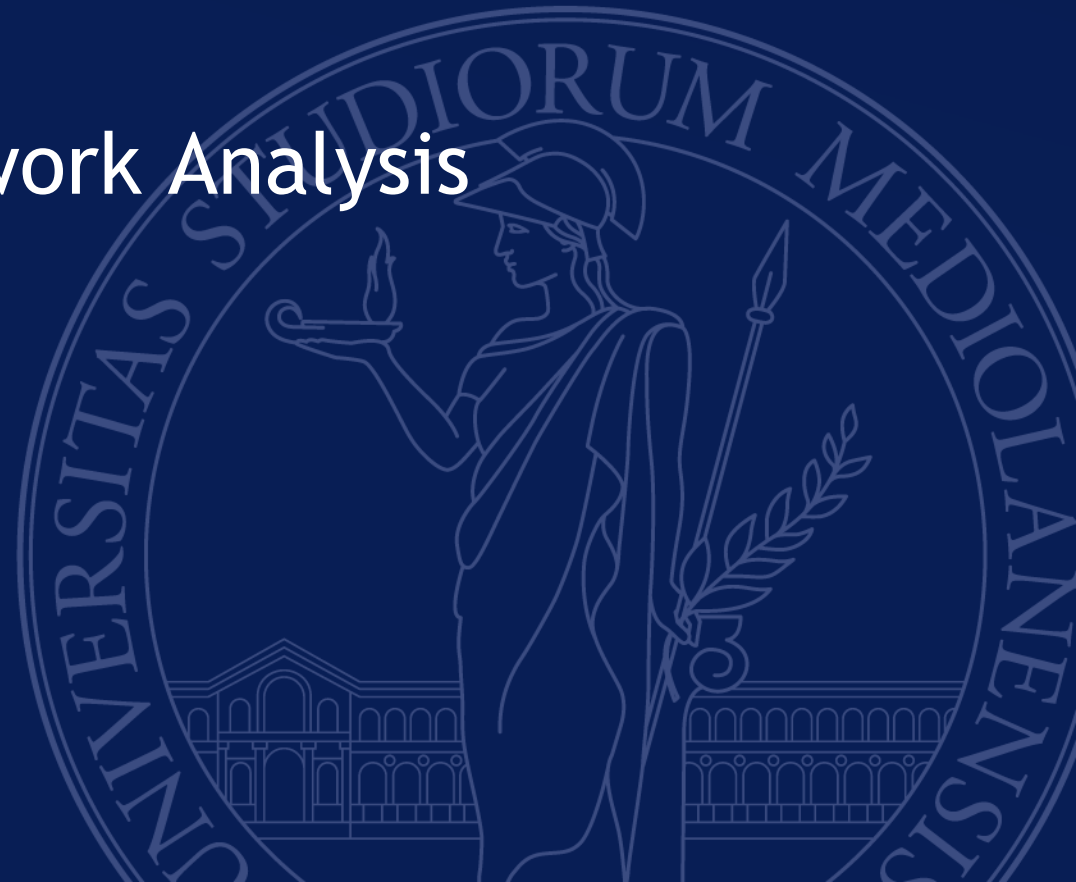


ANY QUESTIONS?



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Social Network Analysis



The social side of the Web

Social Media Landscape 2019



<https://fredcavazza.net/2020/04/21/panorama-des-medias-sociaux-2020/>

FredCavazza.net



Statistics on social media

Vinco's blog: vincos.it

World maps of the first and second ranked social networks on <https://vincos.it/world-map-of-social-networks/>

<https://wearesocial.com>

<https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media>

<https://wearesocial.com/it/digital-2020-italia>

<https://blog.hootsuite.com/social-media-statistics-for-social-media-managers/>



Sociological analysis

In 1974, Blau defined the field of *sociology* as follows:
... Social structures are defined by their parameters—
the criteria underlying the differentiation among people
and governing social interaction ...

The initial focus on the individual



A network perspective

In the 1930s, a new perspective on human data was developed: *sociometry*

instead of only looking at attributes of single persons or aggregating measures of groups of persons

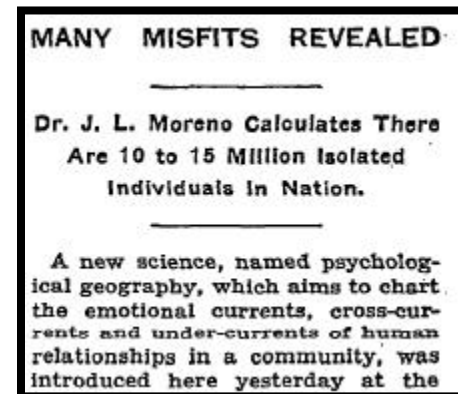
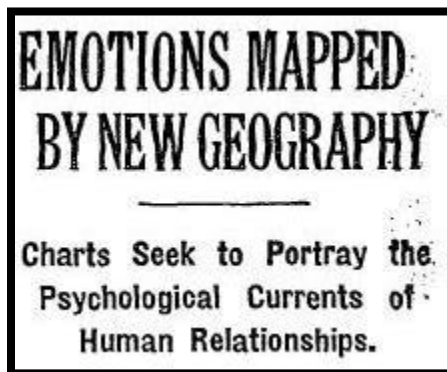
take into account who is connected to whom.



Early social network analysis

In 1933 Moreno displays the first sociogram at a meeting of the Medical Society of the state of New York

- article in NYT
- interests: effect of networks on e.g. disease propagation



Preceded by studies of (pre)school children in the 1920's

Source: The New York Times (April 3, 1933, page 17)



Social network analysis

Wasserman-Faust:

«...Focus on relationships among social entities, and on the patterns and implications of these relationships....

...The fundamental difference between a social network explanation and a non-network explanation of a process is the inclusion of concepts and information on relationships among units in a study...

...The network perspective differs in fundamental ways from standard social and behavioral science ... Rather than focusing on attributes of autonomous individual units, the social network perspective views characteristics of the social units as arising out of structural or relational processes or focuses on properties of the relational systems themselves...

...Relational ties among actors are primary and attributes of actors are secondary...»

Beyond Google Insights, Facebook analytics, ...



Social Network Analysis (SNA)

Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory [Wikipedia]

Social network analysis is both an approach to understanding social structure and a method of analysis which can be applied to other domains, such as web networks, biological networks, economic networks, financial networks, ...

Complex networks theory

Network Science



NETWORKS AT THE HEART OF COMPLEX SYSTEMS

*“I think the next century
will be the century
of complexity.”*

Stephen Hawking

January 23, 2000`

Network Science: Introduction



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Behind each complex system there is a **network**, that defines the interactions between the component.

We will never understand complex system unless we map out and understand the networks behind them.



The network describing the interactions between genes, proteins, and metabolites integrate the processes behind living streams.

The wiring diagram capturing the connections between neural cells hold the key to our understanding of brain functions.

The sum of all professional, friendship, and family ties is the fabric of the society.

Trade networks maintain our ability to exchange goods and services, being responsible for the material prosperity. They also play a key role in the spread of financial and economics crises.

Networks are at the heart of some of the most revolutionary technologies of the 21st century, empowering everything from Google to Facebook, CISCO, and Twitter.

At the end, networks permeate science, technology, and nature



Despite amazing the diversity in form, size, nature, age, and scope present in real networks, most networks observed in nature, society, and technology are driven by common organizing principles.

In other words once we disregard the nature of the components and their interactions, the obtained networks appear to be more similar than different from each other.

NETWORK SCIENCE



Introduction to network science

[“Networks are everywhere” with Albert-László Barabási](#)

Introduction: first 4 minuted

An award-winning documentary, *Connected*, by Australian filmmaker Annamaria Talas, has brought the field to our TV screen, being broadcasted all over the world and winning several prestigious prizes

<https://www.youtube.com/watch?v=2rzxAyY7D7k>

The documentary introduction on:

<https://www.youtube.com/watch?v=zK1Cb9qj3qQ>



Why didn't network science emerge
two hundred years ago?



TWO FORCES HELPED THE EMERGENCE OF NETWORK SCIENCE

Network Science: Introduction



To describe the behavior of a system consisting of hundreds to billions of interacting components, we first need a map of the system's wiring diagram.

In the past, we either lacked the tools to map these networks out, or it was difficult to keep track of the huge amount of data behind these maps.

The emergence of the Internet, offering effective and fast data sharing methods, together with cheap digital storage, fundamentally changed this, allows us to collect, assemble, share, and analyze data pertaining to real networks.



Movie Actor Network, 1998;
World Wide Web, 1999.
C elegans neural wiring diagram 1990
Citation Network, 1998
Metabolic Network, 2000;



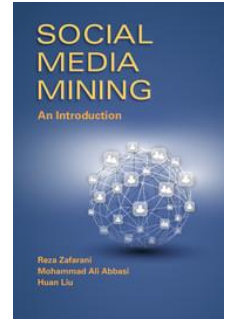
References

Reza Zafarani, Mohammad Ali Abbasi, Huan Liu

Social Media Mining: An Introduction

A Textbook by Cambridge University Press

<http://www.socialmediamining.info/>



Albert-László Barabási

Network Science

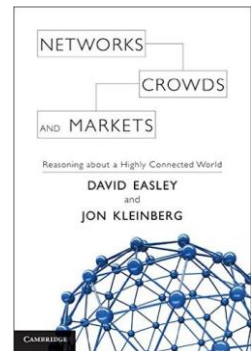
<http://barabasi.com/networksciencebook/>



D. Easley, J. Kleinberg

Networks, Crowds, and Markets: Reasoning About a Highly Connected World

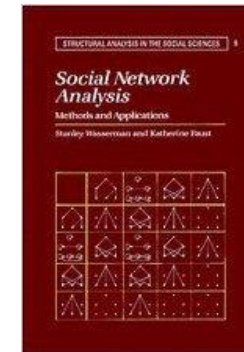
<http://www.cs.cornell.edu/home/kleinber/networks-book/>



Wasserman, Stanley and Katherine Faust. 1994.

Social Network Analysis: Methods and Applications.

Cambridge: Cambridge University Press.

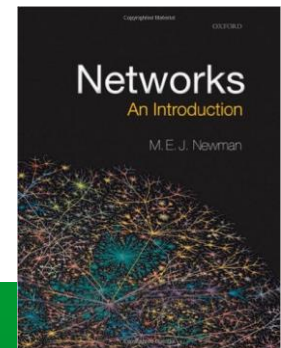


Newman, M.E.J.

Networks: An Introduction.

Oxford University Press. 2010.

<https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199206650.001.0001/acprof-9780199206650>



Networks and Graphs

A-L. Barabási , Network Science , available online, 2015.

D. Easley and J. Kleinberg, Networks, Crowds and Markets , Cambridge Univ Press, 2010 (also available online).

M.E.J. Newman, Networks - An introduction , Oxford Univ Press, 2010.

S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications, Cambridge Univ Press, 1994.

Zafarani et al., Social Media Mining, Cambridge University press, 2014

Wasserman

NETWORKS

Network: definition

A network consists of a finite set of actors (*nodes*) and the relations (*links, ties, edges*) defined on them.

In network science relation ties among actors are primary and attributes of actors are secondary..

From complex systems to networks

The choice of the proper network representation determines our ability to use network theory successfully.

In some cases there is a unique, unambiguous representation.

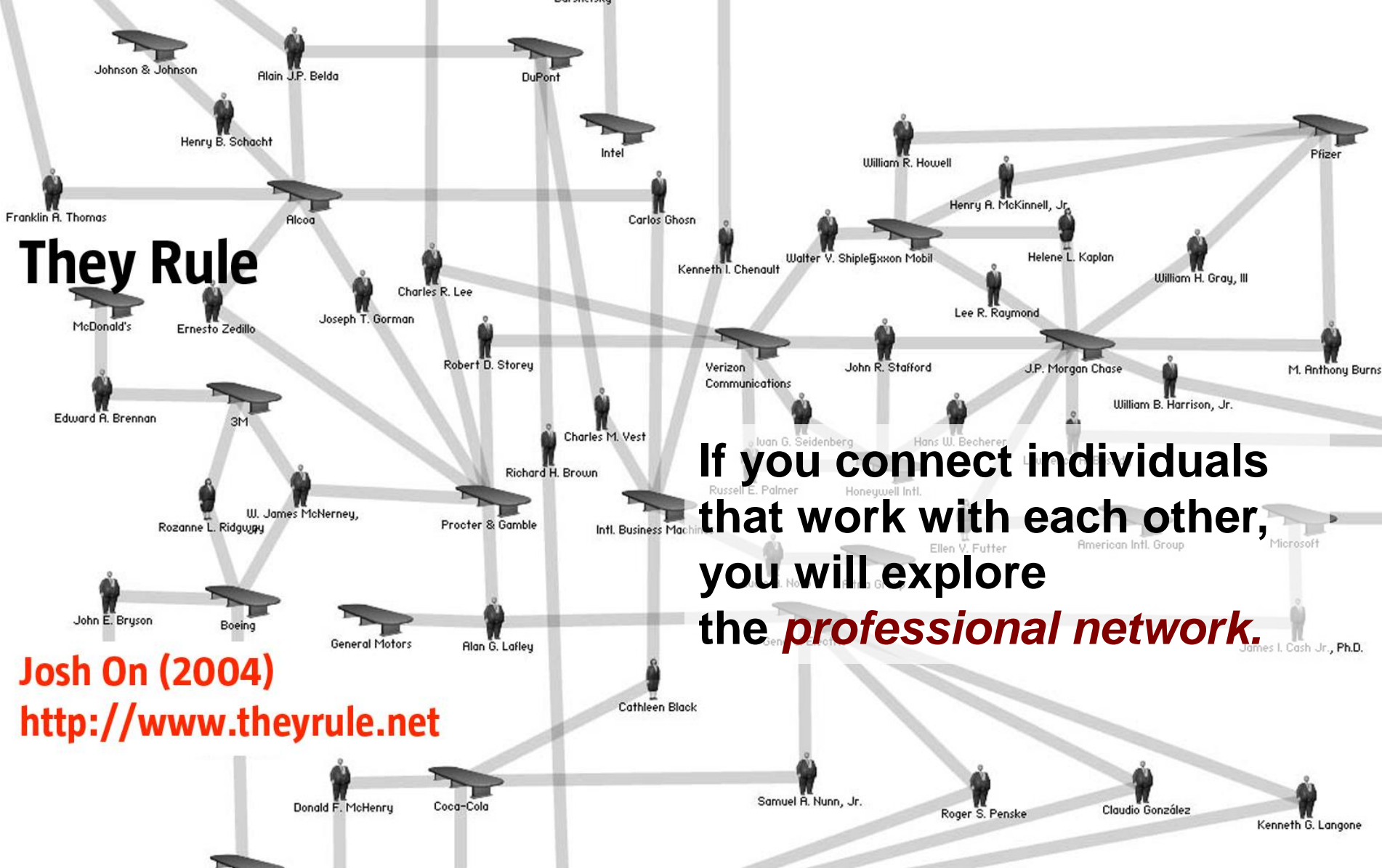
In other cases, the representation is by no means unique.

For example, the way we assign the links between a group of individuals will determine the nature of the question we can study.

They Rule

If you connect individuals that work with each other, you will explore the *professional network*.

Josh On (2004)
<http://www.theyrule.net>



Which network?

If you connect individuals based on their first name (*all Peters connected to each other*), you will be exploring what?

It is a network, nevertheless.

Examples of networks

NETWORK

Internet

WWW

Power Grid

Mobile Phone Calls

Email

Science Collaboration

Actor Network

Citation Network

E. Coli Metabolism

Protein Interactions

NODES

Routers

Webpages

Power plants, transformers

Subscribers

Email addresses

Scientists

Actors

Paper

Metabolites

Proteins

LINKS

Internet connections

Links

Cables

Calls

Emails

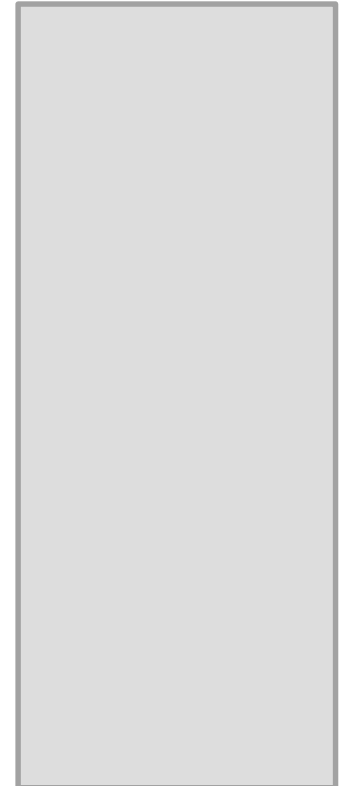
Co-authorship

Co-acting

Citations

Chemical reactions

Binding interactions



Directed and undirected networks

The links of a network can be *directed* or *undirected*. Some systems have directed links, like the WWW, whose uniform resource locators (URL) point from one web document to the other, or phone calls, where one person calls the other. Other systems have undirected links, like romantic ties: if I date Janet, Janet also dates me, or like transmission lines on the power grid, on which the electric current can flow in both directions.

NETWORK

Internet

WWW

Power Grid

Mobile Phone Calls

Email

Science Collaboration

Actor Network

Citation Network

E. Coli Metabolism

Protein Interactions

NODES

Routers

Webpages

Power plants, transformers

Subscribers

Email addresses

Scientists

Actors

Paper

Metabolites

Proteins

LINKS

Internet connections

Links

Cables

Calls

Emails

Co-authorship

Co-acting

Citations

Chemical reactions

Binding interactions

Directed or
undirected ?

Directed and undirected networks

NETWORK

Internet

WWW

Power Grid

Mobile Phone Calls

Email

Science Collaboration

Actor Network

Citation Network

E. Coli Metabolism

Protein Interactions

NODES

Routers

Webpages

Power plants, transformers

Subscribers

Email addresses

Scientists

Actors

Paper

Metabolites

Proteins

LINKS

Internet connections

Links

Cables

Calls

Emails

Co-authorship

Co-acting

Citations

Chemical reactions

Binding interactions

DIRECTED UNDIRECTED

Undirected

Directed

Undirected

Directed

Directed

Undirected

Undirected

Directed

Directed

Undirected

NETWORKS AND GRAPHS

Graphs

The mathematical representation of a network is a graph

System	Network	Graphs
Actors	Nodes	Vertices
Interactions	Links	Edges, Ties

The two terms are often used interchangeably.

$G(N,L)$ or $G(V,E)$

Graphs

The mathematical representation of a network is a (*graph*).

A graph is an ordered pair $G = (V, E)$ comprising a set V of vertices, with a set E of edges, which are 2-element subsets of V (i.e., an edge is associated with a pair of the vertices).

A set of vertices $V = \{v_1, v_2, \dots, v_N\}$

A set of edges $E = \{e_1, e_2, \dots, e_L\}$

where an edge is a pair of vertices $e_k = (v_i, v_j)$,
called (*end-points*)

Networks and Graphs

We will use the following notation to denote both the network and the relative graph: $G = (N, L)$ where $N = \{n_1, n_2, \dots, n_N\}$ is the set of N nodes of the network and $L = \{l_1, l_2, \dots, l_L\}$ is the set of L link of the network.

Network size: cardinality of L , number of links of the network.

Network order: cardinality of V , number of nodes of the networks.

Note: network size is often used for the number of nodes, too.

Undirected graph

In an undirected graph, a link is an unordered pair of vertices (n_i, n_j) . Note that (n_i, n_j) and (n_j, n_i) are the same edge.

Two nodes n_i e n_j are *adjacent* if (n_i, n_j) exists in L .

Two links are *consecutive* if they share an end-point.

Directed graph or digraph

Edges can have directions. A directed edge is sometimes called an arc.
In a directed graph a relation from node n_i to node n_j , is an ordered pair of nodes $l_k = (n_i, n_j)$.

The two links (n_i, n_j) and (n_j, n_i) are different.

A directed graph is a graph whose link are directed.

n_i is called *origin, sender*

n_j is called *terminus, receiver*

Directed graph

In directed graph the arc direction has to be taken into account

n_i is *adjacent to* n_j if there is $(n_i, n_j) \in L$.

n_j *adjacent from* n_i if there is $(n_i, n_j) \in L$.

Consecutive links:



yes



No

Graph drawing

Graphs are represented visually by drawing a dot or circle for every vertex, and drawing an arc between two vertices if they are connected by an edge. If the graph is directed, the direction is indicated by drawing an arrow.

A graph drawing should not be confused with the graph itself (the abstract, non-visual structure) as there are several ways to structure the graph drawing.

Example

A is friend of B, D and F,

B is friend of C,

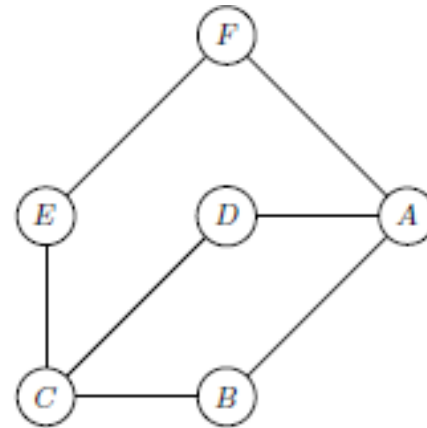
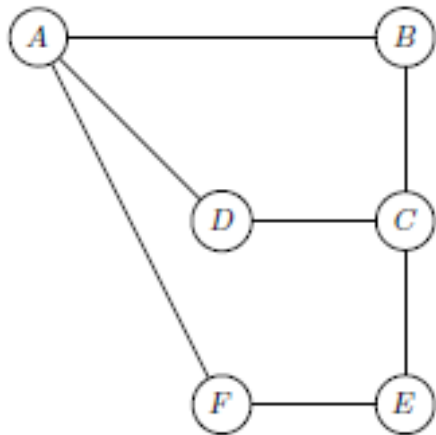
C is friend of D and E is friend of F.

Network: $G = (N, L)$

where

$N = \{A, B, \dots, F\}$

$L = \{(A, B), (A, D), (A, F), (B, C), (C, D), (C, E), (E, F)\}$.



Example

A invites B, D and F,

B invites C,

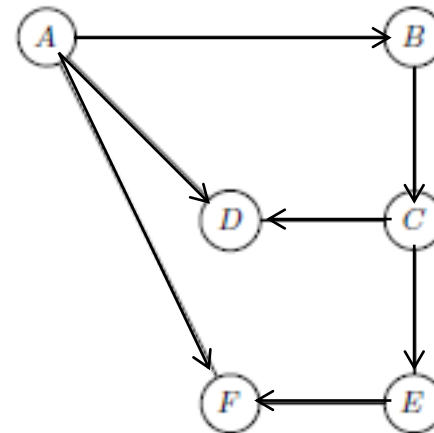
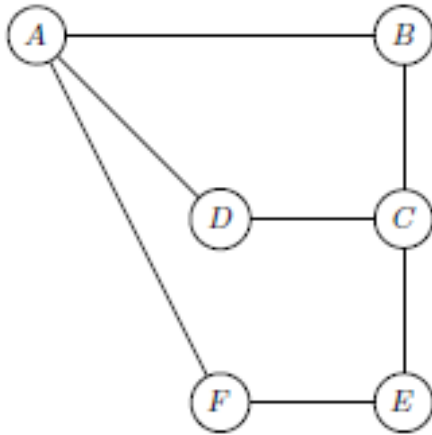
C invites D and E invites F.

Network: $G = (N, L)$

where

$N = \{A, B, \dots, F\}$

$L = \{(A, B), (A, D), (A, F), (B, C), (C, D), (C, E), (E, F)\}$.



Networks and graphs

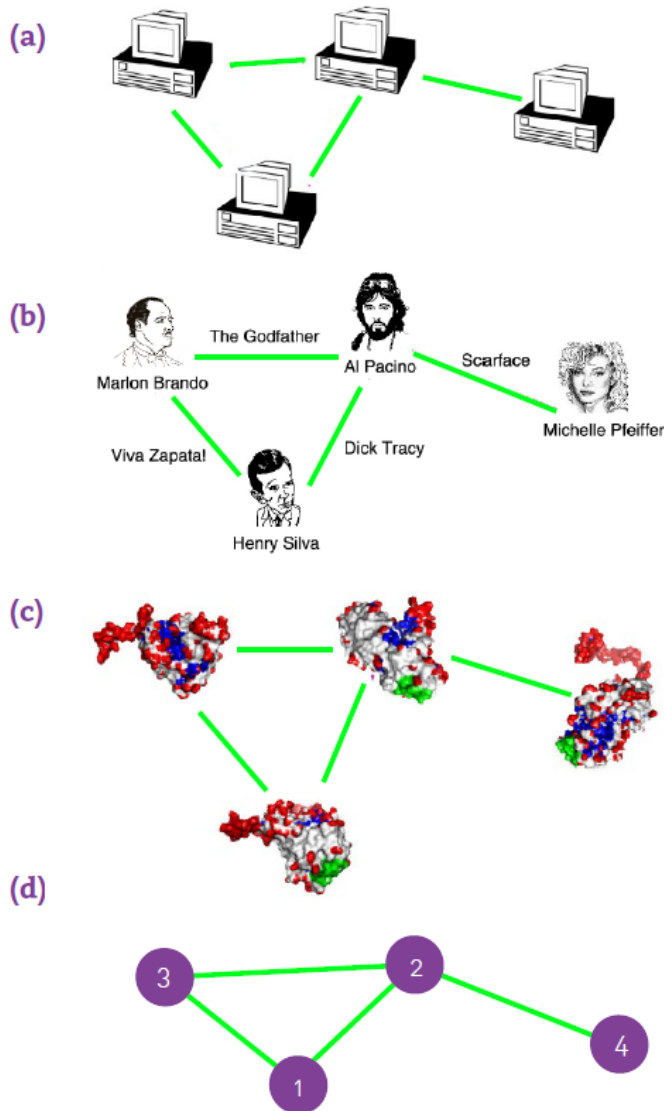


Figure 2.2
Different Networks, Same Graph

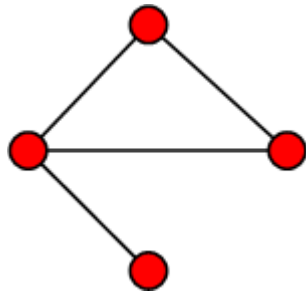
The figure shows a small subset of (a) the Internet, where routers (specialized computers) are connected to each other; (b) the Hollywood actor network, where two actors are connected if they played in the same movie; (c) a protein-protein interaction network, where two proteins are connected if there is experimental evidence that they can bind to each other in the cell. While the nature of the nodes and the links differs, these networks have the same graph representation, consisting of $N = 4$ nodes and $L = 4$ links, shown in (d).

$$N = \{1, 2, 3, 4\}$$
$$L = \{(1, 2), (1, 3), (2, 3), (2, 4)\}$$

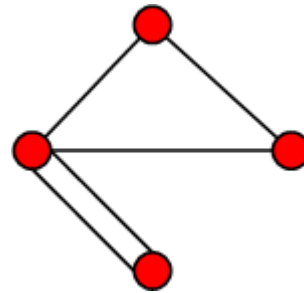
Simple graph

A *loop* is an edge between the same node (n_i, n_i) .

A simple graph is an undirected graph containing no graph loops or multiple edges



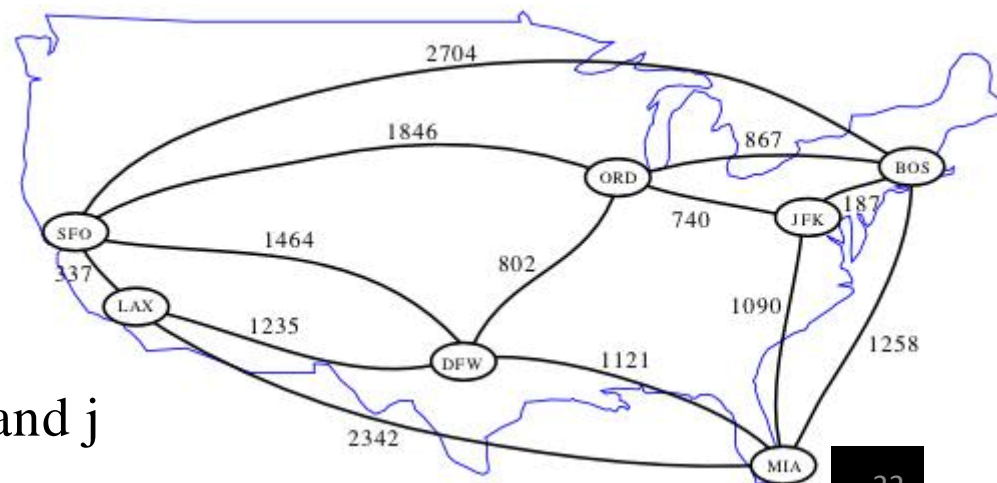
Simple graph



Multigraph

Weighted networks

- A weighted graph is one where edges are associated with weights
 - For example, a graph could represent a map where nodes are cities and edges are routes between them
 - The weight associated with each edge could represent the distance between these cities



$G(V, E, W)$

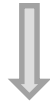
$$A_{ij} = \begin{cases} w, w \in \mathbb{R} \\ 0, \text{There is no edge between } i \text{ and } j \end{cases}$$

Networks and graphs

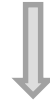
Complex system



Network



Graph



Wiring diagram

GRAPH MATHEMATICAL REPRESENTATION

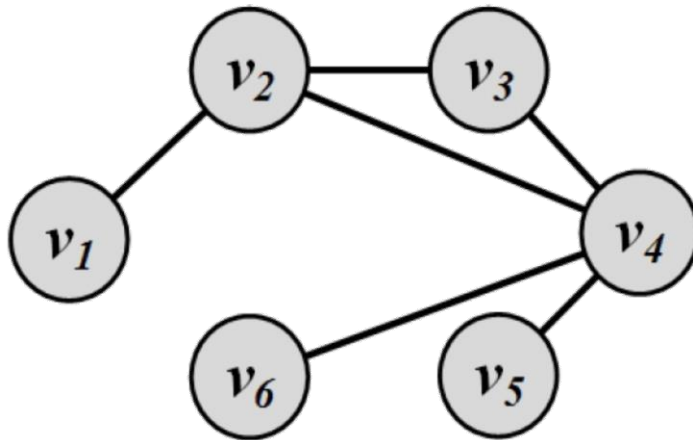
Graph mathematical representations

- Edge List
- Adjacency List
- Adjacency Matrix

Note: we are not speaking about efficient data structure

Edge List

- List of nodes and links



(v_1, v_2)

(v_2, v_3)

(v_2, v_4)

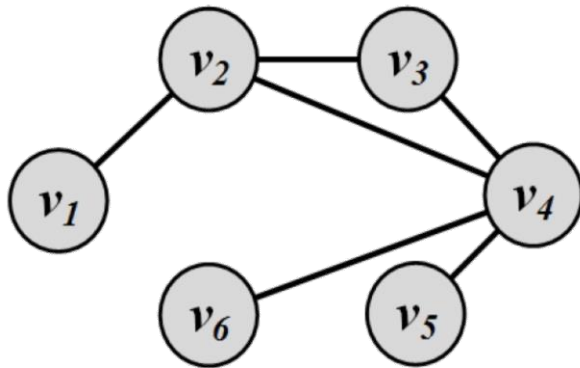
(v_3, v_4)

(v_4, v_5)

(v_4, v_6)

Adjacency list

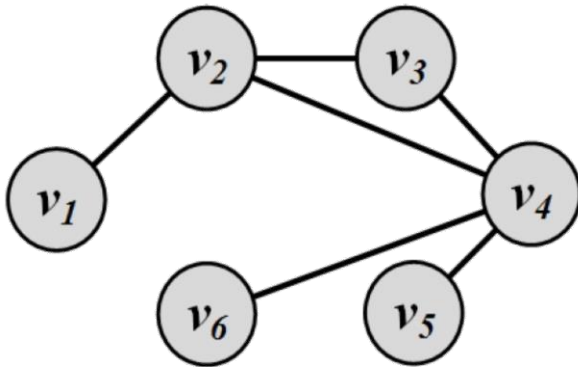
- For each node, the list of nodes which is connected to



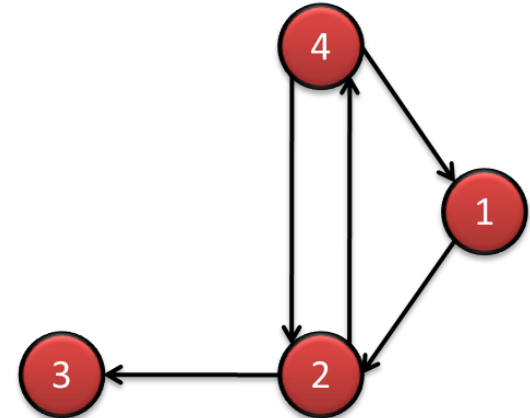
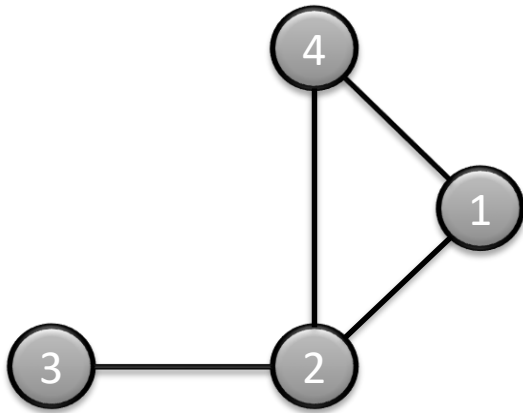
Node	Connected To
v_1	v_2
v_2	v_1, v_3, v_4
v_3	v_2, v_4
v_4	v_2, v_3, v_5, v_6
v_5	v_4
v_6	v_4

Adjacency matrix

$$A_{ij} = \begin{cases} 1, & \text{if there is a link between nodes } v_i \text{ and } v_j \\ 0, & \text{otherwise} \end{cases}$$



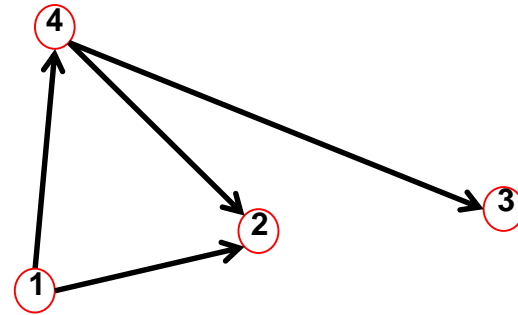
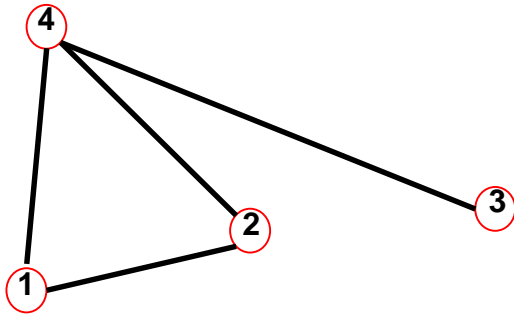
	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	1	0	0	0	0
v_2	1	0	1	1	0	0
v_3	0	1	0	1	0	0
v_4	0	1	1	0	1	1
v_5	0	0	0	1	0	0
v_6	0	0	0	1	0	0



- The adjacency matrix for directed graphs is not symmetric ($A \neq A^T$)
 - ($A_{ij} \neq A_{ji}$)
- The adjacency matrix for undirected graphs is symmetric ($A = A^T$)

Directed graph

$$A_{ij} = \begin{cases} 1, & \text{if there is a link from node } v_j \text{ to node } v_i \\ 0, & \text{otherwise} \end{cases}$$

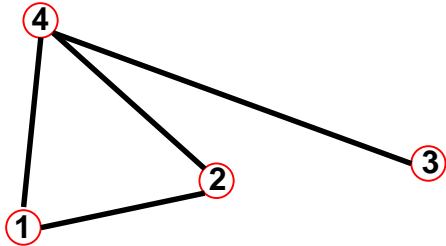


$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

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

Adjacency

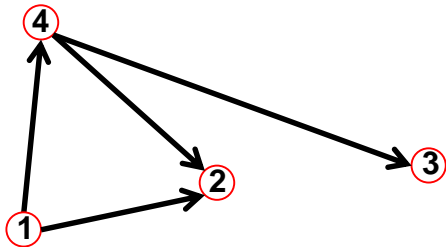
Undirected



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

$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

Directed



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

$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

Row:
in-links
Column:
out-links



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Simple networks

Weighted, (un)directed network

Affiliation networks

Bipartite networks

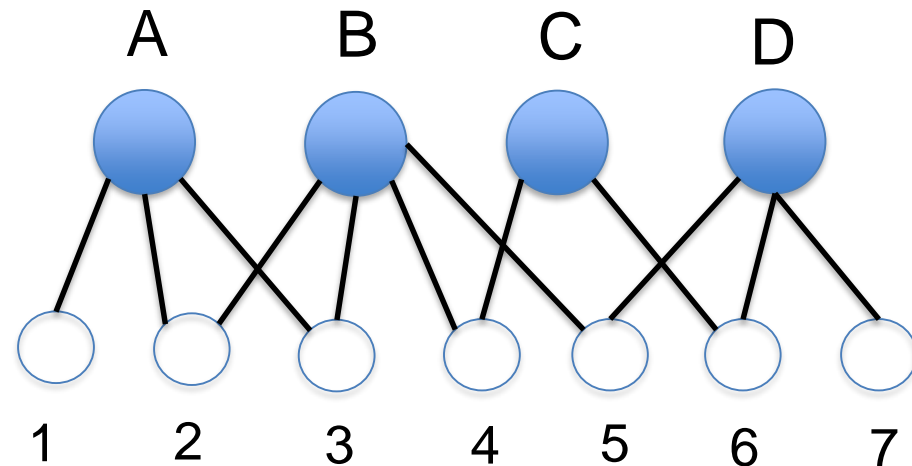
Networks

- We will consider networks which have:
 - No loops
 - No multiple edges
- We will consider:
 - Directed and undirected networks
 - Weighted and unweighted networks



A special case: Affiliation networks

- They have two types of nodes:
 - Actors
 - Groups
- Representation by bipartite (two-mode) networks
- Links connect actors to groups
 - No links between actors
 - No links between groups
- Actors are connected via co-membership of groups



The incidence matrix is a rectangular matrix $g \times n$

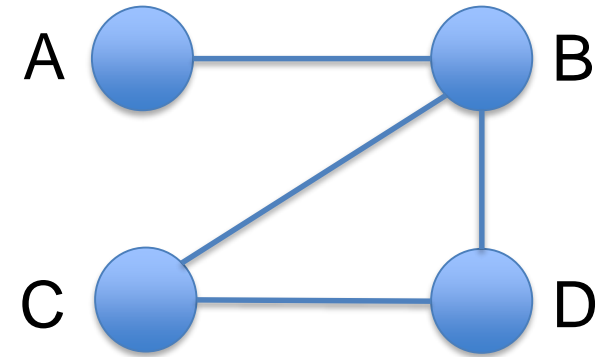
$$B_{ij} = \begin{cases} 1 & \text{if node } j \text{ belongs to group } i \\ 0 & \text{otherwise} \end{cases}$$



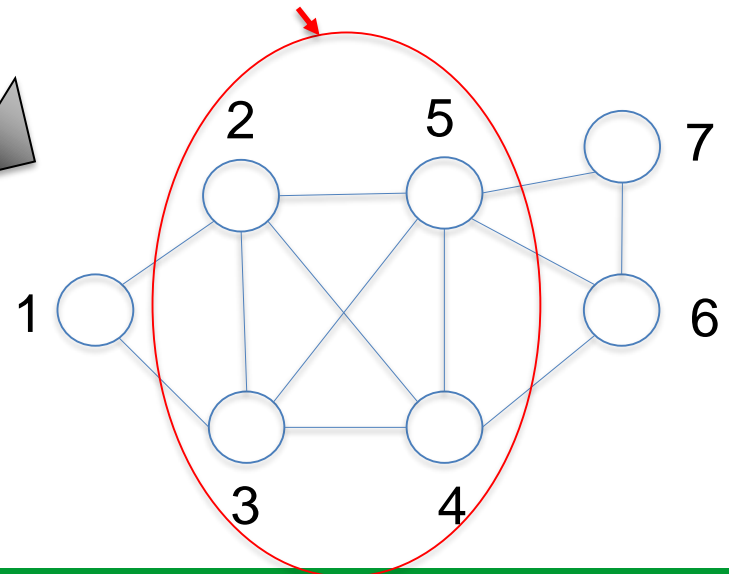
One-mode projections

Projection onto groups

Nodes are groups, two nodes are connected by a link if they share an actor



Nodes sharing a group form a clique

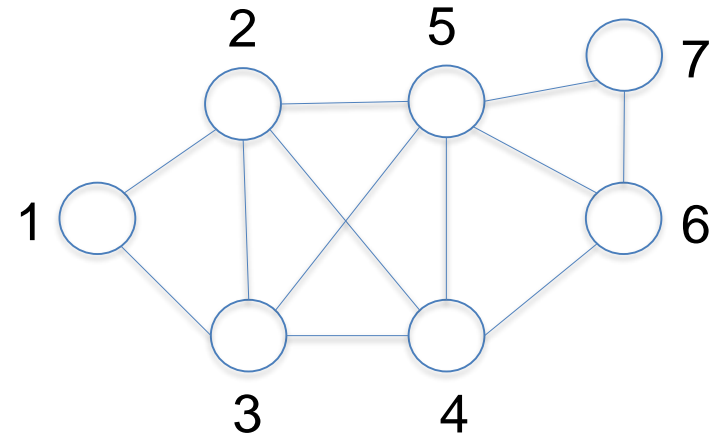
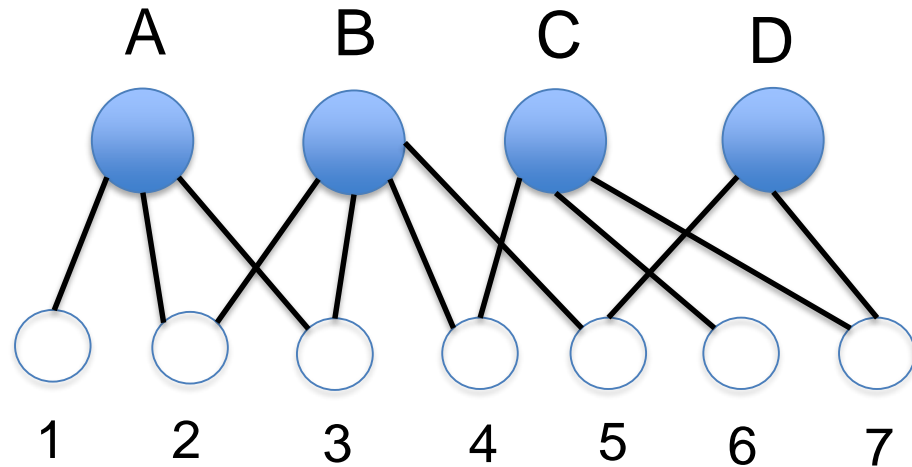


Projection onto actors

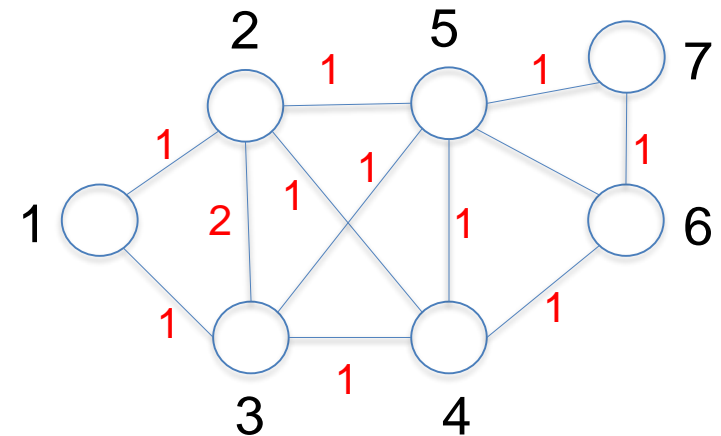
Nodes are actors, two actors are connected if they share a group



One-mode weighted projection

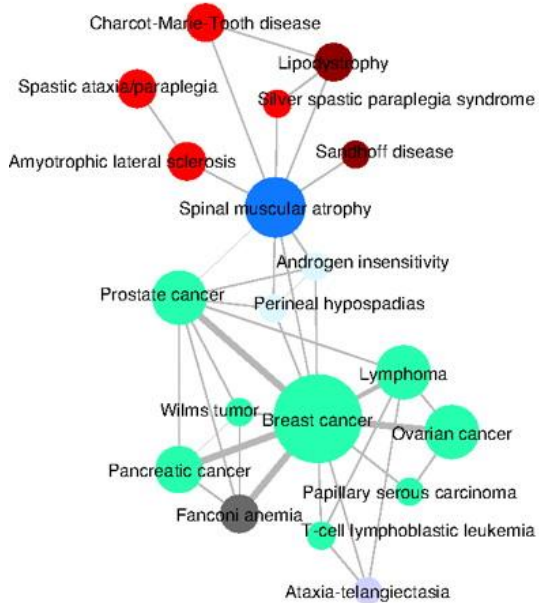


Information loss:
how many groups two nodes share
Projection weighted: give each link a
weight equal to the number of
common groups



Human disease network

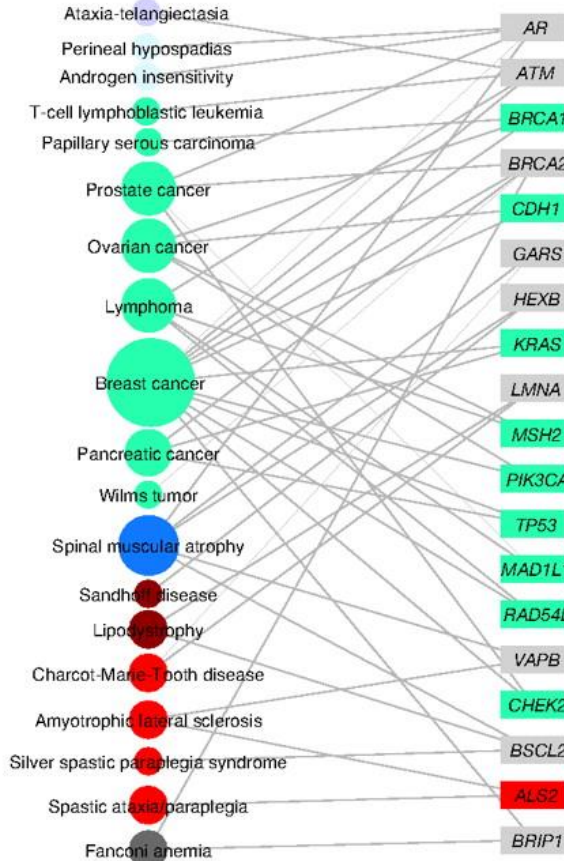
Human Disease Network (HDN)



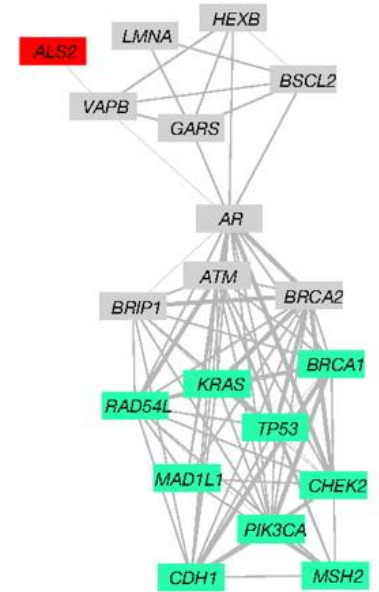
DISEASOME

disease phenome

disease genome



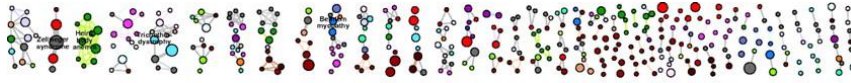
Disease Gene Network (DGN)



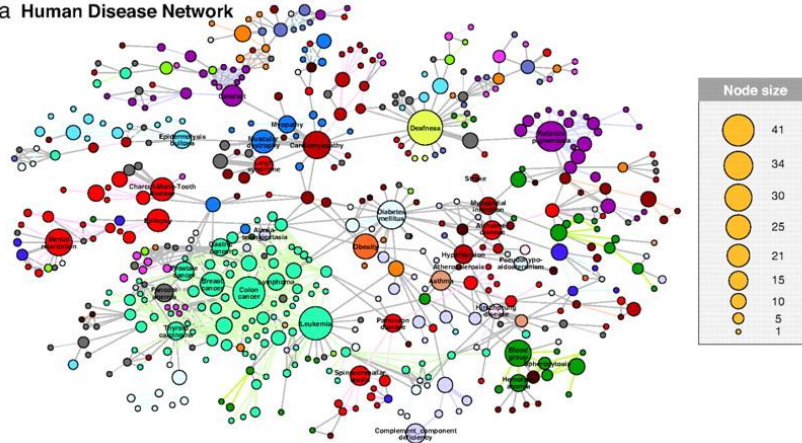
Barabasi's book: 2.7



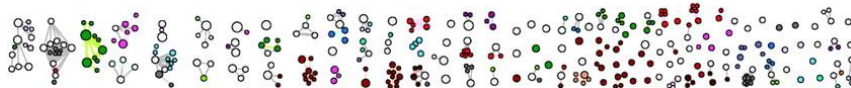
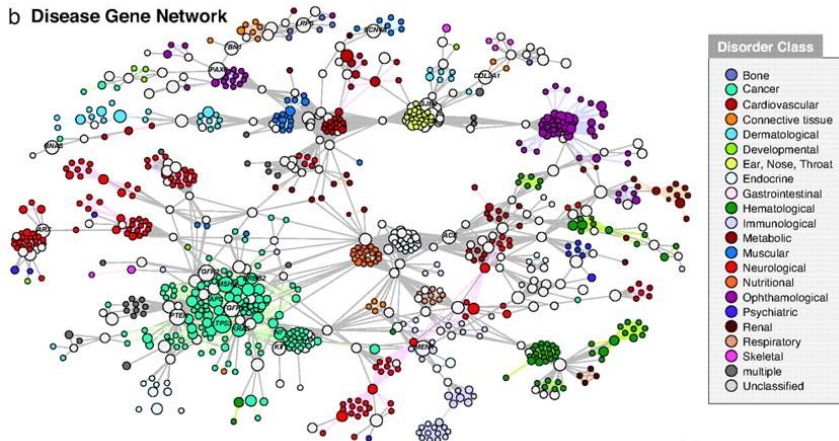
Human disease network



a Human Disease Network



b Disease Gene Network



Barabasi's book: 2.7

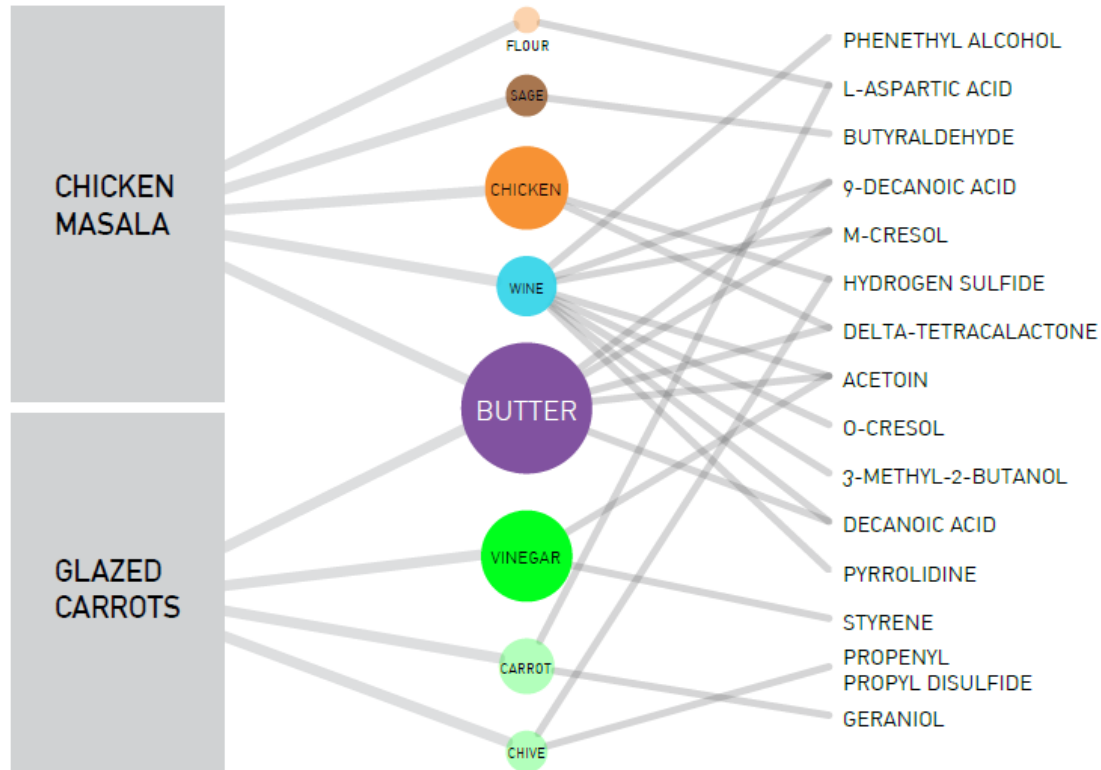


Tripartite networks

(a) RECIPES

INGREDIENTS

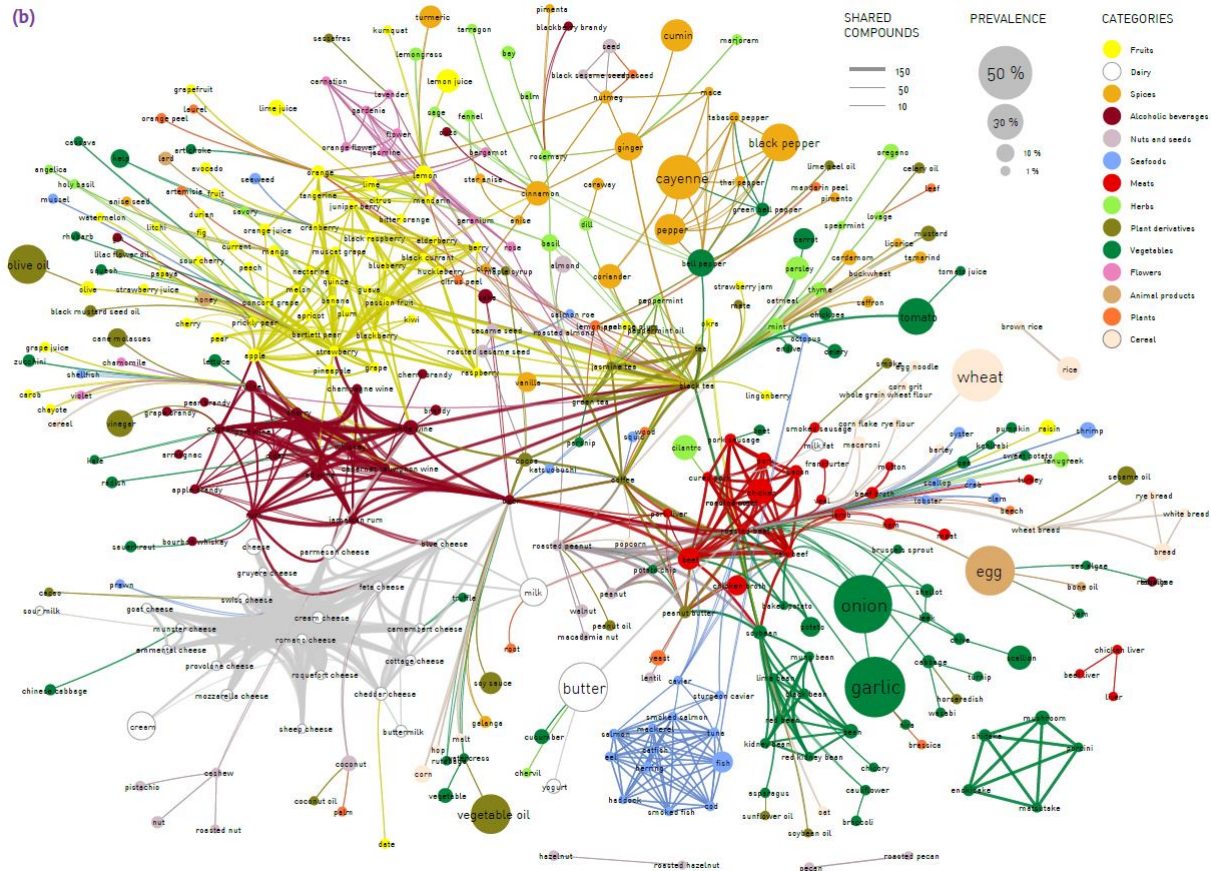
COMPOUNDS



Barabasi's book: 2.7



Tripartite networks



Barabasi's book: 2.7



Credits

M.E.J. Newman
Networks – An Introduction
Oxford University Press
Section 6.6

Barabasi
Network Science
Section 2.7





UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Social Networks Analysis

Density and Average Degree

Degree and Density

Consider a network $G(N,L)$ where we know only:

N: number of nodes

L: number of links

[No information on where the links are]

What can we measure?

Local property

Global property

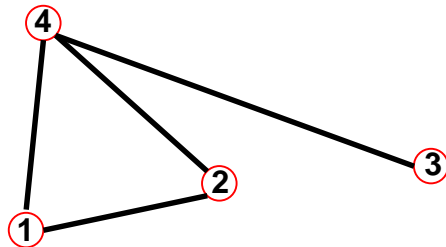


Undirected Networks

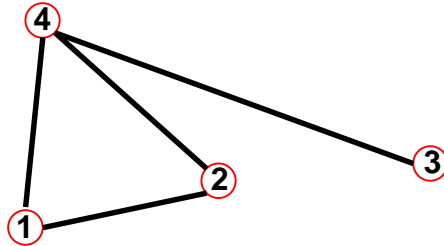
Degree

Node degree: Number of links connected to it

We will denote the degree of node i by d_i or k_i



Degree



For an undirected network of N nodes the degree can be written in terms of the adjacency matrix as:

$$d_i = \sum_j A_{ij}$$

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



Degree

Node degree: Number of links of a node

Do we know anything about node degree from N and L ?

Average degree



Average Degree

Average degree: average number of links per node

$$\langle d \rangle = \frac{\sum_{i=1}^N d_i}{N}$$



Average Degree

Each link in an undirected network has two nodes as end-points. If there are L links then there are $2L$ end-points of links. But the number of end-points is also equal to the sum of the degrees of all nodes.

$$\sum_{i=1}^N d_i = 2L$$

$$\langle d \rangle = \frac{\sum_{i=1}^N d_i}{N}$$

Combining the two :

$$\langle d \rangle = \frac{2L}{N}$$

Is it a local or a global property?



Average degree

$$\langle d \rangle = \frac{\sum_{i=1}^N d_i}{N} \qquad \sum_{i=1}^N d_i = 2L$$

$$\langle d \rangle = \frac{2L}{N}$$

A local property mediated on the global network.

A function of N and L only.

No need to know where the links are in the networks.



Density (connectance)

Average degree:

a local property mediated on the global network

What about a global property?

Does the network have few/many links given the number of nodes? The density is related to the total number of links built by the nodes

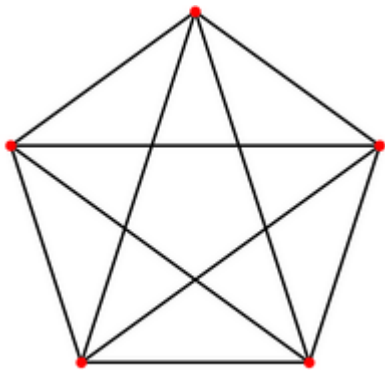
How to define the density of the network?



Density: definition

The density of a network is the fraction of all possible links that are actually present.

The density of a network is the ratio of the number of links L to the number of possible links in a network with N nodes and is given by



?

HINT: compute the number of links in a complete graph of N nodes. Start by thinking node per node



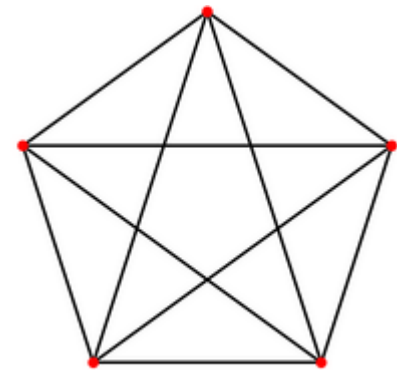
Density

Numerator: L

Denominator:

the maximum possible number number of links
in a network of N nodes is:

$$\binom{N}{2} = \frac{N(N-1)}{2}$$



Density

The density of a network is the ratio of the number of links L to the number of possible links in a network with N nodes given by

$$\Delta = \frac{L}{N(N-1)/2} = \frac{2L}{N(N-1)}$$



Density and Average Degree

$$\Delta = \frac{L}{N(N-1)/2} = \frac{2L}{N(N-1)}$$

$$\langle d \rangle = \frac{2L}{N}$$

The average degree is inversely proportional to the number of nodes

The density is inversely proportional to the square of the number of nodes



Density and Average Degree

The degree is related to the number of links of a single node (local)

The density is related to the number of links of the whole network (global)

Which is the relation between the density and the average degree?



Density and Average Degree

$$\langle d \rangle = \frac{\sum_{i=1}^N d_i}{N}$$

$$\sum_{i=1}^N d_i = 2L$$

$$\langle d \rangle = \frac{2L}{N} \quad \Delta = \frac{2L}{N(N-1)}$$

$$\Delta = \frac{\langle d \rangle}{N-1}$$

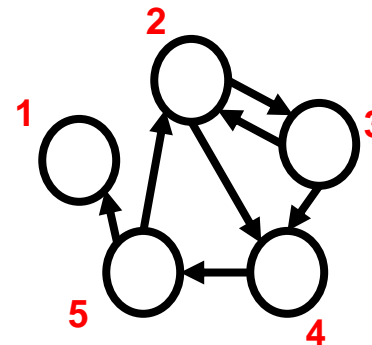


Directed Networks

Degree, in-degree and out-degree

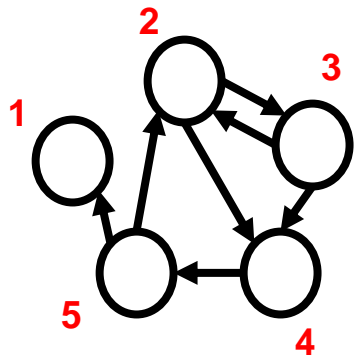
In-degree: number of in-going links of a node (j,i)

Out-degree: number of out-going links of a node (i,j)



- Out-degree: expansiveness
- In-degree: popularity





$$A_{ij} = \begin{cases} 1, & \text{if there is an edge from } j \text{ to } i \\ 0, & \text{otherwise} \end{cases}$$

• Outdegree = $\sum_{i=1}^n A_{ij}$

The outdegree of node 3 is 2, sum of the elements of the third column

$$\sum_{i=1}^n A_{i3}$$

From 5 to 1
↓

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

■ Indegree = $\sum_{j=1}^n A_{ij}$

The indegree of node 3 is 1, sum of the elements of the third row

$$\sum_{j=1}^n A_{3j}$$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Indegree and outdegree

Write the definition of average in- and out- degree

$$\langle d^{in} \rangle = \frac{\sum_{i=1}^N d_i^{in}}{N}$$
$$\langle d^{out} \rangle = \frac{\sum_{i=1}^N d_i^{out}}{N}$$

Are they related?



Indegree and outdegree

$$\langle d^{in} \rangle = \frac{\sum_{i=1}^N d_i^{in}}{N}$$
$$\langle d^{out} \rangle = \frac{\sum_{i=1}^N d_i^{out}}{N}$$

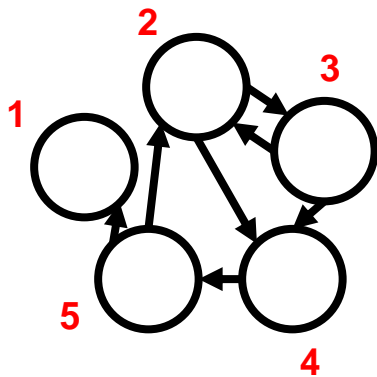
Are they related?

The average in-degree is equal to the average out-degree



Directed Networks

Compute the average in-degree and out-degree as a function of the number of nodes N and the number of links L



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

$$\langle d^{in} \rangle = \langle d^{out} \rangle = \frac{L}{N}$$



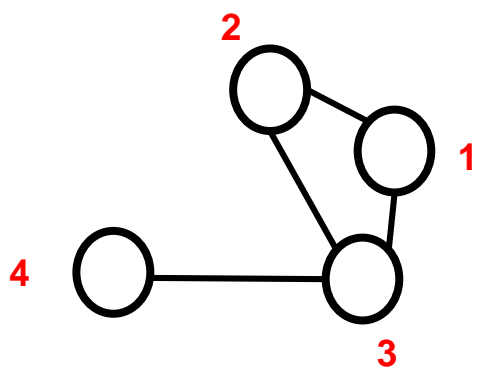
Density: definition

The density of a network is the fraction of all possible links that are actually present.

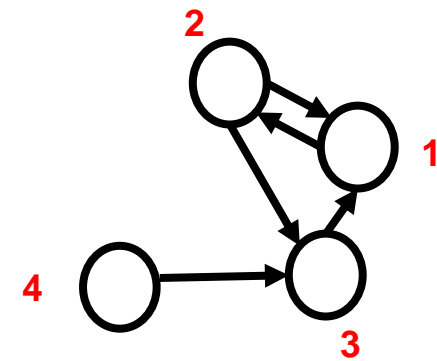
The density of a network is the ratio of the number of links L to the number of possible links in a network with N nodes and is given by

?





Undirected network



Directed network

$$\langle d \rangle = \frac{2L}{N}$$

$$L_{max} = \frac{N(N-1)}{2}$$

$$\Delta = \frac{L}{N(N-1)/2} = \frac{2L}{N(N-1)}$$

$$\Delta = \frac{\langle d \rangle}{N-1}$$

$$\langle d^{in} \rangle = \langle d^{out} \rangle = \frac{L}{N}$$

$$L_{max} = N(N-1)$$

$$\Delta = \frac{L}{N(N-1)}$$

$$\Delta = \frac{\langle d^{in} \rangle}{N-1} = \frac{\langle d^{out} \rangle}{N-1}$$



Real networks are sparse



Real networks are sparse

$$L \ll L_{\max}$$

or

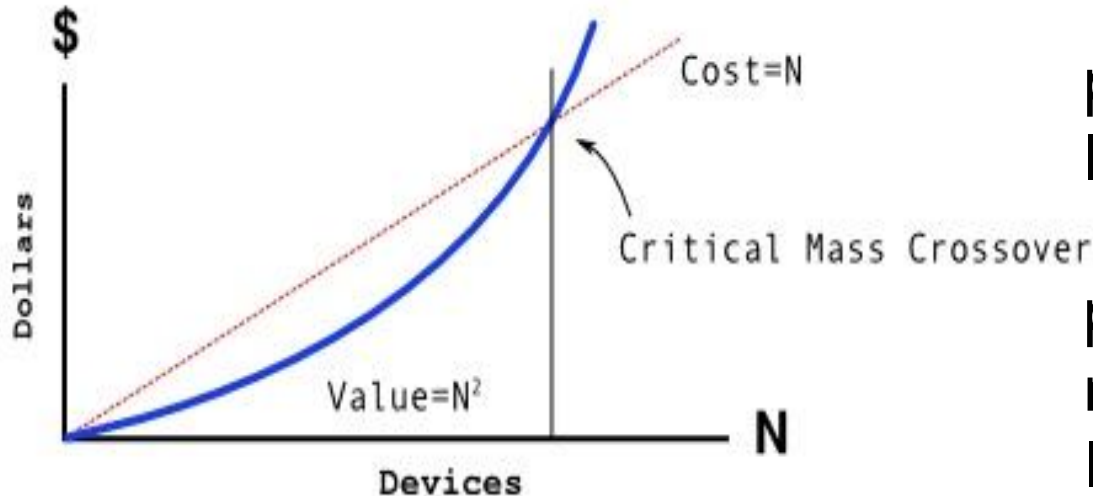
$$\langle d \rangle \ll N-1.$$

Compute the average degree and the density

Undirected network	N	L
Internet	1.92E+05	6.09E+05
Mobile phone calls	3.66E+04	9.18E+04
e-mail	5.72E+04	1.04E+05
Actor network	7.02E+05	2.94E+07
protein network	2018	2930
Facebook2011	7.21E+08	6.90E+09
Twitter2009	4.16E+07	1.4E+09
Youtube	1.10E+06	2.90E+06



Metcalfe's law: the Internet boom of 2000



Two fundamental problems with Metcalfe's law:

- While all links are possible, in real networks not all links are present. Indeed, most real networks are sparse, which means that only a very small fraction of the links are present.

- Not all links are of equal value.

Value: proportional to the square of the number of its consumers

Costs would grow only linearly.

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



Graph densification

«...Most of real networks densify over time, with the number of edges growing super-linearly in the number of nodes...»

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data* 1, 1, Article 2 (March 2007).
DOI: <https://doi.org/10.1145/1217299.1217301>



Credits

Reza Zafarani, Mohammad Ali Abbasi, Huan Liu

Social Media Mining: An Introduction

A Textbook by Cambridge University Press

Albert-László Barabási

Network Science

2.3.1, 2.3.2,

Newman, M.E.J.

Networks: An Introduction.

Oxford University Press. 2010.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data* 1, 1, Article 2 (March 2007).

DOI: <https://doi.org/10.1145/1217299.1217301>





UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Social Network Analysis





UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Network Models

Random networks

*Degree distribution in random
networks*

Generative network models

When we analyse or mine a network we measure the structure of the network with mathematical, statistical and computational methods for making sense of the data we get from our measurements.

This is a data-driven approach.

Why do we need network models?



Network models

1. If we know a network has some particular property, what effects will that have on the overall behavior of the system?

To get a feel for these effects we build mathematical models, i.e. mathematical models of networks. The properties of these networks can be calculated analytically, or at least numerically.



Network models

2. A large part of understanding what properties measured in a network are interesting depends on having an appropriate reference point by which to distinguish interesting from non-interesting.

Random network models represent the conventional reference point (null model).

Compare the network with the observed property to networks without it by create artificial networks with and without that property and compare them.



Network Models

3. Network models allow us to identifying the mechanism of the system that produces an empirically observed pattern.

That allows us to better understand and predict networks and to immediately understand the nature of a new network when we see that pattern



Network Models

A network (or graph) model is a mathematical model of networks in which some specific parameters are fixed, but the network is random in all the other respects.

The aim is to build models that reproduce some or all properties of real-world networks.



Random networks

The Erdos-Renyi network model

Assumption

From a modeling perspective a network is a relatively simple object, consisting of only nodes and links.

The real challenge, however, is to decide where to place the links between the nodes so that we reproduce a system.

In this respect the philosophy behind a completely random network is simple: we assume that this goal is best achieved by placing the links randomly between the nodes.

Links are created randomly



Random networks: models

Two definitions of a random network:

$G(N, L)$ Model

N labeled nodes are connected with L randomly placed links.

$G(N, p)$ Model

Each pair of N labeled nodes is connected with probability p .

$G(N, L)$ model fixes the total number of links L

$G(N, p)$ model fixes the probability p that two nodes are connected

Which one?



Random networks: models

$G(N, L)$ Model

N labeled nodes are connected with L randomly placed links.

$G(N, p)$ Model

Each pair of N labeled nodes is connected with probability p .

Compute the average degree

$G(N, L)$ model

the average degree of a node is simply $\langle d \rangle = 2L/N$

$G(N, p)$ model?

Seems to be more complicated but ...



Random networks: models

While in the $G(N, L)$ model the average degree of a node is simply $\langle d \rangle = 2L/N$, other network characteristics are easier to calculate in the $G(N, p)$

Asymptotically the two models are equivalent

Random network, Erdos-Renyi model: $G(N, p)$



Random networks

A random network consists of N nodes where each node pair is connected with probability p .

To construct a random network we follow these steps:

- 1) Start with N isolated nodes.
- 2) Select a node pair among the $N(N-1)/2$ for undirected networks or $N(N-1)$ directed networks and generate a random number r between 0 and 1.

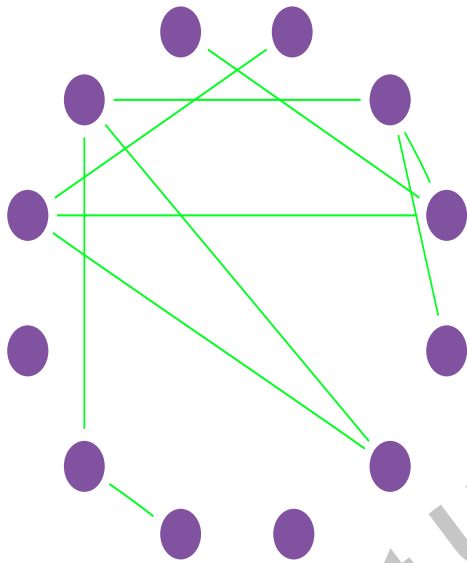
If $r \leq p$, connect the selected node pair with a link, otherwise leave them disconnected.

- 3) Repeat step (2) for each of the node pairs.

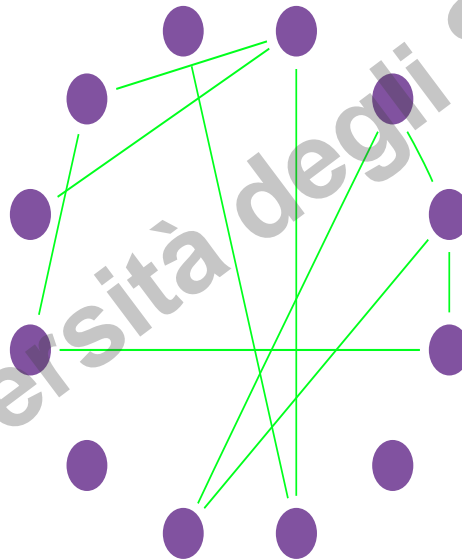


Erdos - Renyi model

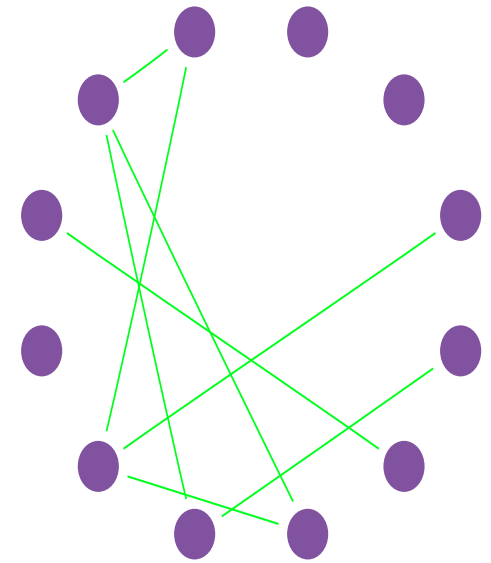
$p=1/6$
 $N=12$



L=8



L=10

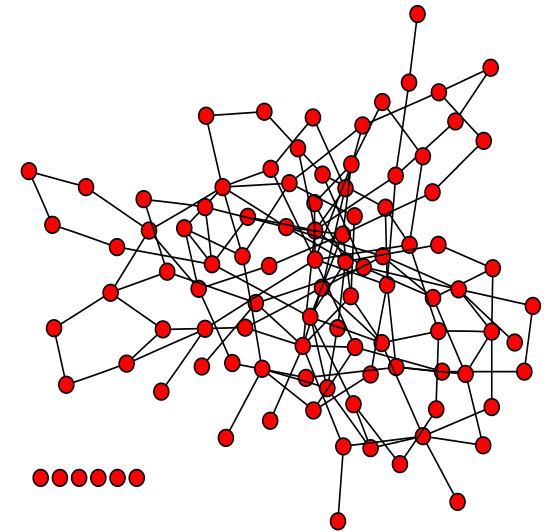
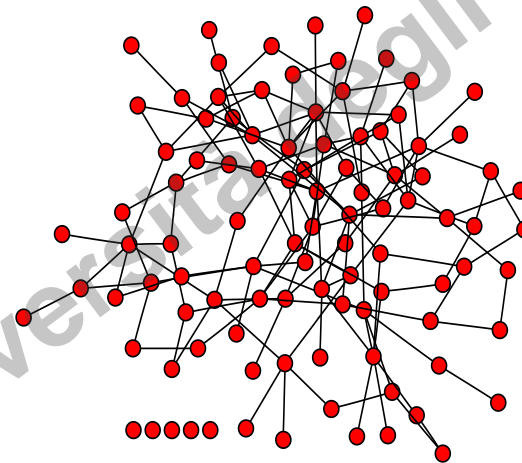
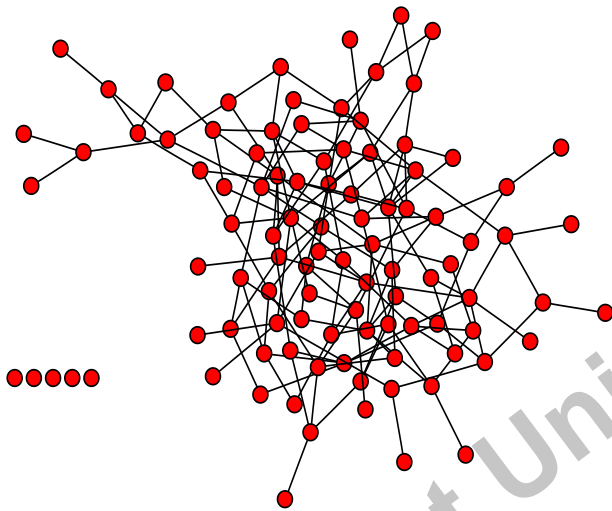


L=7



Erdos - Renyi model

$p=0.03$
 $N=100$



Erdos-Renyi model: L

$P(L)$: the probability to have exactly L links in a network of N nodes and probability p :

$$P(L) = \binom{\binom{N}{2}}{L} p^L (1-p)^{\binom{N(N-1)}{2} - L}$$

The maximum number of links in a network of N nodes.

Probability that L of the attempts to connect all potential pairs have resulted in a link

Number of different ways we can choose L links among all potential links.

Binomial distribution...



Erdos-Renyi model: L

$P(L)$: the probability to have a network of exactly L links

$$P(L) = \binom{\binom{N}{2}}{L} p^L (1-p)^{\binom{N}{2}-L}$$

•The average number of links $\langle L \rangle$ in a random graph

$$\langle L \rangle = \sum_{L=0}^{\binom{N}{2}} L P(L) = p \frac{N(N-1)}{2}$$

$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1).$$

•The variance

$$S^2 = p(1-p) \frac{N(N-1)}{2}$$



How to choose N and p for comparison with the network under study

Real network: N, L

Random network: N, p
 N as the real network

$p?$



How to choose N and p for comparison with a real network

Real network: N, L

Random network: N, p

N as the real network

p such that

$\langle L_{\text{random}} \rangle$ in the random network is equal to the
number of links in the real network

$$\langle L_{\text{random}} \rangle = L$$

$$p = ?$$



How to choose N and p for comparison with a real network

Hint:

$$\Delta = \frac{L}{N(N-1)/2} = \frac{2L}{N(N-1)}$$

$$L = \Delta N(N-1)/2$$

$$\langle L \rangle = \sum_{L=0}^{N(N-1)/2} LP(L) = p \frac{N(N-1)}{2}$$

L

=

<L>



How to choose N and p for comparison with a real network

Real network: N, L

Random network: N, p

N as the real network

p such that

$\langle L_{\text{random}} \rangle$ in the random network is equal to the
number of links in the random network

$$\langle L_{\text{random}} \rangle = L$$

$$p = \Delta$$



An ensemble of networks

The random network model is not defined in terms of a single randomly generated network, but as an ensemble of networks.

When one talks about the properties of random networks, one typically means the average properties of the ensemble.

Some properties can be calculated analytically (as the average number of links), others generating an ensemble of networks with the same parameters and computing the average of the property on them.



Random networks

Random networks are a very useful model to compare with the real-world networks behavior.

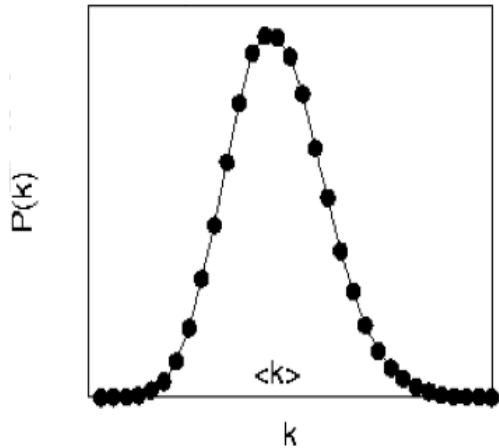
When we study a phenomenon at the real network, we can use a random model to realize if the phenomenon carries information or if it is random.



DEGREE DISTRIBUTION

Copyright Università degli Studi di Milano

Random networks: degree distribution



Probability that a randomly selected node has degree k

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

Select k nodes from $N-1$

probability of having k links

probability of missing $N-1-k$ links

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

$$S_k^2 = p(1-p)(N-1)$$

$$\frac{S_k}{\langle k \rangle} = \left[\frac{1-p}{p} \frac{1}{(N-1)} \right]^{1/2} \approx \frac{1}{(N-1)^{1/2}}$$

As the network size increases, the distribution becomes increasingly narrow—we are increasingly confident that the degree of a node is in the vicinity of $\langle k \rangle$.



Poisson distribution

$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k} \quad \langle k \rangle = p(N-1)$$

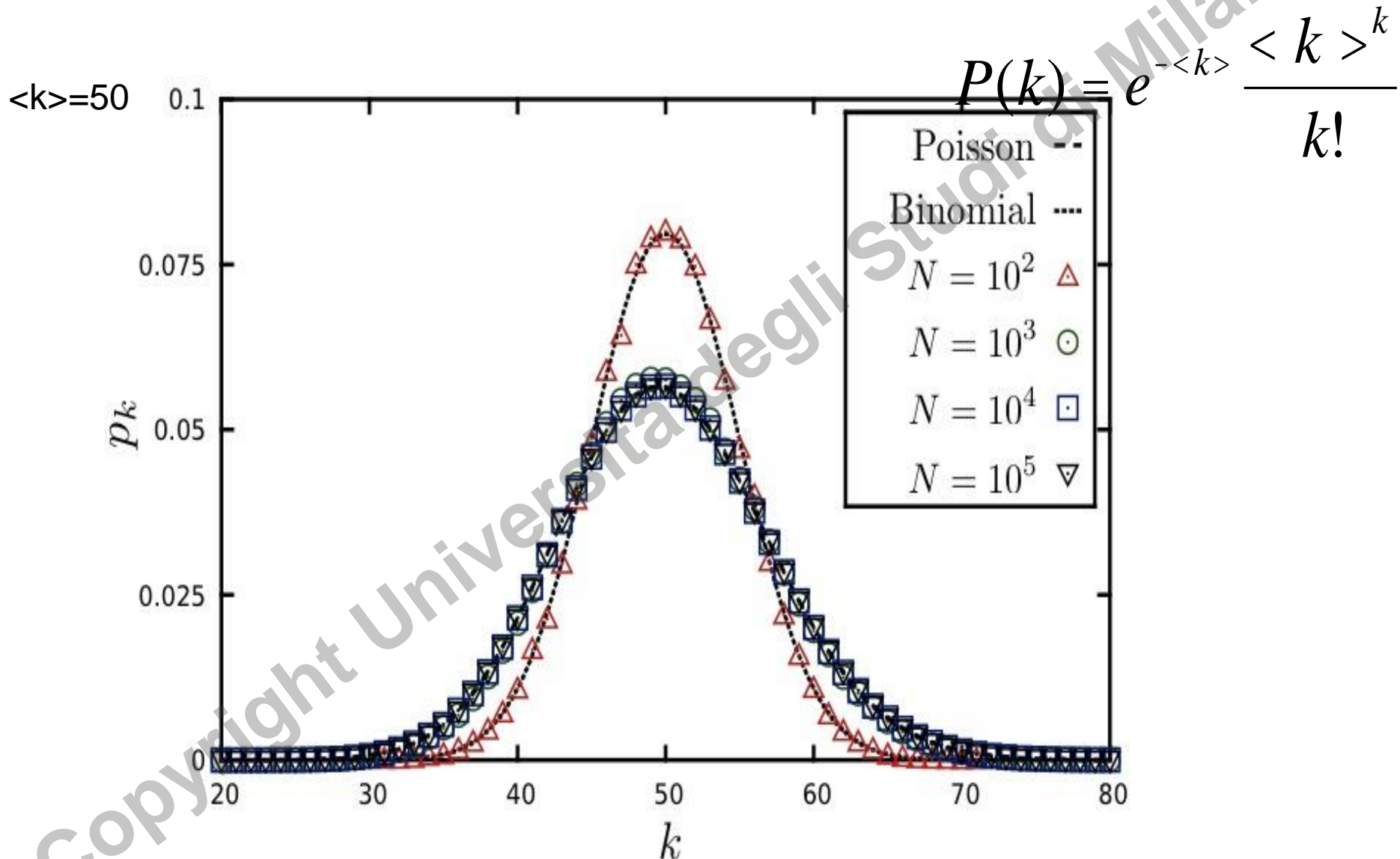
For large N and small k ,
the degree distribution can be approximated by the Poisson distribution:

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

That's why it is also called Poisson random model



Random networks: degree distribution



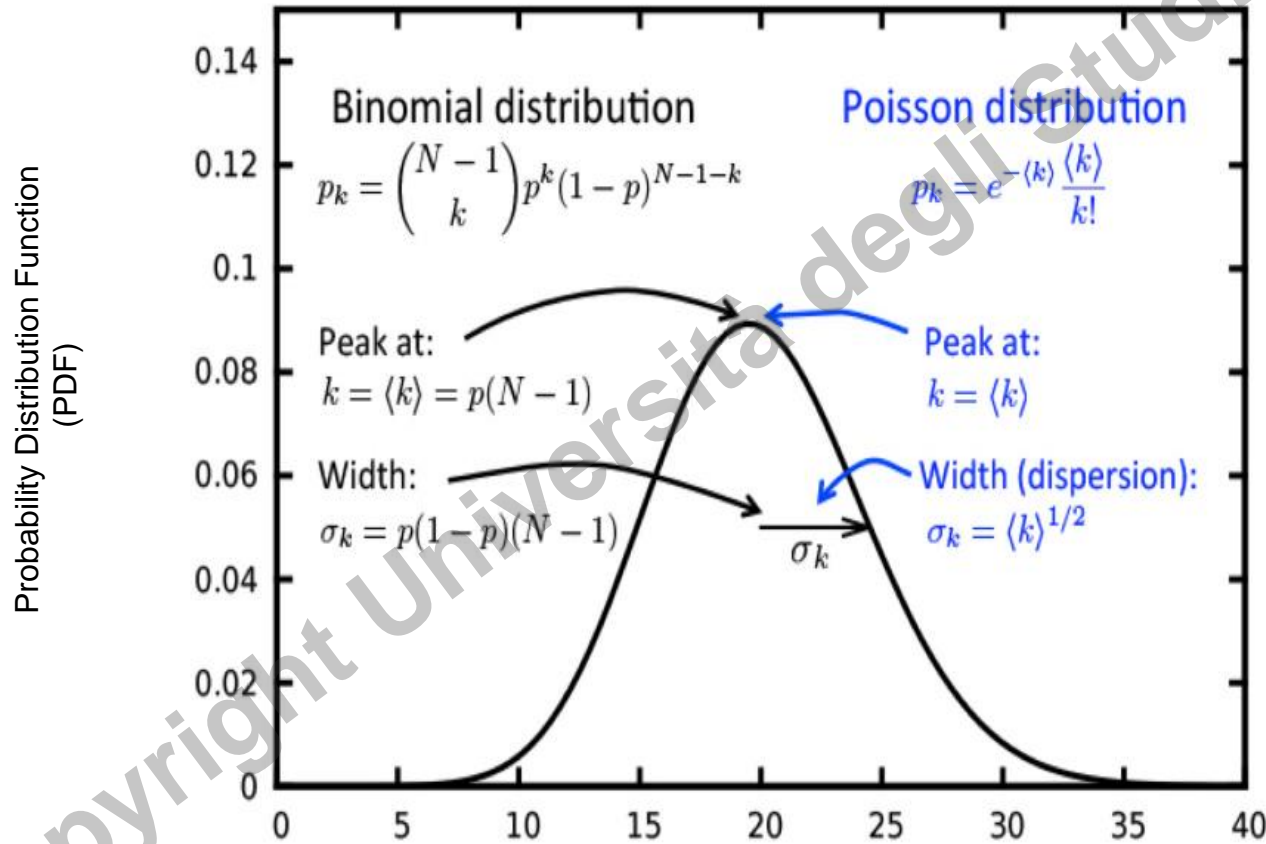
Random networks: degree distribution

Exact Result

-binomial distribution-

Large N limit

-Poisson distribution-



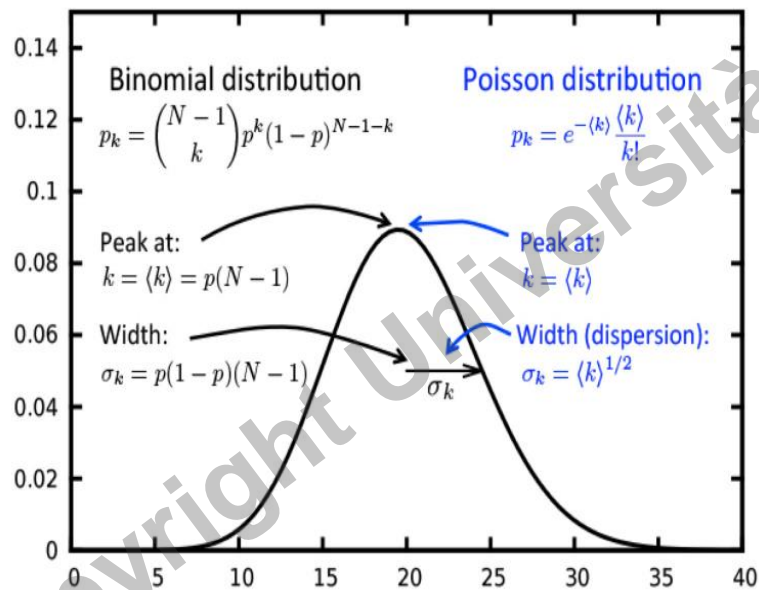
It does not depend on N



Random networks: degree distribution

How big the differences are between the node degrees in a random networks?

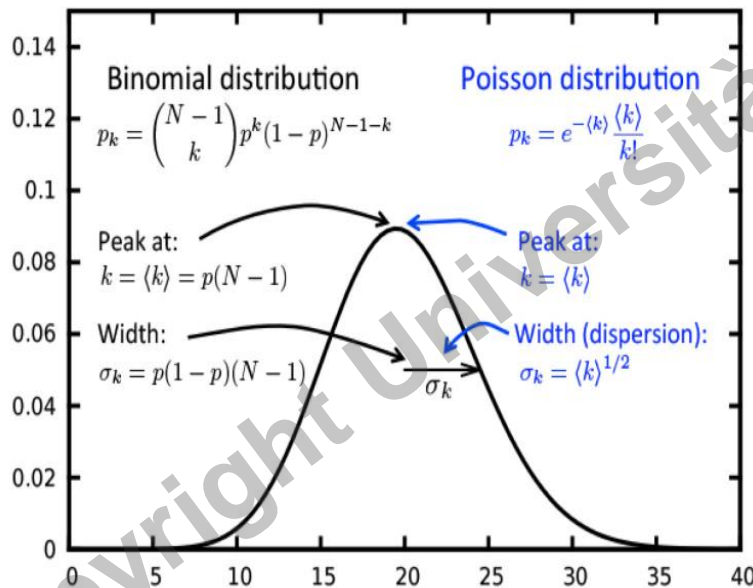
Can high-degree nodes coexist with small-degree nodes?



Random networks: degree distribution

How big the differences are between the node degrees in a random networks?

Can high-degree nodes coexist with small-degree nodes?



Example:

Sociologists estimate that a typical person knows about 1000 individuals on a first-name basis:

$$\langle k \rangle = 1000$$

Human society: $N = 10^9$

Which is the number of friend of a typical individual?



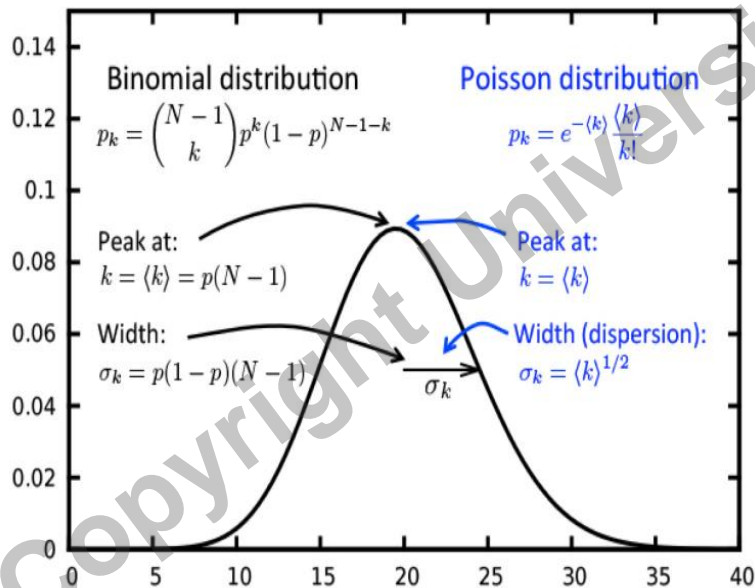
Random networks: degree distribution

How big the differences are between the node degrees in a random networks?

Can high-degree nodes coexist with small-degree nodes?

Sociologists estimate that a typical person knows about 1000 individuals on a first-name basis: $\langle k \rangle = 1000$

Human society: $N = 10^9$



$$\langle k \rangle \pm \sigma_k \quad \sigma_k = \langle k \rangle^{1/2}$$
$$\sigma_k = 31.62.$$

The number of friends a typical individual has is between 968 and 1032, a narrow window



Random networks: degree distribution

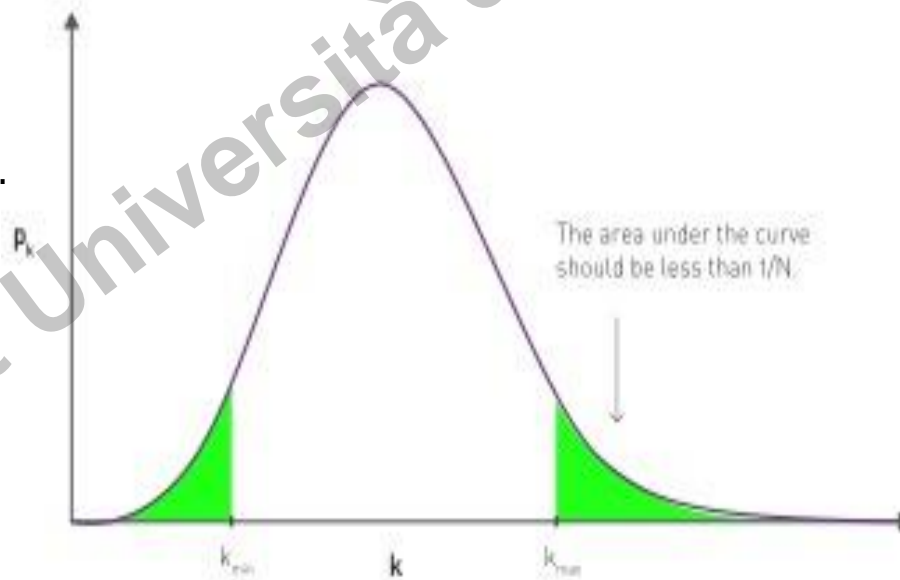
We define k_{\max} such that in a network of N nodes we have at most one node with degree higher than k_{\max}

$$N[1 - P(k_{\max})] \approx 1.$$

$$1 - P(k_{\max}) = 1 - e^{-\langle k \rangle} \sum_{k=0}^{k_{\max}} \frac{\langle k \rangle^k}{k!} = e^{-\langle k \rangle} \sum_{k=k_{\max}+1}^{\infty} \frac{\langle k \rangle^k}{k!} \approx e^{-\langle k \rangle} \frac{\langle k \rangle^{k_{\max}+1}}{(k_{\max}+1)!},$$

$$NP(k_{\min}) \approx 1.$$

$$P(k_{\min}) = e^{-\langle k \rangle} \sum_{k=0}^{k_{\min}} \frac{\langle k \rangle^k}{k!}.$$



$$\langle k \rangle = 1000, \quad N = 10^9$$

$$k_{\min} = 816$$

$$k_{\max} = 1185$$



Random networks: degree distribution

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

- The most connected individual has degree $k_{\max} \sim 1,185$
- The least connected individual has degree $k_{\min} \sim 816$
- The number of friends a typical individual has is between 968 and 1032

The probability to find an individual with degree $k > 2,000$ is 10^{-27} . Hence the chance of finding an individual with 2,000 acquaintances is so tiny that such nodes are virtually inexistent in a random society.

- a random society would consist of mainly average individuals, with everyone with roughly the same number of friends.
- It would lack outliers, individuals that are either highly popular or reclusive, no hubs



Credits

Source:

Albert-László Barabási

Network Science

Chapter 3

Newman, M.E.J.

Networks: An Introduction.

Oxford University Press. 2010.

Reza Zafarani, Mohammad Ali Abbasi, Huan Liu

Social Media Mining: An Introduction

A Textbook by Cambridge University Press



Network Science

Class 4: Scale-free property

Albert-László Barabási

with Roberta Sinatra

www.BarabasiLab.com

Introduction

Copyright Università degli Studi di Milano

WORLD WIDE WEB

Nodes: **WWW documents**

Links: **URL links**

N: around 10^{12} the largest network, even larger than human brain (N^{11})

crawler: collects all URL's found in a document and follows them recursively

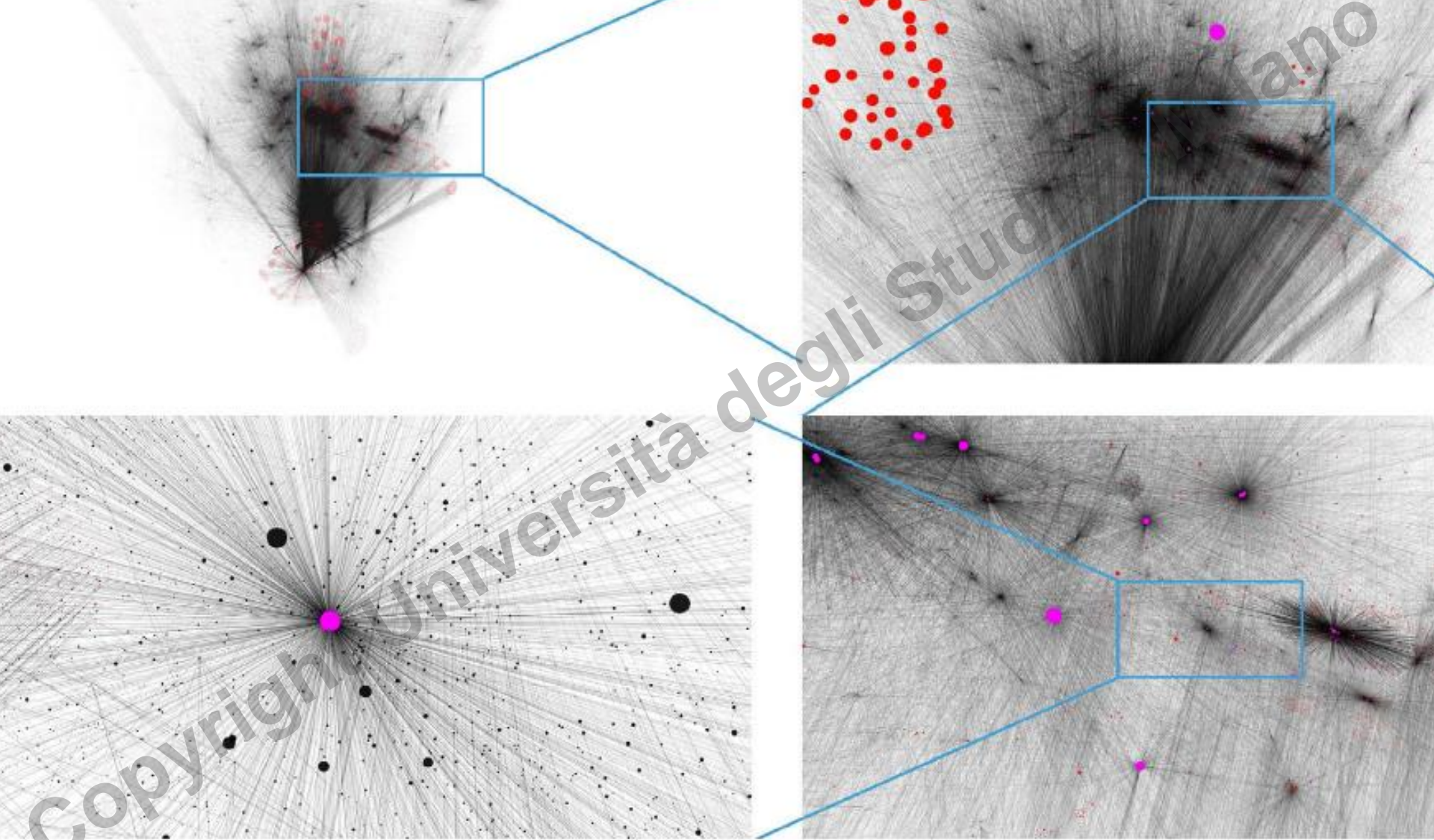
1998

Hawoong Jeong (Barabasi lab) maps out nd.edu:

300000 documents

1.5 million links

<http://barabasi.com/networksciencebook/resources/chapter4.html>

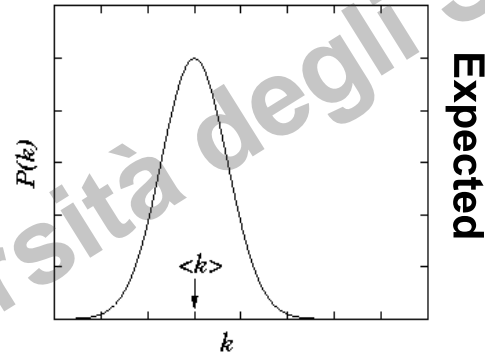


$$\langle k_{\text{in}} \rangle = \langle k_{\text{out}} \rangle = 4.60$$
$$\sigma(k) = 2.14$$

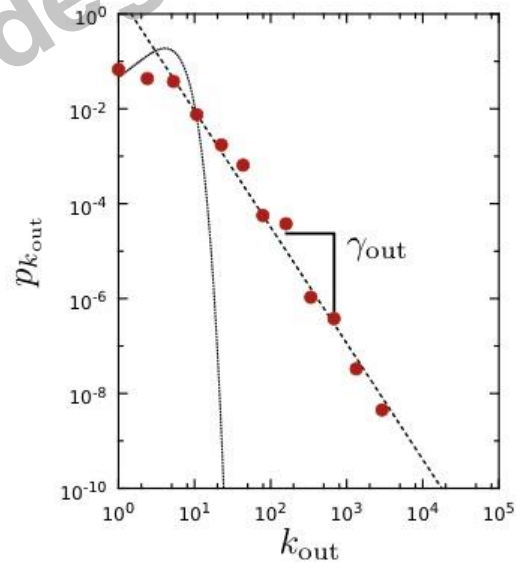
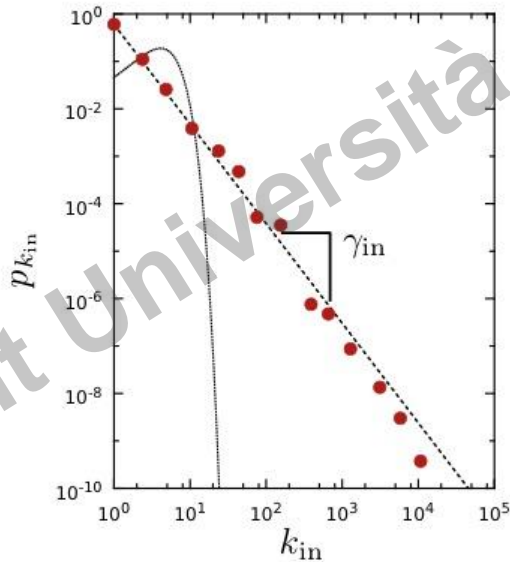
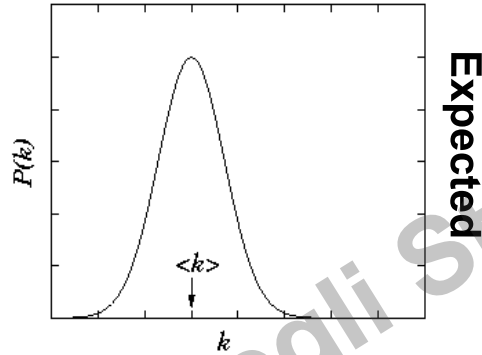
$$P(k > 10) \sim 10^{-3}$$

$$P(k > 20) \sim 10^{-8}$$

$$P(k = 100) \sim 10^{-94}$$



$$\langle k_{in} \rangle = \langle k_{out} \rangle = 4.60$$
$$\sigma(k) = 2.14$$

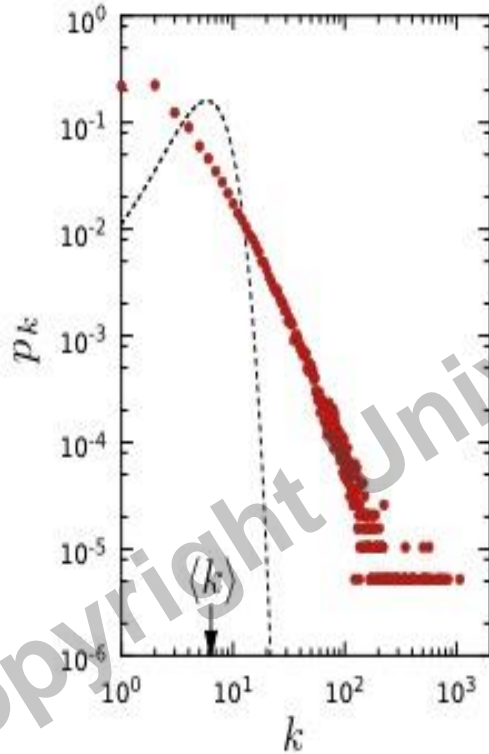


R. Albert, H. Jeong, A-L Barabasi, *Nature*, 401 130 (1999).

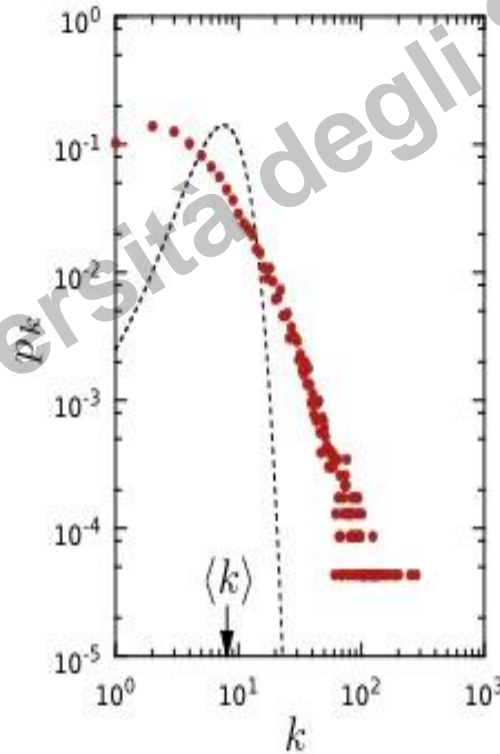
FACING REALITY: Degree distribution of real networks

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

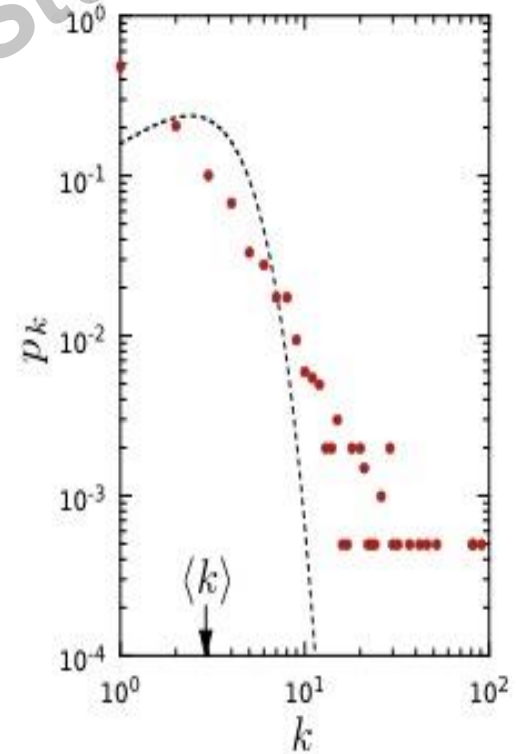
Internet



Science Collaboration



Protein Interactions

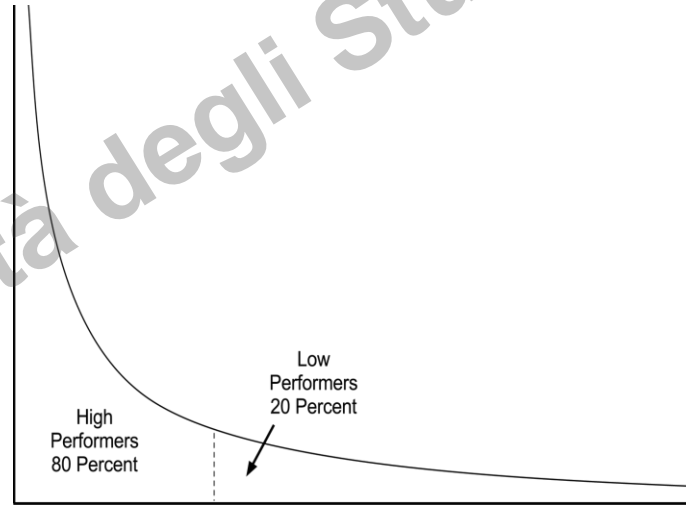


Copyright Università degli Studi di Milano

80/20 RULE



80 percent of money is earned by only 20 percent of the population



Vilfredo Federico Damaso Pareto (1848 – 1923), Italian economist, political scientist and philosopher, who had important contributions to our understanding of income distribution and to the analysis of individuals choices. A number of fundamental principles are named after him, like Pareto efficiency, Pareto distribution (another name for a power-law distribution), the Pareto principle (or 80/20 law).

Vilfredo Pareto, a 19th century economist, noticed that in Italy a few wealthy individuals earned most of the money, while the majority of the population earned rather small amounts. He connected this disparity to the observation that incomes follow a power law, representing the first known report of a power law distribution [3]. His finding entered the popular literature as the 80/20 rule: roughly 80 percent of money is earned by only 20 percent of the population. The 80/20 emerges in many areas, like management, stating that 80 percent of profits are produced by only 20 percent of the employees or that 80 percent of decisions are made during 20 percent of meeting time. They are present in networks as well: 80 percent of links on the Web point to only 15 percent of webpages; 80 percent of citations go to only 38 percent of scientists; 80 percent of links in Hollywood are connected to 30 percent of actors [4]. Typically all quantities obeying the 80/20 rule follow a power law distribution. During the 2009 economic crisis power laws have gained a new meaning: the Occupy Wall Street Movement highlighted the fact that in the US 1% of the population earns a disproportionate 15% of the total US income. This 1% effect, a signature of a profound income disparity, is again a natural consequence of the power law nature of the income distribution.

Discrete vs. Continuum formalism

Discrete Formalism

As node degrees are always positive integers, the discrete formalism captures the probability that a node has exactly k links:

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

$$\sum_{k=1}^{\infty} p_k = 1.$$

$$C \sum_{k=1}^{\infty} k^{-\gamma} = 1 \quad C = \frac{1}{\sum_{k=1}^{\infty} k^{-\gamma}} = \frac{1}{\zeta(\gamma)},$$

$$p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}$$

INTERPRETATION:

$$p_k$$

Continuum Formalism

In analytical calculations it is often convenient to assume that the degrees can take up any positive real value:

$$p(k) = Ck^{-\gamma}.$$

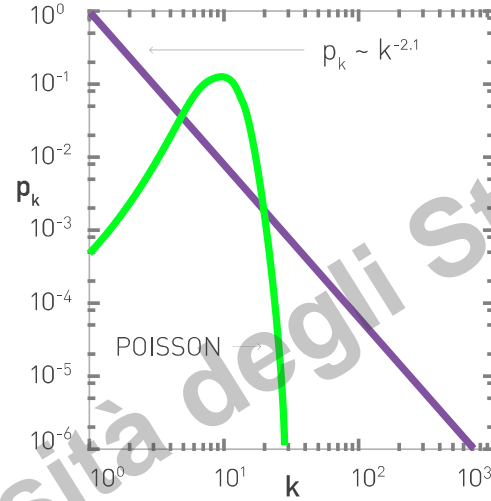
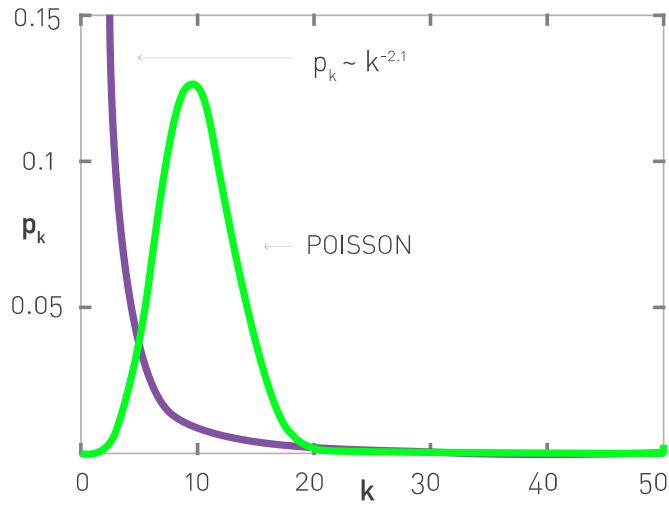
$$\int_{k_{\min}}^{\infty} p(k)dk = 1$$

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

$$p(k) = (\gamma - 1)k_{\min}^{\gamma-1}k^{-\gamma}.$$

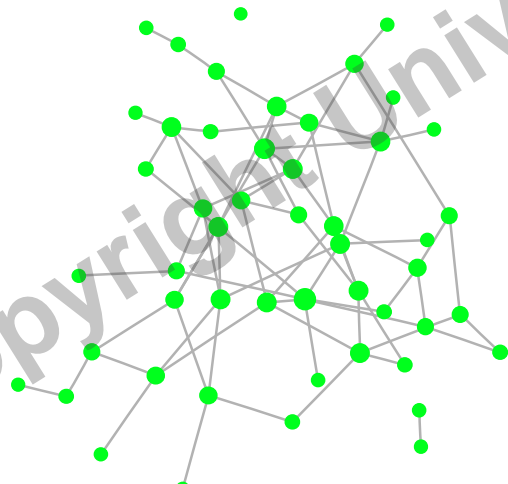
$$\int_{k_1}^{k_2} p(k)dk$$

The difference between a power law and an exponential distribution

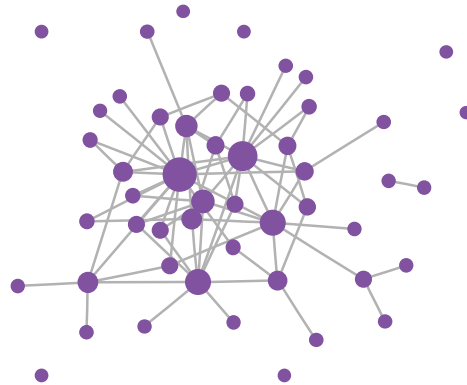


Note the difference for:
small k
 k around $\langle k \rangle$
large k

(c)



(d)



Scale-free networks

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

The difference between a power law and an exponential distribution: hubs

Let us use the WWW to illustrate the properties of the high- k regime.
The probability to have a node with $k \sim 100$ is

- About $p_{100} \simeq 10^{-30}$ in a Poisson distribution
- About $p_{100} \simeq 10^{-4}$ if p_k follows a power law.
- Consequently, if the WWW (10^{12} nodes) were to be a random network, according to the Poisson prediction we would expect 10^{-18} $k > 100$ degree nodes, or none.
- For a power law degree distribution, we expect about 10^8 $k > 100$ degree nodes

The size of the biggest hub

All real networks are finite \rightarrow let us explore its consequences.

\rightarrow We have an expected maximum degree, k_{\max}

Estimating k_{\max}

$$\int_{k_{\max}}^{\infty} P(k) dk \gg \frac{1}{N}$$

Why: the probability to have a node larger than k_{\max} should not exceed the prob. to have one node, i.e. $1/N$ fraction of all nodes

$$\int_{k_{\max}}^{\infty} P(k) dk = (g-1)k_{\min}^{g-1} \int_{k_{\max}}^{\infty} k^{-g} dk = \frac{(g-1)}{(-g+1)} k_{\min}^{g-1} \left[k^{-g+1} \right]_{k_{\max}}^{\infty} = \frac{k_{\min}^{g-1}}{k_{\max}^{g-1}} \approx \frac{1}{N}$$

$$k_{\max} = k_{\min} N^{\frac{1}{g-1}}$$

The size of the largest hub

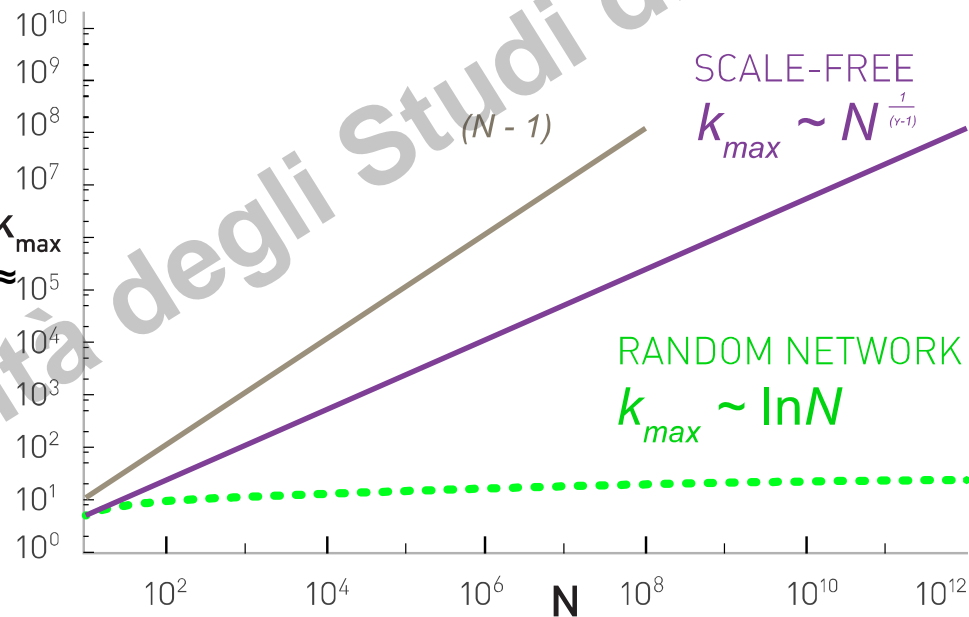
To illustrate the difference in the maximum degree of an exponential and a scale-free network let us return to the WWW sample of [Image 4.1](#), consisting of $N \approx 3 \times 10^5$ nodes.

As $k_{min} = 1$, if the degree distribution were to follow an exponential, (4.17) predicts that the maximum degree should be $k_{max} \approx 13$.

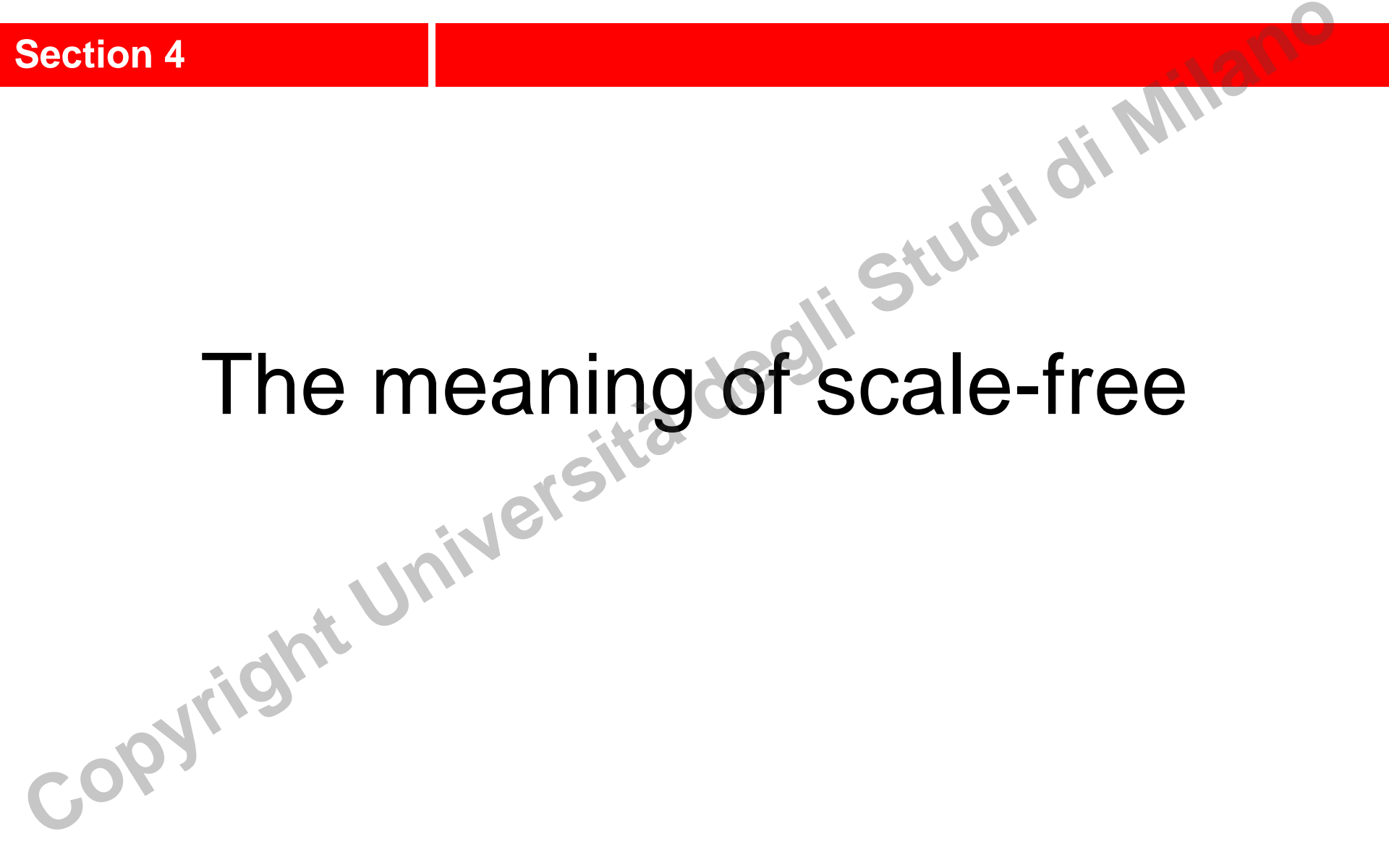
In a scale-free network of similar size and $\gamma = 2.1$, (4.18) predicts $k_{max} \approx 95,000$, a remarkable difference.

Note that the largest in-degree of the WWW map of [Image 4.1](#) is 10,721, which is comparable to k_{max} predicted by a scale-free network.

This reinforces our conclusion that *in a random network hubs are effectively forbidden, while in scale-free networks they are naturally present.*



The meaning of scale-free



Definition:

Networks with a power law tail in their degree distribution are called 'scale-free networks'

Where does the name come from?

- Correlation length diverges at the critical point: the whole system is correlated!
- **Scale invariance:** there is no characteristic scale for the fluctuation (**scale-free behavior**).
- **Universality:** exponents are independent of the system's details.

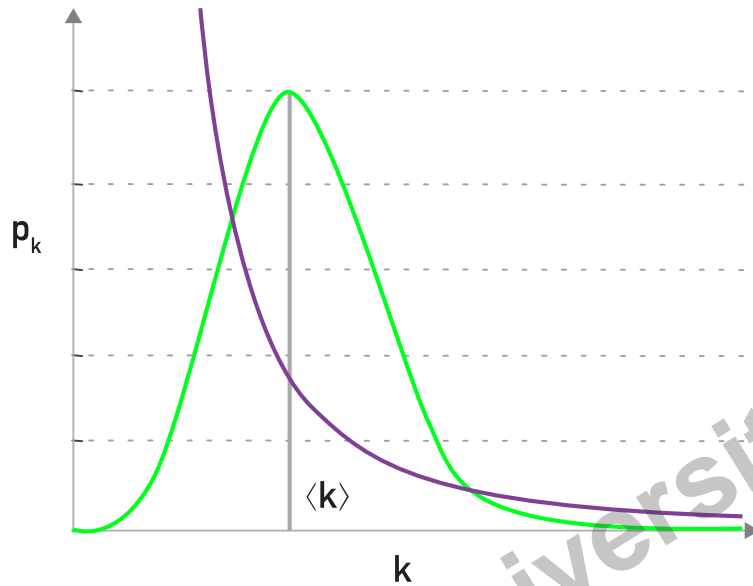
DIVERGENCE OF THE HIGHER MOMENTS

Network	Size	$\langle k \rangle$	κ	γ_{out}	γ_{in}
WWW	325 729	4.51	900	2.45	2.1
WWW	4×10^7	7		2.38	2.1
WWW	2×10^8	7.5	4000	2.72	2.1
WWW, site	260 000				1.94
Internet, domain*	3015–4389	3.42–3.76	30–40	2.1–2.2	2.1–2.2
Internet, router*	3888	2.57	30	2.48	2.48
Internet, router*	150 000	2.66	60	2.4	2.4
Movie actors*	212 250	28.78	900	2.3	2.3
Co-authors, SPIRES*	56 627	173	1100	1.2	1.2
Co-authors, neuro.*	209 293	11.54	400	2.1	2.1
Co-authors, math.*	70 975	3.9	120	2.5	2.5
Sexual contacts*	2810			3.4	3.4
Metabolic, <i>E. coli</i>	778	7.4	110	2.2	2.2
Protein, <i>S. cerev.</i> *	1870	2.39		2.4	2.4
Ythan estuary*	134	8.7	35	1.05	1.05
Silwood Park*	154	4.75	27	1.13	1.13
Citation	783 339	8.57			3
Phone call	53×10^5	3.16		2.1	2.1
Words, co-occurrence*	460 902	70.13		2.7	2.7
Words, synonyms*	22 311	13.48		2.8	2.8

Many degree exponents are smaller than 3

→ $\langle k^2 \rangle$ diverges in the $N \rightarrow \infty$ limit!!!

The meaning of scale-free



Random Network

Randomly chosen node: $k = \langle k \rangle \pm \langle k \rangle^{1/2}$

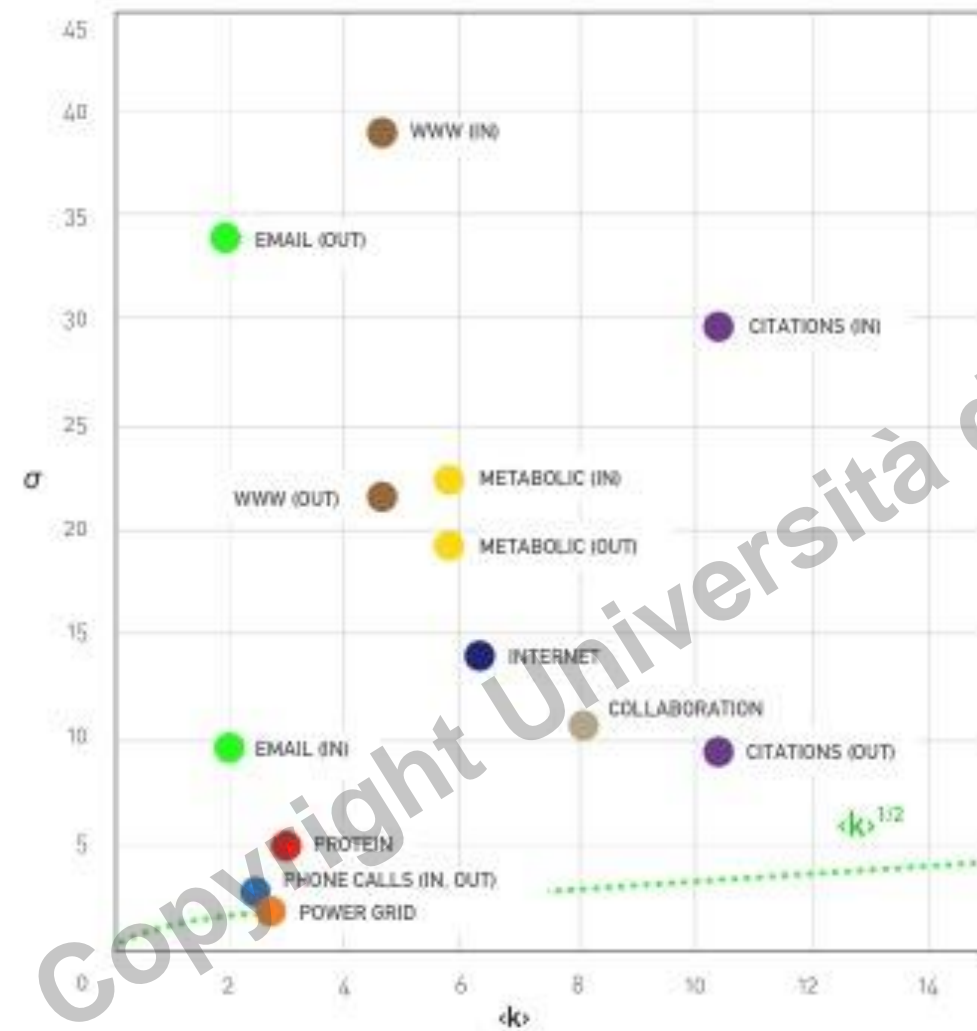
Scale: $\langle k \rangle$

Scale-Free Network

Randomly chosen node: $k = \langle k \rangle \pm \infty$

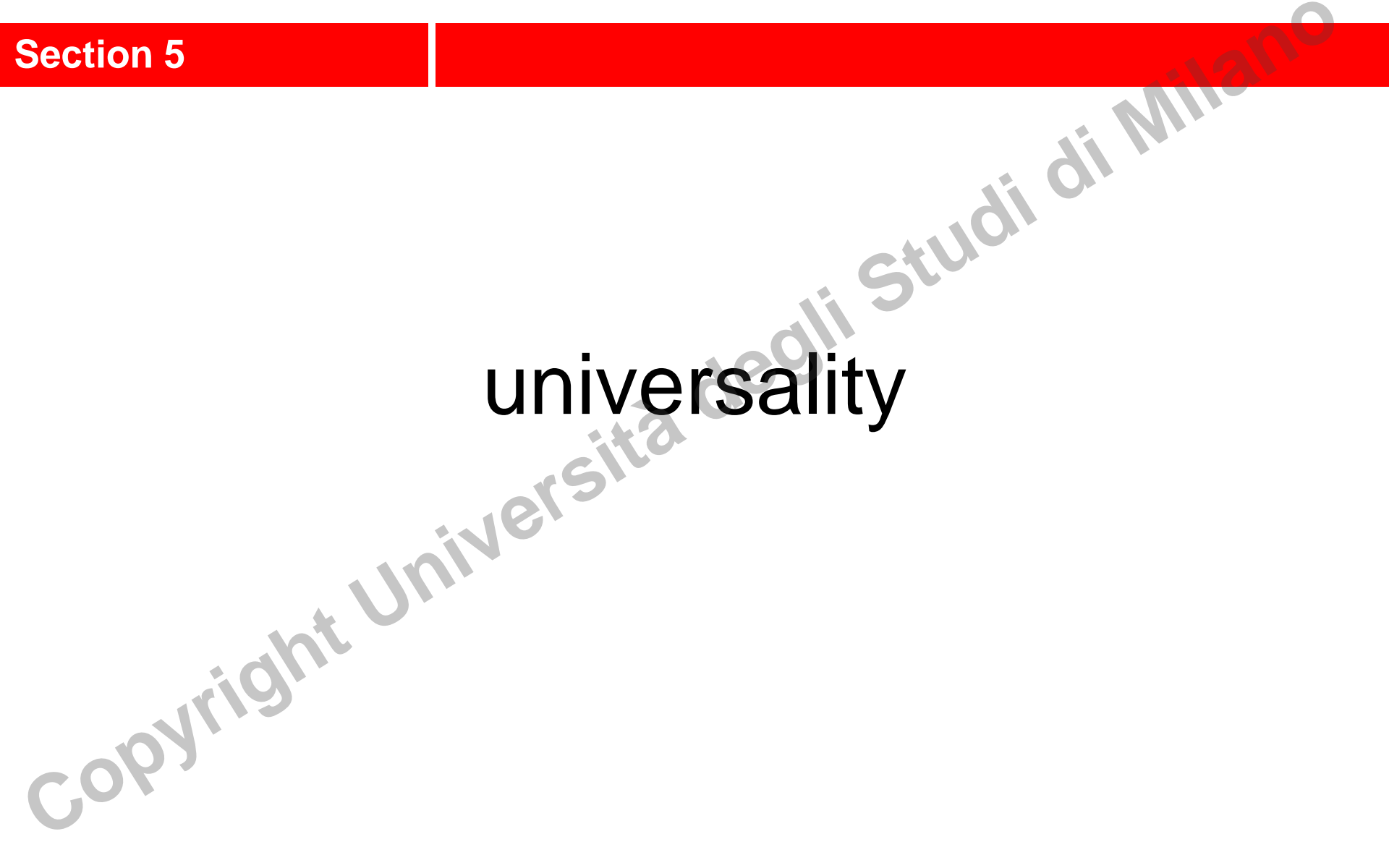
Scale: none

The meaning of scale-free



$$k = \langle k \rangle \pm \sigma_k$$

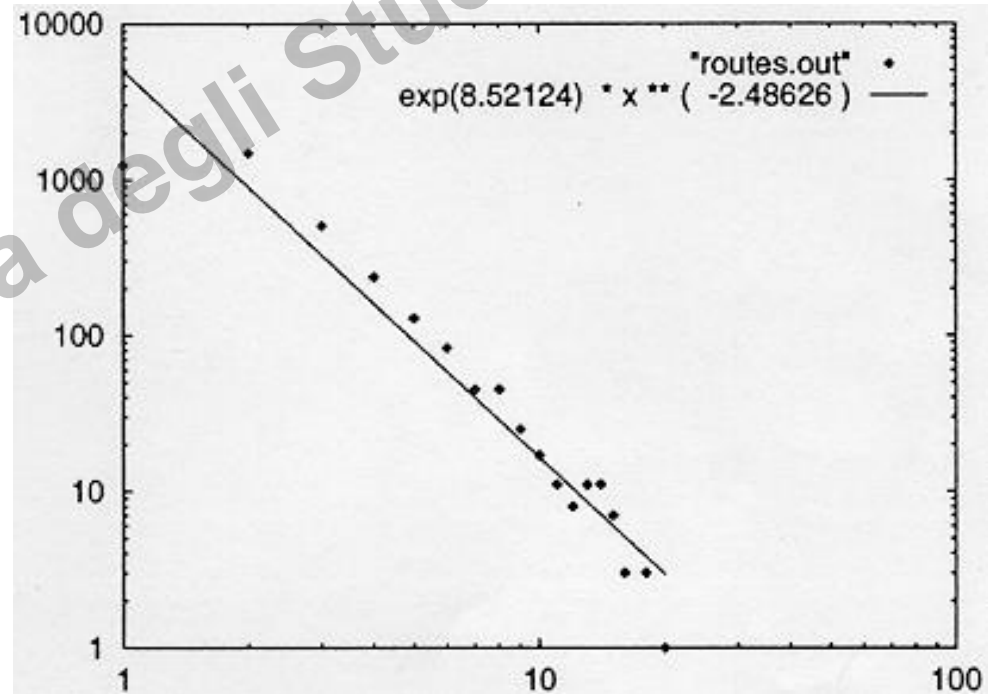
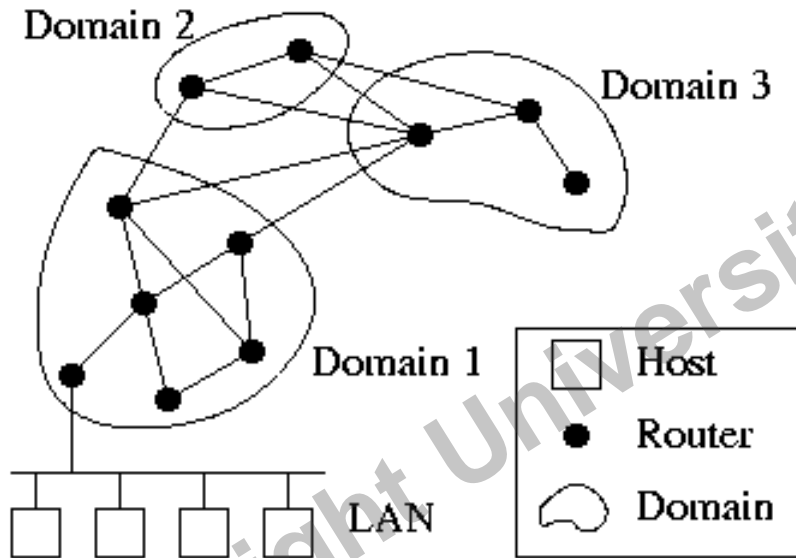
universality



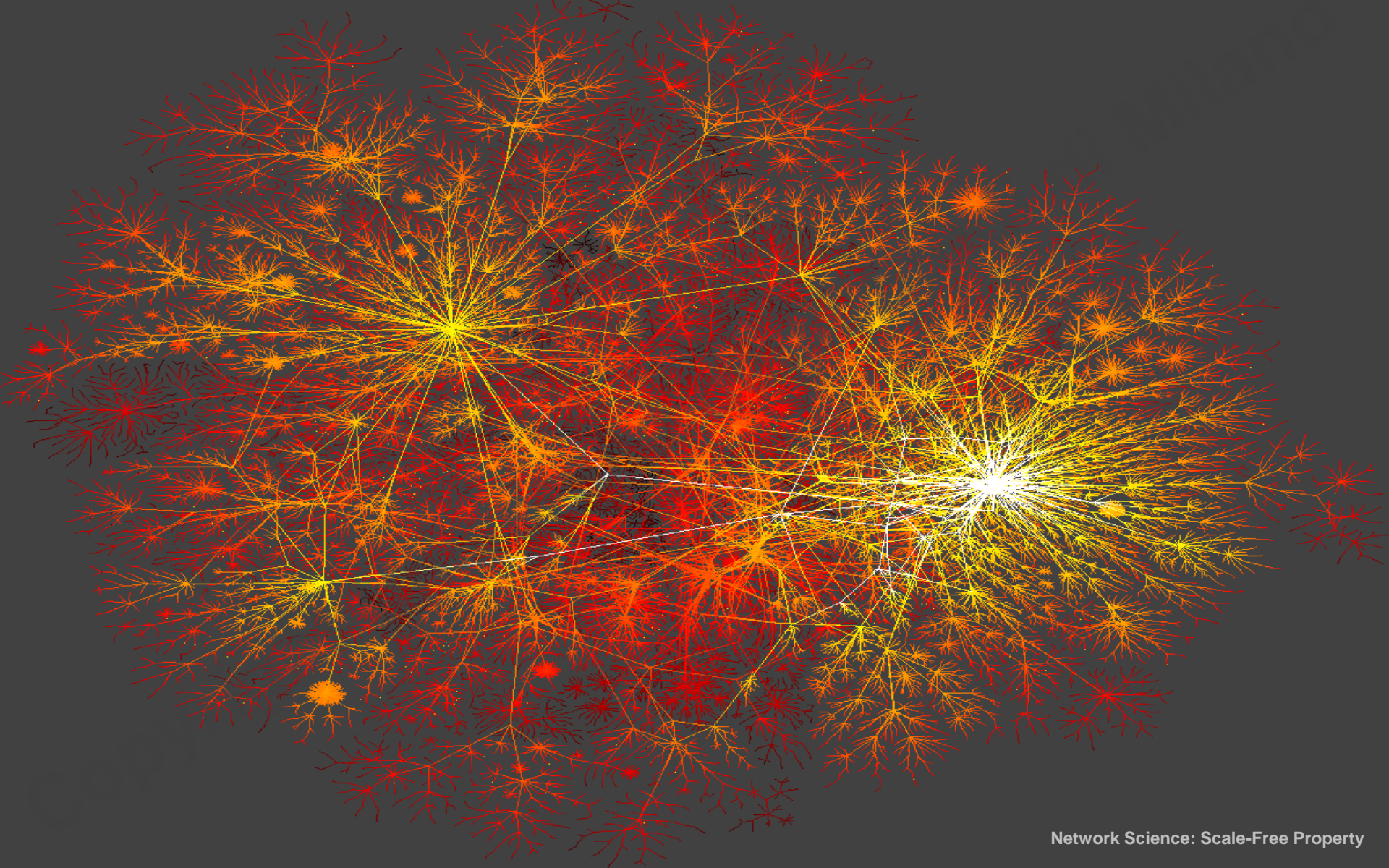
INTERNET BACKBONE

Nodes: computers, routers

Links: physical lines



(Faloutsos, Faloutsos and Faloutsos, 1999)

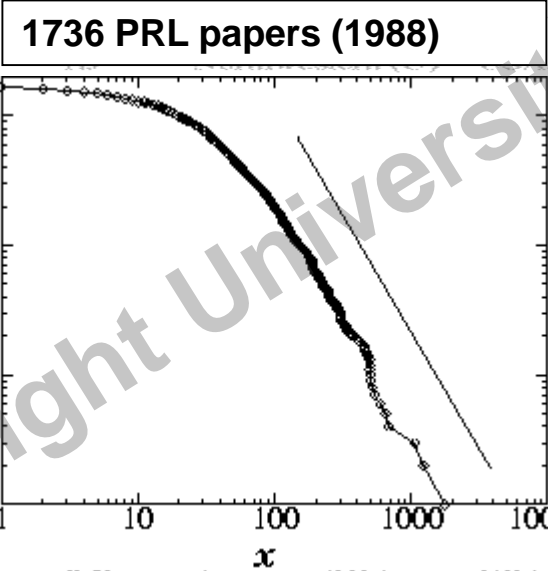
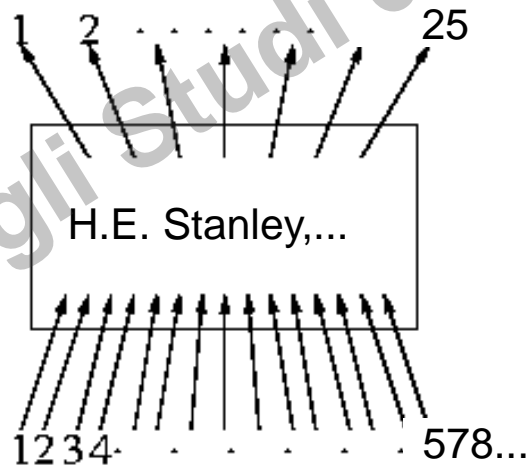


SCIENCE CITATION INDEX

Out of over 500,000 Examined
(see <http://www.sst.nrel.gov>)

Nodes: papers
Links: citations

Author	Institute	Country	Field	avg. cites	total art.	total cites	rank by total cit.
Witten	Princeton (U)	USA, NJ	High-energy (P)	168	138	23735	1
Essler	UCSB (U)	USA, CA	Semie				2
Cava	Bell Labs (I)	USA, NJ	Supern				3
Batlogg	Bell Labs (I)	USA, NJ	Supern				4
Floog	Max-Planck (NL)	Germany	Semie				5
Ellis	Euro Nuclear Cent.	Switzerland	Astroph				6
Fisk	Florida State (U)	USA, FL	Solid S				7
Cardona	Max Planck (NL)	Germany	Semie				8
Nanopoulos	Texas A&M (U)	USA, TX	High-e				9
Heeger	UCSB (U)	USA, CA	Polym				10
Lee*							11
Suzuki*							12
Anderson							13
Suzuki*							14
Freeman							15
Tani							16
Mull							17
Schn							18
Chen							19
Mork							19
Mille							21
Chu				44	213	9453	22
Bedn				85	85	9311	23
Cobe				284	284	9311	23
Metz				86	108	9300	25
Wasz				57	162	9170	26
Shira				33	269	8841	27
Wieg				85	104	8822	28
Vand				67	129	8686	29
Uchi				28	301	8520	30
Hor				72	119	8512	31
Mmp				111	76	8439	32
Birge				41	286	8375	33
Jorge				50	167	8298	34
Hinks	Argonne (NL)	USA, IL	Supetconductivity (E)	37	223	8263	35



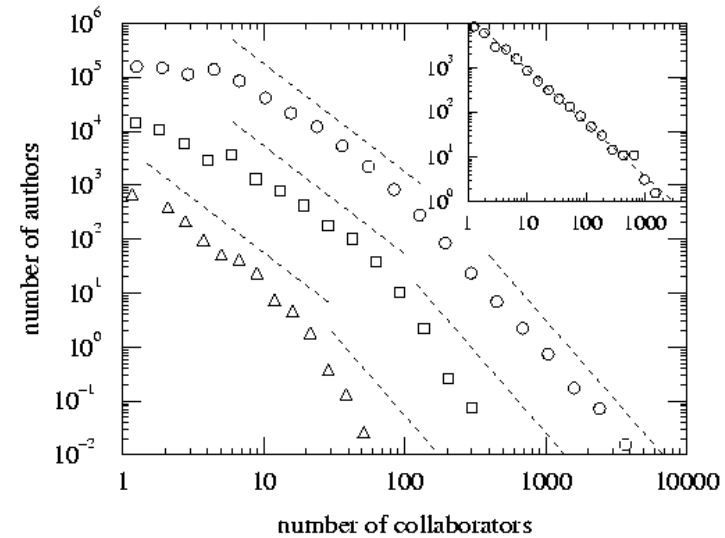
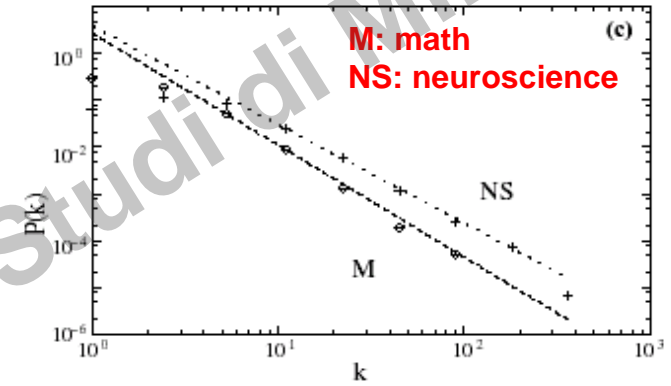
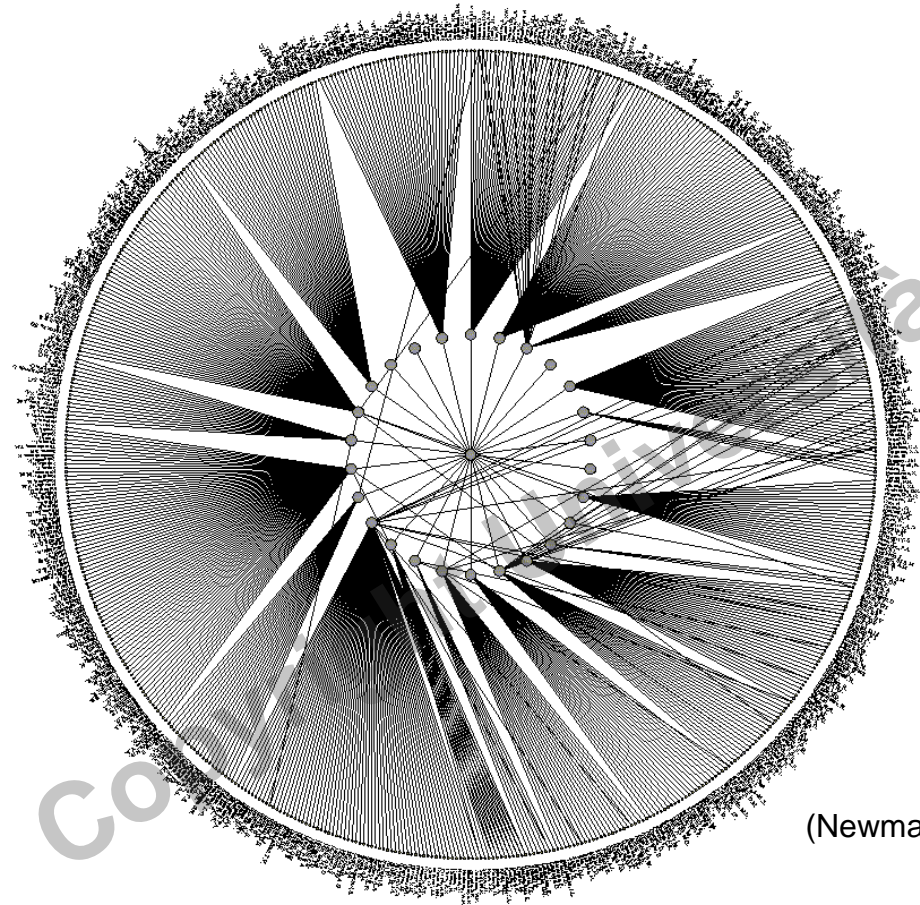
(S. Redner, 1998)

* citation total may be skewed because of multiple authors with the same name

SCIENCE COAUTHORSHIP

Nodes: scientist (authors)

Links: joint publication

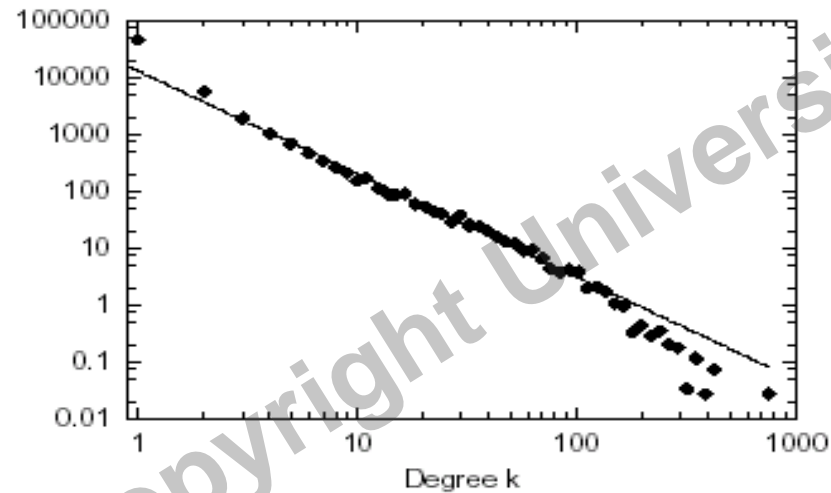


(Newman, 2000, Barabasi et al 2001)

ONLINE COMMUNITIES

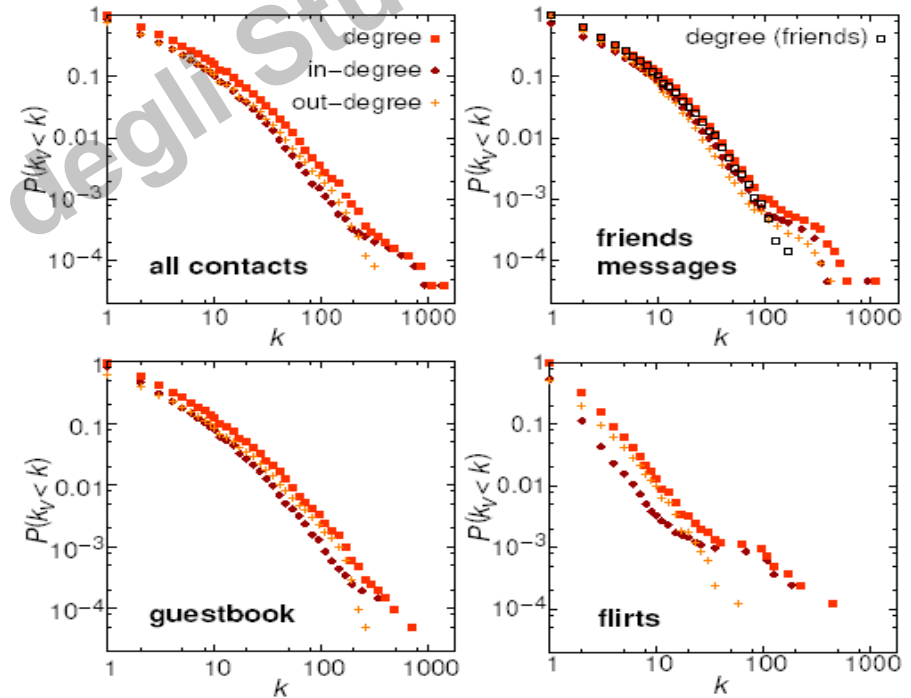
Nodes: online user
Links: email contact

Kiel University log files
112 days, N=59,912 nodes



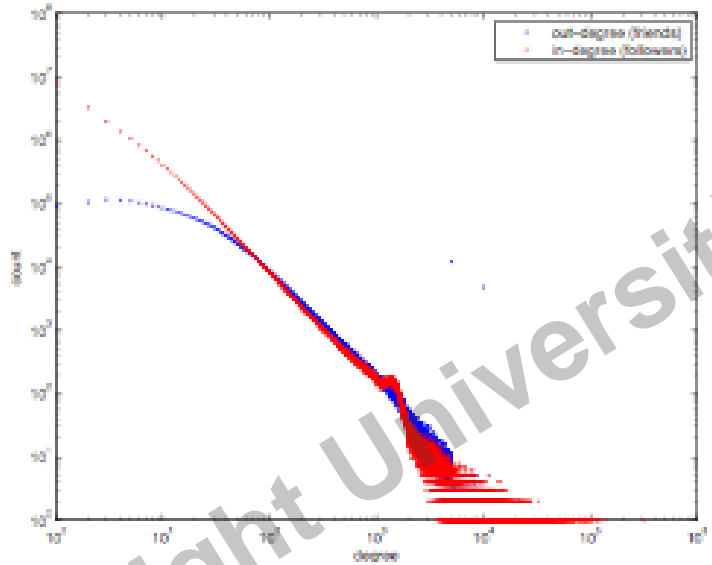
Ebel, Mielsch, Bornholdtz, PRE 2002.

Pussokram.com online community;
512 days, 25,000 users.

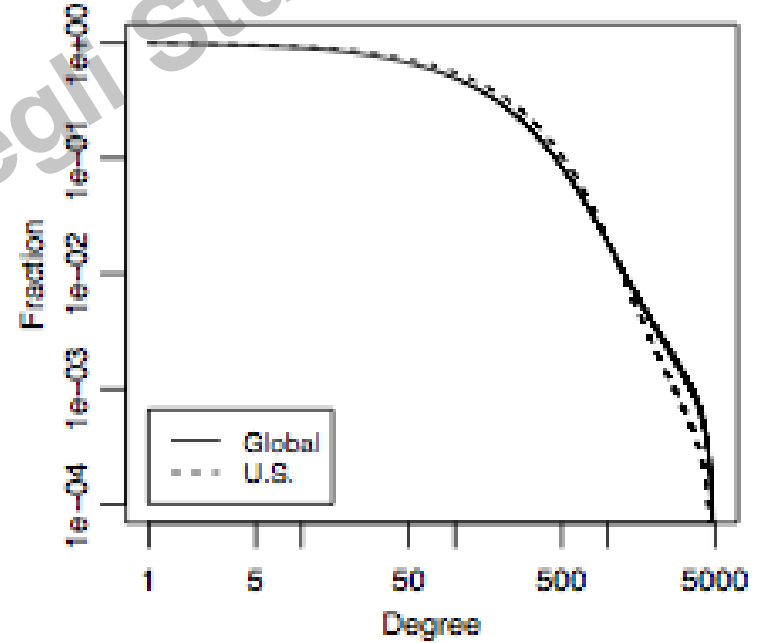


Holme, Edling, Liljeros, 2002.

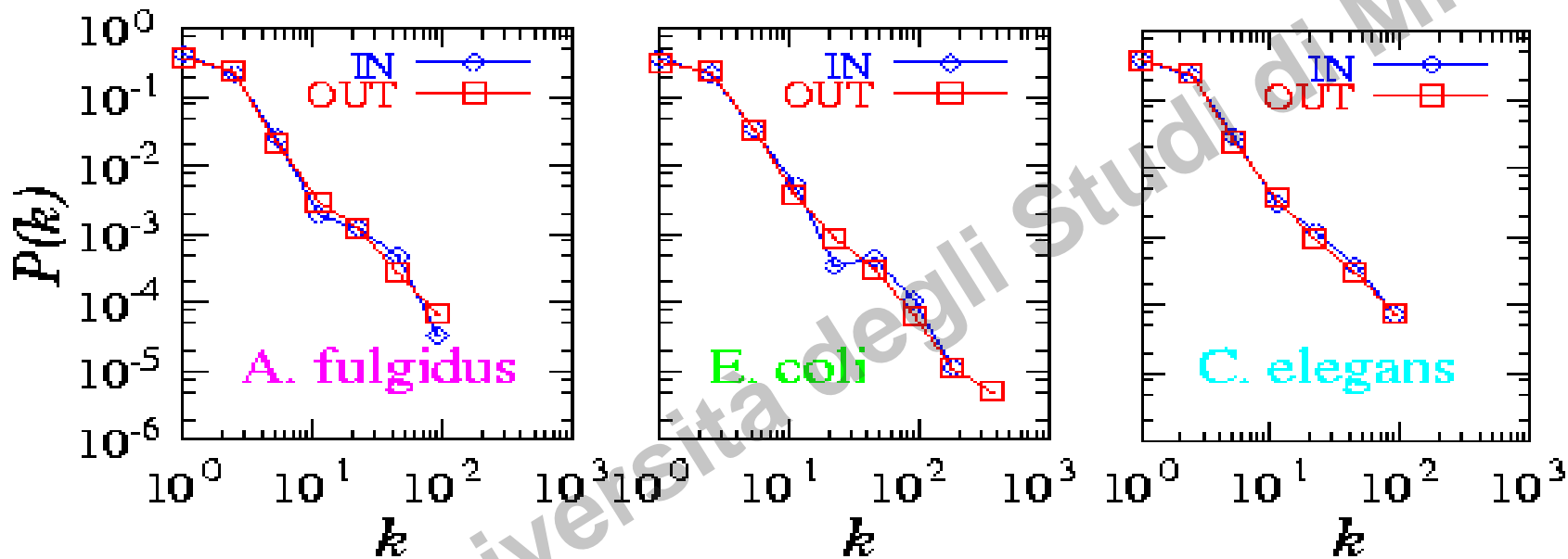
Twitter:



Facebook



METABOLIC NETWORK



Archaea

Bacteria

Eukaryotes

Organisms from all three domains of life are **scale-free!**

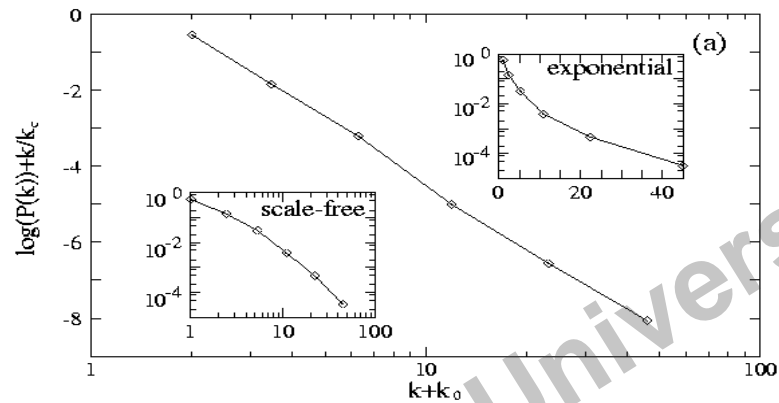
$$P_{in}(k) \approx k^{-2.2}$$

$$P_{out}(k) \approx k^{-2.2}$$

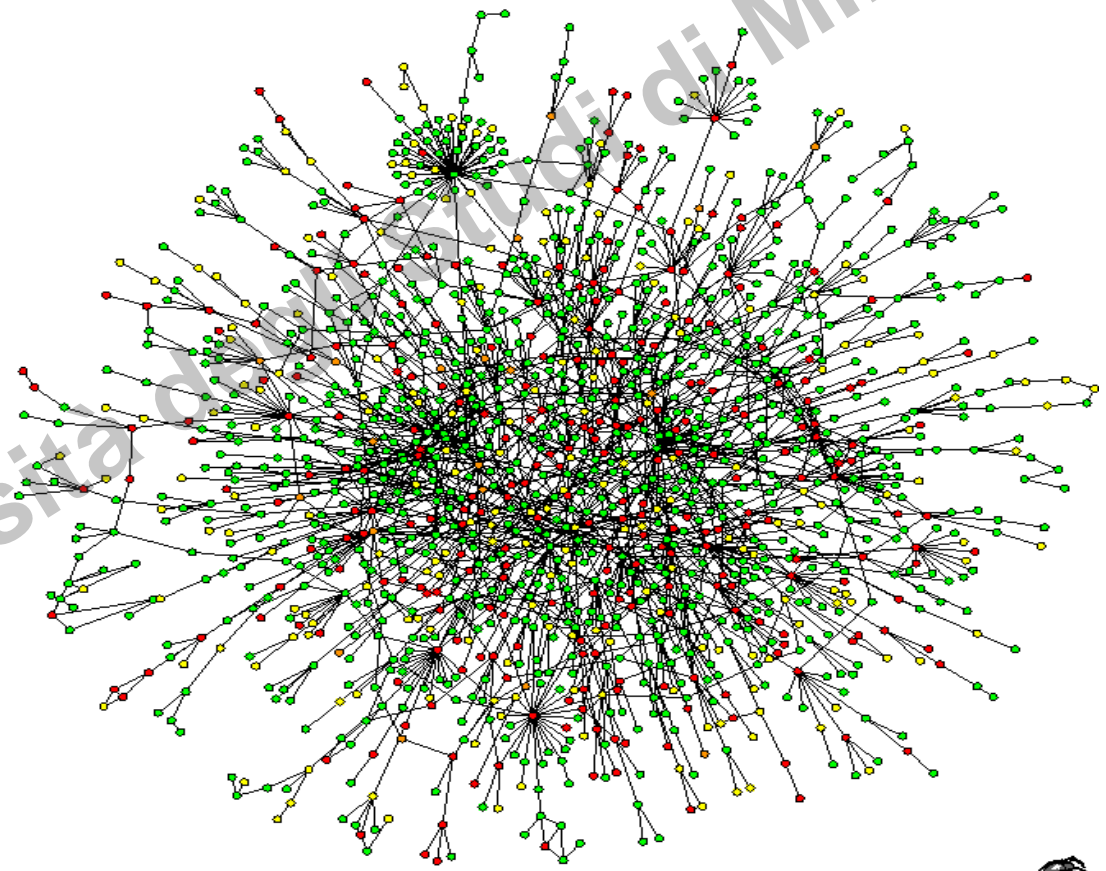
H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi, *Nature*, 407 651 (2000)

TOPOLOGY OF THE PROTEIN NETWORK

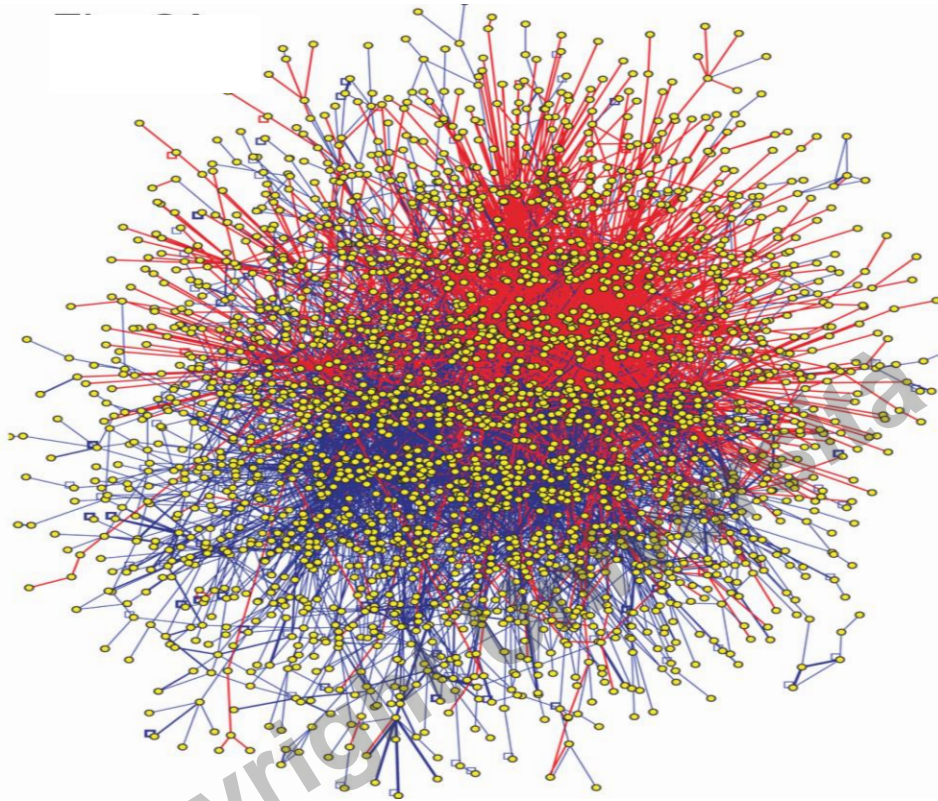
Nodes: proteins
Links: physical interactions-binding



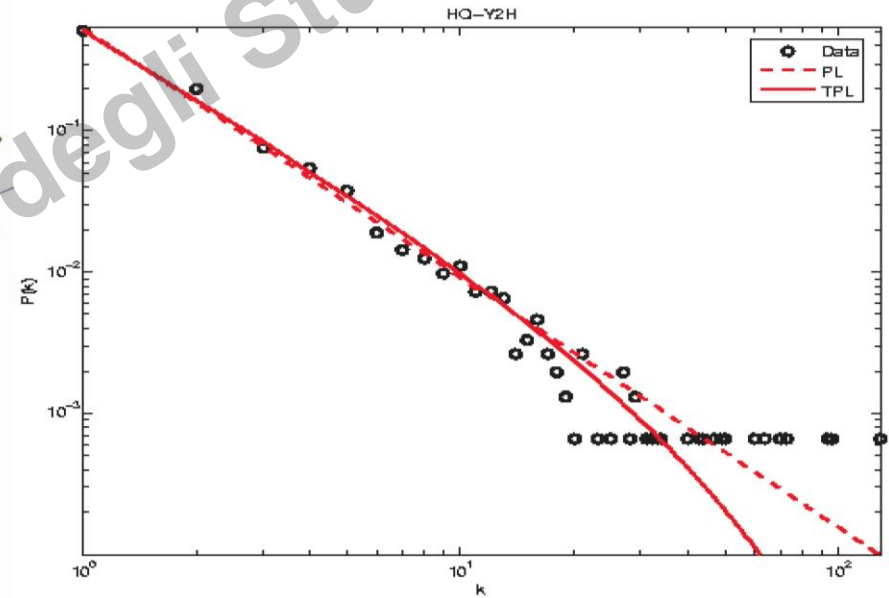
$$P(k) \sim (k + k_0)^{-\gamma} \exp\left(-\frac{k + k_0}{k_\tau}\right)$$



HUMAN INTERACTION NETWORK



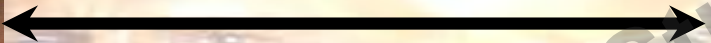
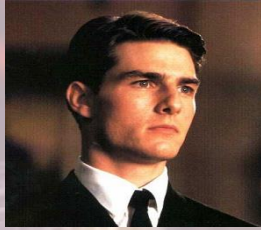
2,800 Y2H interactions
4,100 binary LC interactions
(HPRD, MINT, BIND, DIP, MIPS)



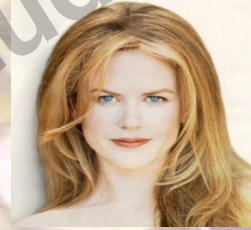
ACTOR NETWORK

Nodes: actors

Links: cast jointly



Days of Thunder (1990)
Far and Away (1992)
Eyes Wide Shut (1999)

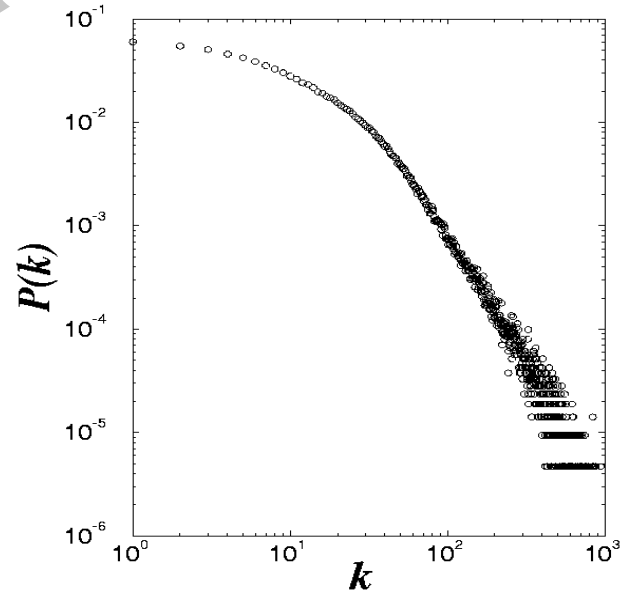


$N = 212,250$ actors

$\langle k \rangle = 28.78$

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

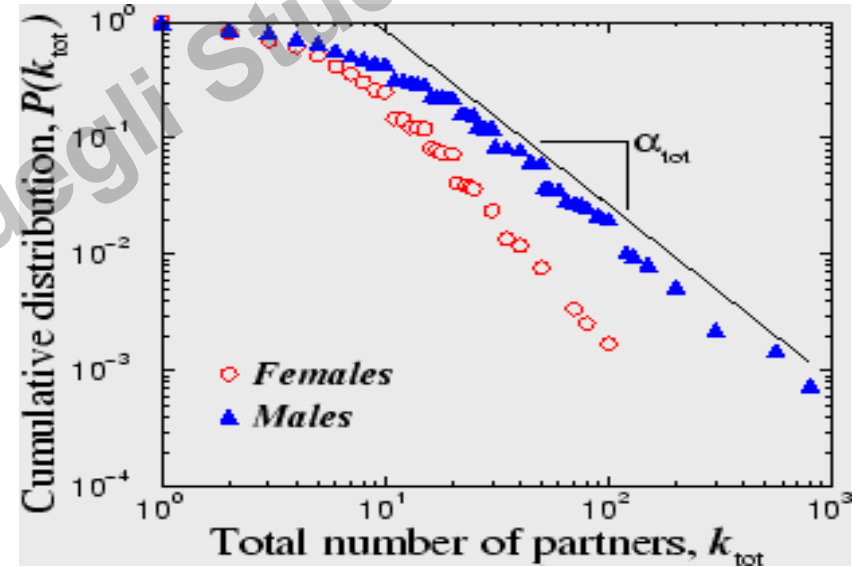
$\gamma = 2.3$





Nodes: people (Females; Males)

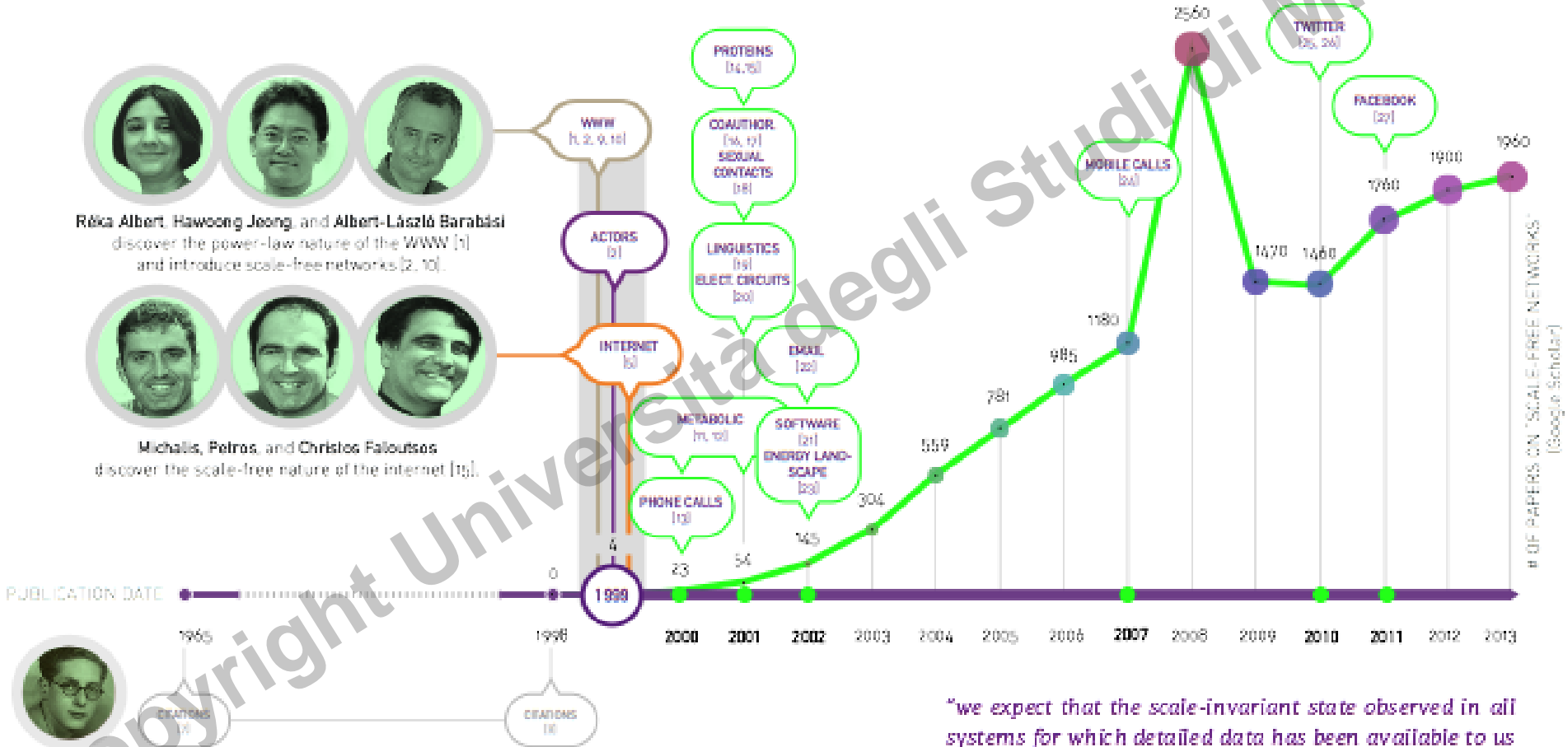
Links: sexual relationships



4781 Swedes; 18-74;
59% response rate.

Liljeros et al. Nature 2001

TIMELINE: SCALE-FREE NETWORKS



Réka Albert, Hawoong Jeong, and Albert-László Barabási
 discover the power-law nature of the WWW [1] and introduce scale-free networks [2, 10].

Michalis Petros, and Christos Faloutsos
 discover the scale-free nature of the internet [15].

Derek de Solla Price (1902 - 1983)
 discovers that citations follow a power-law distribution [7], a finding later attributed to the scale-free nature of the citation network [2].

"we expect that the scale-invariant state observed in all systems for which detailed data has been available to us is a generic property of many complex networks, with applicability reaching far beyond the quoted examples."

Barabási and Albert, 1999

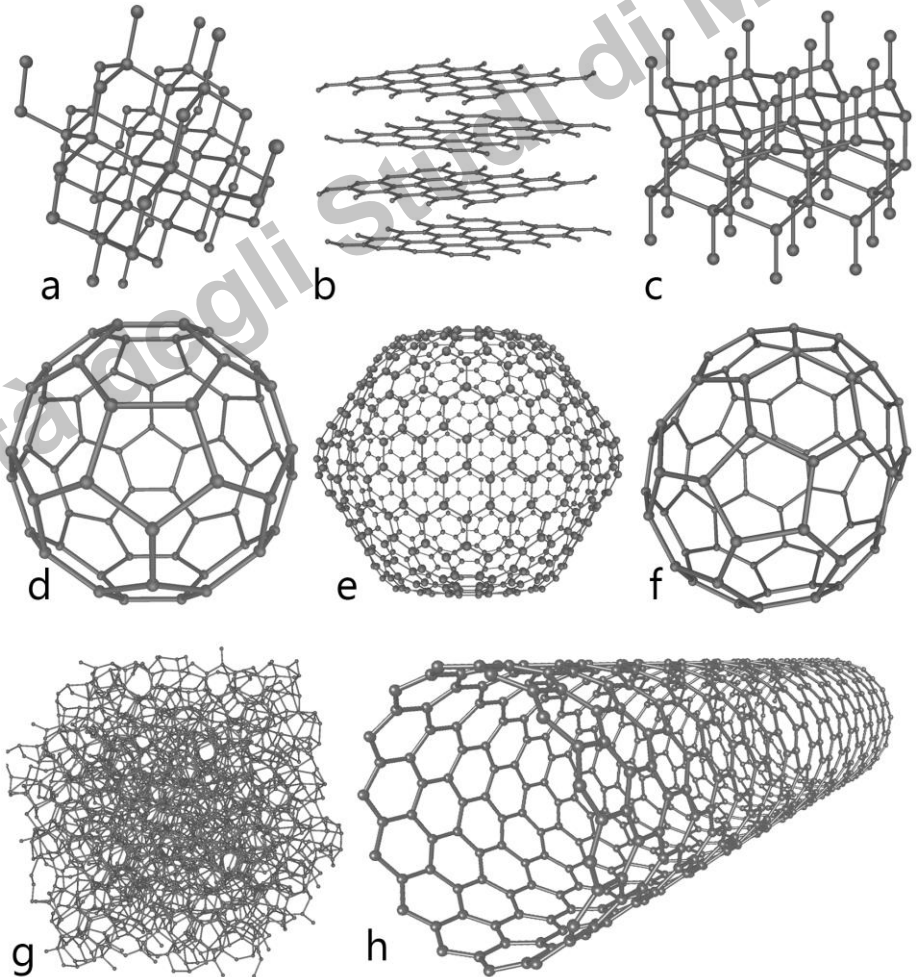
OF PAPERS ON "SCALE-FREE NETWORKS" (Google Scholar)

Not all networks are scale-free

- Networks appearing in material science, like the network describing the bonds between the atoms in crystalline or amorphous materials, where each node has exactly the same degree.

- The neural network of the *C.elegans* worm.

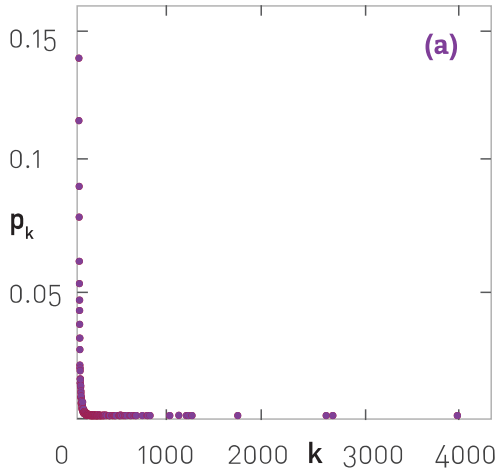
- The power grid, consisting of generators and switches connected by transmission lines



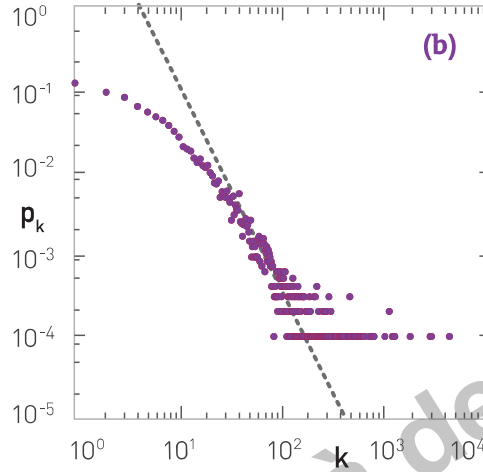
PLOTTING POWER LAWS

Copyright Università degli Studi di Milano

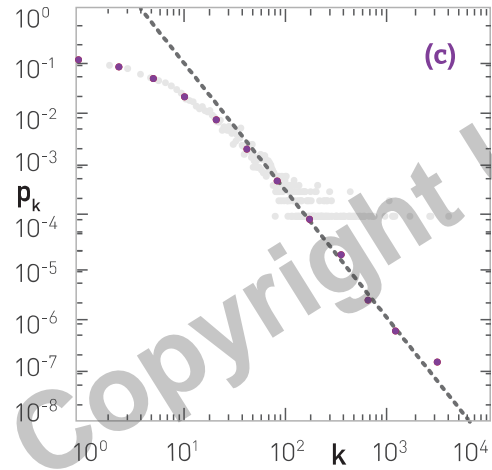
LINEAR SCALE



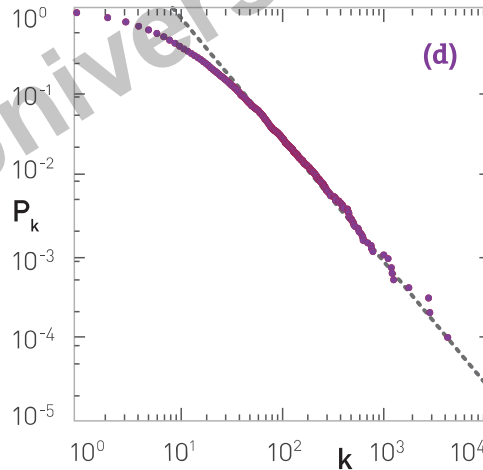
LINEAR BINNING



LOG-BINNING



CUMULATIVE



$$P(k) = N_k / N$$

Use a Log-Log Plot

Avoid Linear Binning

Use Logarithmic Binning

Use Cumulative Distribution

Random network model

First drawback:

The random network model is characterized by a Poisson degree distribution, in contrast to power-law distribution as seen in real networks.

In a random networks all vertices are alike, while real networks are characterized by a small number of vertices with very large degree while most vertices maintain a very low degree.

Credits

Albert-László Barabási

Network Science

Chapter 4.1 – 4.5, 4.11, 4.12, 4.13

Copyright Università degli Studi di Milano

POWER LAW (ce mni)

$$x > 0$$

$$\alpha > 0$$

$$\text{PDF: } \begin{cases} p(x) \propto x^{-\alpha} \\ \log p(x) \propto -\alpha \log x \end{cases}$$

$$p(x) = Cx^{-\alpha}$$

$$p(x) = (\alpha - 1)x^{-\alpha}$$

$$\begin{aligned} \text{CDF} \quad F(x) &= 1 - x^{-\alpha+1} \\ &= 1 - x^{-(\alpha-1)} \\ &= P(X \leq x) \end{aligned}$$

$$\begin{aligned} \text{CCDF} \quad P(X > x) &= 1 - P(X \leq x) \\ &= x^{-(\alpha-1)} \end{aligned}$$

$$\log \text{CCDF} = -(\alpha - 1) \log x$$

$$\text{MEDIA:} \begin{cases} \text{non definita} & \alpha \leq 2 \\ \frac{\alpha-1}{\alpha-2} & \alpha > 2 \end{cases}$$

$$\text{VARIANZA} \begin{cases} \text{non def} & \alpha \in (2, 3] \\ \frac{\alpha-1}{(\alpha-2)^2(\alpha-3)^2} & \alpha > 3 \end{cases}$$

$$2 < \alpha < 3$$

Media finita

Varianza non def.

α oppure γ

ALTERNATIVA

$$\begin{array}{l} \text{PDF} \quad \approx x^{-(\eta+1)} \\ \text{CCDF} \quad \approx x^{-\eta} \end{array}$$

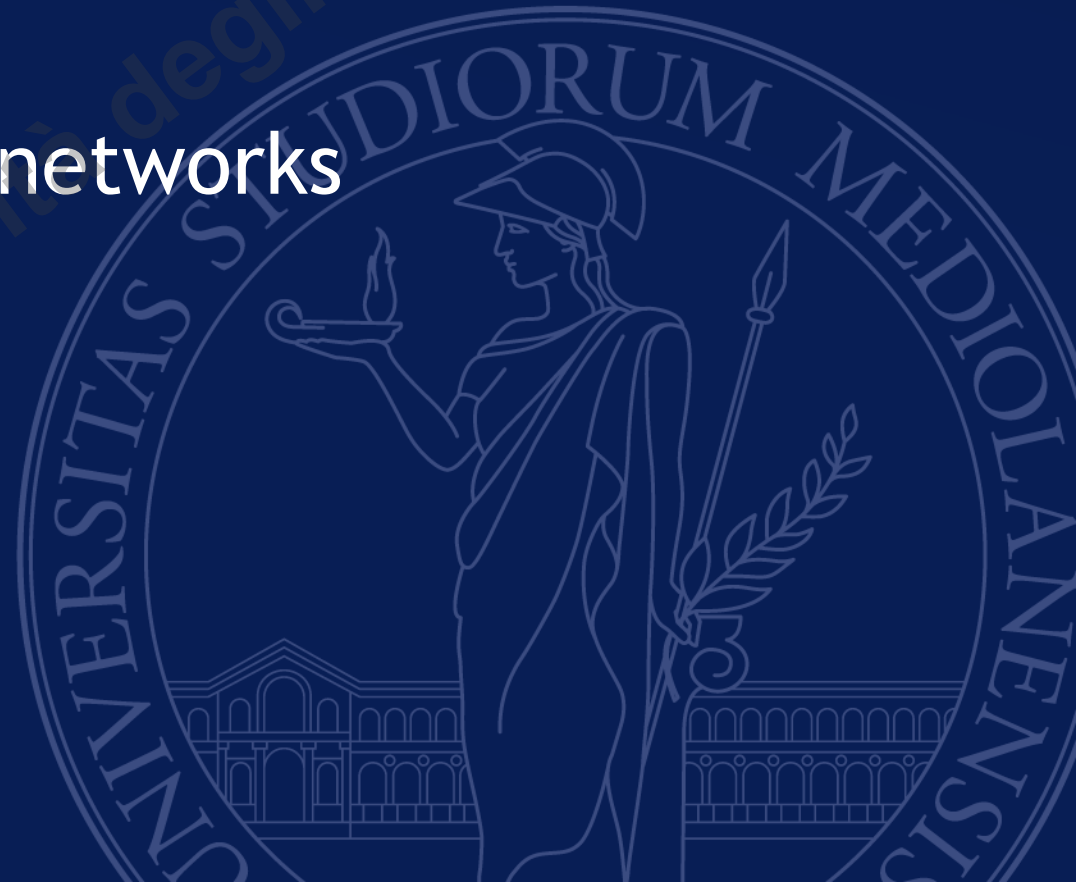
$$\alpha = \eta + 1$$



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Scale Free networks

Examples



Facebook

- Ugander, Johan & Karrer, Brian & Backstrom, Lars & Marlow, Cameron. (2011). The Anatomy of the Facebook Social Graph. arXiv preprint. 1111.4503.
 - Degree distribution
Pag 3, Figure 1



Twitter

- Seth A. Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. 2014. Information network or social network?: the structure of the twitter follow graph. In Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion). ACM, New York, NY, USA, 493-498. DOI: <https://doi.org/10.1145/2567948.2576939>

- Degree distribution

Chapter 3.1, Figure 1, Table 1



Web

- Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the Web. *Comput. Netw.* 33, 1-6 (June 2000), 309-320.
DOI=[http://dx.doi.org/10.1016/S1389-1286\(00\)00083-9](http://dx.doi.org/10.1016/S1389-1286(00)00083-9)
 - Degree distribution
Chapter 2.2.1, Figure 1-4



Mobile communication networks

- Structure and tie strengths in mobile communication networks,
J. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.
-L. Barabási

Proceedings of the National Academy of Sciences May
2007, 104 (18) 7332- 7336; DOI: 10.1073/pnas.0610245104

- Degree distribution: Figure 1a

- Calling, texting, and moving: multidimensional interactions of mobile
phone users

Matteo Zignani, Christian Quadri, Sabrina Gaito & Gian Paolo Rossi
Computational Social Networks volume 2, Article number: 13 (2015)

- Degree distribution: Figure 5



Network Science

Class 5: BA model

Albert-László Barabási

With

Roberta Sinatra and Sean P. Cornelius

www.BarabasiLab.com

Hubs represent the most striking difference between a random and a scale-free network. Their emergence in many real systems raises several fundamental questions:

- Why does the random network model of Erdős and Rényi fail to reproduce the hubs and the power laws observed in many real networks?
- Why do so different systems as the WWW or the cell converge to a similar scale-free architecture? (Different type of nodes, links, history and purpose)

To understand why so different systems converge to a similar architecture we need to first uncover the *mechanism* responsible for the emergence of the scale-free property

Given the major differences between the systems that display the scale-free property, the explanation must be *simple* and *fundamental*.

Why are hubs and power laws absent in random networks?

There are *two hidden assumptions* of the Erdios-Renyi model, that are violated in real networks

Growth and preferential attachment

BA MODEL: Growth

ER model:

the number of nodes, N , is fixed (static models)

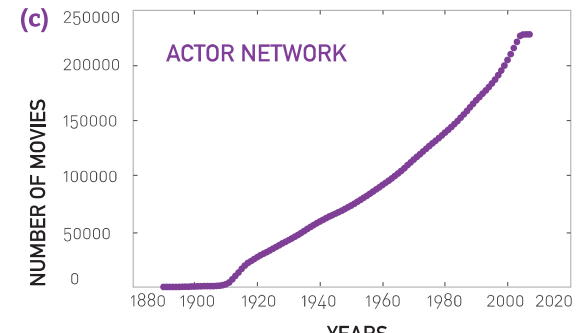
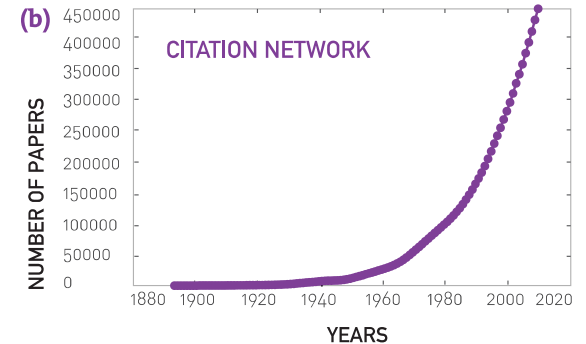
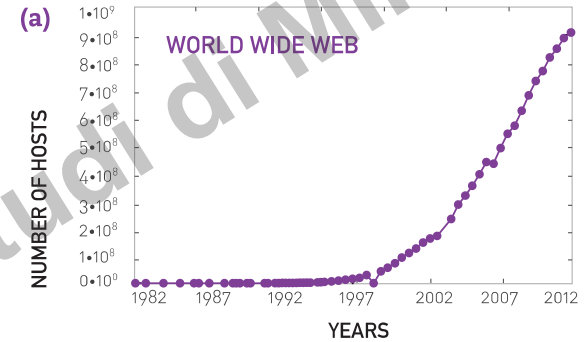
networks expand through the addition of new nodes

In 1991 the www had a single node, today the Web has over a trillion (10^{12}) documents

The protein interaction network may appear to be static, yet it is not. The number of genes in a human cell has grown from a few to over 20000 in four billion years

If we wish to model these networks, we cannot resort to a static model. Our modeling approach must instead acknowledge that networks are the product of a steady growth process.

Barabási & Albert, *Science* **286**, 509 (1999)



ER model: links are added randomly to the network

New nodes prefer to connect to the more connected nodes

Rich-gets-richer phenomenon

BA MODEL: Preferential attachment

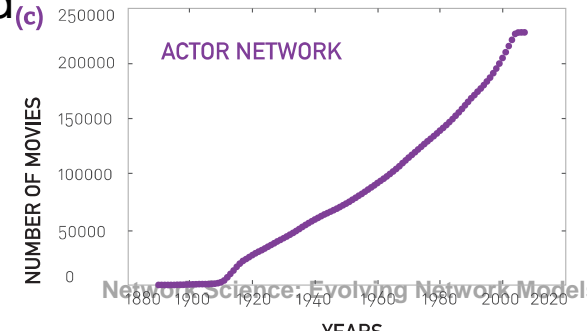
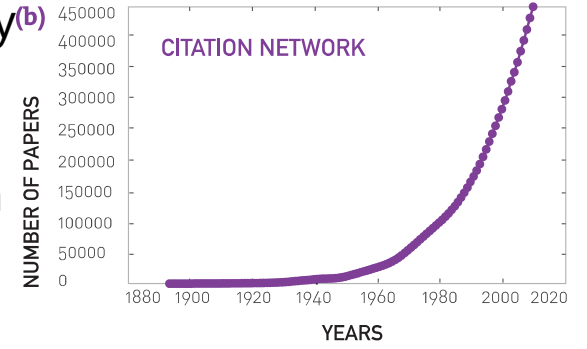
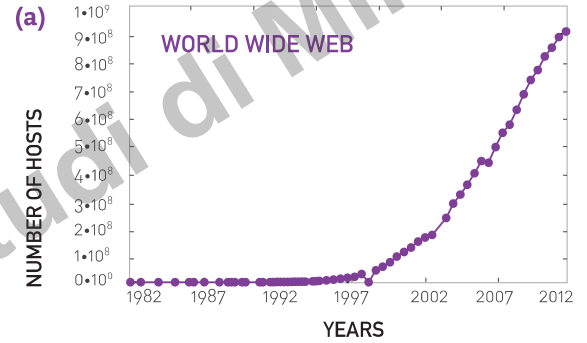
We are familiar with only a tiny fraction of the trillion or more documents available on the **WWW**. The nodes we know are not entirely random: we all heard about Google and Facebook, but we rarely encounter the billions of less-prominent nodes that populate the Web. As our knowledge is biased towards the more connected nodes, we are more likely to link to a hub than to a node with only few links.

With more than a million scientific **papers** published each year, no scientist can attempt to read them all. The more cited is a paper, the more likely that we will notice it.

Therefore, our citations are biased towards the more cited publications.

The more movies an **actor** has played in, the higher are the chances that he/she will be considered for a new role

Barabási & Albert, *Science* **286**, 509 (1999)



Section 2: Growth and Preferential Attachment

The random network model differs from real networks in two important characteristics:

Growth: While the random network model assumes that the number of nodes is fixed (time invariant), real networks are the result of a growth process that continuously increases.

Preferential Attachment: While nodes in random networks randomly choose their interaction partner, in real networks new nodes prefer to link to the more connected nodes.

The Barabási-Albert model

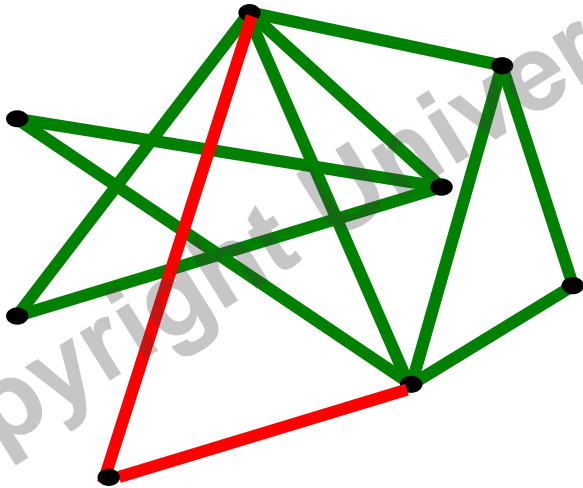
Origin of SF networks: Growth and preferential attachment

(1) Networks continuously expand by the addition of new nodes

WWW : addition of new documents

(2) New nodes prefer to link to highly connected nodes.

WWW : linking to well known sites



Barabási & Albert, *Science* **286**, 509 (1999)

GROWTH:

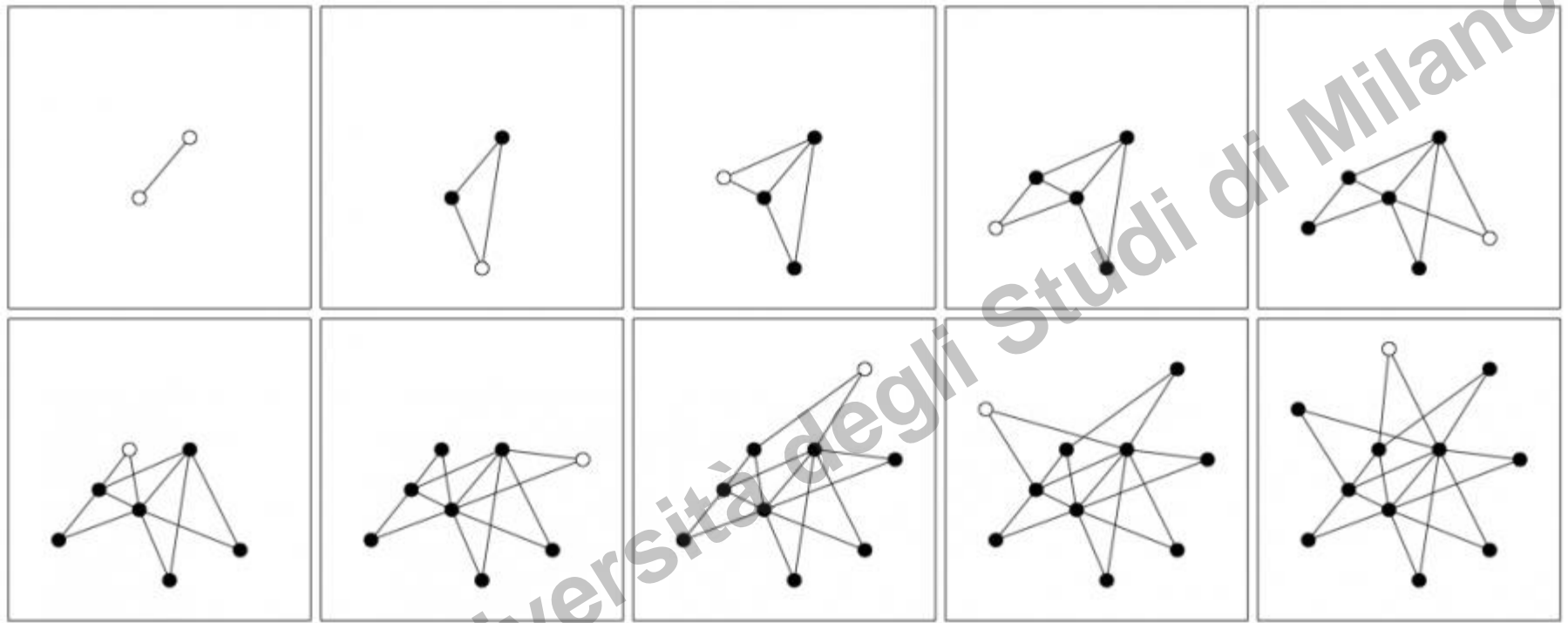
add a new node at each time step with k links that connect the new node to k nodes already in the network

PREFERENTIAL ATTACHMENT:

the probability that a node connects to a node with k links is proportional to k .

$$P(k_i) = k_i / \sum_j k_j$$

Preferential attachment is a probabilistic mechanism: A new node is free to connect to *any* node in the network. However, if a new node has a choice between a degree-two and a degree-four node, it is twice as likely that it connects to the degree-four node.



It shows nine subsequent steps of the Barabasi-Albert model.

Empty circles: newly added nodes deciding where to connect their two links using preferential attachment.

A few nodes gradually turn into hubs

Barabási-Albert model

The definition of the Barabási-Albert model leaves many mathematical details open:

It does not specify the precise initial configuration of the first m_0 nodes.

It does not specify whether the k links assigned to a new node are added one by one, or simultaneously. This leads to potential mathematical conflicts: If the links are truly independent, they could connect to the same node i , resulting in multi-links and loops.

The first mathematical model was introduced by Bollobas et al.:

Linearized chord diagram model

Time in networks

As we compare the predictions of the network models with real data, we have to decide how to measure *time* in networks. Real networks evolve over rather different time scales:

World Wide Web

The first webpage was created in 1991. Given its trillion documents, the WWW added a node each millisecond (10^3 sec)

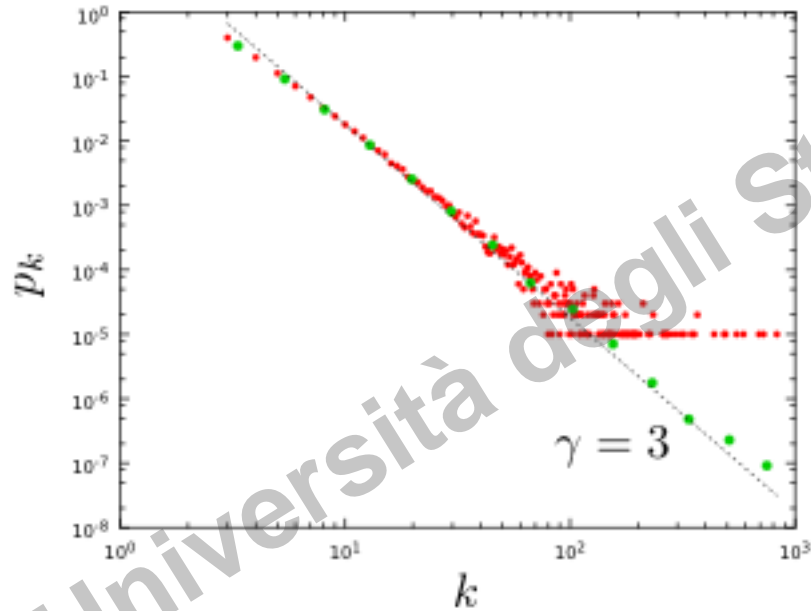
Cell

The cell is the result of 4 billion years of evolution. With roughly 20,000 genes in a human cell, on average the cellular network added a node every 200,000 years ($\sim 10^{13}$ sec).

Given these enormous time-scale differences, it is impossible to use real time to compare the dynamics of different networks. Therefore, in network theory we use *event time*, advancing our time-step by one each time when there is a change in the network topology.

For example, in the Barabási-Albert model the addition of each new node corresponds to a new time step, hence $t=N$.

In other models time is also advanced by the arrival of a new link or the deletion of a node. If needed, we can establish a direct mapping between event time and the physical time.



Red: linear binning
Green: log binning

The degree distribution of a network generated by the Barabási-Albert model. The plot shows both the linearly-binned (red symbols) as well as the log-binned version (green symbols). The straight line is added to guide the eye and has slope -3 , corresponding to the resulting network's degree exponent.

Do we need both growth and preferential attachment?

YEP

The absence of preferential attachment leads to a growing network with a stationary but exponential degree distribution

The absence of growth leads to the loss of stationarity, forcing the network to converge to a complete graph

The BA model is only a minimal model.



Founded six years after birth of the World Wide Web, Google was a latecomer to search. By the late 1990s Alta Vista and Inktomi, two search engines with an early start, have been dominating the search market. Yet Google, the third mover, soon not only became the leading search engine, but acquired links at such an incredible rate that by 2000 became the most connected node of the Web as well [1]. But its status didn't last: in 2011 Facebook, with an even later start, took over as the Web's biggest hub.

Bibliography

Emergence of Scaling in Random Networks
BY ALBERT-LÁSZLÓ BARABÁSI, RÉKA ALBERT
SCIENCE 15 OCT 1999 : 509-512

Albert-László Barabási
Network Science
Chapter 3



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

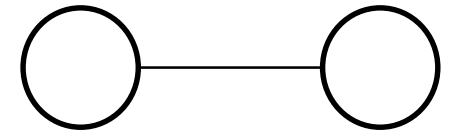
Connectivity
Social Networks Analysis



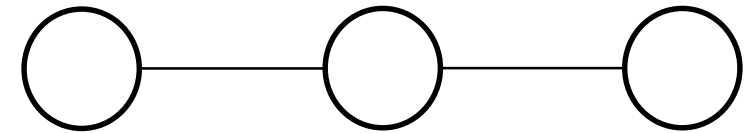
Connectivity

How nodes are connected via a sequence of links in a network

Two nodes are **adjacent** if they are connected via a link.



Two links are **incident**, if they share an end-point



An edge in a graph can be traversed when one starts at one of its end-nodes, moves along the edge, and stops at its other end-node.



Path

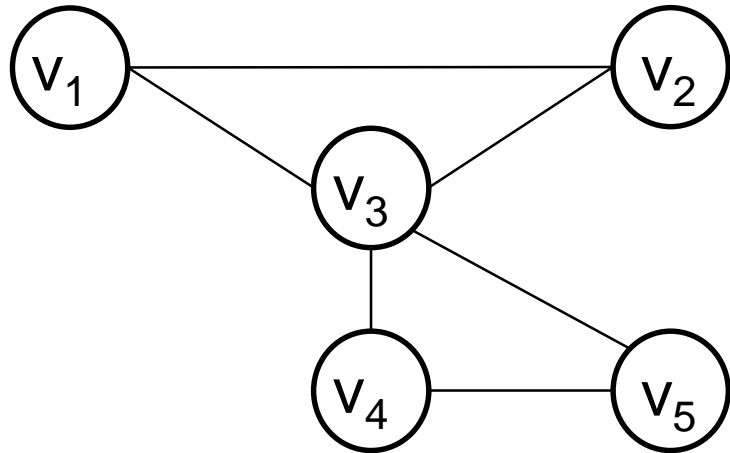
Walk: a sequence of incident links visited one after another

$\{(v_1, v_3), (v_3, v_4), (v_4, v_5), (v_5, v_3), (v_3, v_2)\}$

Path: a walk where nodes and links are distinct

$\{(v_1, v_3), (v_3, v_4), (v_4, v_5)\}$ [Alternatively in simple graph: $\{v_1, v_3, v_4, v_5\}$]

Path length: the number of links visited in the path

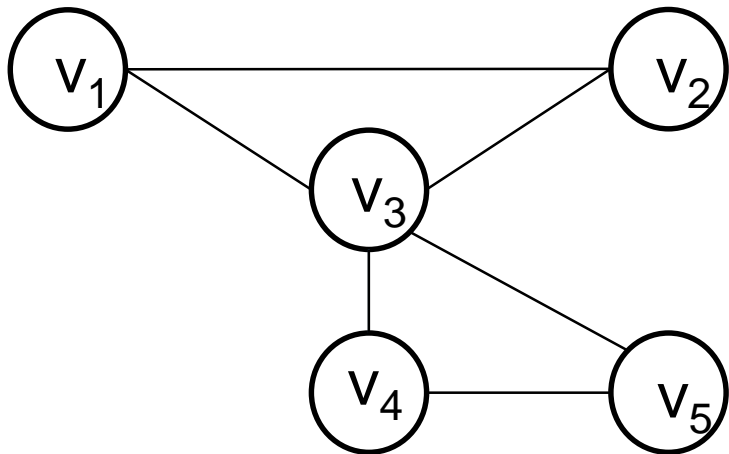


A node v_i is connected to node v_j (or reachable from v_j) if it is adjacent to it or there exists a path from v_i to v_j .



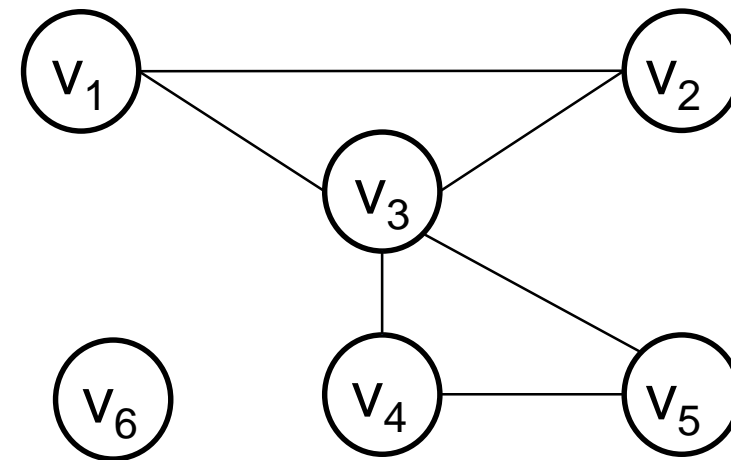
Connected graph

A graph is *connected*, if there exists a path between *any* pair of nodes in it



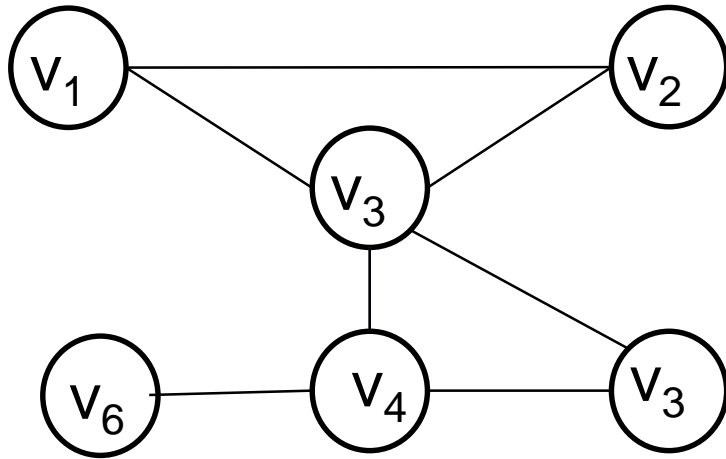
A graph is *disconnected*, if it is not connected.

[It exists at least a pair of nodes which are not connected]

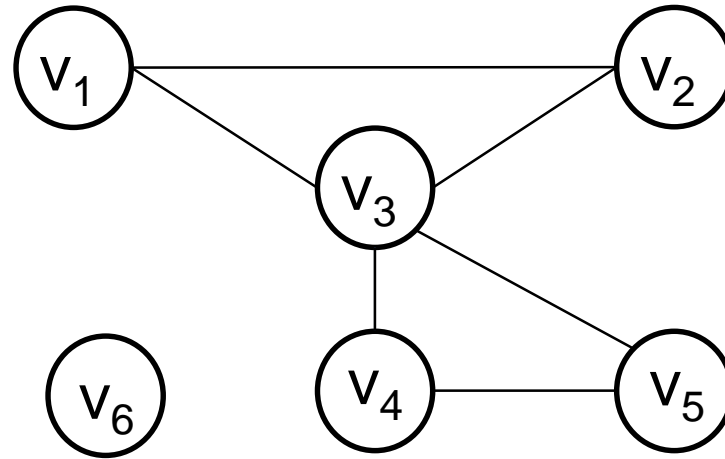


Connected components: intuition

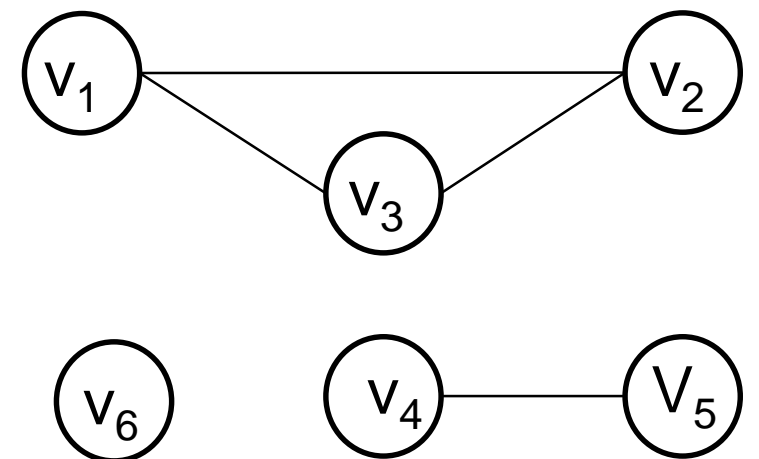
Subgroups of nodes, with no connections



1 connected components



2 connected components



3 connected components

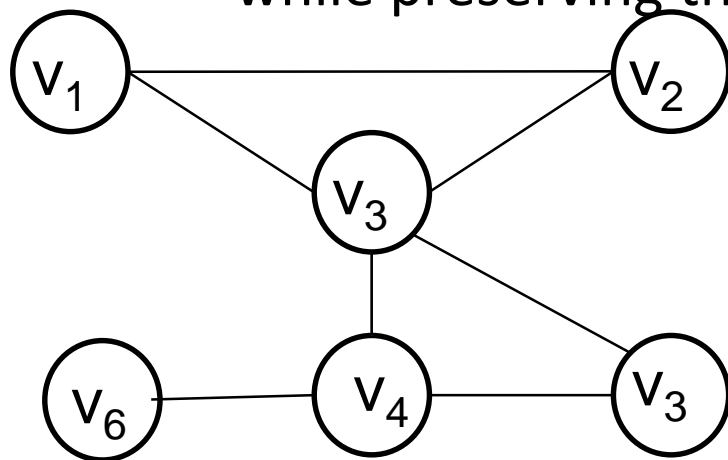


Connected components: definition

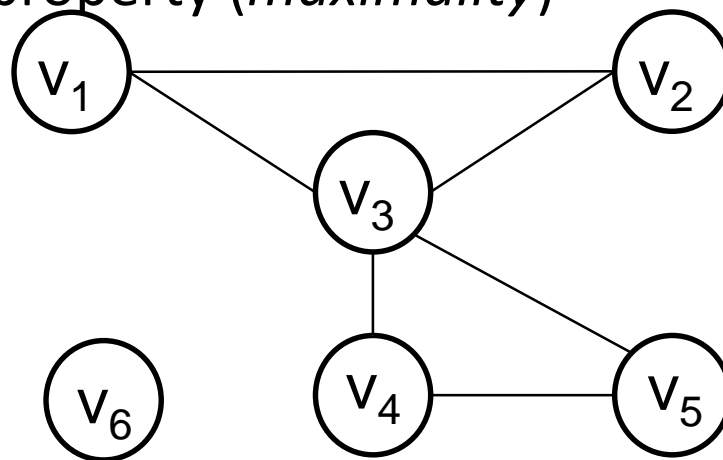
A connected component is a subgraph of a network such that there exists at least one path from each member of that subgraph to each other member,

and

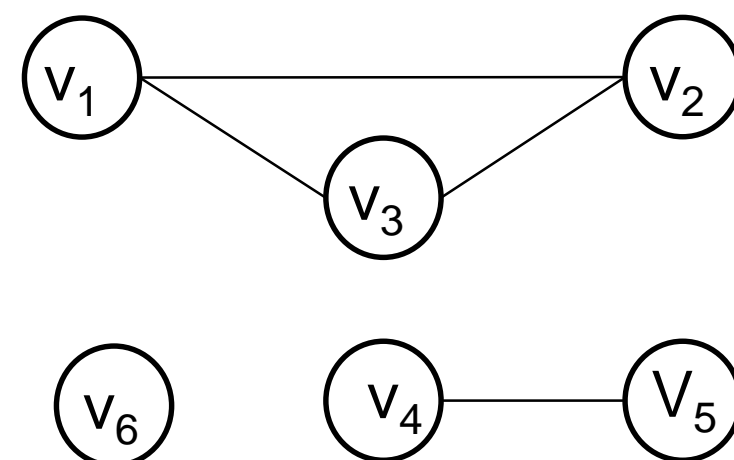
no other vertex in the network can be added to the subgraph while preserving this property (*maximality*)



1 connected components



2 connected components



3 connected components



Connected component: definition

A connected component is a subgraph of a network such that

It is a node-generated subgraph, i.e. the subsets of vertices and all edges that are between them

there exists at least one path from each member of that subgraph to each other member,

There is a path between all pair of vertices in the component

Each node of the component is reachable from any other node of the component

no other vertex in the network can be added to the subgraph while preserving this property

There is no path between a node in the component and any other not in the component (maximality)



Connected components in social networks

There is typically a very large component that fills most of the network - usually more than half and not infrequently over 90% - while the rest of the network is divided into a large number of small components.

There are some networks for which the largest component fills the entire network such as the Internet, communication networks, transportation networks, power grids
In these cases there is always a good specific reason.



Connected Components in social networks

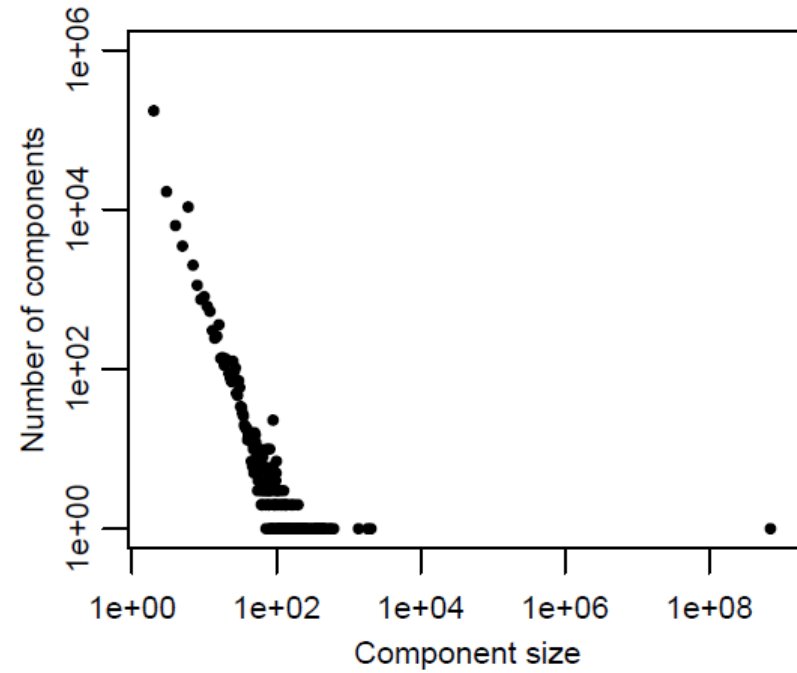


Figure 3. Component size distribution. The fraction of components with a given component size on a log-log scale. Most vertices (99.91%) are in the largest component.

Ugander et al., The Anatomy of the Facebook Social Graph
, 2011.



Connect Components in social networks

Can a network have two or more large components that fill a sizable fraction of the entire graph?

Usually the answer is no.

The argument is that if a network of n nodes was divided into two large components of about $n/2$ nodes each, then there would be $n^2/4$ possible pairs of nodes such that one node was in one large component and the other node in the other large component.

It is highly unlikely that not one such pair would be connected.



Connected components in directed networks

a directed graph is strongly connected if there exists a directed path between any pair of nodes

a directed graph is weakly connected if there exists a path between any pair of nodes, without following the edge directions



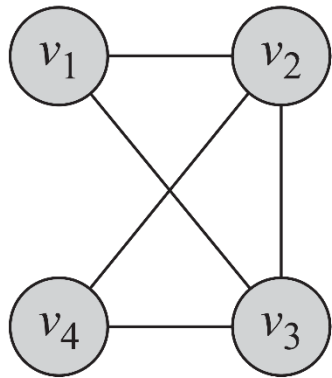
Connected components in directed networks

In a directed graph, a **strongly connected component** is a *maximal* subset of nodes such that each can reach and is reachable from all the others along a *directed* path

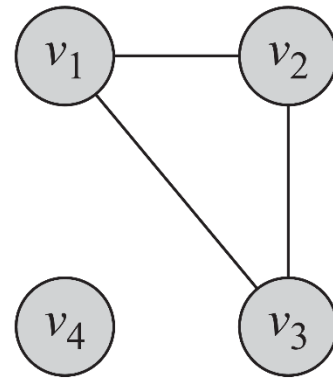
In a directed graph, a **weakly connected component** is a *maximal* subset of nodes such that each can reach and is reachable from all the others along an *undirected* path



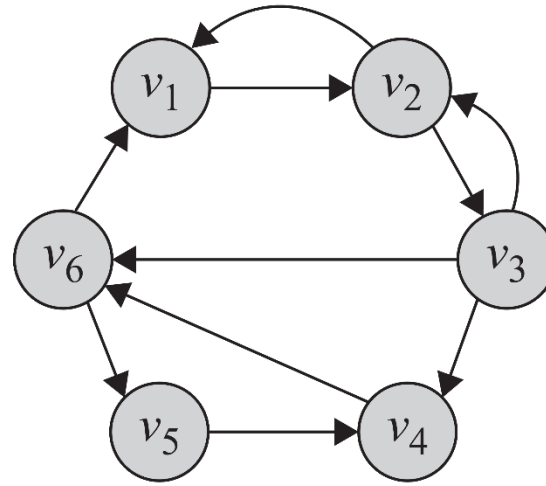
Connectivity in directed networks



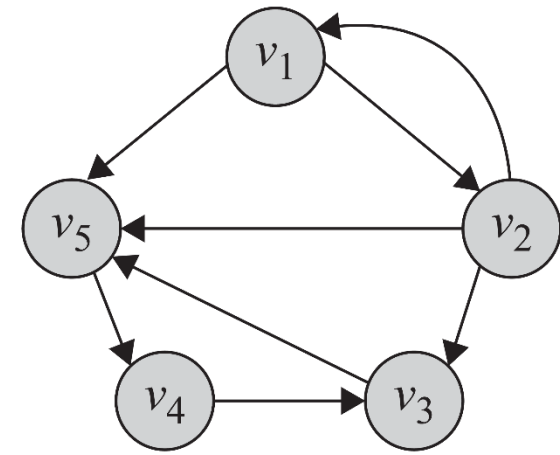
(a) Connected



(b) Disconnected



(c) Strongly connected



(d) Weakly connected



Connected components in real-world directed networks

There is typically one large strongly connected component and a selection of small ones.

The largest strongly connected component in the Web fills about a quarter of the network



Credits

Reza Zafarani, Mohammad Ali Abbasi, Huan Liu
Social Media Mining: An Introduction
A Textbook by Cambridge University Press
Chapter 2.4

Newman, M.E.J.
Networks: An Introduction.
Oxford University Press. 2010.
Chapters 6.11, 8.1

Albert-László Barabási
Network Science



RANDOM-REAL NETWORKS CONNECTED COMPONENTS

Connected components in real-world networks

Real-world networks: giant component
and power-law connected components
size distribution



Growing a random network

Starting with N isolated nodes, the links are added gradually through a random process.

This corresponds to a gradual increase of p , with striking consequences on the network topology.

To quantify this process, we first inspect how the size of the largest connected cluster within the network, N_G , varies with the average degree $\langle k \rangle$.



Connected components in random networks

N_G : number of nodes in the giant component

Two extreme cases:

$p=0 \rightarrow$ disconnected nodes, $\langle k \rangle = 0$, $N_G = 1$

$p=1 \rightarrow$ fully connected, $\langle k \rangle = N-1$, $N_G = N$

One would expect that the largest component grows gradually from $N_G = 1$ to $N_G = N$

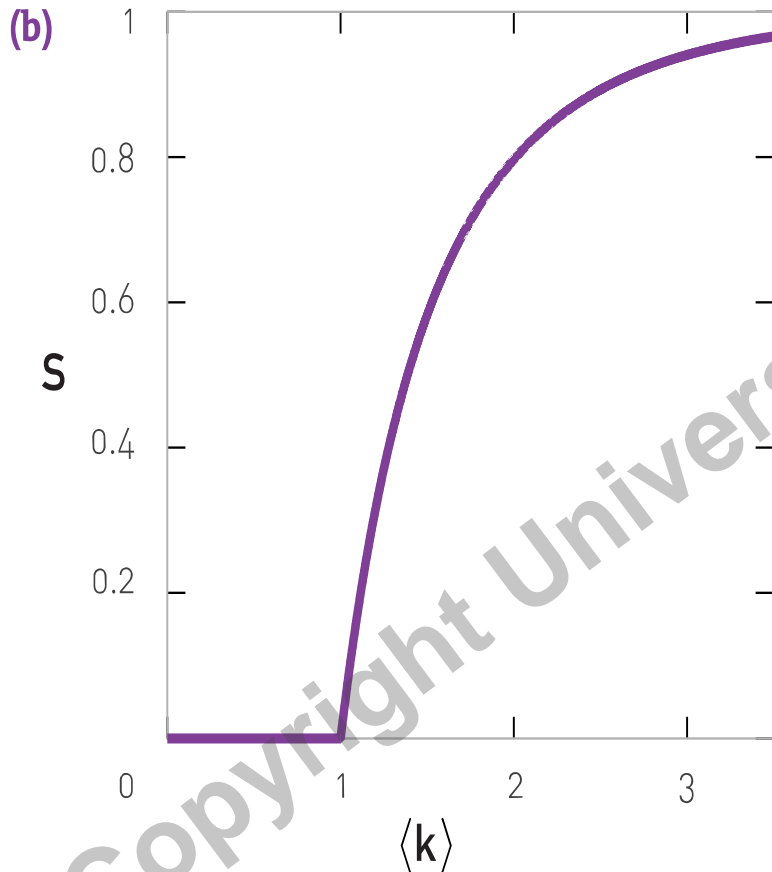


Size of the giant component

$$S = N_G/N$$

S: fraction of nodes in the largest connected component

One would expect that the largest component grows gradually
Yet, this is not the case.



N_G/N remains zero for small $\langle k \rangle$, indicating the lack of a large connected component.

Once $\langle k \rangle$ exceeds a critical value, N_G/N increases, signaling the rapid emergence of a large component that we call the *giant component*.

Erdős and Renyi in their classical 1959 paper predicted that the condition for the emergence of the giant component is:

$$\langle k \rangle = 1$$



Size of the giant component

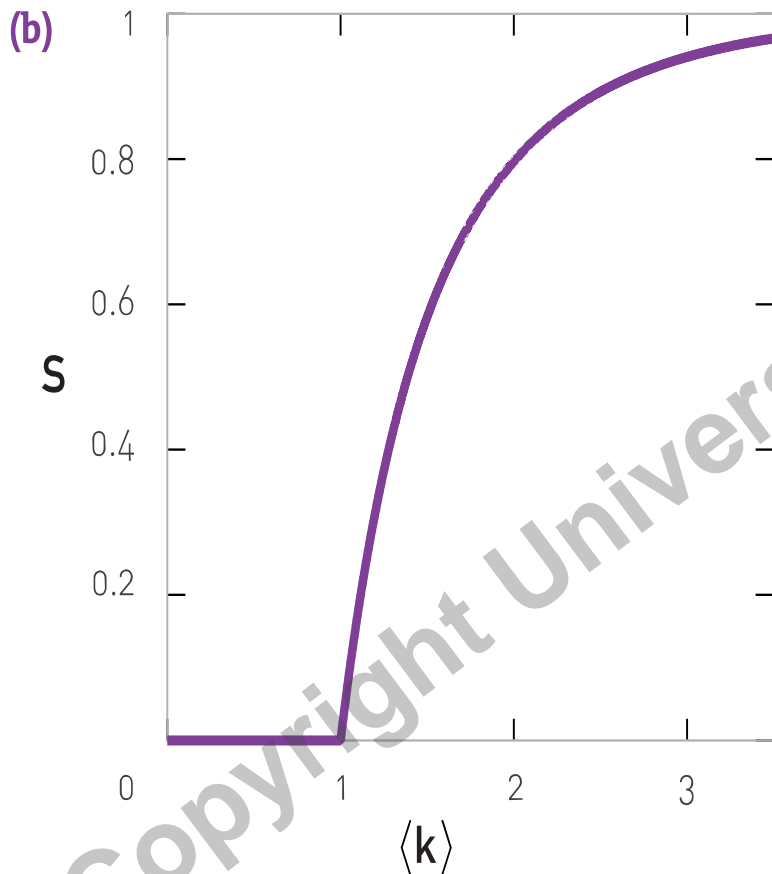
$$S = N_G / N$$

S : fraction of nodes in the largest connected components

In other words, we have a giant component if and only if each node has on average more than one link.

The fact that we need at least one link per node to observe a giant component is not unexpected. Indeed, for a giant component to exist, each of its nodes must be linked to at least one other node.

It is somewhat counterintuitive, however, that one link is *sufficient* for its emergence.



In this section we introduce the argument, proposed independently by Solomonoff and Rapoport [11], and by Erdős and Rényi [2], for the emergence of giant component at $\langle k \rangle = 1$ [33].

Let us denote with $u = 1 - N_c/N$ the fraction of nodes that are not in the giant component (GC), whose size we take to be N_c . If node i is part of the GC, it must link to another node j , which must also be part of the GC. Hence if i is *not* part of the GC, that could happen for two reasons:

- There is no link between i and j (probability for this is $1-p$).
- There is a link between i and j , but j is not part of the GC (probability for this is pu).

Therefore the total probability that i is not part of the GC via node j is $1-p+pu$. The probability that i is not linked to the GC via any other node is therefore $(1-p+pu)^{N-1}$, as there are $N-1$ nodes that could serve as potential links to the GC for node i . As u is the fraction of nodes that do not belong to the GC, for any p and N the solution of the equation

$$u = (1-p+pu)^{N-1} \quad (3.30)$$

provides the size of the giant component via $N_c = N(1-u)$. Using $p = \langle k \rangle / (N-1)$ and taking the logarithm of both sides, for $\langle k \rangle \ll N$ we obtain

$$\ln u = (N-1) \ln \left[1 - \frac{\langle k \rangle}{N-1} (1-u) \right] \approx (N-1) \left[-\frac{\langle k \rangle}{N-1} (1-u) \right] = -\langle k \rangle (1-u), \quad (3.31)$$

where we used the series expansion for $\ln(1+x)$.

Taking an exponential of both sides leads to $u = \exp[-\langle k \rangle (1-u)]$. If we denote with S the fraction of nodes in the giant component, $S = N_c / N$, then $S = 1-u$ and (3.31) results in

$$S = 1 - e^{-\langle k \rangle S}. \quad (3.32)$$

Giant component
(Barabasi's book, Section 3.14 –
Advanced topics 3.C)

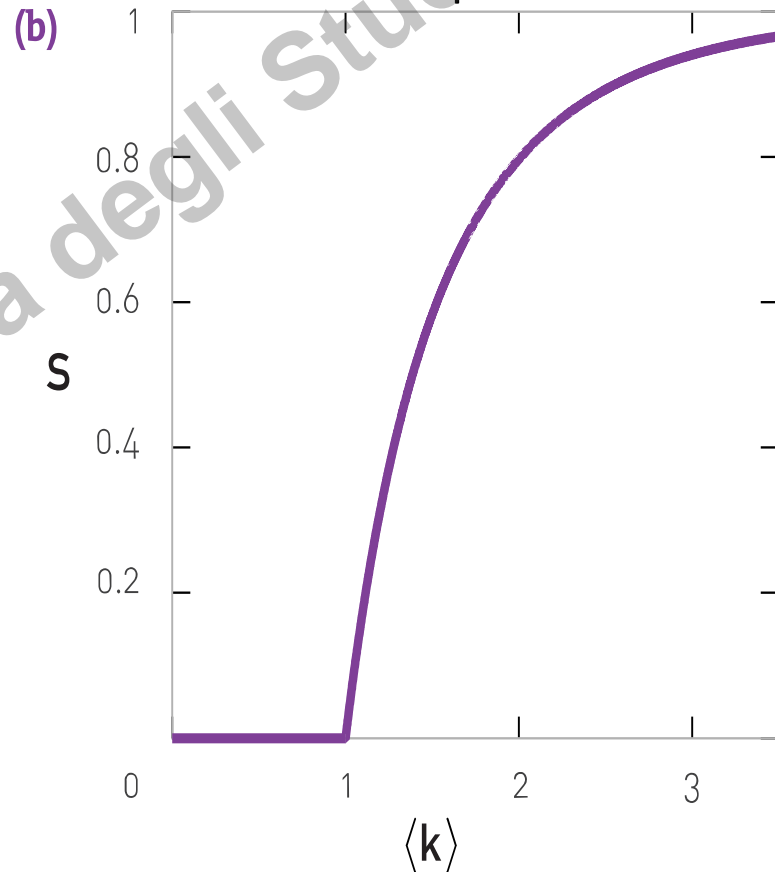
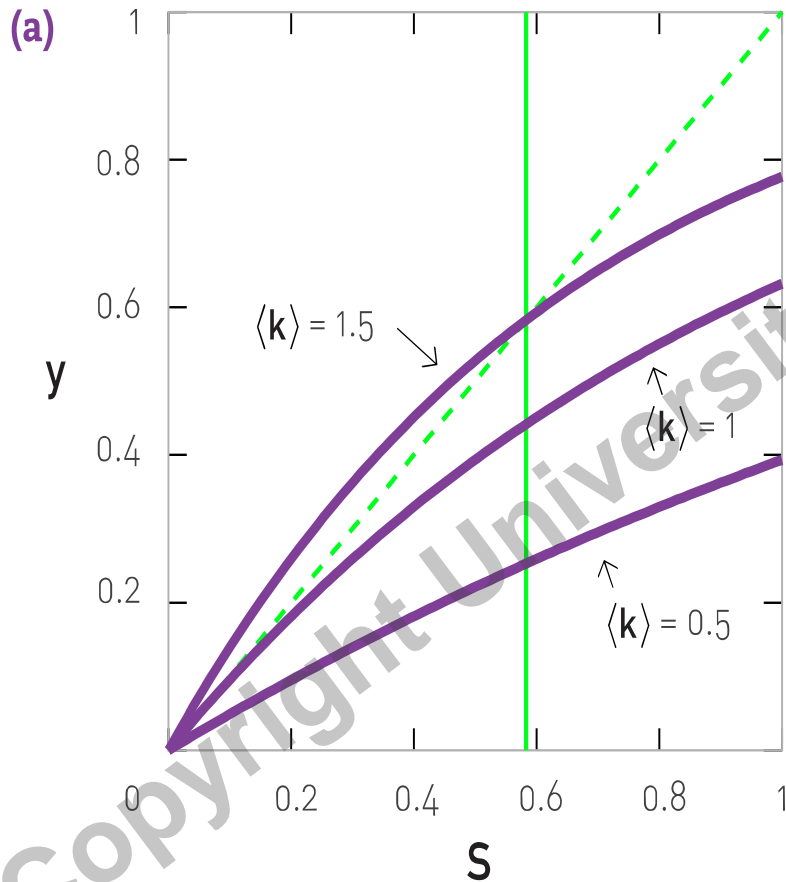


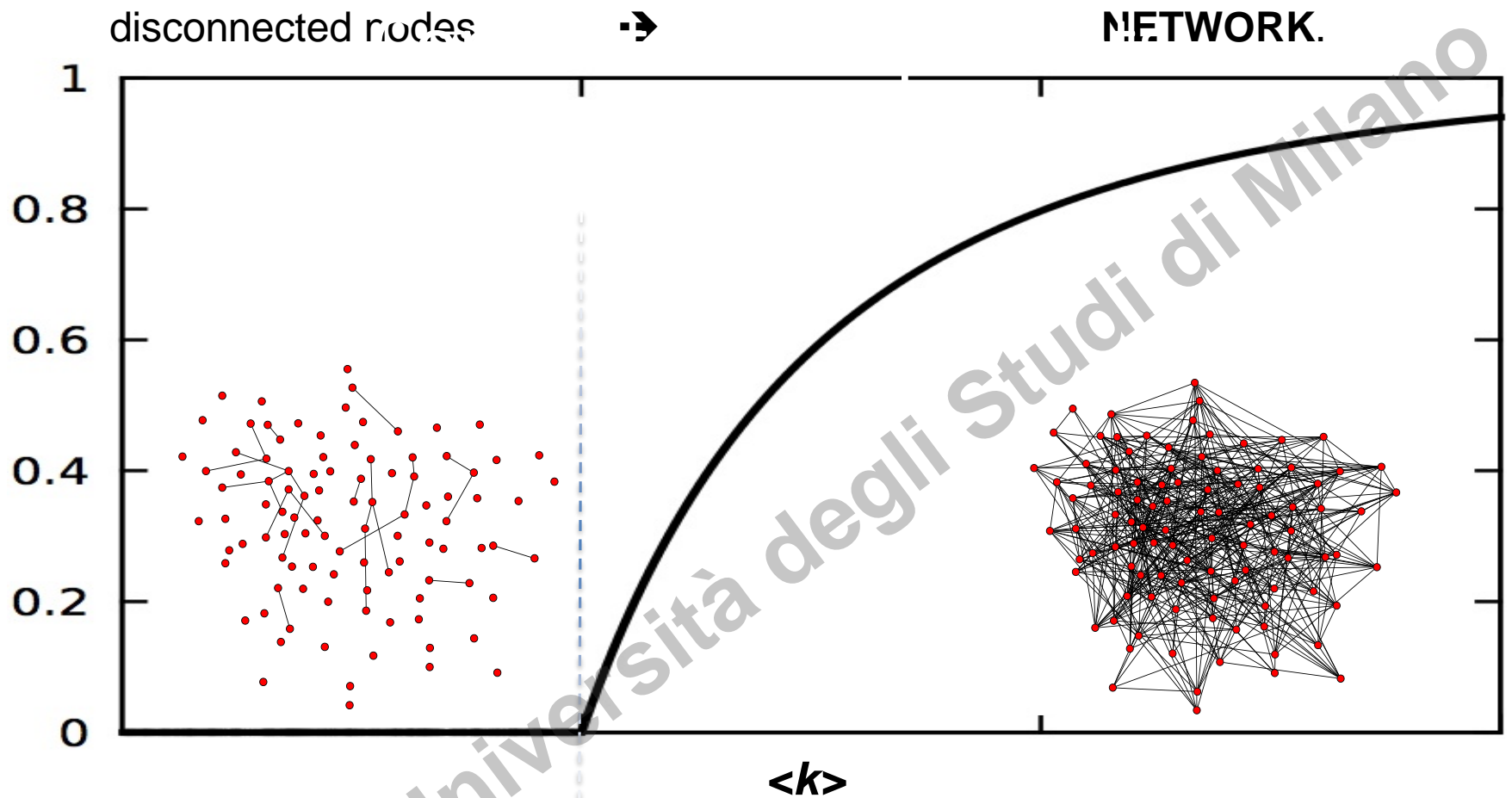
Size of the giant component (3.14)

$$S = 1 - e^{-\langle k \rangle S} \quad (3.32)$$

$$S = N_G/N$$

S: fraction of nodes in the largest connected components



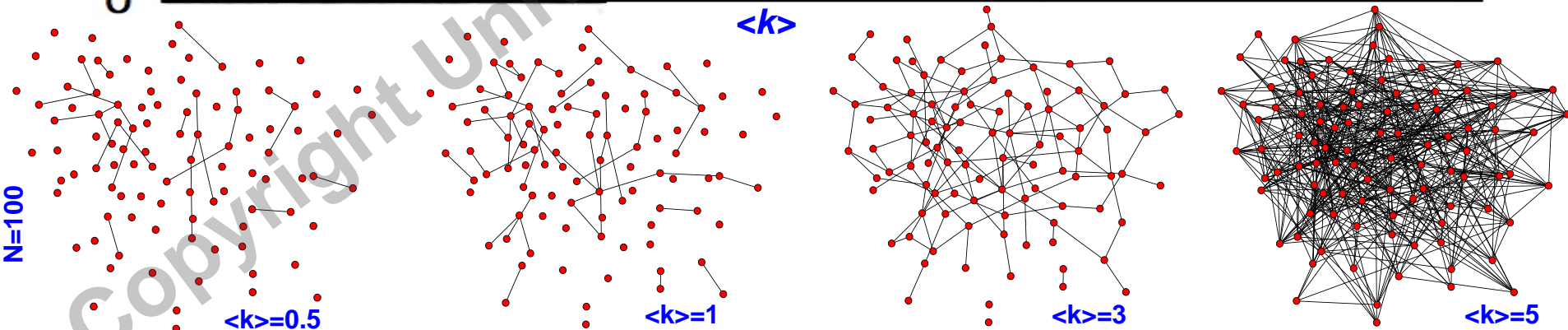
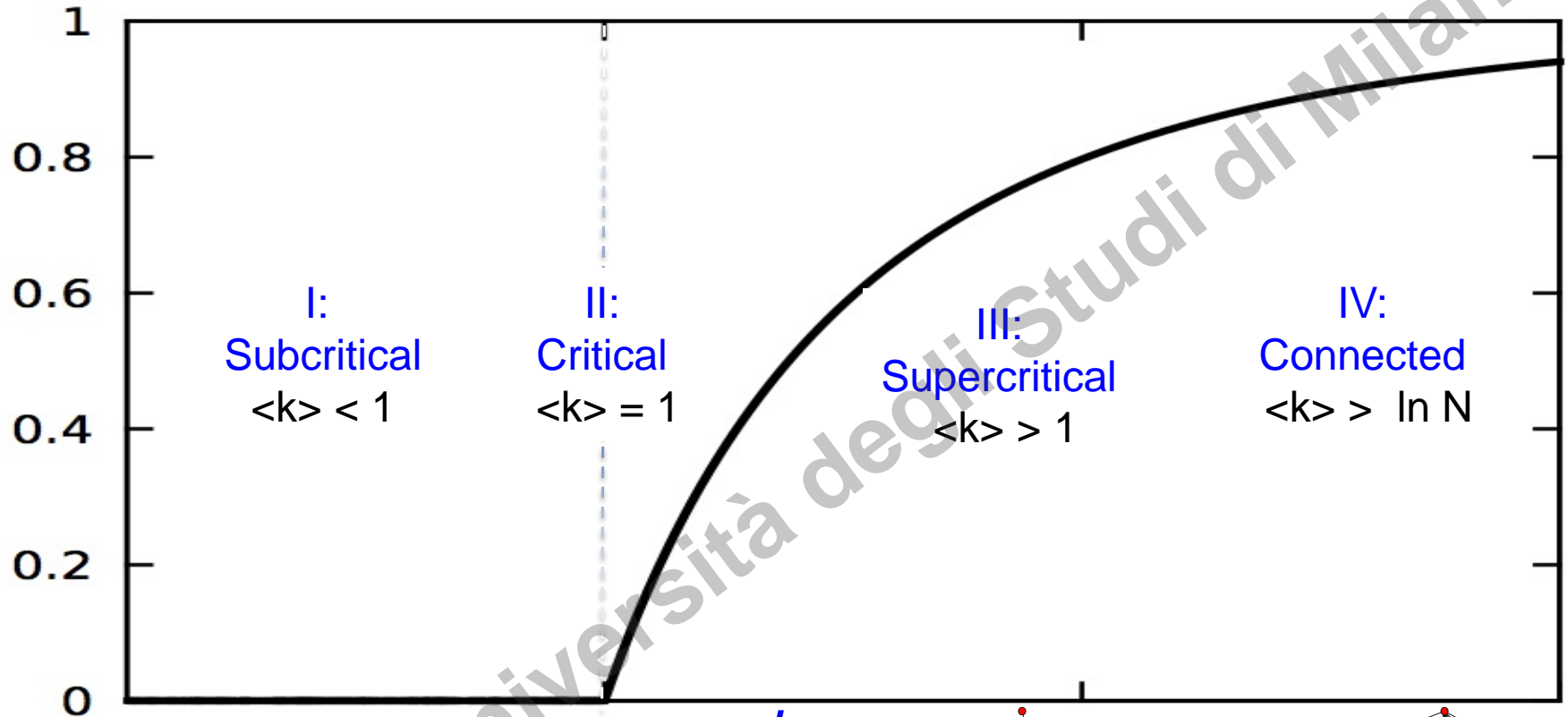


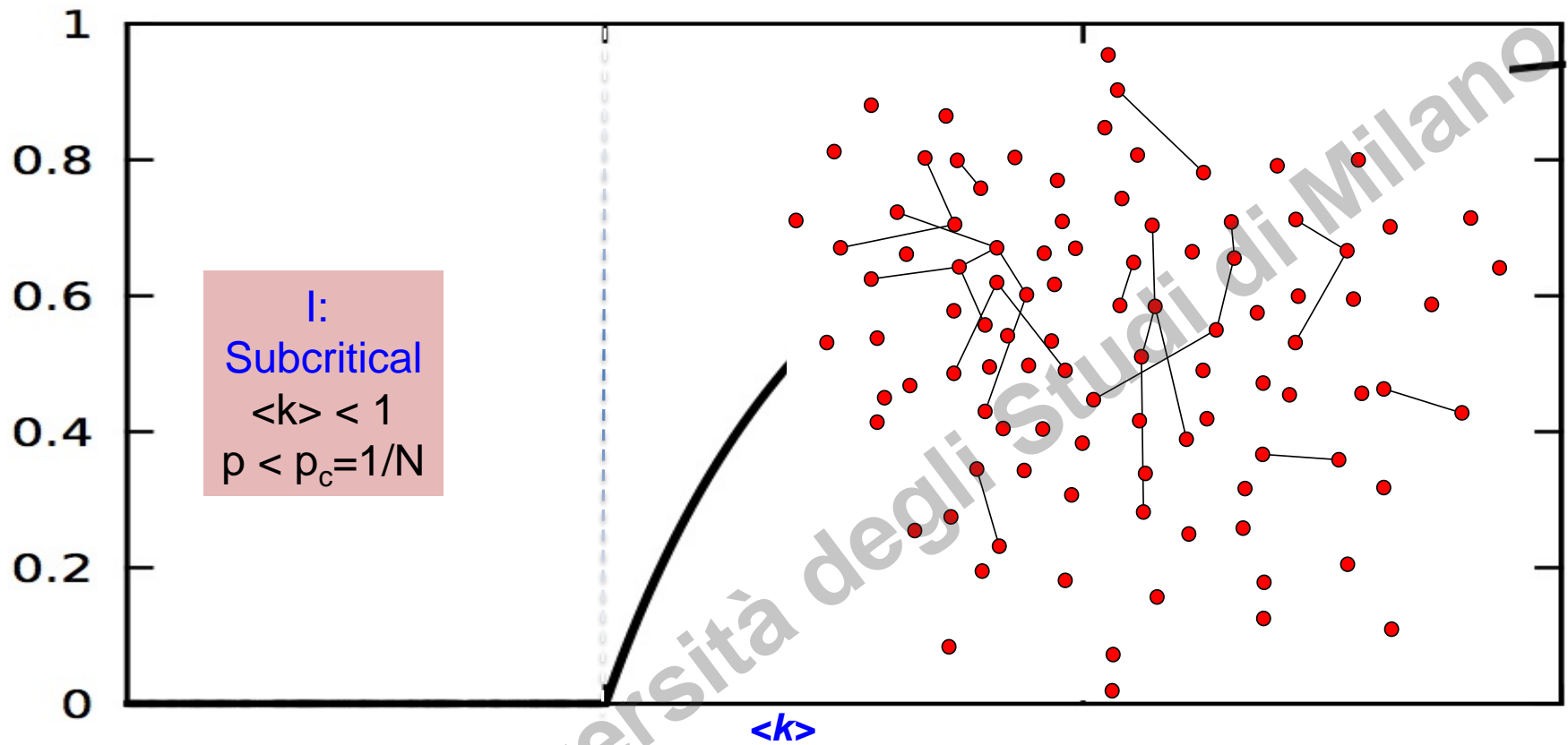
Erdos and Renyi (1959): the condition for the emergence of a giant component is $\langle k \rangle = 1$.

It is evident that one link per node is necessary, but counterintuitive that it is also sufficient.



Four distinct regimes





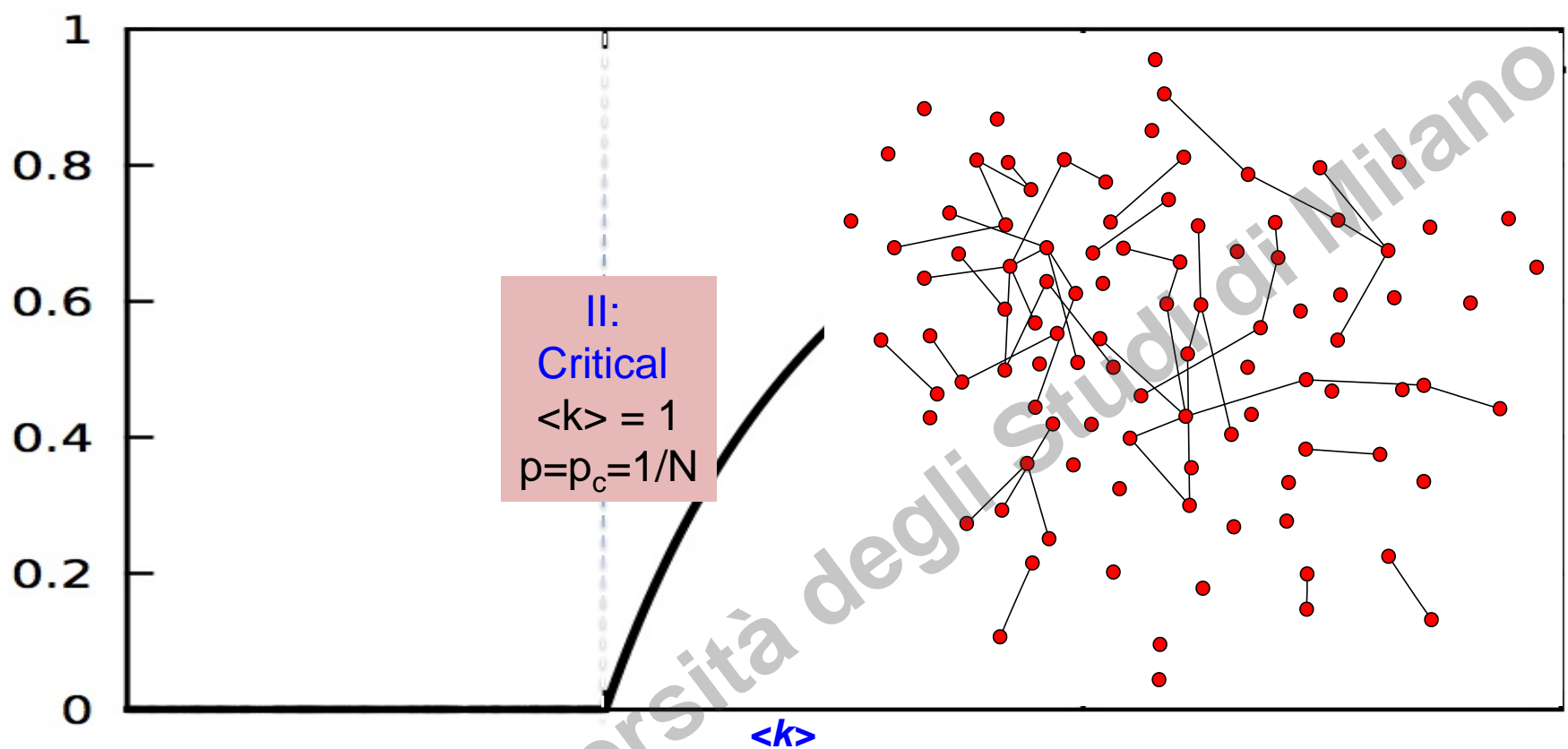
The network consists of numerous tiny components, whose size follows the exponential distribution. Hence these components have comparable sizes, lacking a clear winner that we could designate as a giant component.

No giant component.

Isolated clusters, cluster size distribution is exponential

The largest cluster is a tree, its size $\sim \ln N$. Hence N_G/N is vanishing





At this point the relative size of the largest component is still zero

Unique giant component: $N_G \sim N^{2/3}$

→ contains still a vanishing fraction of all nodes, $N_G/N \sim N^{-1/3}$

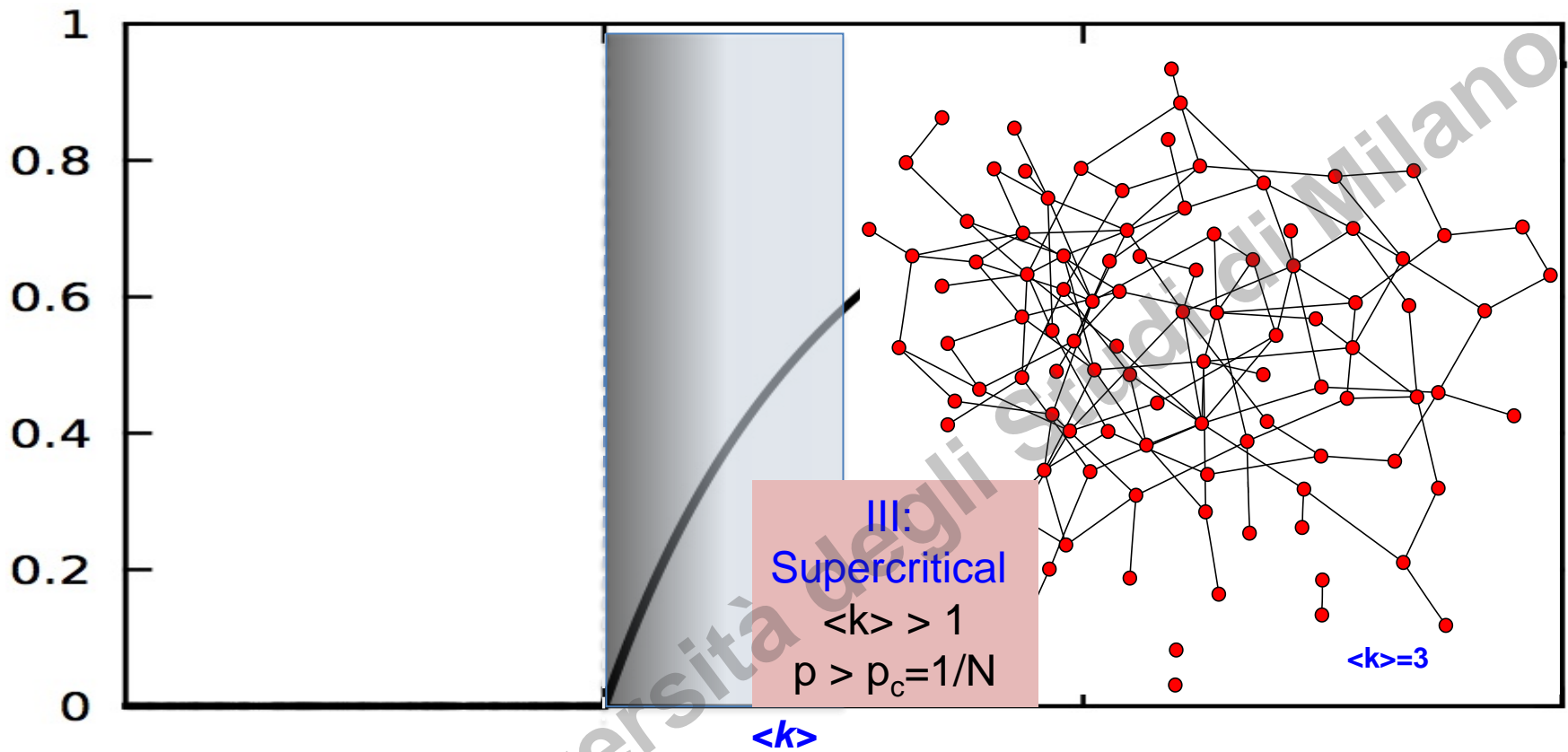
→ Numerous small components which are trees.

Cluster size distribution: $p(s) \sim s^{-3/2}$

A jump in the cluster size:

$N = 7 \cdot 10^9 \rightarrow \ln N \sim 22$; $N^{2/3} \sim 3,659,250$





The giant component contains a finite fraction of the nodes.

Unique giant component: $N_G \sim (p - p_c)N$

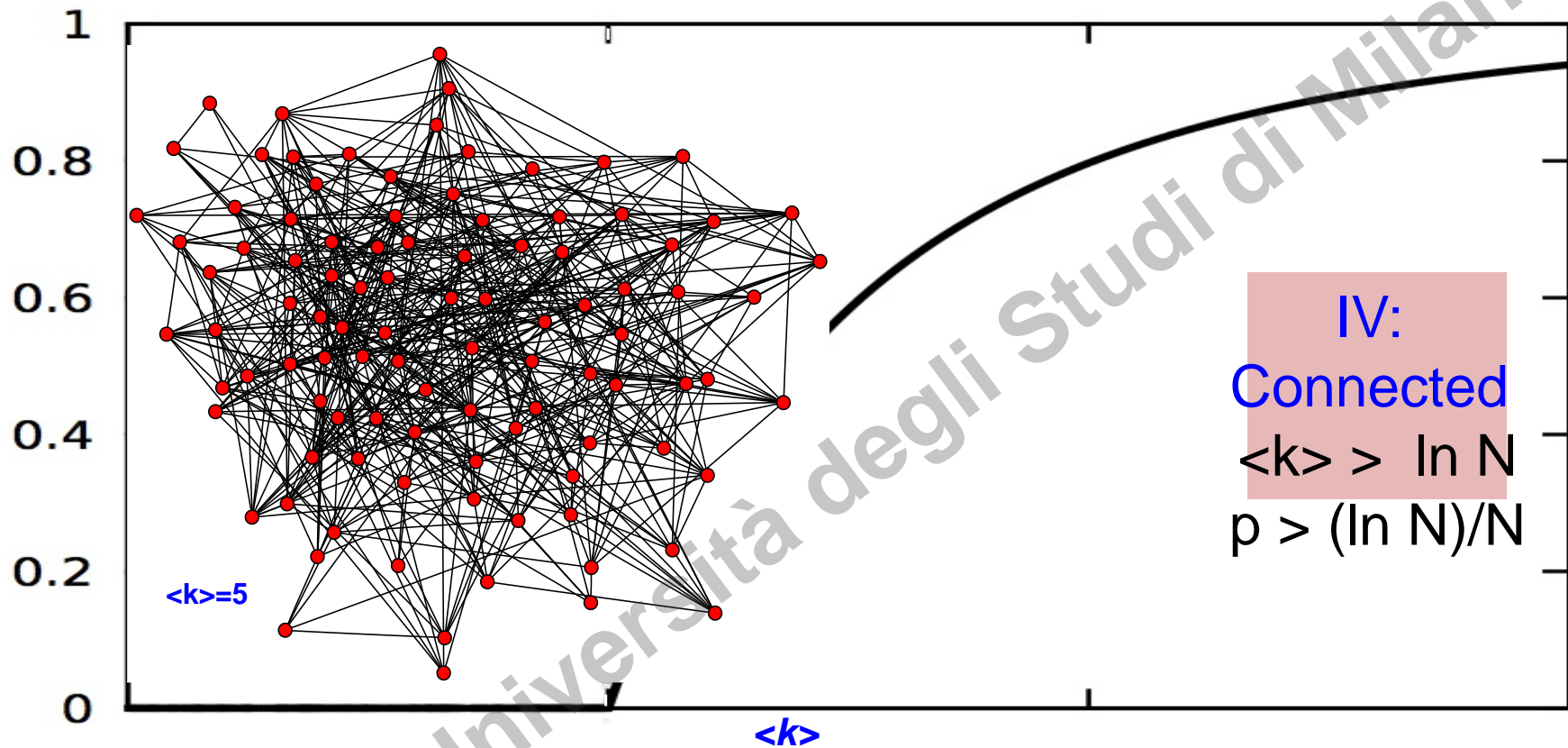
→ Non vanishing

Cluster size distribution: exponential

$$p(s) \sim s^{-3/2} e^{-\langle k \rangle s + (s-1) \ln \langle k \rangle}$$

The supercritical regime lasts until all nodes are absorbed by the giant component.





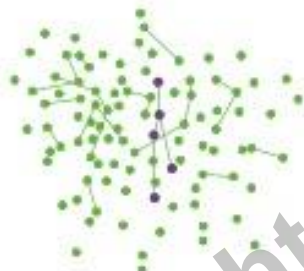
Only one cluster: $N_G = N$

→ GC is dense.

Cluster size distribution: None



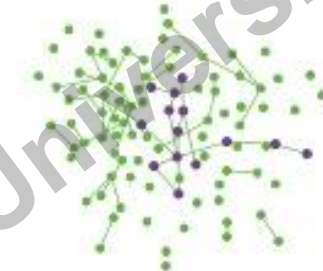
Summary



$\langle k \rangle < 1$

(b) Subcritical Regime

- No giant component
- Cluster size distribution: $p_s \sim s^{-2} e^{-s}$
- Size of the largest cluster: $N_g \sim \ln N$
- The clusters are trees



$\langle k \rangle = 1$

(c) Critical Point

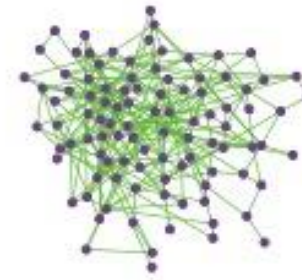
- No giant component
- Cluster size distribution: $p_s \sim s^{-3/2}$
- Size of the largest cluster: $N_g \sim N^{1/2}$
- The clusters may contain loops



$\langle k \rangle > 1$

(d) Supercritical Regime

- Single giant component
- Cluster size distribution: $p_s \sim s^{-2} e^{-s}$
- Size of the giant component: $N_g \sim (p - p_c)N$
- The small clusters are trees
- Giant component has loops



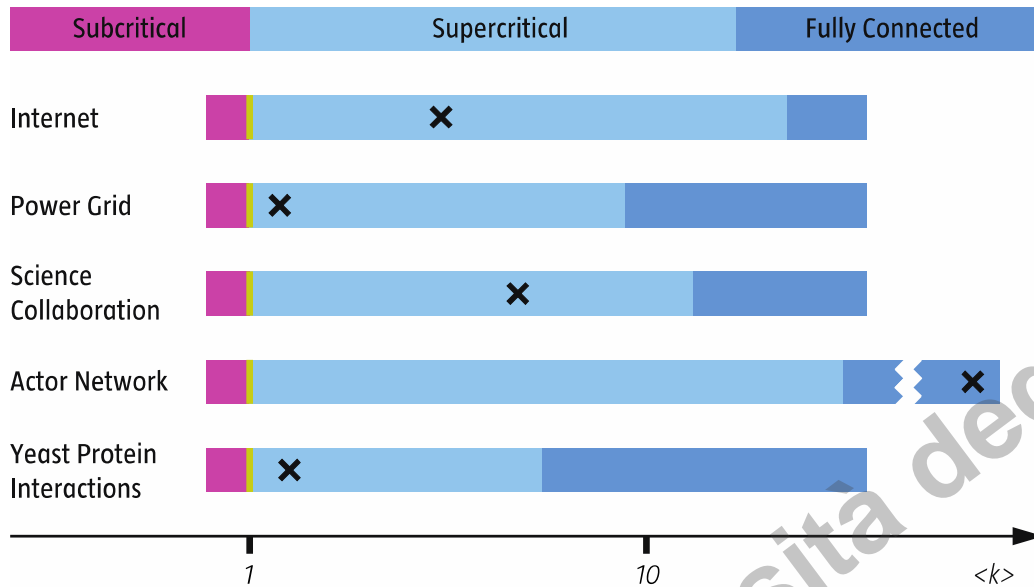
$\langle k \rangle \geq \ln N$

(e) Connected Regime

- Single giant component
- No isolated nodes or clusters
- Size of the giant component: $N_g = N$
- Giant component has loops



Connected components



Network	N	L	$\langle k \rangle$	$\ln N$
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	186,936	8.08	10.04
Actor Network	212,250	3,054,278	28.78	12.27
Yeast Protein Interactions	2,018	2,930	2.90	7.61

Supercritical: not fully connected

Internet: we should have routers that, being disconnected from the giant component, are unable to communicate with other routers.

Power grid: some consumers should not get powered

Fully connected

Social media: no individual disconnected



Albert-László Barabási
Network Science
Chapter 3.6

Newman, M.E.J.
Networks: An Introduction.
Oxford University Press. 2010.
Chapter 12.5

Reza Zafarani, Mohammad Ali Abbasi, Huan Liu
Social Media Mining: An Introduction
A Textbook by Cambridge University Press





UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Connected components

Examples



Facebook

- Ugander, Johan & Karrer, Brian & Backstrom, Lars & Marlow, Cameron. (2011). The Anatomy of the Facebook Social Graph. arXiv preprint. 1111.4503.
 - Connected components
Pag 3, Figure 1



Twitter

- Seth A. Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. 2014. Information network or social network?: the structure of the twitter follow graph. In Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion). ACM, New York, NY, USA, 493-498. DOI: <https://doi.org/10.1145/2567948.2576939>
 - Connected components
Chapter 3.2, Figure 2



Web

- Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the Web. *Comput. Netw.* 33, 1-6 (June 2000), 309-320.
DOI=[http://dx.doi.org/10.1016/S1389-1286\(00\)00083-9](http://dx.doi.org/10.1016/S1389-1286(00)00083-9)
 - Connected components
Chapter 2.2.2, 2.2.3, Figure 5,6



Mobile communication networks

- Structure and tie strengths in mobile communication networks,
J. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.
-L. Barabási

Proceedings of the National Academy of Sciences May
2007, 104 (18) 7332- 7336; DOI: 10.1073/pnas.0610245104

- Weakly largest connected component: 84%

- Calling, texting, and moving: multidimensional interactions of mobile
phone users

Matteo Zignani, Christian Quadri, Sabrina Gaito & Gian Paolo Rossi
Computational Social Networks volume 2, Article number: 13 (2015)

- Weakly largest connected component: 90%





UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Centrality measures



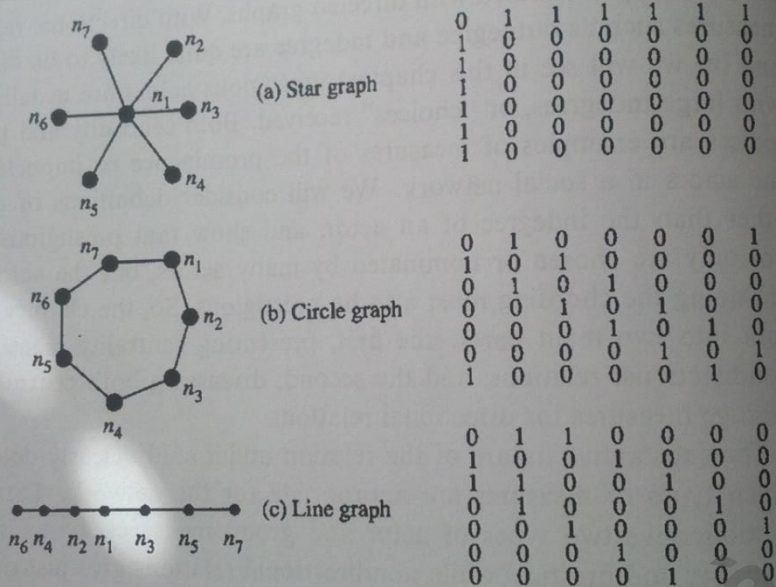


Fig. 5.1. Three illustrative networks for the study of centrality and prestige

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Three artificial graphs that highlight the differences among centrality measures.

One node in the star completely outranks the others, while the other themselves are interchangeable.

All nodes in the circle are interchangeable.

In the line graph centrality decreases from that for n_1 , to n_2 and n_3 , and so on up to n_6 and n_7 .

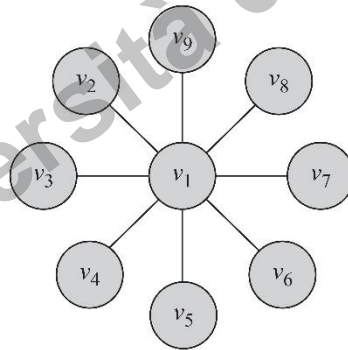
Degree centrality

Degree Centrality in undirected networks

- The degree centrality ranks nodes with more connections higher in terms of centrality

$$C_d(v_i) = d_i$$

- d_i is the degree (number of adjacent edges) for vertex v_i



In this graph degree centrality for vertex v_1 is $d_1 = 8$ and for all others is $d_j = 1, j \neq 1$



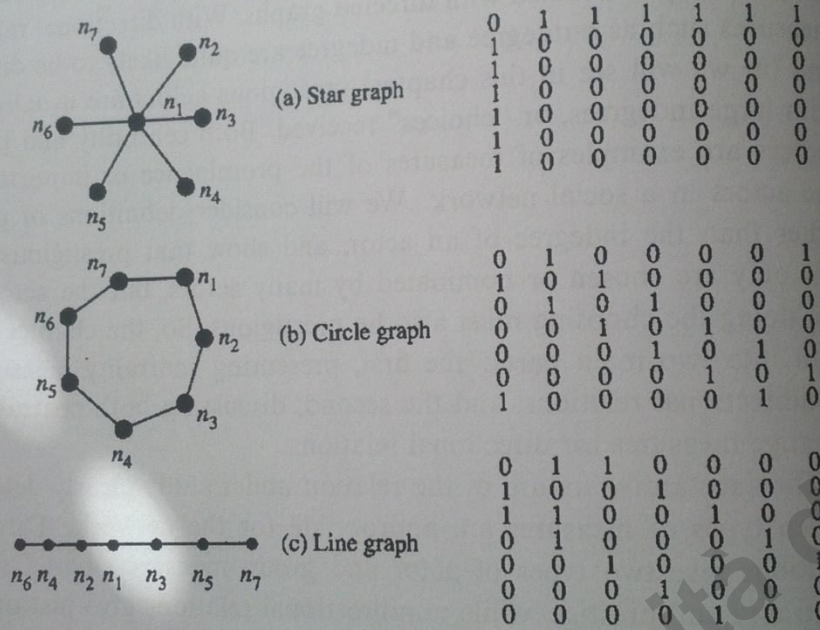


Fig. 5.1. Three illustrative networks for the study of centrality and prestige

177
is,
ac
o
“
v
c

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Star: $C_d(1) = 6, C_d(\text{other nodes}) = 0$

Circle: $C_d(\text{all nodes}) = 2$

Line: $C_2(1) = 2, C_b(2,3) = 2, C_b(4,5) = 2, C_b(6,7) = 1$

Normalized Degree Centrality

The degree centrality does not allow for centrality values to be compared across networks.

- Normalized by the maximum possible degree

$$C_d^{norm}(v_i) = \frac{d_i}{n - 1}$$

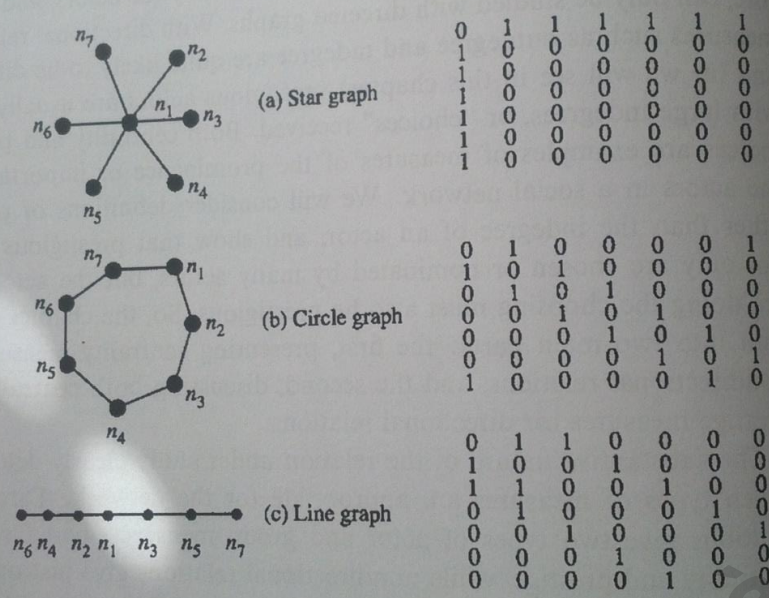
- Normalized by the maximum degree. Issue: outlier

$$C_d^{max}(v_i) = \frac{d_i}{\max_j d_j}$$

- Normalized by the degree sum

$$C_d^{sum}(v_i) = \frac{d_i}{\sum_j d_j} = \frac{d_i}{2|E|}$$





0	1	1	1	1	1	1	1
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0

0	1	0	0	0	0	1
1	0	1	0	0	0	0
0	1	0	1	0	0	0
0	0	1	0	1	0	0
0	0	0	1	0	1	0
0	0	0	0	1	0	1
1	0	0	0	0	1	0

0	1	1	0	0	0	0
1	0	0	1	0	0	0
1	1	0	0	1	0	0
0	1	0	0	0	1	0
0	0	1	0	0	0	1
0	0	0	1	0	0	0
0	0	0	0	1	0	0

Fig. 5.1. Three illustrative networks for the study of centrality and prestige

$$C_d^{norm}(v_i) = \frac{d_i}{n-1}$$

$$C_d^{max}(v_i) = \frac{d_i}{\max_j d_j}$$

$$C_d^{sum}(v_i) = \frac{d_i}{\sum_j d_j} = \frac{d_i}{2|E|}$$

	STAR		CIRCLE	LINE			
	Node 1	Others	All	Node 1	Nodes 2,3	Nodes 4,5	Nodes 6,7
C ^{degree}	6	1	2	2	2	2	1
C ^{norm}	6/6=1	1/6	2/6	2/6	2/6	2/6	1/6
C ^{max}	6/6=1	1/6	2/2=1	2/2	2/2	2/2	1/2
C ^{sum}	6/12	1/12	2/14	2/12	2/12	2/12	1/12

Degree Centrality in Directed Graphs

- In directed graphs, we can either use the in-degree, the out-degree, or the combination as the degree centrality value:

$$C_d(v_i) = d_i^{in} \quad (\textit{prestige}),$$

$$C_d(v_i) = d_i^{out} \quad (\textit{gregariousness}),$$

$$C_d(v_i) = d_i^{in} + d_i^{out}.$$

d_i^{out} is the number of outgoing links for vertex v_i



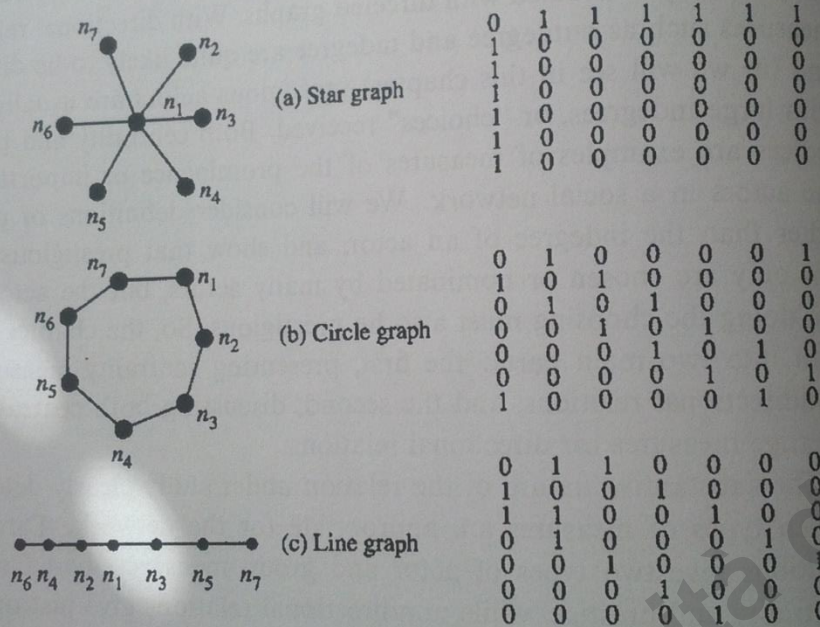


Fig. 5.1. Three illustrative networks for the study of centrality and prestige

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Can the degree centrality fully represent the different aspects of the concept of centrality?

What about nodes n_1 and n_4 in the line graph having the same centrality?

Betweenness centrality

Undirected and directed
networks

Idea

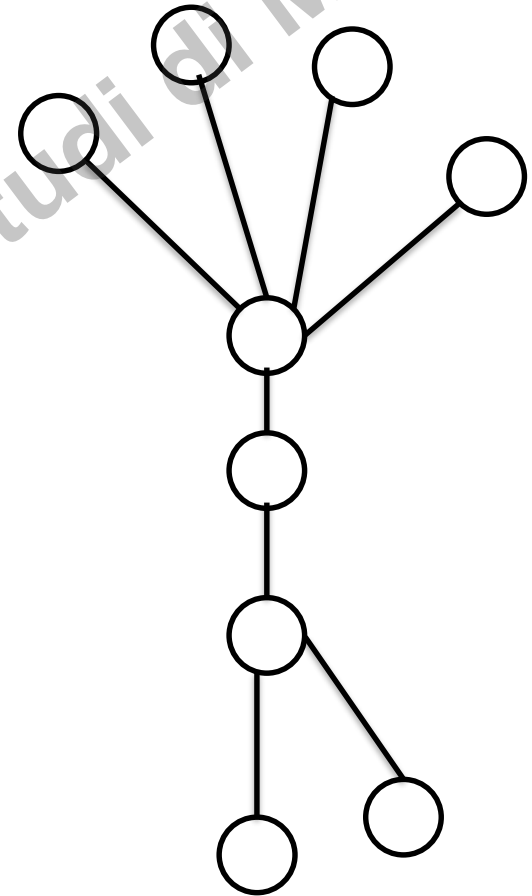
To measure the extent to which a node lies on paths between other nodes, i.e. how much a node falls between others, while the degree centrality measures how well-connected a node is.

Nodes with a high betweenness centrality have control over information flowing in the network.

Example: Internet

Example: Organization networks

Example: Grid networks



Assumptions

Consider information, rumours, news, etc. flowing within a network as they are passed from one person to another.

Let's assume that:

- All pair of nodes (connected by a path) exchange the same amount of information per time unit
- All information flow on the shortest paths

Asymptotically,

how many information will pass through each node?



Note

In the next 3 slides the definition of betweenness centrality and the results of computation on the star, circle and line graphs are presented.

For a step by step lesson please refer to the pdf and mp4 files named Betweenness



Betweenness Centrality

We define the betweenness centrality as:

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

$\sigma_{st}(v_i)$ the number of shortest paths from s to t that pass through v_i

σ_{st} the number of shortest paths from vertex s to t – also known as information pathways.

Note that the path from s to t is different from the path from t to s , even in undirected networks.

The definition holds for both undirected and directed networks (rarely used in directed networks)



Normalizing Betweenness Centrality

- In the best case, node v_i is on all shortest paths from s to t , hence,

$$\frac{\sigma_{st}(v_i)}{\sigma_{st}} = 1$$

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}} = \sum_{s \neq t \neq v_i} 1 = 2 \binom{n-1}{2} = (n-1)(n-2).$$

Therefore, the maximum value is $2 \binom{n-1}{2}$

Normalized betweenness centrality:

$$C_b^{\text{norm}}(v_i) = \frac{C_b(v_i)}{2 \binom{n-1}{2}}.$$



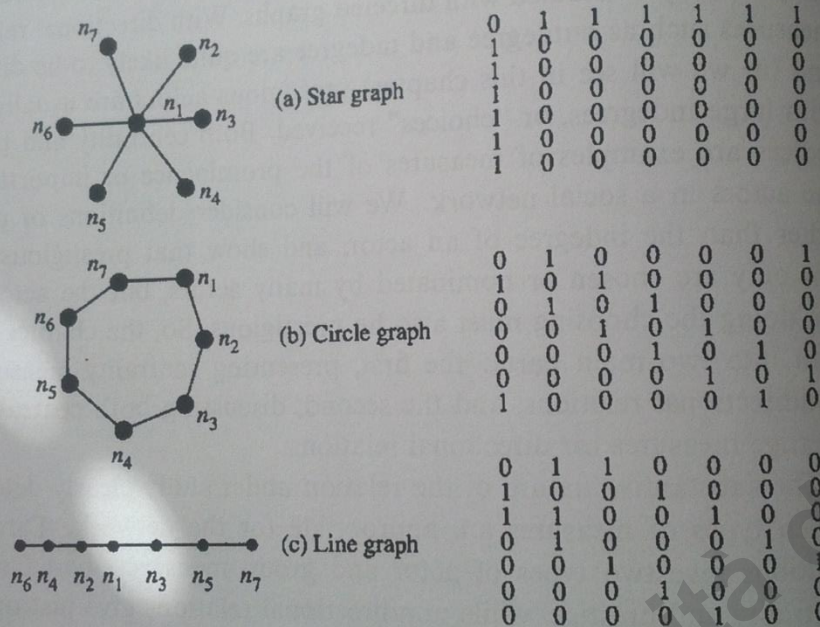


Fig. 5.1. Three illustrative networks for the study of centrality and prestige

Examples

See files named Examples_betweenness for a step by step solution

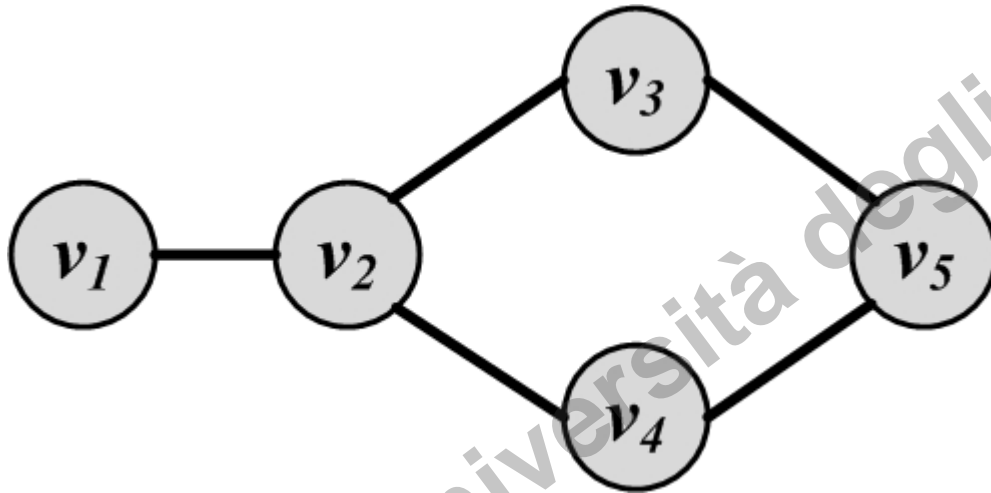
Star: $C_b(1) = 1, C_b(\text{other nodes}) = 0$

Circle: $C_b(\text{all nodes}) = 1/5$

Line: $C_b(1) = 9/15, C_b(2,3) = 8/15, C_b(4,5) = 5/15, C_b(6,7) = 0$

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Betweenness Centrality Example



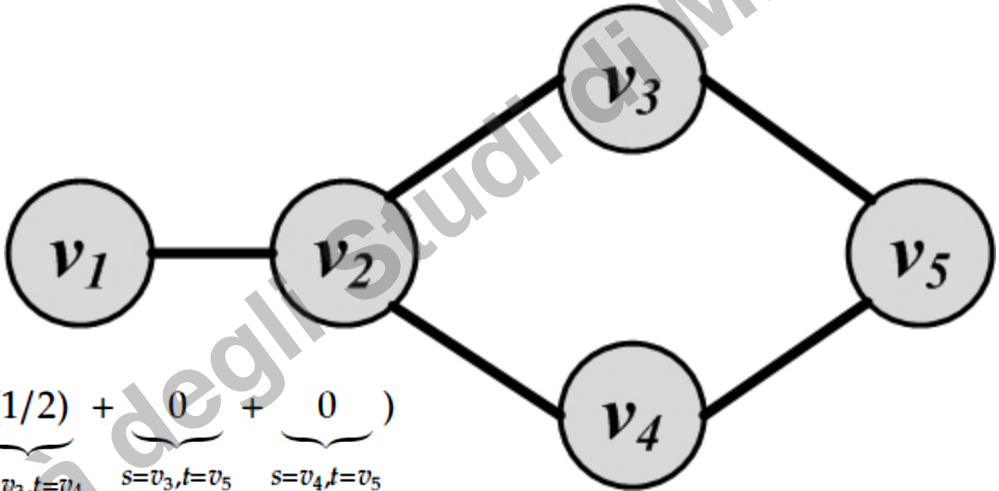
Exercise:

See files

Exercise_betweenness
for a step by step
solution



Betweenness Centrality Example



$$C_b(v_2) = 2 \times \left(\underbrace{(1/1)}_{s=v_1, t=v_3} + \underbrace{(1/1)}_{s=v_1, t=v_4} + \underbrace{(2/2)}_{s=v_1, t=v_5} + \underbrace{(1/2)}_{s=v_3, t=v_4} + \underbrace{0}_{s=v_3, t=v_5} + \underbrace{0}_{s=v_4, t=v_5} \right)$$

$$= 2 \times 3.5 = 7,$$

$$C_b(v_3) = 2 \times \left(\underbrace{0}_{s=v_1, t=v_2} + \underbrace{0}_{s=v_1, t=v_4} + \underbrace{(1/2)}_{s=v_1, t=v_5} + \underbrace{0}_{s=v_2, t=v_4} + \underbrace{(1/2)}_{s=v_2, t=v_5} + \underbrace{0}_{s=v_4, t=v_5} \right)$$

$$= 2 \times 1.0 = 2,$$

$$C_b(v_4) = C_b(v_3) = 2 \times 1.0 = 2,$$

$$C_b(v_5) = 2 \times \left(\underbrace{0}_{s=v_1, t=v_2} + \underbrace{0}_{s=v_1, t=v_3} + \underbrace{0}_{s=v_1, t=v_4} + \underbrace{0}_{s=v_2, t=v_3} + \underbrace{0}_{s=v_2, t=v_4} + \underbrace{(1/2)}_{s=v_3, t=v_4} \right)$$

$$= 2 \times 0.5 = 1,$$



Closeness centrality

Undirected and directed networks

Closeness Centrality

- The intuition is that influential and central nodes can quickly reach other nodes
- These nodes should have a smaller average shortest path length to other nodes
- We define the closeness centrality as:

$$\bar{l}_{v_i} = \frac{1}{n-1} \sum_{v_j \neq v_i} l_{i,j}$$

$$C_c(v_i) = \frac{1}{\bar{l}_{v_i}}$$

that is node v_i 's average shortest path length to other nodes.



consider the inverse.

But, with this definition:

Low values for more central nodes, high values for less central nodes.

Now:

Low values for less central nodes, high values for more central nodes.

Issues: It holds within components, the range of values is small in small-world networks.



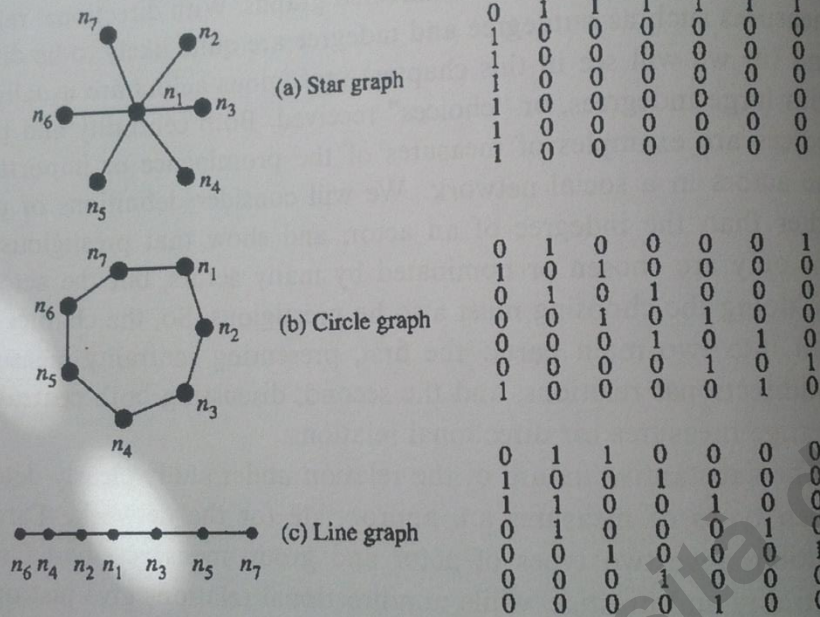


Fig. 5.1. Three illustrative networks for the study of centrality and prestige

Examples

See files named
Examples_closeness

for a step by step
solution

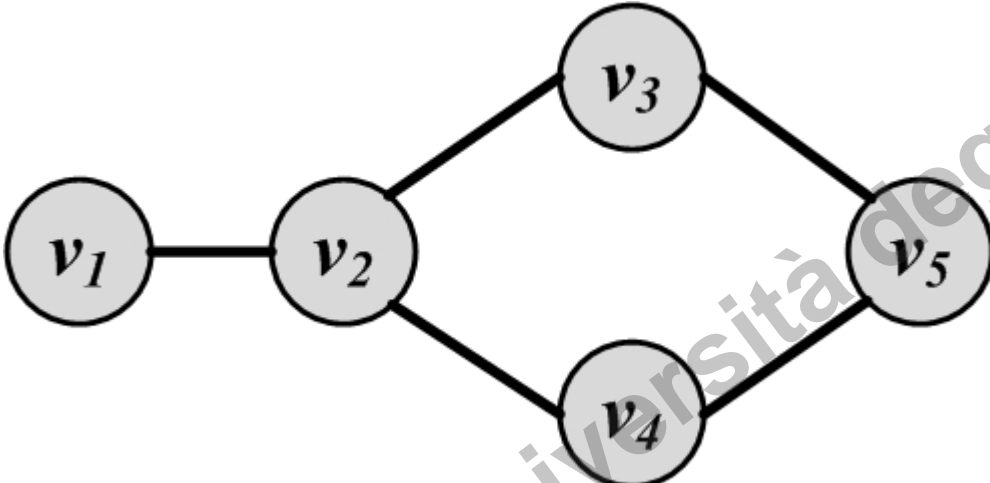
Star: $C_c(1) = 1, C_c(\text{other nodes}) = 6/11$

Circle: $C_c(\text{all nodes}) = 1/2$

Line: $C_c(1) = 1/2, C_c(2,3) = 6/13, C_c(4,5) = 3/8, C_c(6,7) = 6/21$

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Example: Compute Closeness Centrality

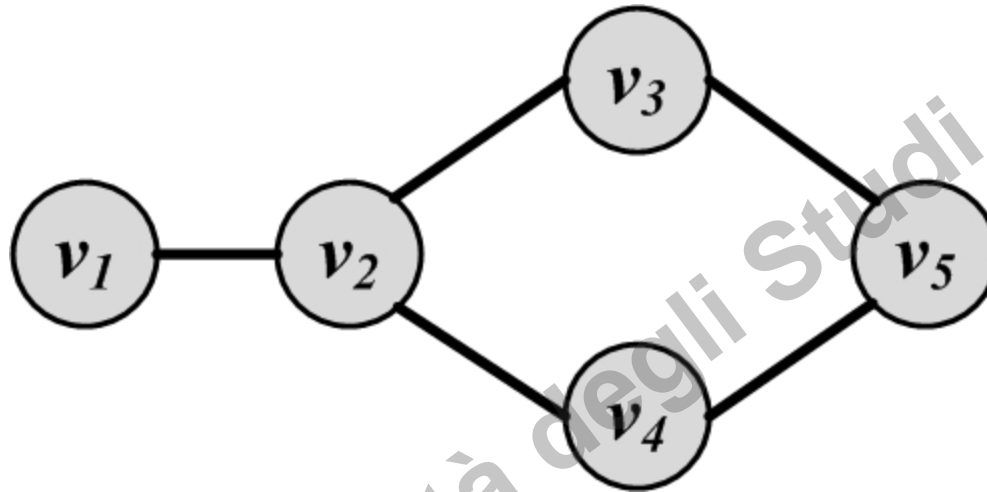


Exercise

See files named Exercise_closeness for a step by step solution



Compute Closeness Centrality



$$C_c(v_1) = 1/((1 + 2 + 2 + 3)/4) = 0.5,$$

$$C_c(v_2) = 1/((1 + 1 + 1 + 2)/4) = 0.8,$$

$$C_c(v_3) = C_b(v_4) = 1/((1 + 1 + 2 + 2)/4) = 0.66,$$

$$C_c(v_5) = 1/((1 + 1 + 2 + 3)/4) = 0.57$$



Eigenvector centrality

Katz centrality

Page rank

(cenni)

Eigenvector Centrality (undirected)

- It is an extension of the degree centrality.
- Not all friends are equivalent. Thus, having more friends does not by itself guarantee that someone is more important, but having more **important friends** provides a stronger signal.
- Eigenvector centrality tries to generalize degree centrality by incorporating the importance of the neighbors.



Eigenvector Centrality (undirected): idea

- Degree centrality: awarding nodes just one point for each friend
- Eigenvector centrality: awarding nodes a score proportional to the sum of the scores of its friends.



Copyright Università degli Studi di Milano



Eigenvector Centrality (undirected)

- Degree centrality: awarding nodes just one point for each friend
- Eigenvector centrality: awarding nodes a score proportional to the sum of the scores of its friends.
- For directed graphs, we can use incoming or outgoing edges

$$c_e(v_i) = \frac{1}{\lambda} \sum_{j=1}^n A_{j,i} c_e(v_j),$$

$c_e(v_i)$: the eigenvector centrality of node v_i
 λ : some fixed constant



Eigenvector Centrality, cont. (undirected)

- Let $\mathbf{C}_e = (C_e(v_1), C_e(v_2), \dots, C_e(v_n^T))$

$$\rightarrow \lambda \mathbf{C}_e = A^T \mathbf{C}_e.$$

- This means that \mathbf{C}_e is an eigenvector of adjacency matrix A and λ is the corresponding eigenvalue
- Which eigenvalue-eigenvector pair should we choose?



Eigenvector Centrality (undirected)

Theorem 3.1 (Perron-Frobenius Theorem). *Let $A \in \mathcal{R}^{n \times n}$ represent the adjacency matrix for a [strongly] connected graph or $A : A_{i,j} > 0$, i.e., a positive n by n matrix. There exists a positive real number (Perron-Frobenius eigenvalue) λ_{max} , such that λ_{max} is an eigenvalue of A and any other eigenvalue of A is strictly smaller than λ_{max} . Furthermore, there exists a corresponding eigenvector $v = (v_1, v_2, \dots, v_n)$ of A with eigenvalue λ_{max} such that $\forall v_i > 0$.*

Therefore, to have positive centrality values, we can compute the eigenvalues of A and then select the largest eigenvalue.

The corresponding eigenvector is \mathbf{C}_e .

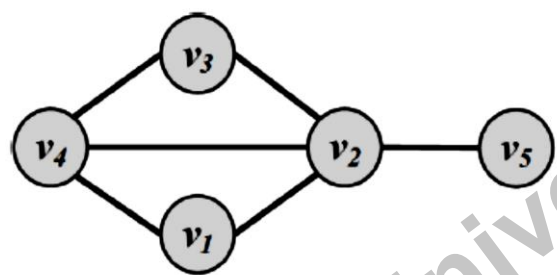
Based on the Perron-Frobenius theorem, all the components of \mathbf{C}_e will be positive, and this vector corresponds to eigenvector centralities for the graph.



Eigenvector Centrality: Example

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$\Rightarrow \lambda = (2.68, -1.74, -1.27, 0.33, 0.00)$
Eigenvalues Vector



$\lambda_{\max} = 2.68 \Rightarrow$

$$C_e = \begin{bmatrix} 0.4119 \\ 0.5825 \\ 0.4119 \\ 0.5237 \\ 0.2169 \end{bmatrix}$$

Based on eigenvector centrality, node v_2 is the most central node.



Eigenvector Centrality (directed)

It can be computed also for undirected networks but some issues arise.

The adjacency matrix is asymmetric \rightarrow it has two sets of eigenvectors, the left eigenvectors and the right eigenvectors.

Which of the two should be used?

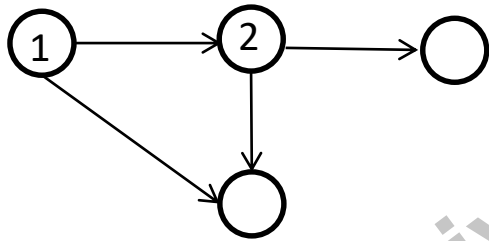
Usually the right eigenvectors because it accounts for the incoming links.



Eigenvector Centrality (directed)

A major problem with eigenvector centrality arises when it deals with directed graphs

Centrality only passes over *outgoing* edges and in special cases such as when a node is in a weakly connected component centrality becomes zero even though the node can have many edge connected to it



Node 1 has only outgoing links and hence has eigenvector centrality zero.

Node 2 has one ingoing link, but it originates at node 1 and hence node B has centrality zero, too.

Mathematically, only nodes in a strongly connected components of two or more nodes can have non-zero eigenvector centrality.



Katz Centrality

- To resolve this problem we add a bias term β to the centrality values for all nodes

$$C_{Katz}(v_i) = \alpha \sum_{j=1}^n A_{j,i} C_{Katz}(v_j) + \beta.$$



Katz Centrality, cont.

$$C_{Katz}(v_i) = \alpha \sum_{j=1}^n A_{j,i} C_{Katz}(v_j) + \beta.$$

Controlling term

Bias term

Rewriting equation in a vector form

$$\mathbf{C}_{Katz} = \alpha \mathbf{A}^T \mathbf{C}_{Katz} + \beta \mathbf{1}$$

vector of all 1's

Katz centrality: $\mathbf{C}_{Katz} = \beta (\mathbf{I} - \alpha \mathbf{A}^T)^{-1} \cdot \mathbf{1}.$



Katz Centrality, cont.

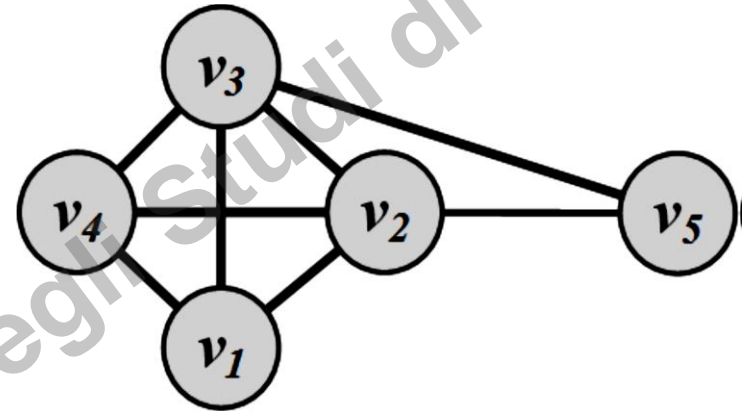
$$\mathbf{C}_{Katz} = \beta(\mathbf{I} - \alpha A^T)^{-1} \cdot \mathbf{1}.$$

- When $\alpha=0$, the eigenvector centrality is removed and all nodes get the same centrality value β
- As α gets larger the effect of β is reduced
- For the matrix $(\mathbf{I} - \alpha A^T)$ to be invertible, we must have
 - $\det(\mathbf{I} - \alpha A^T) \neq 0$
 - By rearranging we get $\det(A^T - \alpha^{-1}\mathbf{I})=0$
 - This is basically the characteristic equation, which first becomes zero when the largest eigenvalue equals α^{-1}
- In practice we select $\alpha < 1/\lambda$, where λ is the largest eigenvalue of A^T



Katz Centrality Example

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} = A^T.$$



- The eigenvalues are $-1.68, -1.0, 0.35, 3.32$
- We assume $\alpha=0.25 < 1/3.32$ $\beta=0.2$

$$C_{Katz} = \beta(\mathbf{I} - \alpha A^T)^{-1} \cdot \mathbf{1} = \begin{bmatrix} 1.14 \\ 1.31 \\ 1.31 \\ 1.14 \\ 0.85 \end{bmatrix}.$$



PageRank

- Problem with Katz centrality: in directed graphs, once a node becomes an authority (high centrality), it passes **all** its centrality along **all** of its out-links
- This is less desirable since not everyone known by a well-known person is well-known
- To mitigate this problem we can divide the value of passed centrality by the number of outgoing links, i.e., out-degree of that node such that each connected neighbor gets a fraction of the source node's centrality



PageRank, cont.

$$C_p(v_i) = \alpha \sum_{j=1}^n A_{ji} \frac{C_p(v_j)}{d_j^{\text{out}}} + \beta.$$

$$\begin{cases} (d_j^{\text{out}} > 0) \\ D = \text{diag}(d_1, d_2, \dots, d_n) \end{cases}$$



$$\mathbf{C}_p = \alpha \mathbf{A}^T \mathbf{D}^{-1} \mathbf{C}_p + \beta \mathbf{1},$$

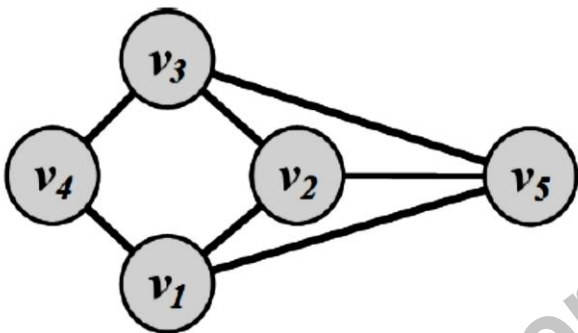


$$\mathbf{C}_p = \beta (\mathbf{I} - \alpha \mathbf{A}^T \mathbf{D}^{-1})^{-1} \cdot \mathbf{1},$$



PageRank Example

- We assume $\alpha=0.95$ and $\beta=0.1$



$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

$$C_p = \beta(\mathbf{I} - \alpha A^T D^{-1})^{-1} \cdot \mathbf{1} = \begin{bmatrix} 2.14 \\ 2.13 \\ 2.14 \\ 1.45 \\ 2.13 \end{bmatrix}.$$



Credits

Reza Zafarani
Social Media Mining
Chapter 2

M.E.J. Newman
Networks
An Introduction
Oxford university press



15_a_Exercise_eigenvector_centralty

domenica 3 maggio 2020 17:44



$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$\lambda I = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}$$

$$A - \lambda I = \begin{bmatrix} -\lambda & 1 & 0 \\ 1 & -\lambda & 1 \\ 0 & 1 & -\lambda \end{bmatrix}$$

$$(A - \lambda I) \underline{c}_e = 0 \quad \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

$$\begin{bmatrix} -\lambda & 1 & 0 \\ 1 & -\lambda & 1 \\ 0 & 1 & -\lambda \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0$$

$$\det(A - \lambda I) = 0$$

$$\det(A - \lambda I) = -\lambda(\lambda^2 - 1) - 1(-\lambda) = -\lambda^3 + 2\lambda$$

$$\det(A - \lambda I) = 0 \Rightarrow \lambda = 0, \lambda = -\sqrt{2}, \lambda = \sqrt{2}$$

$$(A - \lambda I) \underline{c}_e = 0$$

$$\begin{bmatrix} -\sqrt{2} & 1 & 0 \\ 1 & \sqrt{2} & 1 \\ 0 & 1 & -\sqrt{2} \end{bmatrix} \begin{bmatrix} c_e(1) \\ c_e(2) \\ c_e(3) \end{bmatrix} = 0 \quad (c_e(1))^2 + (c_e(2))^2 + (c_e(3))^2 = 1$$

$$\begin{cases} -\sqrt{2} c_e(1) + c_e(2) = 0 \\ c_e(1) - \sqrt{2} c_e(2) + c_e(3) = 0 \\ c_e(2) - \sqrt{2} c_e(3) = 0 \\ c_e^2(1) + c_e^2(2) + c_e^2(3) = 1 \end{cases}$$

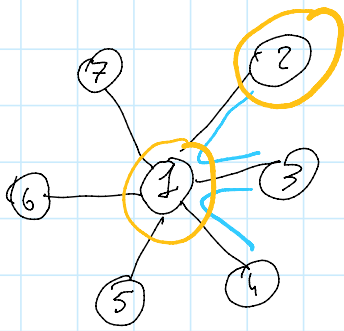
$$\begin{cases} c_e(1) = c_e(3) = \frac{1}{2} \\ c_e(2) = \frac{\sqrt{2}}{2} \end{cases}$$



Examples_betweenness centrality

sabato 2 maggio 2020 18:00

● STAR GRAPH



$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

$\sigma_{st}(v_i)$: number of paths from s to t that pass through v_i

σ_{st} : number of paths from s to t

$$C_b^{\text{norm}}(v_i) = \frac{C_b(v_i)}{(n-1)(n-2)}$$

Node 1

list the node pairs (s, t)

2 3 **1/1 = 1**
 2 4 **1**
 2 5 **1**
 2 6 **1**
 2 7 **1**

3 2 **1**
 3 4 **1**
 3 5 **1**
 3 6 **1**
 3 7 **1**

4 2 **1**
 4 3 **1**
 4 5 **1**
 4 6 **1**
 4 7 **1**

5 2 **1**
 5 3 **1**
 5 4 **1**
 5 6 **1**
 5 7 **1**

6 2 **1**
 6 3 **1**
 6 4 **1**
 6 5 **1**
 6 7 **1**

7 2 **1**
 7 3 **1**
 7 4 **1**
 7 5 **1**
 7 6 **1**

$$C_b(1) = \sum_{s \neq t \neq 1} 1 = 30$$

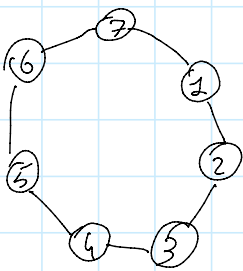
$$C_b^{\text{norm}}(1) = \frac{30}{(7-1)(7-2)} = \frac{30}{6 \cdot 5} = 1$$

$$C_b^{\text{norm}}(1) = \frac{30}{(n-1)(n-2)} = \frac{30}{30} = 1$$

$$C_b(2) = C_b(3) = C_b(4) = C_b(5) = C_b(6) = C_b(7) = 0$$



CIRCLE GRAPH



$$C_b(1)$$

2 3	0
2 4	0
2 5	0
2 6	1/1:1
2 7	1

3 2	0
3 4	0
3 5	0
3 6	0
3 7	1/1

4 2	0
4 3	0
4 5	0
4 6	0
4 7	0

5 2	0
5 3	0
5 4	0
5 6	0
5 7	0

6 2	1/1:1
6 3	0
6 4	0
6 5	0
6 7	0

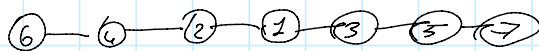
7 2	1/1:1
7 3	1/1:1
7 4	0
7 5	0
7 6	0

$$C_b(1) = 6$$

$$C_b^{\text{norm}}(1) = \frac{6}{30} = \frac{1}{5}$$



LINE GRAPH



Node 1

2 3	1
2 4	0
2 5	1
2 6	0
2 7	1

3 2	1
3 4	1
3 5	0
3 6	1
3 7	0

4 2	0
4 3	1
4 5	1
4 6	0
4 7	1

5 2	1
5 3	0
5 4	1
5 6	1
5 7	0

2	6	0
2	7	1

3	6	1
3	7	0

4	6	0
4	7	1

5	6	1
5	7	0

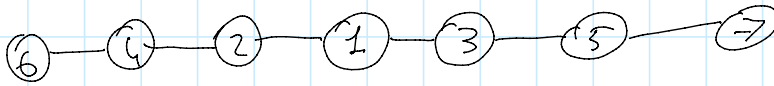
6	2	0
6	3	1
6	4	0
6	5	1
6	7	1

7	2	1
7	3	0
7	4	1
7	5	0
7	6	1

$$C_b(1) = 18$$

$$C_b^{norm}(1) = \frac{18}{30} = \frac{3}{5}$$

Node 2



1	3	0
1	4	1
1	5	0
1	6	1
1	7	0

3	1	0
3	4	1
3	5	0
3	6	1
3	7	0

4	1	1
4	3	1
4	5	1
4	6	0
4	7	1

5	1	0
5	3	0
5	4	1
5	6	1
5	7	0

6	1	1
6	3	1
6	4	0
6	5	1
6	7	1

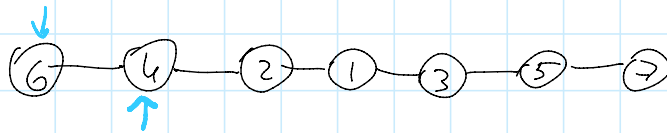
7	1	0
7	3	0
7	4	1
7	5	0
7	6	1

$$C_b(2) = 16$$

$$C_b^{norm}(2) = \frac{16}{30} = \frac{8}{15}$$

$$C_b^{norm}(3) = C_b^{norm}(2) = \frac{8}{15}$$

Node 4



1	2	0
1	3	0
1	5	0
1	6	1
1	7	0

2	1	0
2	3	0
2	5	0
2	6	1
2	7	0

3	1	0
3	2	0
3	5	0
3	6	1
3	7	0

5	1	0
5	2	0
5	3	0
5	6	1
5	7	0

1 7 0

2 7 0

3 7 0

5 7 6

6 1 1

6 2 1

6 3 1

6 5 1

6 7 1

7 1 0

7 2 0

7 3 0

7 5 6

7 6 1

$$C_b(4) = 10$$

$$C_b^{\text{norm}}(4) = \frac{10}{30} = \frac{5}{15} = \frac{1}{3}$$

$$C_b^{\text{norm}}(5) = C_b^{\text{norm}}(4) = \frac{1}{3}$$

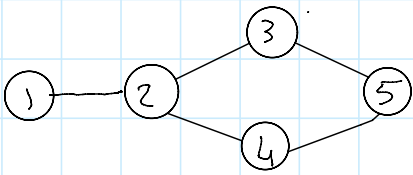
Node
6

$$C_b^{\text{norm}}(6) = C_b^{\text{norm}}(7) = 0$$



Exercise_betweenness centrality

sabato 2 maggio 2020 18:46



$$C_b^{norm}(v_i) = \frac{\sum_{s \neq t \neq v_i} \sigma_{st}(v_i)}{(n-1)(n-2)}$$

N.B. $n=5 \Rightarrow (n-1)(n-2) = 4 \cdot 3 = 12$

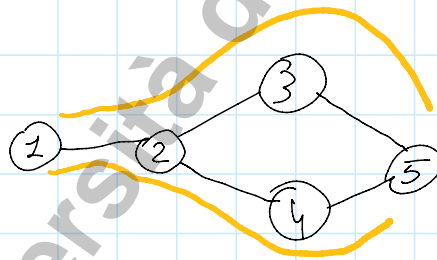
Node 1: $C_b(1) = 0$

Node 2: $C_b(2) = 7 \Rightarrow C_b^{norm}(2) = \frac{7}{4 \cdot 3} = \frac{7}{12}$

1	3	1 1 = 1
1	4	1 1 = 1
1	5	2 2 = 1
3	4	1 2 = 1 2
3	5	0
4	5	0

+ reciprocal paths

Node 3:

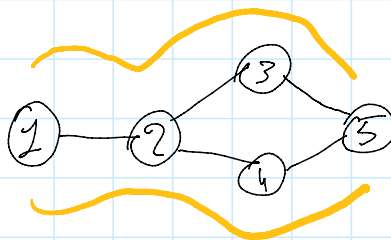


1	2	0
1	4	0
1	5	1 2
2	4	0
2	5	1 2
4	5	0

+ reciprocal

$C_b(3) = 2 \Rightarrow C_b^{norm}(3) = \frac{2}{4 \cdot 3} = \frac{1}{6}$

Node 4:



1	2	0
1	3	0
1	5	1 2
2	3	0
2	5	1 2
3	5	0

$C_b(4) = 2 \Rightarrow C_b^{norm}(4) = \frac{2}{12} + \text{reciprocal}$

$$\begin{array}{cc|c} 2 & 5 & 1|2 \\ 3 & 5 & 0 \end{array}$$

$$C_b(4) = 2 \quad C_b(4) = \frac{2}{12} + \text{reciprocal}$$

Node 5:

$$\begin{array}{cc|c} 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 4 & 0 \\ 2 & 3 & 0 \\ 2 & 4 & 0 \\ 3 & 4 & 1|2 \end{array}$$

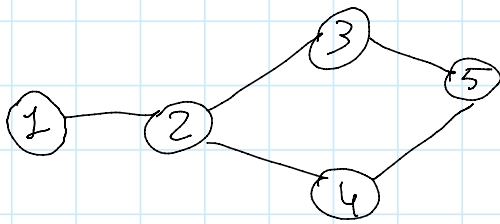
$$C_b(5) = 1 \quad C_b^{(nom)} = \frac{1}{12} + \text{reciprocal}$$



Copyright Università degli Studi di Milano

15_b_Exercise_closeness centrality

domenica 3 maggio 2020 11:16



$$C_c(1) = \left[\frac{1}{4} (1+2+2+3) \right]^{-1} = \frac{4}{8}$$

$$C_c(2) = \left[\frac{1}{4} (1+1+1+2) \right]^{-1} = \frac{4}{5}$$

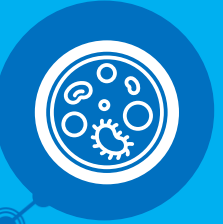
$$C_c(3) = C_c(4) = \left[\frac{1}{4} (2+1+2+1) \right]^{-1} = \frac{4}{6}$$

$$C_c(5) = \left[\frac{1}{4} (3+2+1+1) \right]^{-1} = \frac{4}{7}$$

MA

PageRank

Cheick Tidiane Ba



Ranking

- In the web we care about Ranking: we want to determinate the **importance** of a page or a user in a network.
- **Endogenous Ranking: rank based** on a page's content. Look at terms in web pages to figure out wheter they are relevant for the user's query.
- Issue: **term spamming**
- I can insert a lot of keywords to appear in many searches, obtaining always a high rank

Ranking

- We want an **exogenous** centrality measure
 - Harder to tamper with
 - In theory it is harder to have control on multiple web pages
- Exogenous measures can be divided in
 - **Geometric** Centralities
 - **Spectral** Centralities

Geometric centralities

- Geometric centralities rely on the concept **degree** (number of connected nodes) or **distance** measures
- You have seen already:
 - Degree Centrality
 - Betweenness centrality
 - Closeness centrality

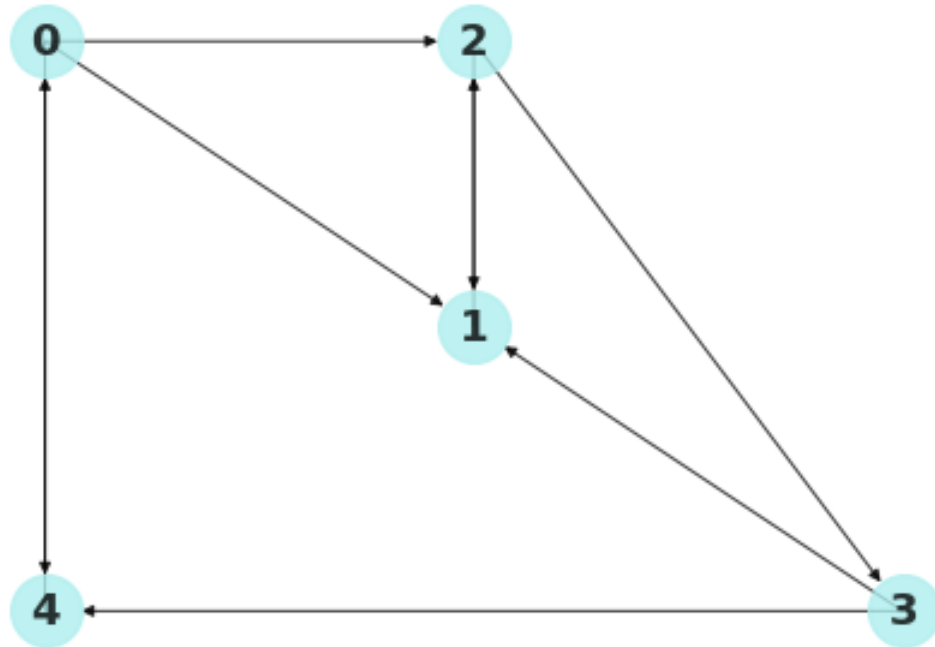
Spectral Ranking

- Spectral rankings are methods based on eigenvectors
- Among them we have:
 - Eigenvector centrality
 - Katz centrality
 - **PageRank**
 - Invented by Larry Page and Sergey Brin, founders of Google
 - Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.
 - The first metric used in Google Search and the reason of Google's success
 - Now is not the only metric, more metrics are considered
 - **HITS (Hyperlink-Induced Topic Search)**
 - also known as hubs and authorities

PageRank

- Simulation of a user's web browsing.
- **Random Surfer** on the web, browses through the WWW network.
- The WWW network is described by an adjacency matrix A , where if $i \rightarrow j$ then $A_{i,j} = 1$
- A **transition matrix** \bar{A} is obtained by dividing each row by its sum.
- A row in the transition matrix \bar{A} describes the probability of moving from a page i to page j .

PageRank – Transition Matrix example



A	0	1	2	3	4
0	0.0	1.0	1.0	0.0	1.0
1	0.0	0.0	1.0	0.0	0.0
2	0.0	1.0	0.0	1.0	0.0
3	0.0	1.0	0.0	0.0	1.0
4	1.0	0.0	0.0	0.0	0.0

\bar{A}	0	1	2	3	4
0	0.0	0.33	0.33	0.0	0.33
1	0.0	0.0	1.0	0.0	0.0
2	0.0	0.5	0.0	0.5	0.0
3	0.0	0.5	0.0	0.0	0.5
4	1.0	0.0	0.0	0.0	0.0

PageRank

- We are interested in the visits on a certain node during the random surfing.
- Given a vector p_t , that expresses the probability to be on a certain page at time t.
 - $p_t = (p_0, p_1, \dots, p_n)$
- We can simulate a user moving to a new page by computing: $p_{t+1} = p_t * \bar{A}$
 - This is a Markov chain
- We keep multiplying till p_{t+1} doesn't change too much with respect to the previous p_t
- The vector p_t contains the PageRank value for each page

PageRank – Computation example

- **Start**

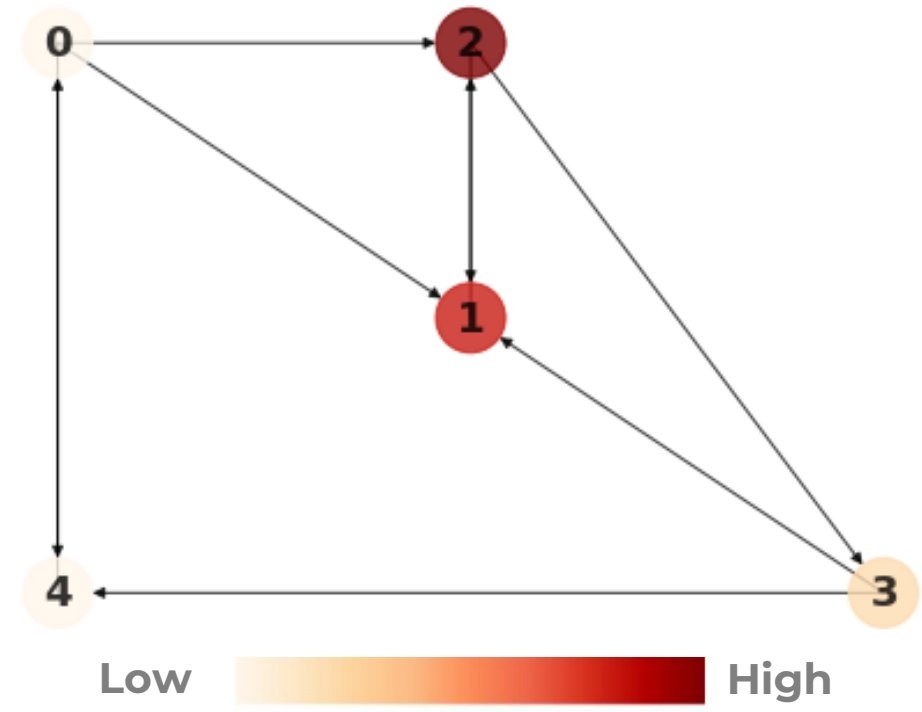
- $p_0 = [0.2 \ 0.2 \ 0.2 \ 0.2 \ 0.2]$

- Transition matrix \bar{A}

- **Calculation: $p_{t+1} = p_t * \bar{A}$**

- 0.362, 0.483, 0.483, 0.181, 0.302
 - 0.294, 0.440, 0.587, 0.235, 0.206
 - 0.205, 0.509, 0.538, 0.293, 0.215
 - ...
 - 0.211, 0.490, 0.561, 0.280, 0.210
 - 0.210, 0.491, 0.560, 0.281, 0.210
 - 0.210, 0.491, 0.561, 0.280, 0.210
 - **0.210, 0.491, 0.561, 0.281, 0.210**

\bar{A}	0	1	2	3	4
0	0.0	0.33	0.33	0.0	0.33
1	0.0	0.0	1.0	0.0	0.0
2	0.0	0.5	0.0	0.5	0.0
3	0.0	0.5	0.0	0.0	0.5
4	1.0	0.0	0.0	0.0	0.0

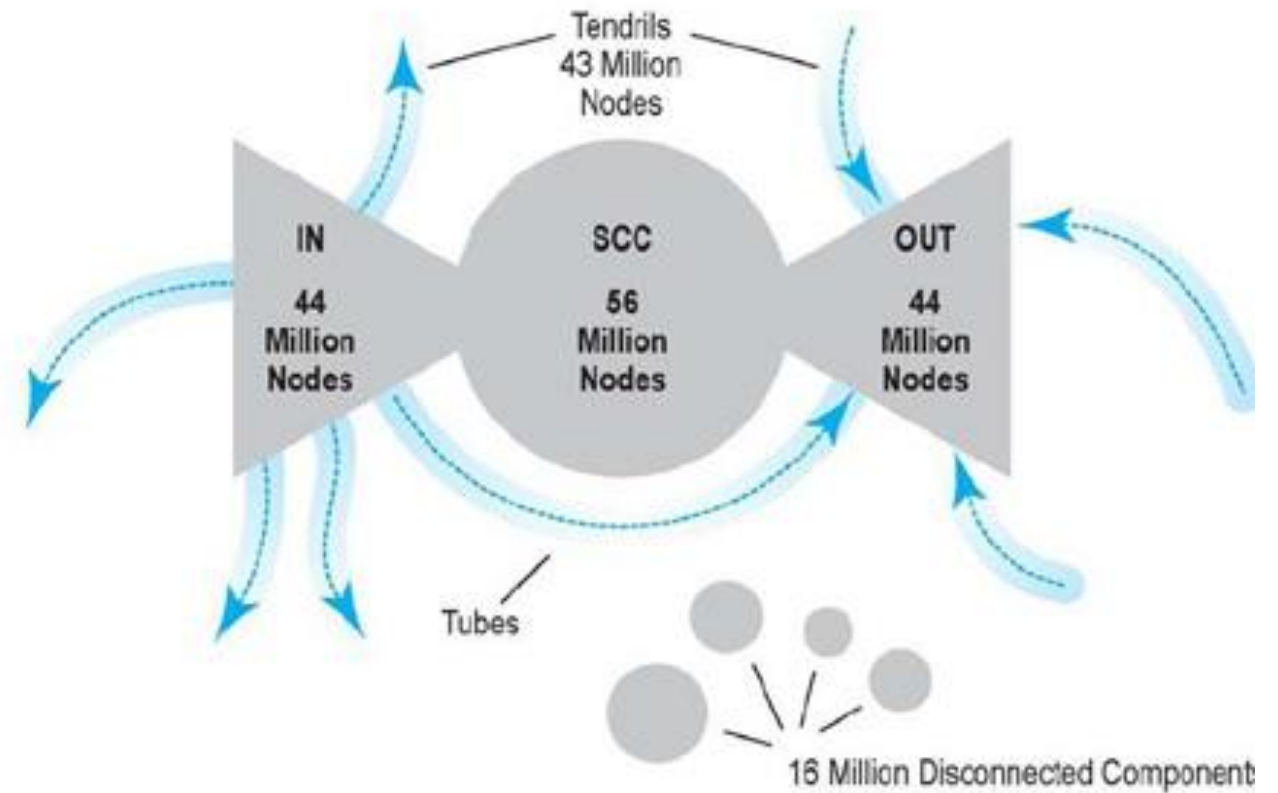


Issues

- **Several theorems and demonstrations from eigenvalue theory and markov chain are available in the literature.**
- **They grant us that we can find a ranking vector p as long as we are not considering:**
 - **Isolated components**
 - **Dangling nodes (nodes without outgoing edges)**
- **These structure are present, we need to address them**

Bowtie

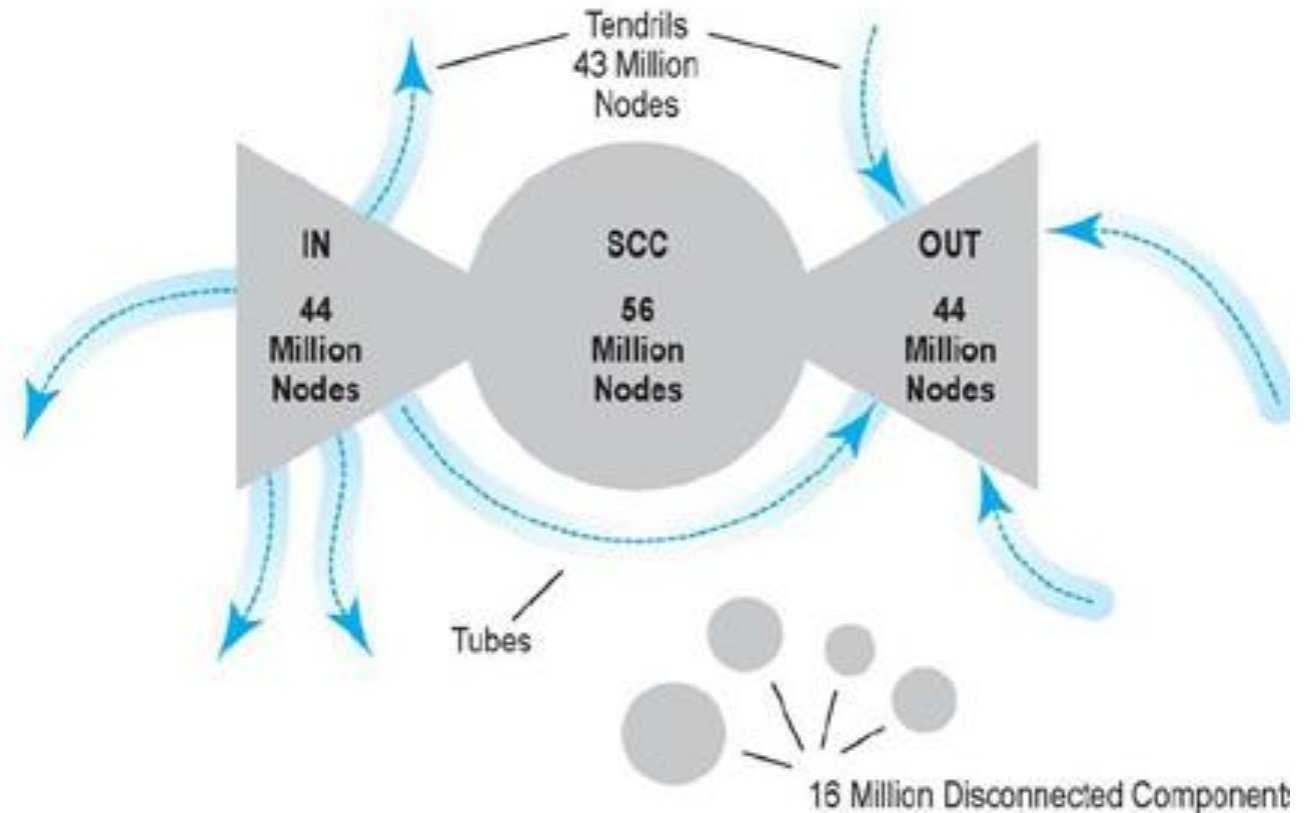
- The Bowtie is still relevant



Source: K. Laudon & C. Trever, E-Commerce
2009 (5th Edition), Prentice Hall.

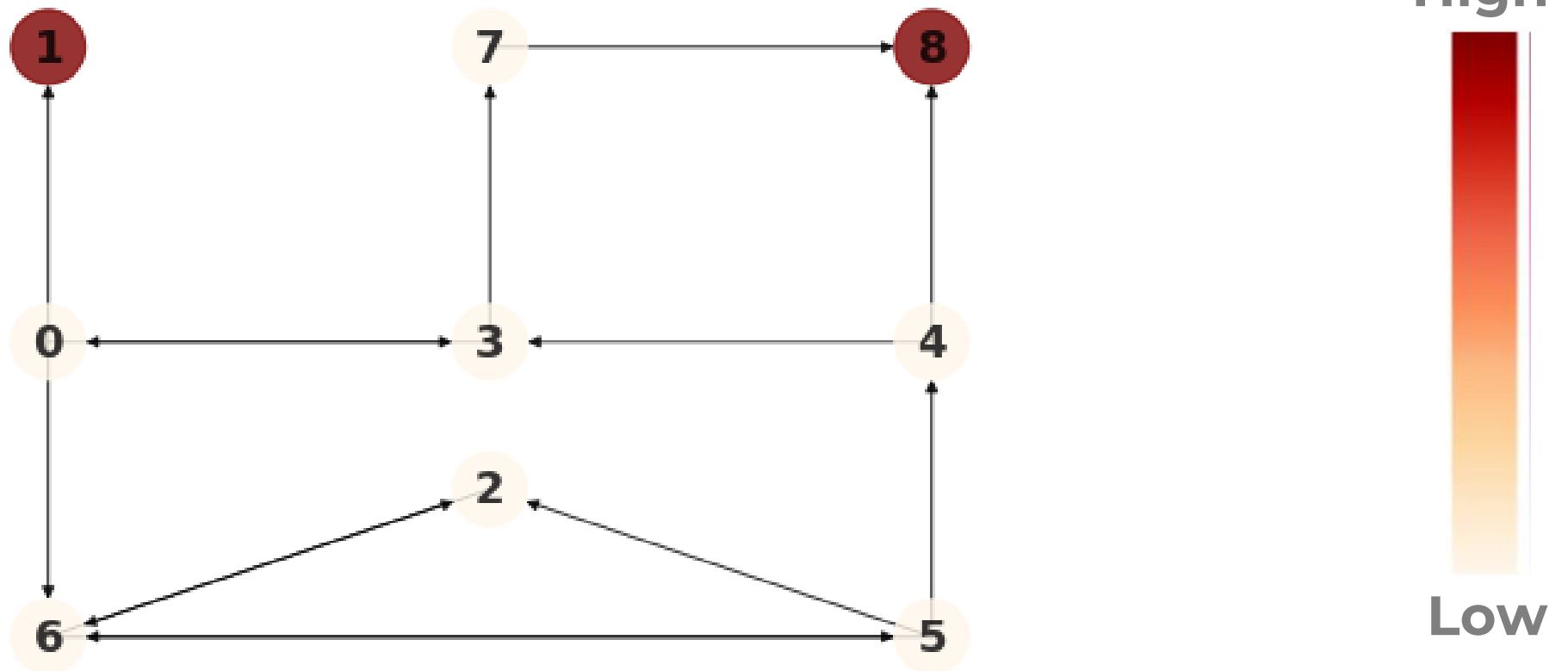
Bowtie

- In Component
 - They have mainly edges towards the scc
- SCC: strongly connected component
- OUT component
 - Reached by other nodes
- Lots of isolated components
- Tubes connect In and Out components



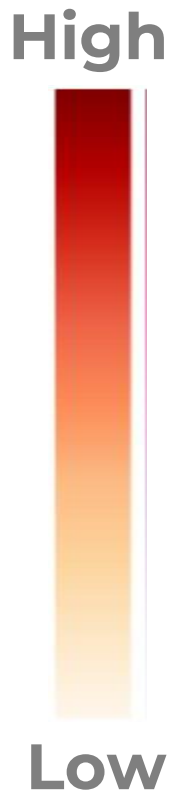
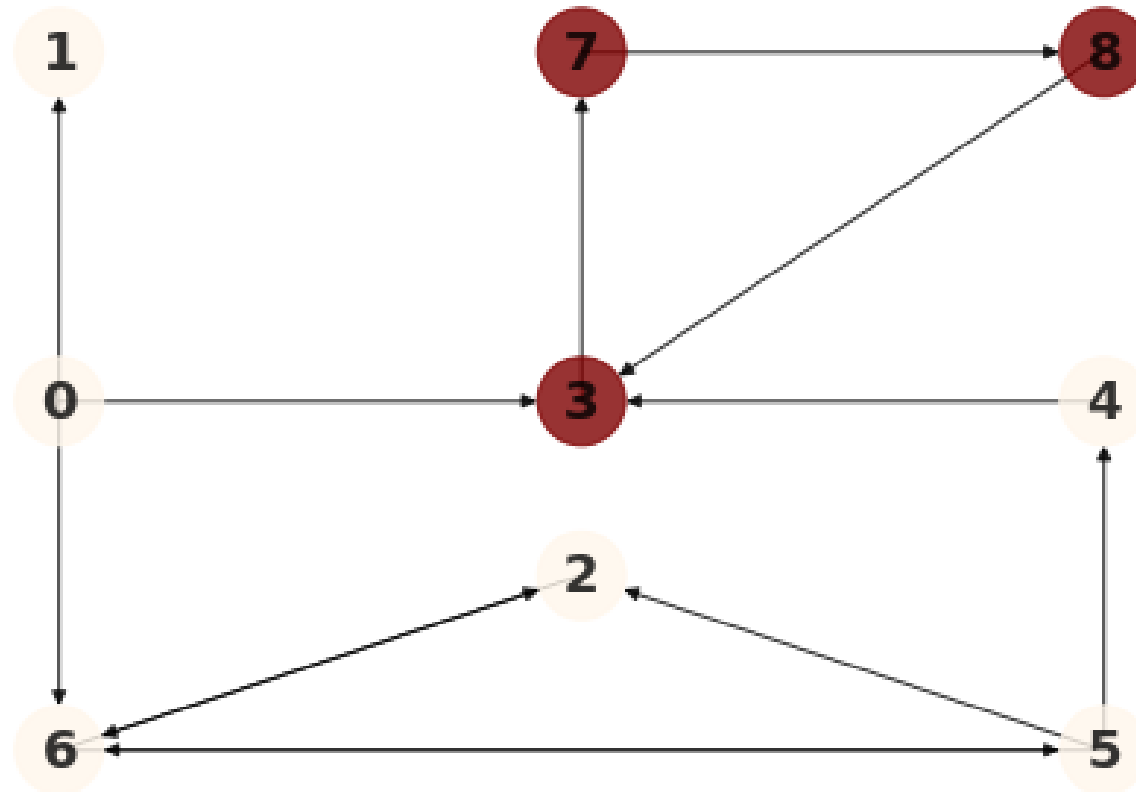
Dangling nodes

- The random surfer on the Internet always ends up reaching a dangling node.
- **The surfer stops** surfing through hyperlinks.



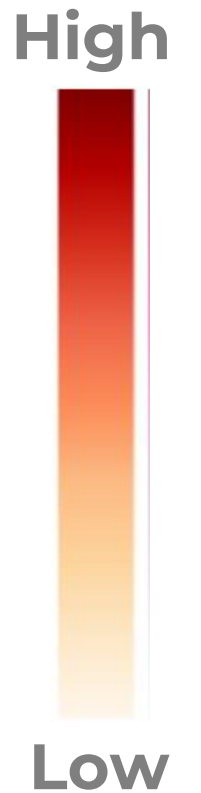
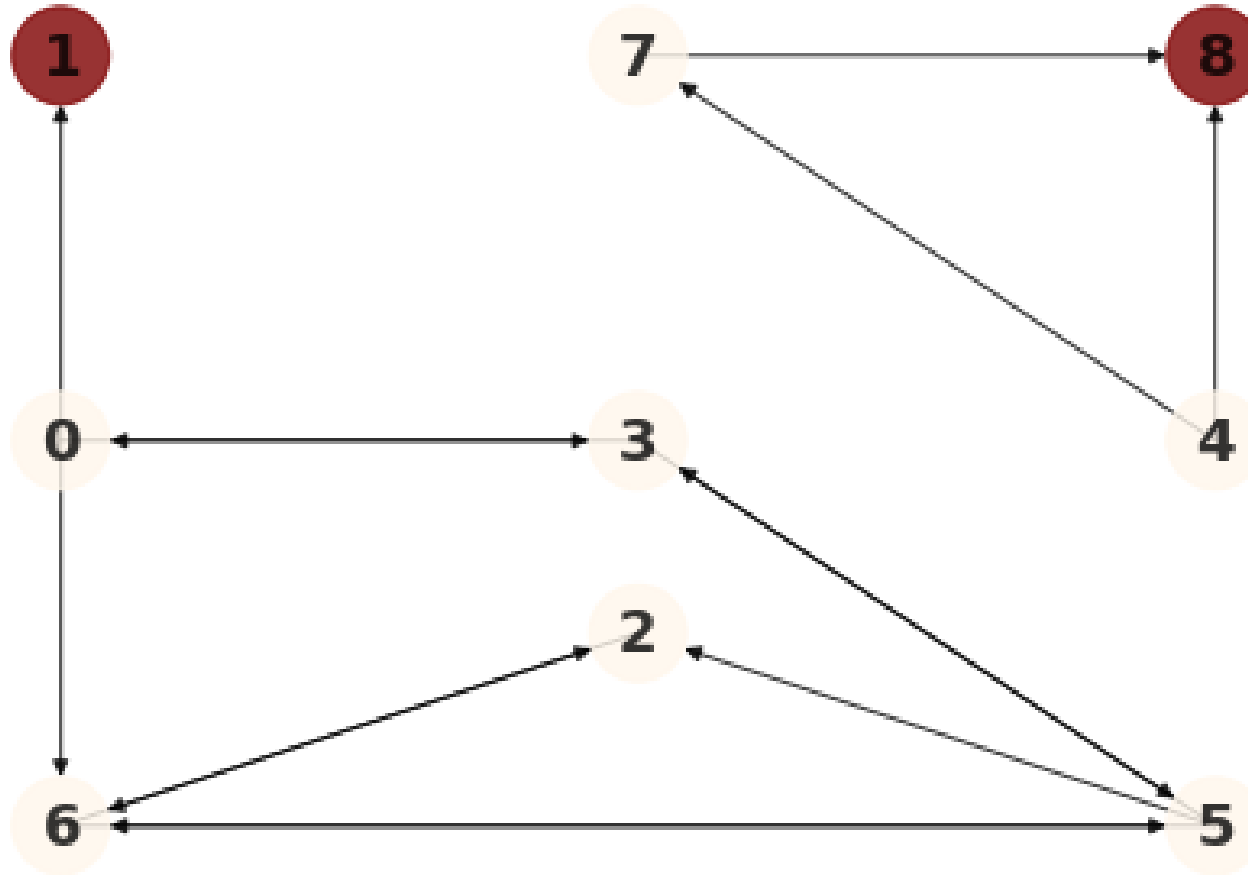
Loops

- The random surfer can get **stuck in loops**
- Values are higher for the nodes in the loop
- These structure are also called **spider traps**



Isolated components

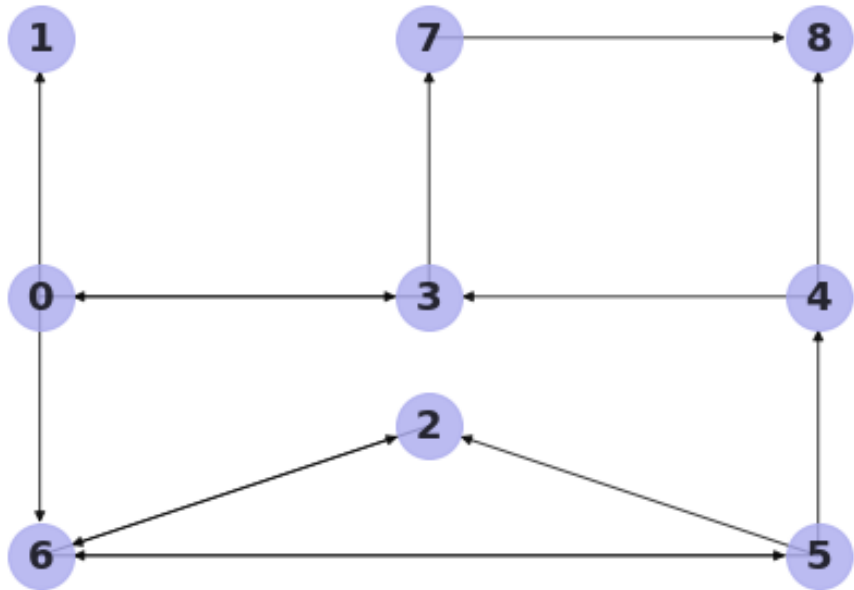
- Surfer stuck like in the previous scenarios



Teleport

- The solution is known as **teleport or tax**
- At each step, the user can:
 - Go one of the following pages with probability α
 - **Teleport** to a random node with probability $1 - \alpha$
- We can express the probability to teleport to a certain page as a distribution.
 - It is just a vector v , with sum equal to 1
 - Usually a uniform distribution
- The final formula becomes:
 - $p_{t+1} = \alpha * \bar{A}p_t + (1 - \alpha) * v$

Teleport – Transition Matrix example



\bar{A}	0	1	2	3	4	5	6	7	8
0	0.0	0.33	0.00	0.33	0.00	0.0	0.33	0.0	0.0
1	0.0	0.00	0.00	0.00	0.00	0.0	0.00	0.0	0.0
2	0.0	0.00	0.00	0.00	0.00	0.0	1.00	0.0	0.0
3	0.0	0.00	0.00	0.00	0.00	0.0	0.00	1.0	0.0
4	0.0	0.00	0.00	1.00	0.00	0.0	0.00	0.0	0.0
5	0.0	0.00	0.33	0.00	0.33	0.0	0.33	0.0	0.0
6	0.0	0.00	0.50	0.00	0.00	0.5	0.00	0.0	0.0
7	0.0	0.00	0.00	0.00	0.00	0.0	0.00	0.0	1.0
8	0.0	0.00	0.00	1.00	0.00	0.0	0.00	0.0	0.0

Teleport – Computation example

- **Basic**

- $p_0 = [0.11, \dots, 0.11]$
- Transition matrix \bar{A}

- **Calculation of $p_{t+1} = p_t * \bar{A}$**

- 0.000, 0.040, 0.099, 0.277, 0.040, 0.059, 0.198, 0.119, 0.119
- 0.000, 0.000, 0.121, 0.161, 0.020, 0.101, 0.121, 0.282, 0.121
- ...
- 0.000, 0.000, 0.000, 0.197, 0.000, 0.000, 0.000, 0.298, 0.180
- 0.000, 0.000, 0.000, 0.180, 0.000, 0.000, 0.000, 0.197, 0.298
- **0.000, 0.000, 0.000, 0.298, 0.000, 0.000, 0.000, 0.180, 0.197**

- **With Teleport**

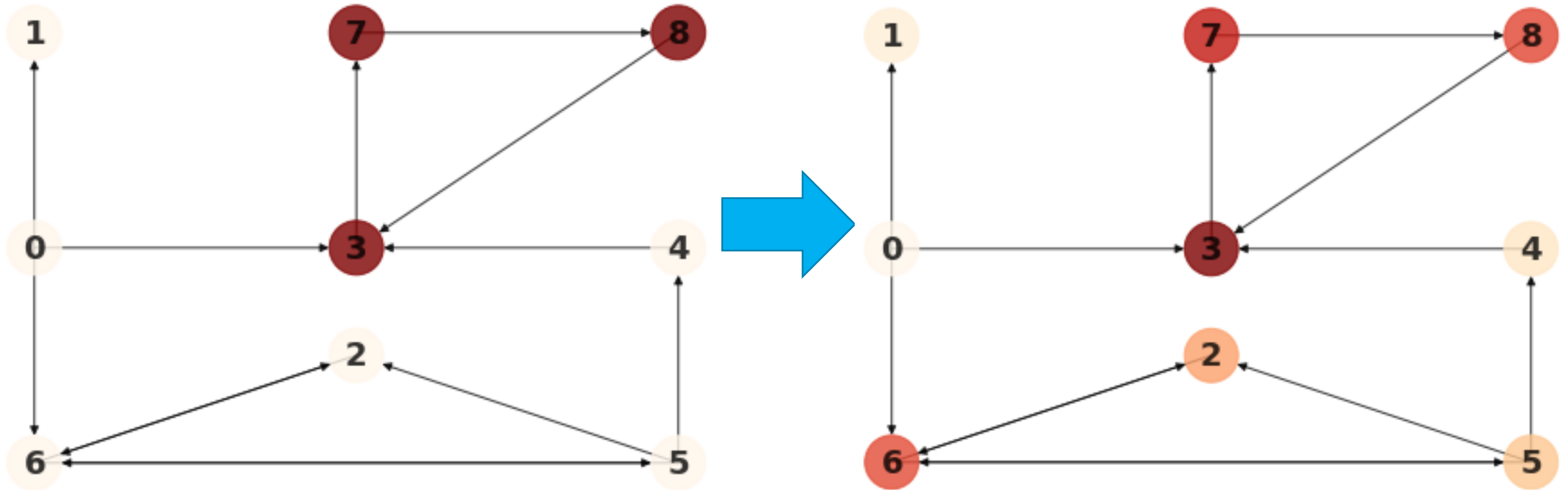
- $p_0 = [0.11, \dots, 0.11]$
- Transition matrix \bar{A}
- $v = [0.11, \dots, 0.11]$
- $\alpha = 0.5$

- **Calculation of**

$$p_{t+1} = \alpha * \bar{A}p_t + (1 - \alpha) * v$$

- 0.067, 0.106, 0.165, 0.343, 0.106, 0.126, 0.264, 0.185, 0.185
- 0.067, 0.082, 0.184, 0.278, 0.095, 0.156, 0.221, 0.298, 0.191
- ...
- 0.067, 0.081, 0.169, 0.297, 0.097, 0.139, 0.221, 0.260, 0.237
- 0.067, 0.081, 0.169, 0.298, 0.097, 0.139, 0.221, 0.260, 0.236
- **0.067, 0.081, 0.169, 0.298, 0.097, 0.139, 0.221, 0.261, 0.236**

Loops with teleport

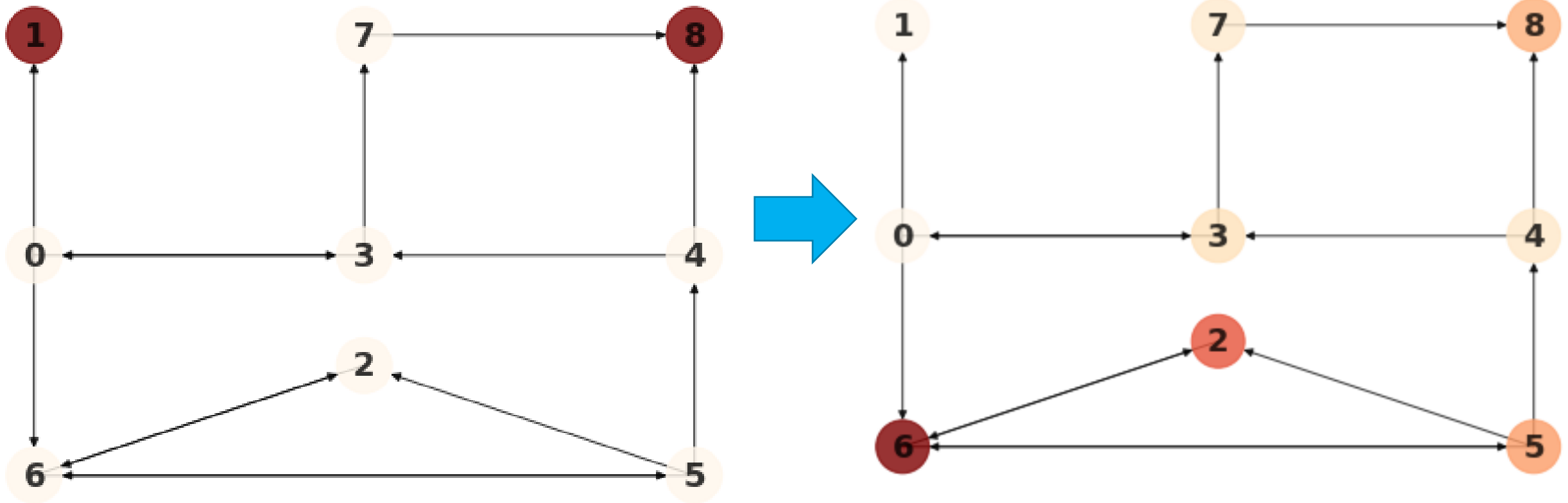


Low



High

Dangling nodes with teleport

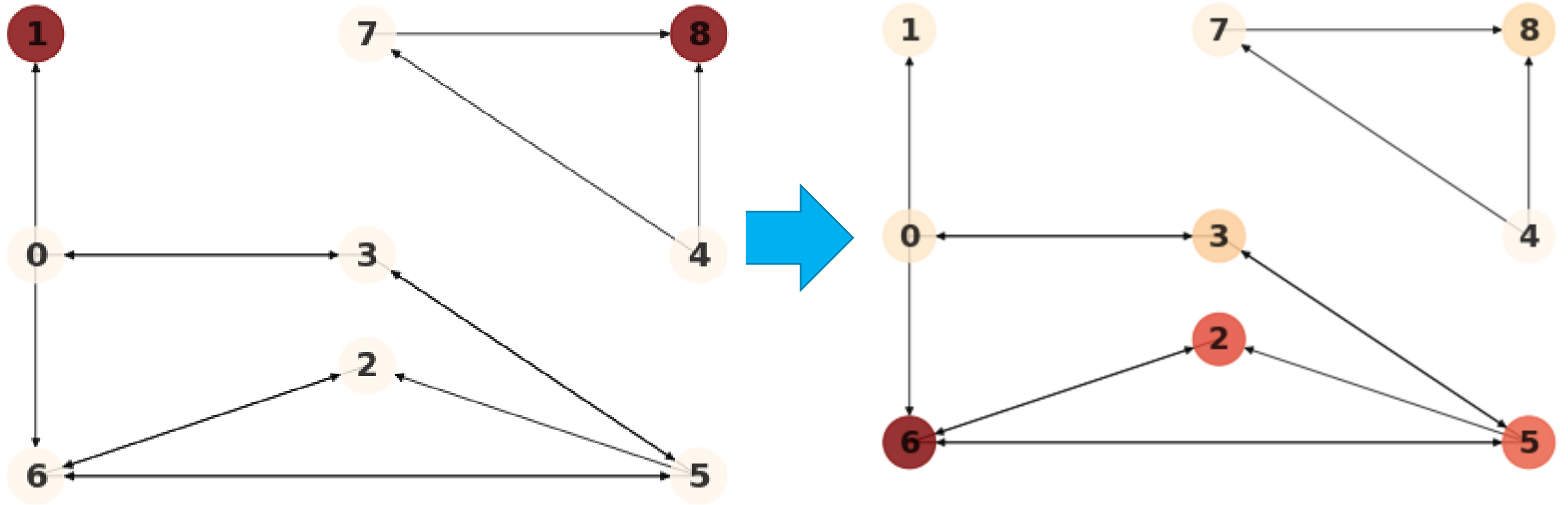


Low



High

Isolated components with teleport



Low



High

For the final project

- PageRank can be computed directly in Gephi
 - Implemented version deals with the all the issues

The screenshot shows the 'Context' menu in Gephi. It displays the following information: Nodes: 10469, Edges: 178115, and Undirected Graph. Below this, there are tabs for 'Filters' and 'Statistics'. Under the 'Statistics' tab, the 'Network Overview' section is expanded, showing a list of metrics with 'Run' buttons: Average Degree, Avg. Weighted Degree, Network Diameter, Graph Density, HITS, Modularity, PageRank, and Connected Components.

The screenshot shows the 'Page Rank settings' dialog box. It has a title bar with a close button. The main content area is titled 'PageRank' and contains the following text: 'Ranks nodes "pages" according to how often a user following links will non-randomly reach the node "page".' Below this, there are two radio buttons: 'Directed' (unselected) and 'Undirected' (selected). To the right of the 'Undirected' radio button, there is a text input field for 'Probability (p):' with the value '0.85'. Below this, there is a text input field for 'Epsilon:' with the value '0.001'. A small text description below the epsilon field reads: 'Stopping criterion, the smaller this value, the longer convergence will take.' At the bottom, there is a checkbox for 'Use edge weight' which is currently unchecked. At the bottom right, there are 'OK' and 'Cancel' buttons.

The screenshot shows the 'Appearance' dialog box in Gephi. It has a title bar with a close button. Below the title bar, there are tabs for 'Nodes' and 'Edges'. The 'Nodes' tab is selected. Below the tabs, there are icons for 'Unique', 'Partition', and 'Ranking'. The 'Ranking' tab is selected. Below the tabs, there is a dropdown menu with the text '--Choose an attribute'. Below the dropdown menu, there is a list of attributes: '--Choose an attribute', 'Degree', and 'PageRank'. The 'PageRank' attribute is highlighted in blue. At the bottom right, there is an 'Apply' button.

For the final project

- **Methods for Pagerank computation are also available in Networkx**
 - `pagerank(G[, alpha, personalization, ...])`
 - `pagerank_numpy(G[, alpha, personalization, ...])`
 - `pagerank_scipy(G[, alpha, personalization, ...])`
- **Same algorithm, difference in computation times**
- **Docs available here:**
 - https://networkx.org/documentation/stable/reference/algorithms/link_analysis.html?highlight=pagerank

Thank you for the attention!

For any question send an email at
cheick.ba@unimi.it





UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Transitivity
Global and Local
Clustering coefficient
in undirected networks

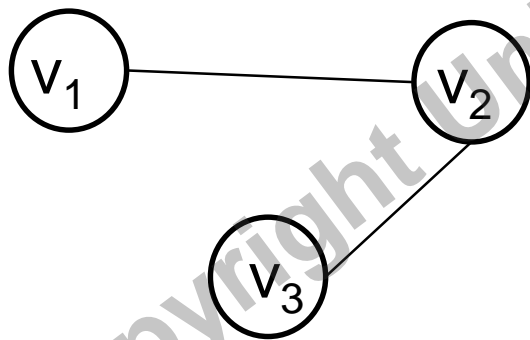
Transitivity

Mathematic representation:

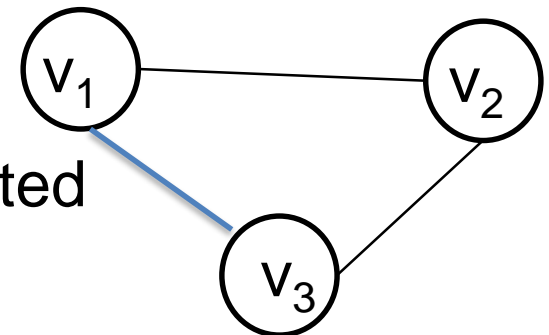
- For a transitive relation R: $aRb \wedge bRc \rightarrow aRc$

Networks:

- the transitive relation R: connected by a link
- If v_1, v_2 are connected and v_2, v_3 are connected



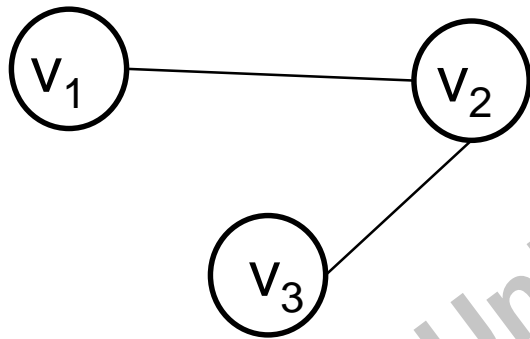
v_1, v_3 are connected



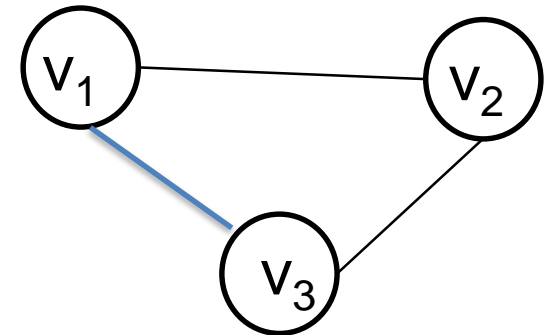
Transitivity

Social Networks:

- the transitive relation R : friendship
- If v_1, v_2 are friends and v_2, v_3 are friends



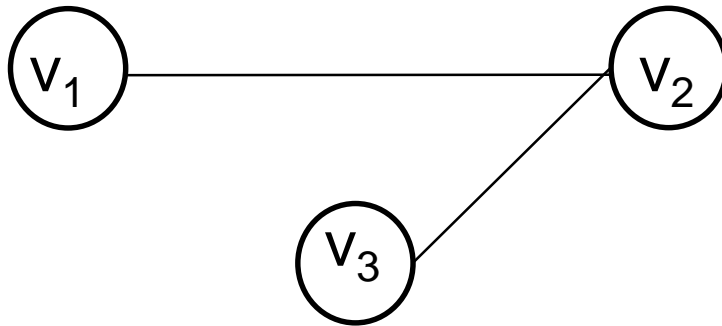
v_1, v_3 are friends



***Transitivity is when
a friend of my friend is my friend***



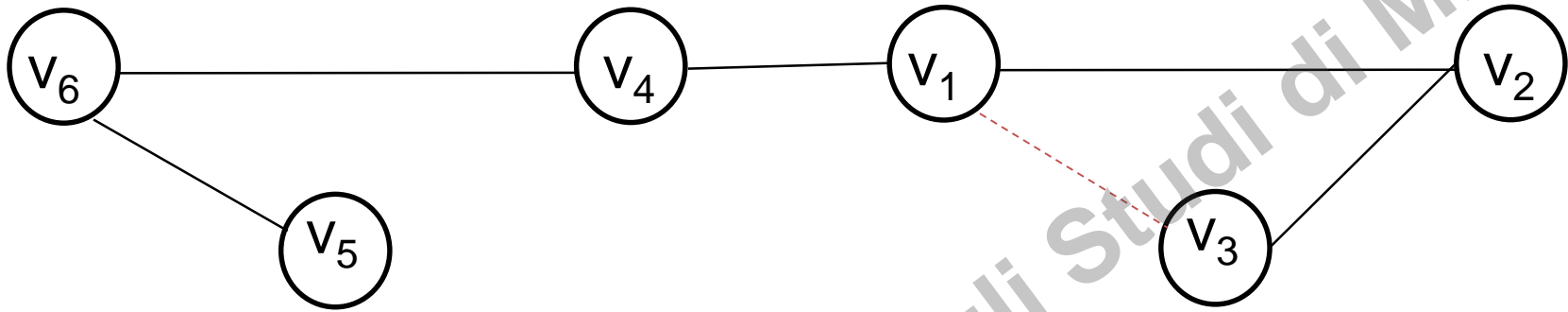
Transitivity



- Perfect transitivity only occurs in networks where each component is a fully connected graph or clique (a subgraph in which all nodes are connected to all others)
- Perfect transitivity is a useless concept in social networks as it never occurs



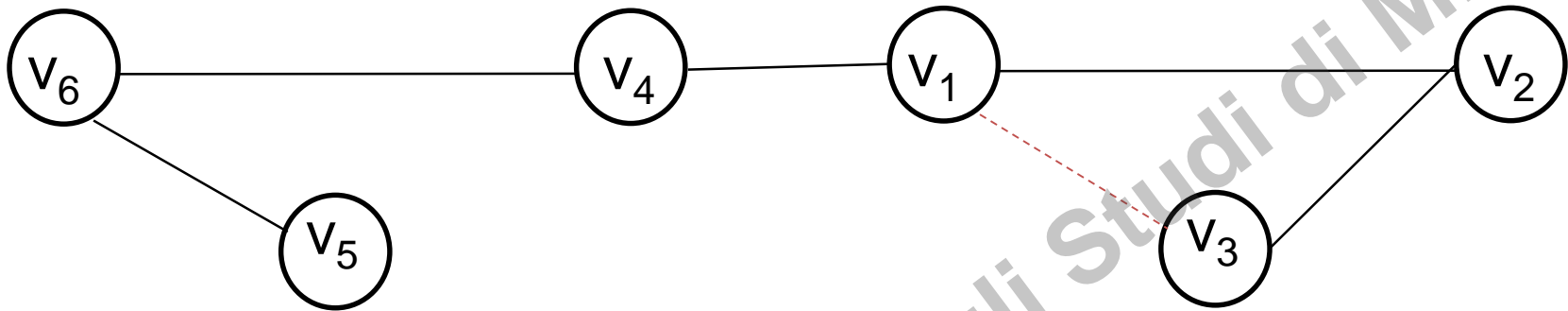
Transitivity



- Partial transitivity is much more useful
- The friend of my friend is not guaranteed to be my friend
- But is far more likely to be my friend than any other node in the network.
- v_1 is more likely to be friend of v_3 than v_5
- Is v_1 more likely to be friend of v_3 than v_6 ?



Transitivity



- Partial transitivity is much more useful
- The friend of my friend is not guaranteed to be my friend
- But is far more likely to be my friend than any other node in the network.
- v_1 is more likely to be friend of v_3 than v_5
- v_1 has the same probability to be friend of v_3 and v_6

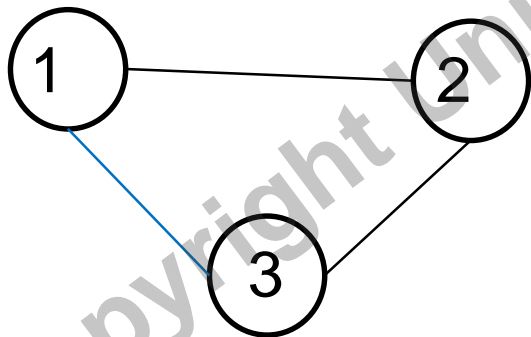


Global Clustering Coefficient Measure based on paths

We want to quantify the level of transitivity of a network

We can measure it by counting the paths of length two and check whether the third edge exists

$$C = \frac{|\text{Paths of Length 2 that have the third edge}|}{|\text{Paths of Length 2}|}$$



A path of length 2 which has the third link is called *closed path* as it forms a loop of length 3. [Closed paths are also called *closed triad in social networks*.]



Global Clustering Coefficient

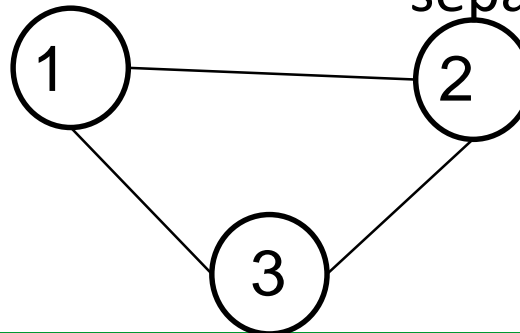
$$C = \frac{|\text{Paths of Length 2 that have the third edge}|}{|\text{Paths of Length 2}|}$$

Note that paths, also closed paths, have a direction in undirected network, too.

Two paths that traverse the same links but in opposite direction are counted separately.

Path of length 2	Third edge
213	32
312	23
123	31
321	13
132	21
231	12

$$C=6/6$$



Global Clustering Coefficient

$$C = \frac{|\text{Paths of Length 2 that have the third edge}|}{|\text{Paths of Length 2}|}$$

Path of length 2

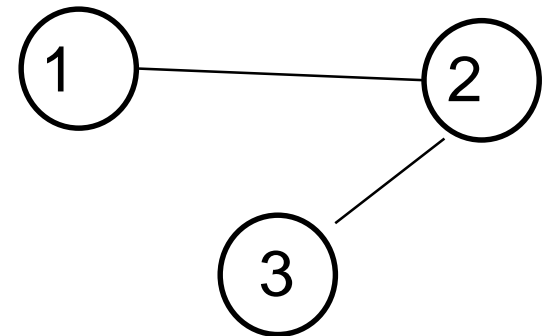
123

321

Third edge

-

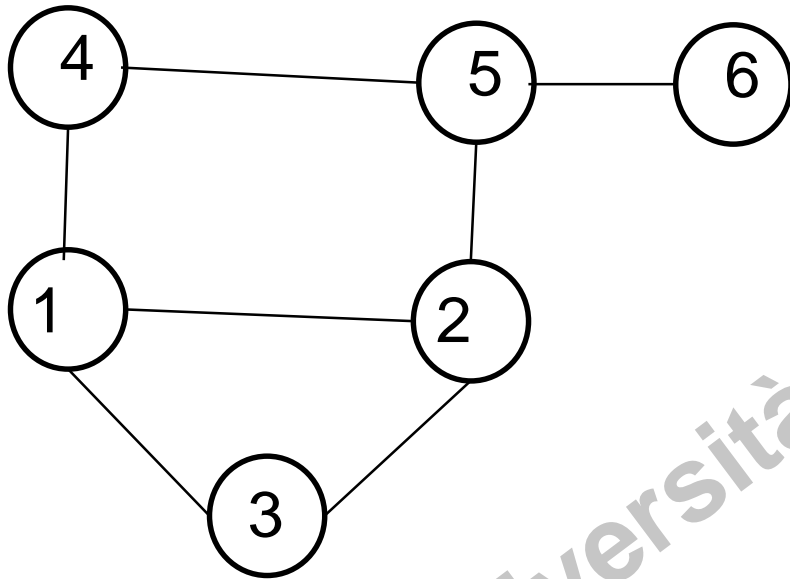
-



$$C=0/2=0$$



Example



$$C = 6/22 = 3/11$$

Path of length 2	Third edge
213 312	yes
214 412	no
314 413	no
123 321	yes
125 521	no
325 523	no
132 231	yes
145 541	no
254 452	no
256 652	no
456 654	no

[Note: you could divide both the numerator and the denominator by two, by considering paths in one direction only]



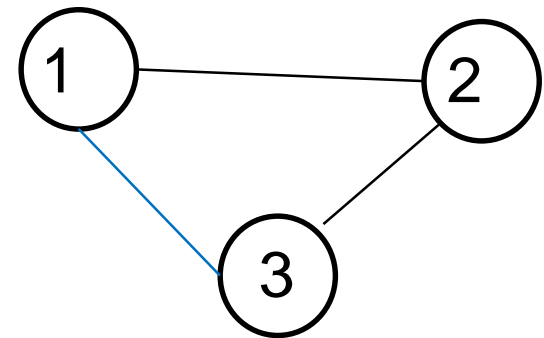
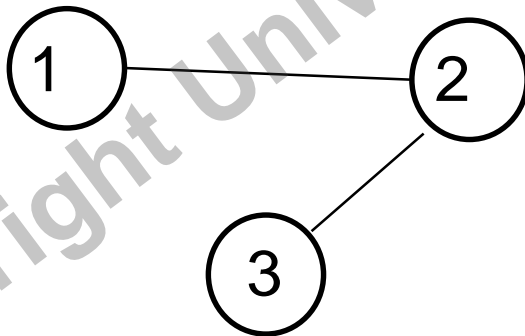
Global Clustering Coefficient

Measure based on triples

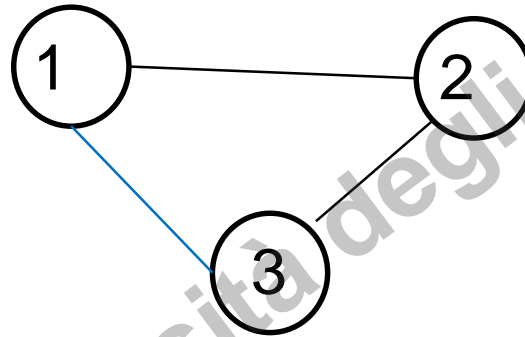
If we have a path $\{1,2,3\}$ (and $\{3,2,1\}$) of length 2, it is also true to say that nodes 1 and 3 have a common neighbour: node 2.

If the triad $\{1,2,3\}$ is closed, nodes 1 and 3 are themselves friends.

The clustering coefficient can be thought as the **fraction of pairs of people with a common friend who are themselves friend.**



Clustering coefficient



a friend of my friend is my friend

pairs of people with a common friend who are themselves friend.



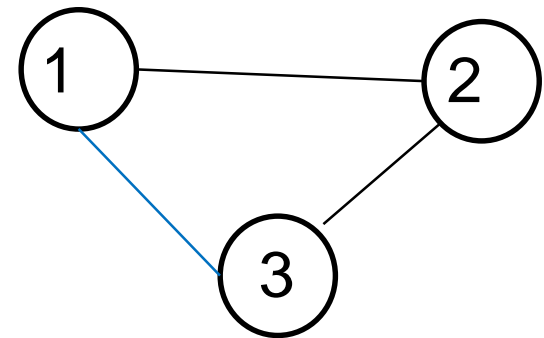
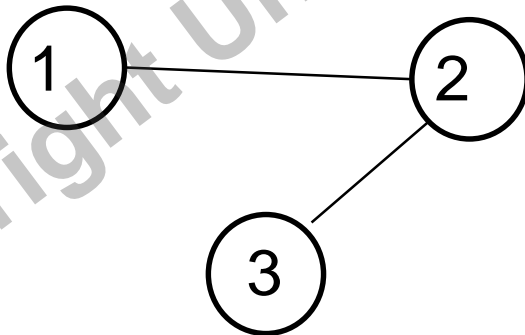
Global Clustering Coefficient

Measure based on triplets

The clustering coefficient can be thought as the fraction of pairs of people with a common friend who are themselves friend.

We can also define the global clustering coefficient based on the concept of (connected) triplets of nodes.

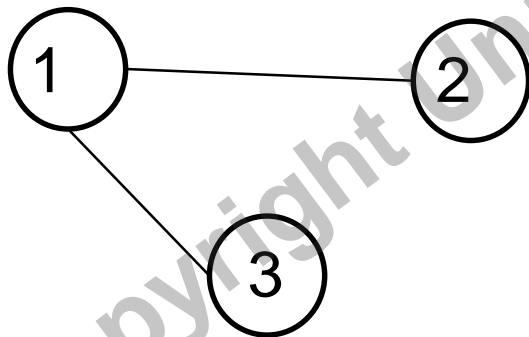
A connected triplet consists of three nodes $\{v_1, v_2, v_3\}$, that are connected by the two links (v_1, v_2) and (v_2, v_3) . The third link (v_1, v_3) can be present (closed triplet) or not (open triplet).



Global Clustering Coefficient Measure based on triples

Triples (open)

$\{2,1,3\}$ [with links $(2,1)$ and $(1,3)$]

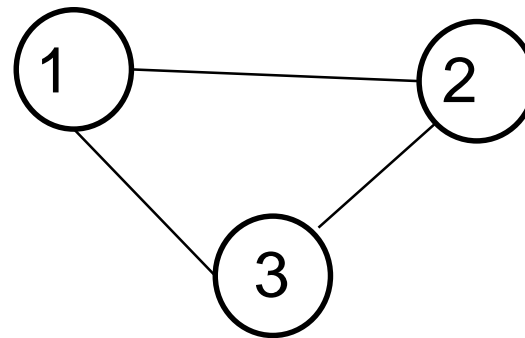


Triples (closed)

$\{2,1,3\}$ [with links $(2,1)$ and $(1,3)$]

$\{1,2,3\}$ [with links $(1,2)$ and $(2,3)$]

$\{1,3,2\}$ [with links $(1,3)$ and $(3,2)$]



Global Clustering Coefficient

Measure based on triples

The global clustering coefficient is the number of closed triplets over the total number of triplets (both open and closed):

$$\frac{\text{number of closed triplets}}{\text{total number of triplets}}$$



Global Clustering Coefficient

Measure based on triples

Triples (open)

213

$$C=0/2=0$$

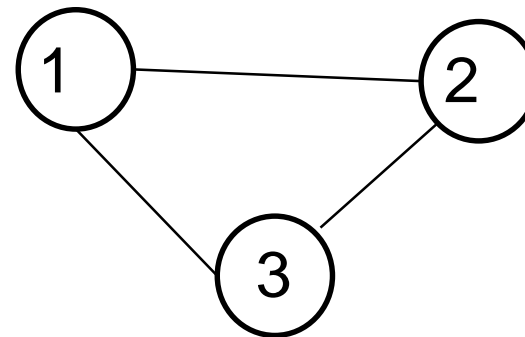
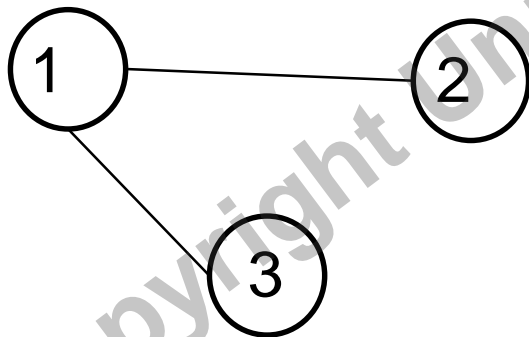
Triples (closed)

213

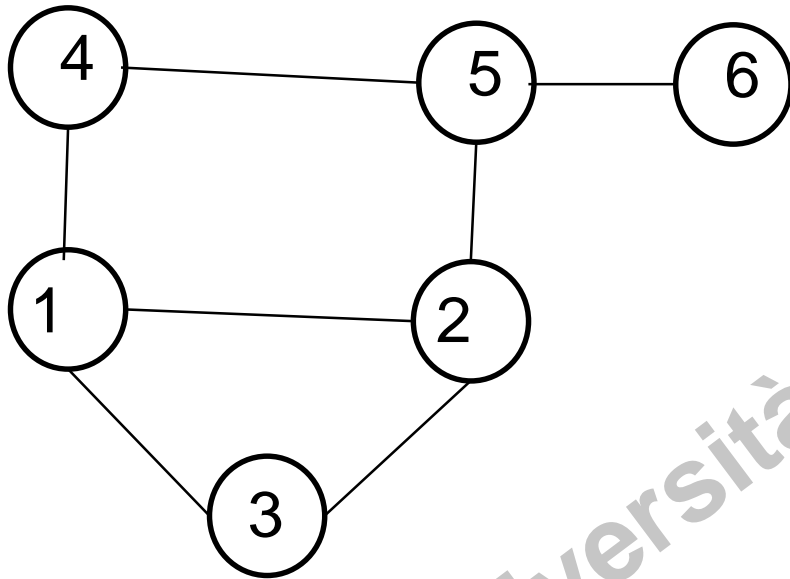
123

132

$$C=3/3=1$$



Example



$C=3/11$

Triplets

213

Closed?

yes

214

no

314

no

123

yes

125

no

325

no

132

yes

145

no

254

no

256

no

456

no



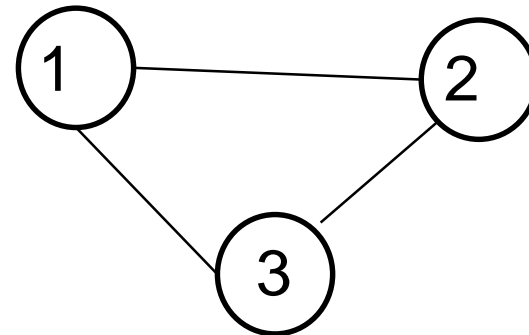
Global Clustering Coefficient Measure based on triangles

A triangle consists of six paths.

213, 312, 123, 321, 132, 231

A triangle consists of three triples, one centered on each of the nodes.

213, 123, 231



Global Clustering Coefficient

Measure based on triangles

The global clustering coefficient is the number of closed paths of length 2 (or 6 x triangles) over the total number of paths of length 2

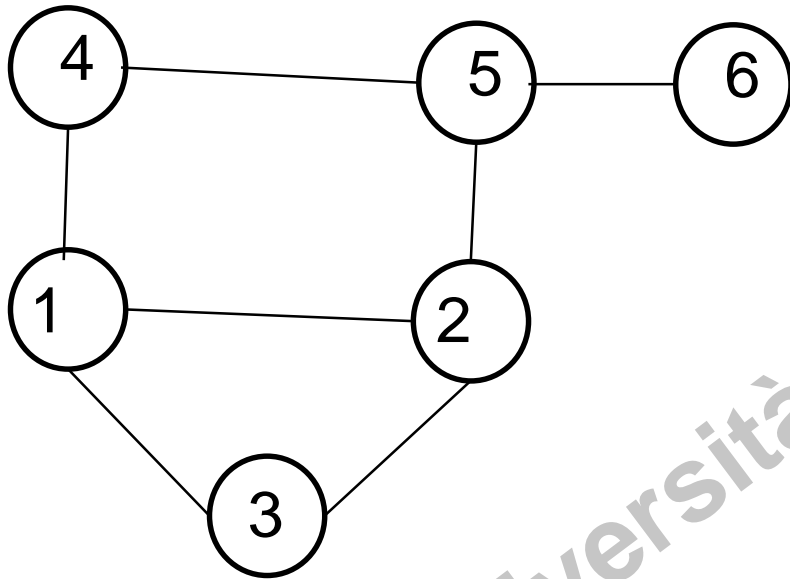
$$\frac{6 * \text{number of triangles}}{\text{number of paths of length 2}}$$

The global clustering coefficient is the number of closed triplets (or 3 x triangles) over the total number of triplets

$$\frac{3 * \text{number of triangles}}{\text{number of triplets}}$$



Example



$$C = 1 * 6 / 22 = 3 / 11$$

Path of length 2

213 312

214 412

314 413

123 321

125 521

325 523

132 231

145 541

254 452

256 652

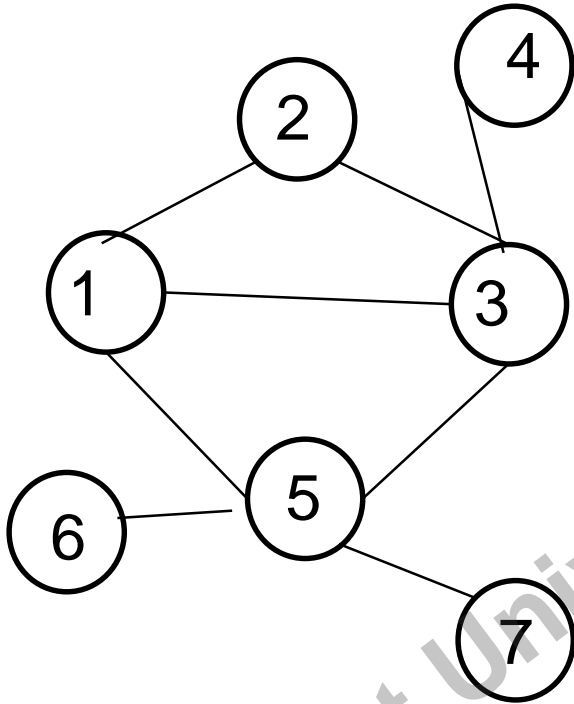
456 654

Triangles

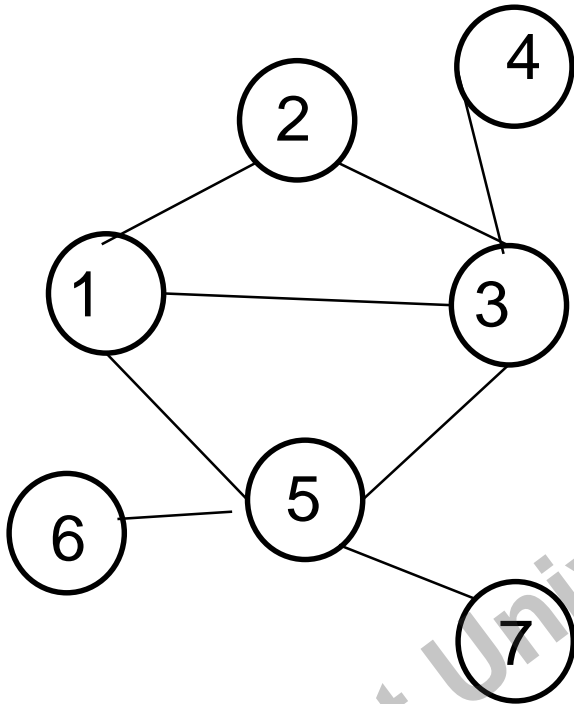
123



Exercise



Exercise



$$C = 12/32 = 3/8$$

Path of length 2

213 312

215 512

315 513

123 321

132 231

134 431

135 531

234 432

235 532

435 534

153 351

156 651

157 751

356 653

357 753

657 756

Third edge

yes

no

yes

yes

yes

no

yes

no

no

no

yes

no

no

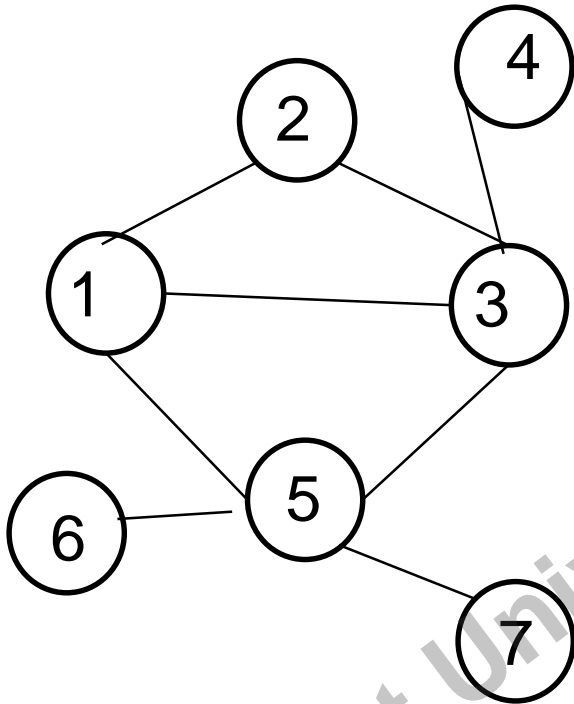
no

no

no



Exercise



$$C = 6/16 = 3/8$$

Triplets

213

Closed?

yes

215

no

315

yes

123

yes

132

yes

134

no

135

yes

234

no

235

no

435

no

153

yes

156

no

157

no

356

no

357

no

657

no



Exercise

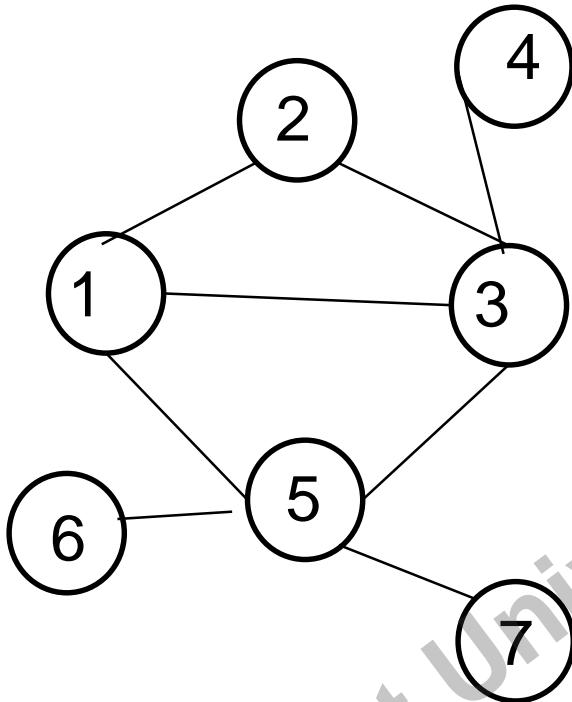
Number of triangles: 2

$$C = \frac{6 * \text{number of triangles}}{\text{number of paths of length 2}}$$

$$C = 6 * 2 / 32 = 12 / 32 = 3 / 8$$

$$C = \frac{3 * \text{number of triangles}}{\text{number of triplets}}$$

$$C = 3 * 2 / 16 = 6 / 16 = 3 / 8$$



$$C = 6 / 16 = 3 / 8$$



Local Clustering Coefficient

- Local clustering coefficient measures transitivity at the node level
- Commonly employed for undirected graphs, it computes how strongly neighbors of a node v (nodes adjacent to v) are themselves connected

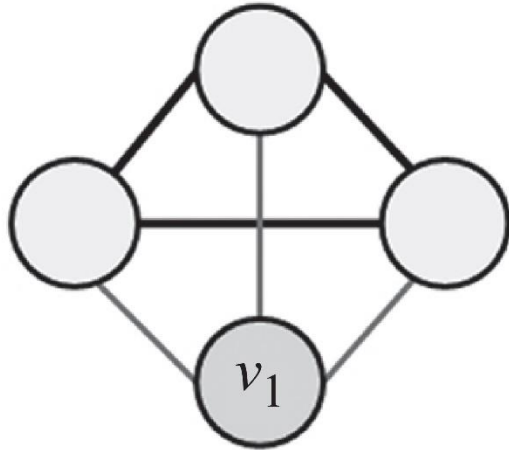
$$C(v_i) = \frac{\text{number of pairs of neighbors of } v_i \text{ that are connected}}{\text{number of pairs of neighbors of } v_i}.$$

In an undirected graph, the denominator can be rewritten as:

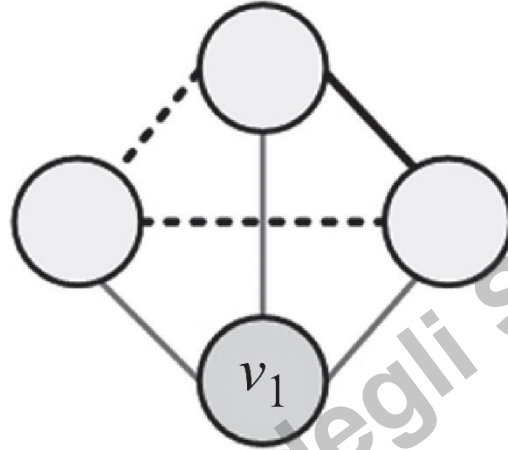
$$\binom{d_i}{2} = d_i(d_i - 1)/2,$$



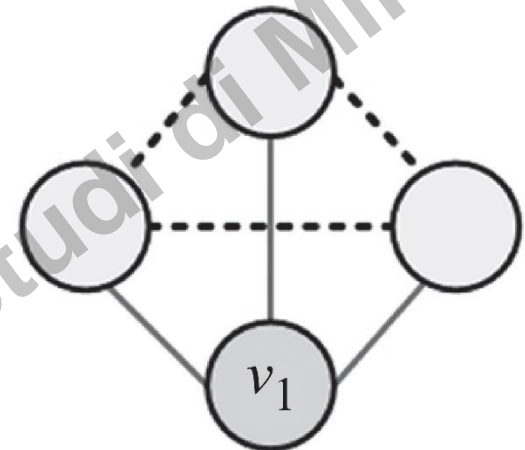
Local Clustering Coefficient:



$$C(v_1) = 1$$



$$C(v_1) = 1/3$$

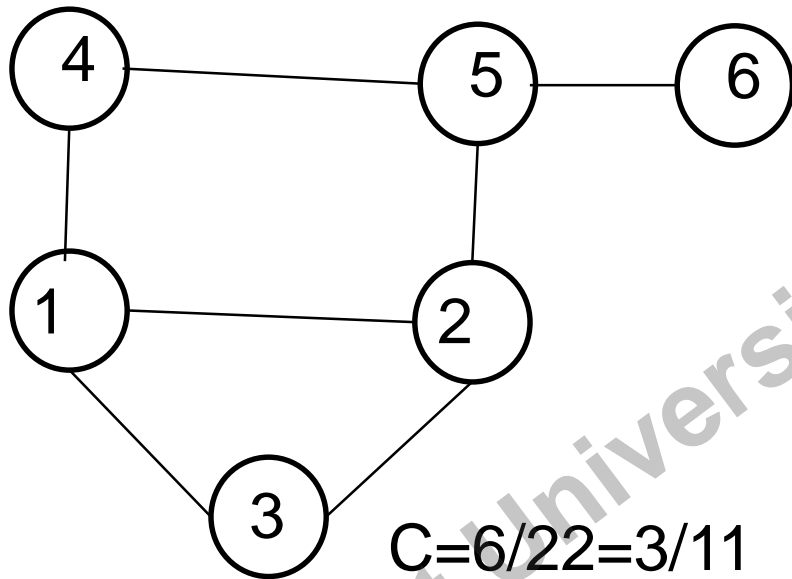


$$C(v_1) = 0$$

- Thin lines depict connections to neighbors
- Dashed lines are the missing connections among neighbors
- Solid lines indicate connected neighbors
 - When none of neighbors are connected $C=0$
 - When all neighbors are connected $C=1$



Example



Node	Pairs of friends	Open or closed triad	Clustering coeff
1	23	closed	$c=1/3$
	24	open	
	34	open	
2	13	closed	$c=1/3$
	15	open	
	35	open	
3	12	closed	$c=1/1$
4	15	open	$c=0$
5	24	open	$c=0/3=0$
	26	open	
	46	open	
6	-	-	-



Example

The clustering coefficient distribution therefore is:

c	Frequency
0	2/5
1/3	2/5
1	1/5

The mean clustering coefficient is: $1/3$

Node	Pairs of friends	Open or closed triad	Clustering coeff
1	23	closed	$c=1/3$
	24	open	
	34	open	
2	13	closed	$c=1/3$
	15	open	
	35	open	
3	12	closed	$c=1/1$
4	15	open	$c=0$
5	24	open	$c=0/3=0$
	26	open	
	46	open	
6	-	-	-



Average and Global clustering coefficient

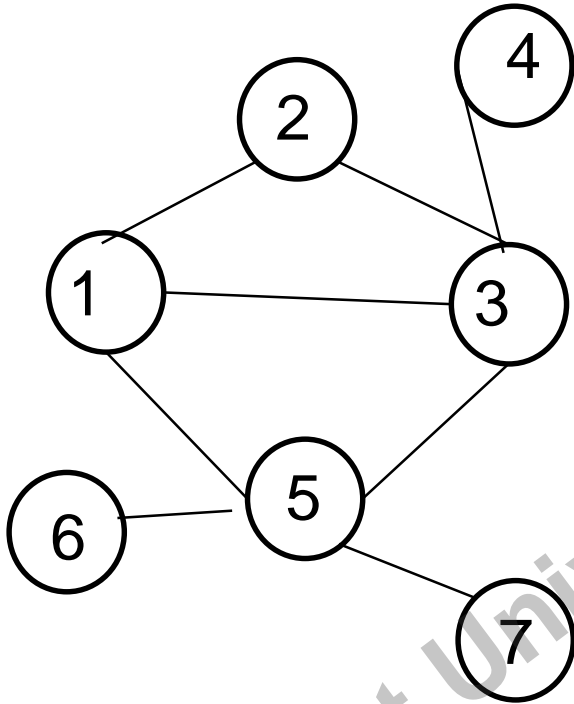
For the previous example, the average clustering is $1/3$ while the global clustering is $3/11$.

These two common measures of clustering can differ. Here the average clustering is higher than the overall clustering, it can also go the other way.

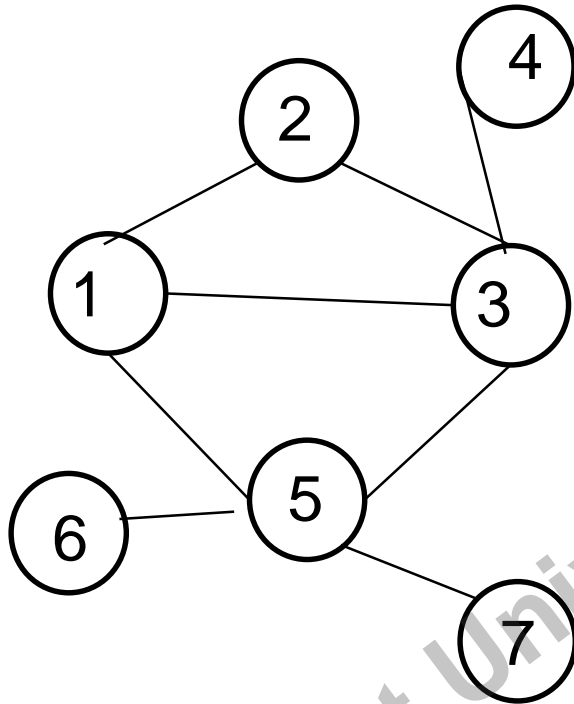
Moreover, it is possible to generate networks where the two measures can produce very different numbers for the same network.



Exercise



Exercise



Pairs of neighbours

Connected?

213

yes

215

no

315

yes

$c(1)=2/3$

123

yes

$c(2)=1/1$

132

yes

134

no

135

yes

234

no

235

no

435

no

$c(3)=2/6$

153

yes

156

no

157

no

356

no

357

no

657

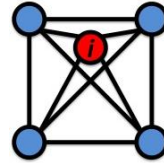
no

$c(5)=1/6$

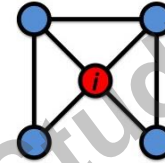


Clustering coefficient in random network

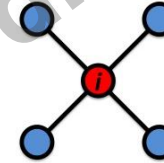
$$C_i \circ \frac{2 \langle L_i \rangle}{k_i(k_i - 1)}$$



$C_i = 1$



$C_i = 1/2$



$C_i = 0$

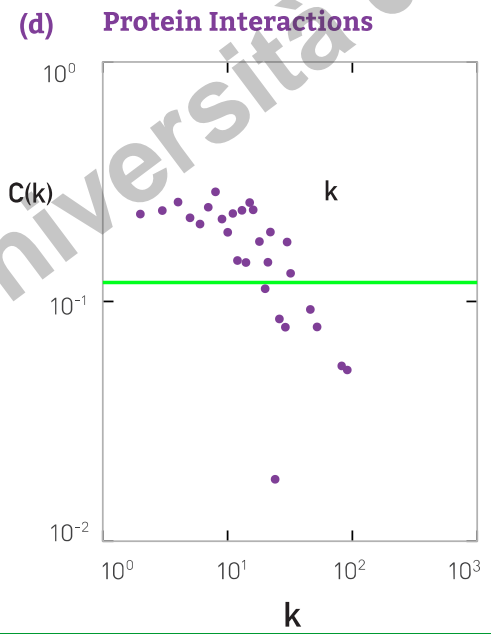
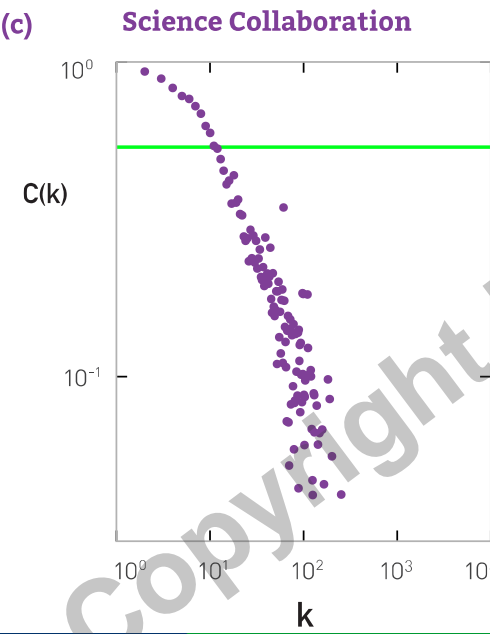
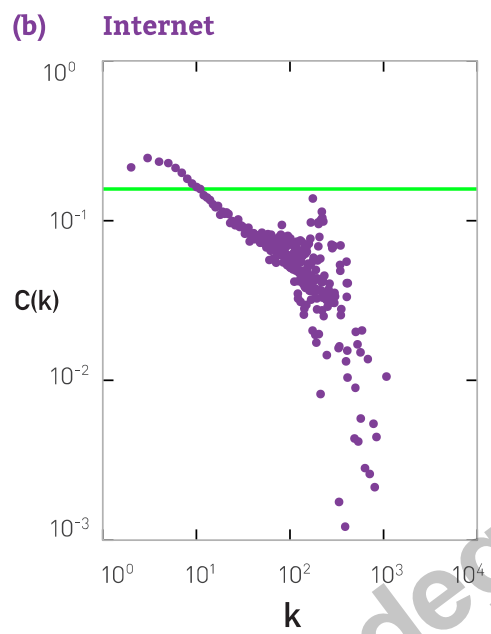
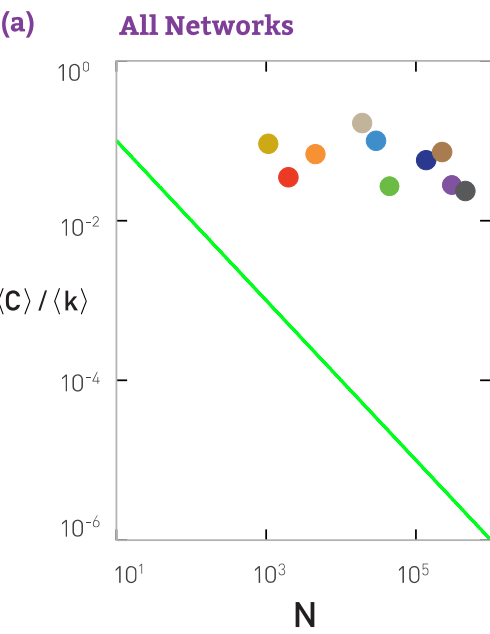
L_i represents the number of links between the k_i neighbors of node i .

Since edges are independent and have the same probability p ,

$$\langle L_i \rangle @ p \frac{k_i(k_i - 1)}{2} \quad \Rightarrow \quad C_i = \frac{2 \langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$

- The clustering coefficient of random graphs is small.
- C is independent of a node's degree k .





Comparing the average clustering coeff. of real networks with the prediction for random networks

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}$$

C decreases with the system size N .

C is independent of a node's degree k .

The random network model does not capture the clustering of real networks.

Instead real networks have a much higher clustering coefficient than expected for a random network of similar N and L .



Clustering coefficient

Random networks

The clustering coefficient of random graphs is small.

For fixed degree C decreases with the system size N .

C is independent of a node's degree k .

$$C_i = \frac{\langle k \rangle}{N}$$

Real networks

A much higher clustering coefficient than expected for a random network of similar N and L .

Independent of N

High-degree nodes tend to have a smaller clustering coefficient than low-degree nodes.



Clustering coefficient

A clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterized by a relatively high density of ties; this likelihood tends to be greater than the average probability of a tie randomly established between two nodes.



Clustering Coefficient

- In real-world networks, friendships are highly transitive, i.e., friends of an individual are often friends with one another
 - These friendships form triads -> high average [local] clustering coefficient
- In May 2011, Facebook had an average clustering coefficient of 0.5 for individuals who had 2 friends.

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
0.081	0.14 (with 100 friends)	0.31	0.33	0.17	0.13



Credits

Reza Zafarani, Mohammad Ali Abbasi, Huan Liu
Social Media Mining: An Introduction
A Textbook by Cambridge University Press
Chapter 3.2.1

Newman, M.E.J.
Networks: An Introduction.
Oxford University Press. 2010.
Chapter 7.9

Albert-László Barabási
Network Science
Chapter 3.9





UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Transitivity
Global and Local
clustering coefficient
in undirected networks



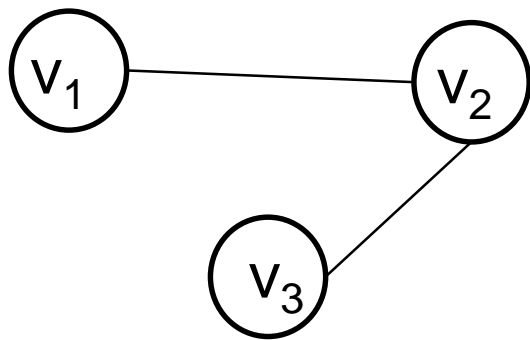
Transitivity

Mathematic representation:

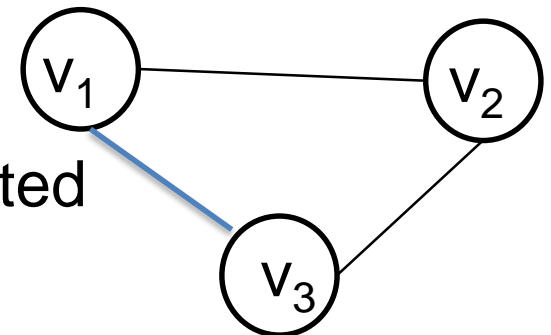
- For a transitive relation R: $aRb \wedge bRc \rightarrow aRc$

Networks:

- the transitive relation R: connected by a link
- If v_1, v_2 are connected and v_2, v_3 are connected



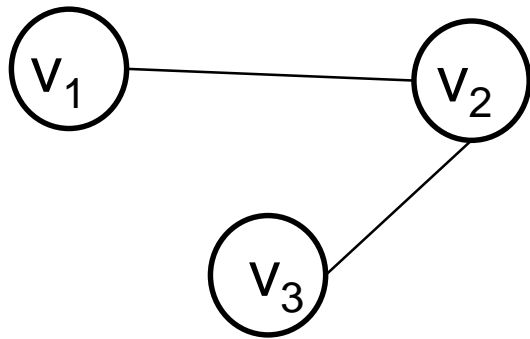
v_1, v_3 are connected



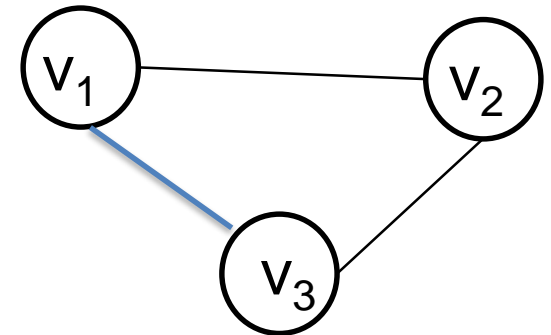
Transitivity

Social Networks:

- the transitive relation R : friendship
- If v_1, v_2 are friends and v_2, v_3 are friends



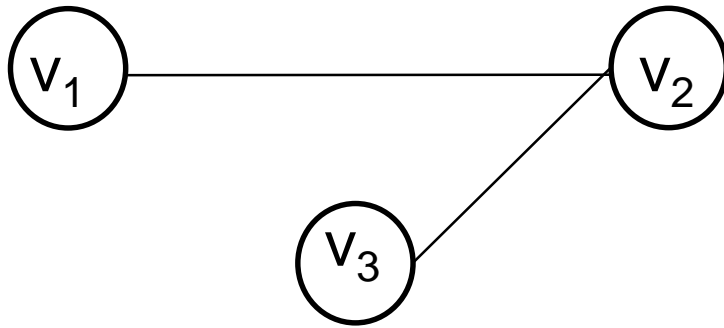
v_1, v_3 are friends



***Transitivity is when
a friend of my friend is my friend***



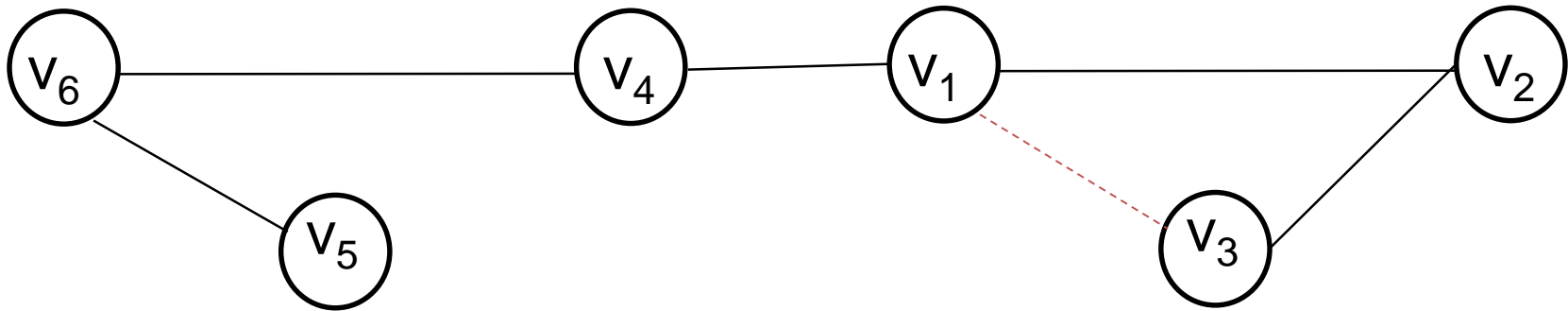
Transitivity



- Perfect transitivity only occurs in networks where each component is a fully connected graph or clique (a subgraph in which all nodes are connected to all others)
- Perfect transitivity is a useless concept in social networks as it never occurs



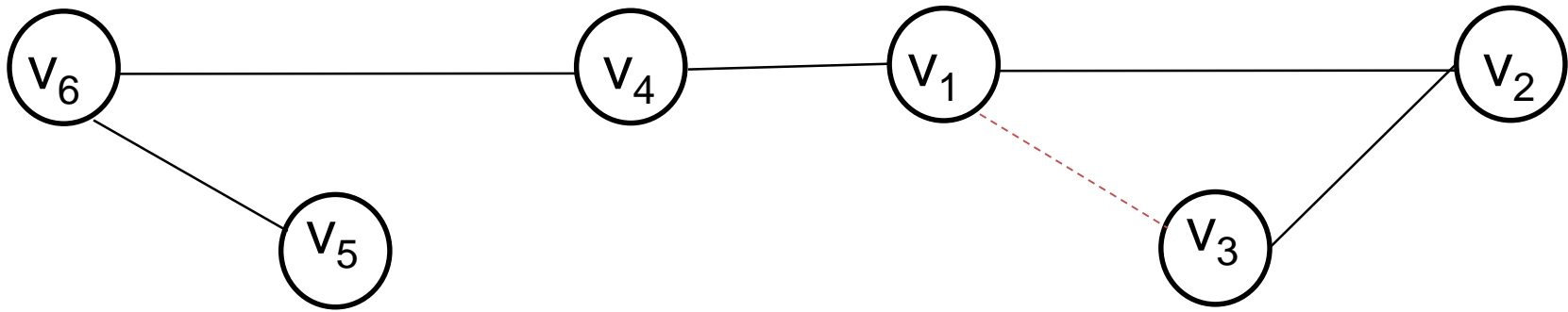
Transitivity



- Partial transitivity is much more useful
- The friend of my friend is not guaranteed to be my friend
- But is far more likely to be my friend than any other node in the network.
- v_1 is more likely to be friend of v_3 than v_5
- Is v_1 more likely to be friend of v_3 than v_6 ?



Transitivity



- Partial transitivity is much more useful
- The friend of my friend is not guaranteed to be my friend
- But is far more likely to be my friend than any other node in the network.
- v_1 is more likely to be friend of v_3 than v_5
- v_1 has the same probability to be friend of v_3 and v_6

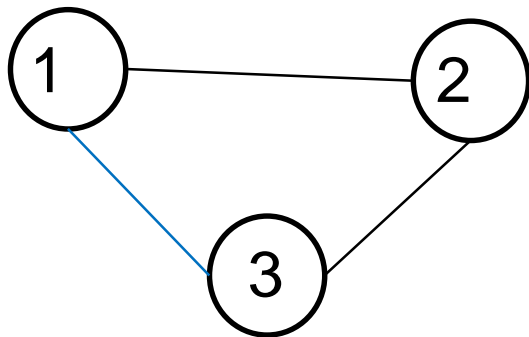


Global Clustering Coefficient Measure based on paths

We want to quantify the level of transitivity of a network

We can measure it by counting the paths of length two and check whether the third edge exists

$$C = \frac{|\text{Paths of Length 2 that have the third edge}|}{|\text{Paths of Length 2}|}$$



A path of length 2 which has the third link is called *closed path* as it forms a loop of length 3. [Closed paths are also called *closed triad in social networks*.]



Global Clustering Coefficient

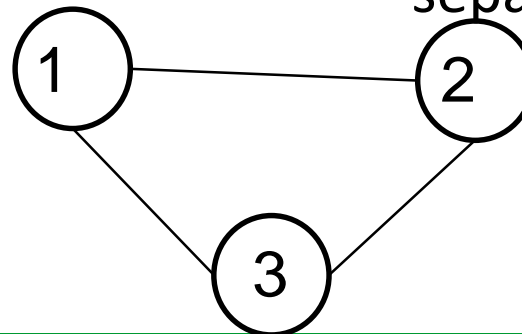
$$C = \frac{|\text{Paths of Length 2 that have the third edge}|}{|\text{Paths of Length 2}|}$$

Note that paths, also closed paths, have a direction in undirected network, too.

Two paths that traverse the same links but in opposite direction are counted separately.

Path of length 2	Third edge
213	32
312	23
123	31
321	13
132	21
231	12

$$C=6/6$$



Global Clustering Coefficient

$$C = \frac{|\text{Paths of Length 2 that have the third edge}|}{|\text{Paths of Length 2}|}$$

Path of length 2

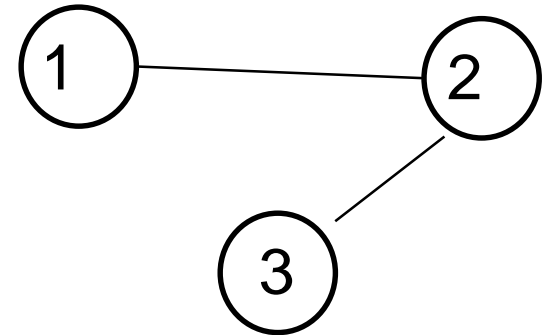
123

321

Third edge

-

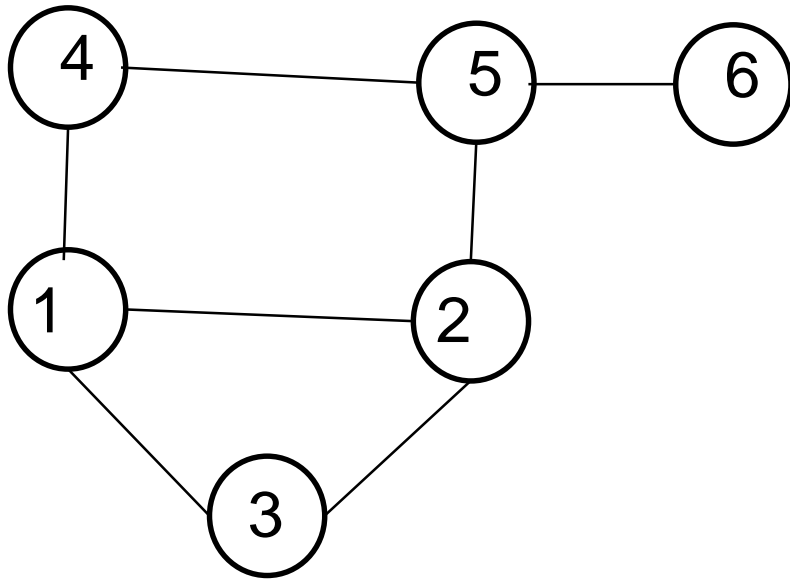
-



$$C = 0/2 = 0$$



Example



$$C = 6/22 = 3/11$$

Path of length 2

2**1**3 312

214 412

314 413

1**2**3 321

125 521

325 523

1**3**2 231

1**4**5 541

2**5**4 452

256 652

456 654

Third edge

yes

no

no

yes

no

no

yes

no

no

no

no

[Note: you could divide both the numerator and the denominator by two, by considering paths in one direction only]



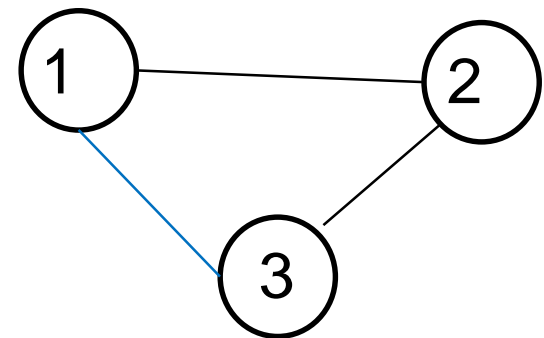
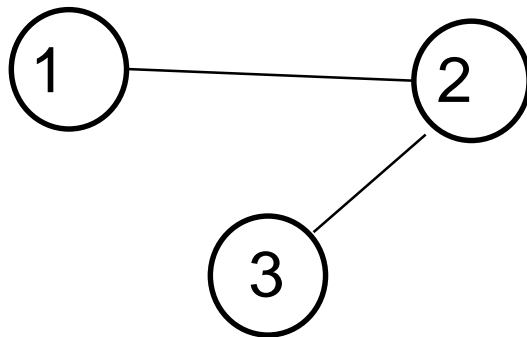
Global Clustering Coefficient

Measure based on triples

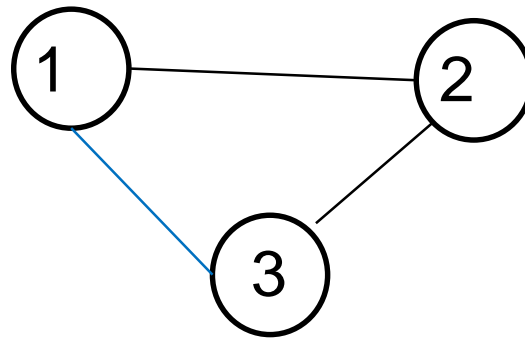
If we have a path $\{1,2,3\}$ (and $\{3,2,1\}$) of length 2, it is also true to say that nodes 1 and 3 have a common neighbour: node 2.

If the triad $\{1,2,3\}$ is closed, nodes 1 and 3 are themselves friends.

The clustering coefficient can be thought as the **fraction of pairs of people with a common friend who are themselves friend.**



Clustering coefficient



a friend of my friend is my friend

pairs of people with a common friend who are themselves friend.

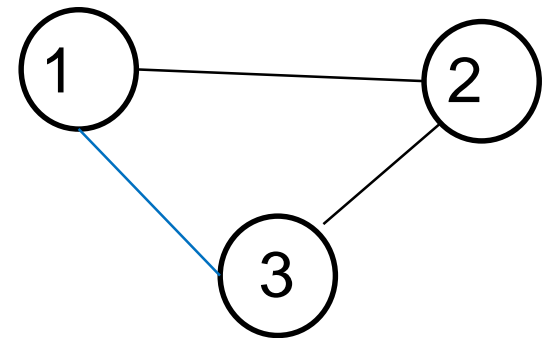
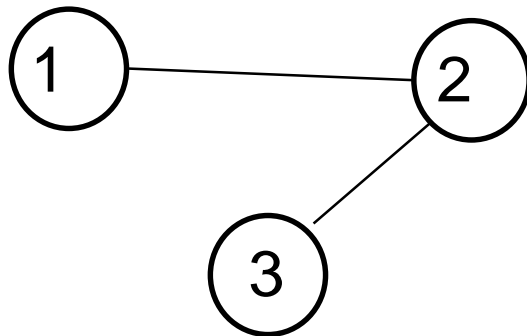


Global Clustering Coefficient Measure based on triplets

The clustering coefficient can be thought as the fraction of pairs of people with a common friend who are themselves friend.

We can also define the global clustering coefficient based on the concept of (connected) triplets of nodes.

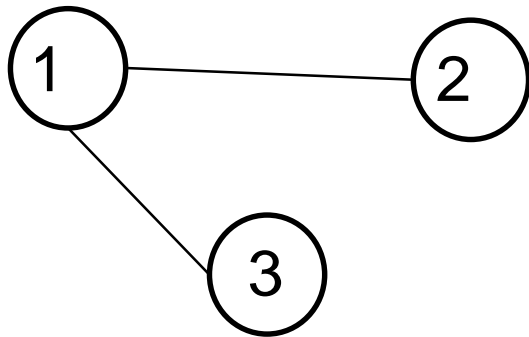
A connected triplet consists of three nodes $\{v_1, v_2, v_3\}$, that are connected by the two links (v_1, v_2) and (v_2, v_3) . The third link (v_1, v_3) can be present (closed triplet) or not (open triplet).



Global Clustering Coefficient Measure based on triples

Triples (open)

$\{2,1,3\}$ [with links $(2,1)$ and $(1,3)$]

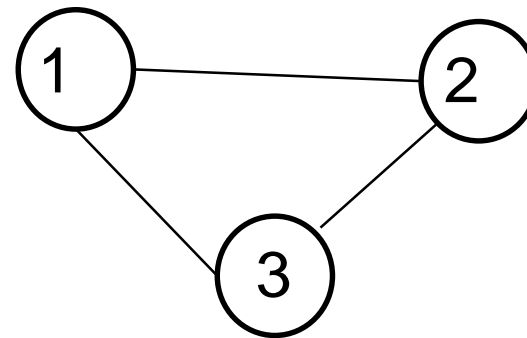


Triples (closed)

$\{2,1,3\}$ [with links $(2,1)$ and $(1,3)$]

$\{1,2,3\}$ [with links $(1,2)$ and $(2,3)$]

$\{1,3,2\}$ [with links $(1,3)$ and $(3,2)$]



Global Clustering Coefficient Measure based on triples

The global clustering coefficient is the number of closed triplets over the total number of triplets (both open and closed):

$$\frac{\text{number of closed triplets}}{\text{total number of triplets}}$$

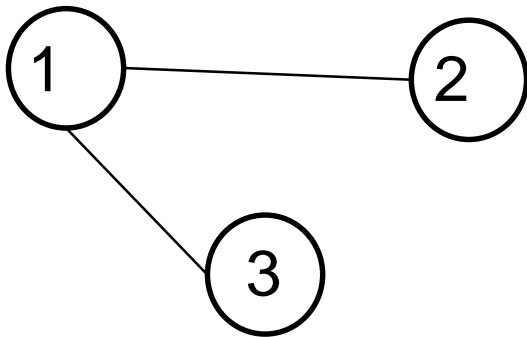


Global Clustering Coefficient Measure based on triples

Triples (open)

213

$$C=0/2=0$$



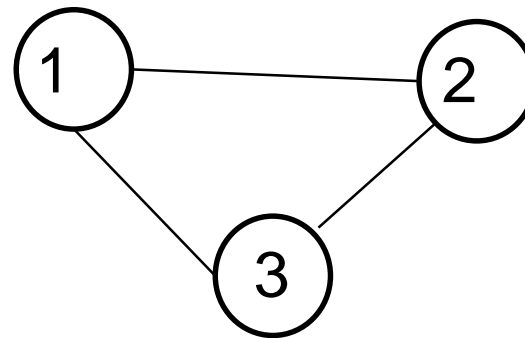
Triples (closed)

213

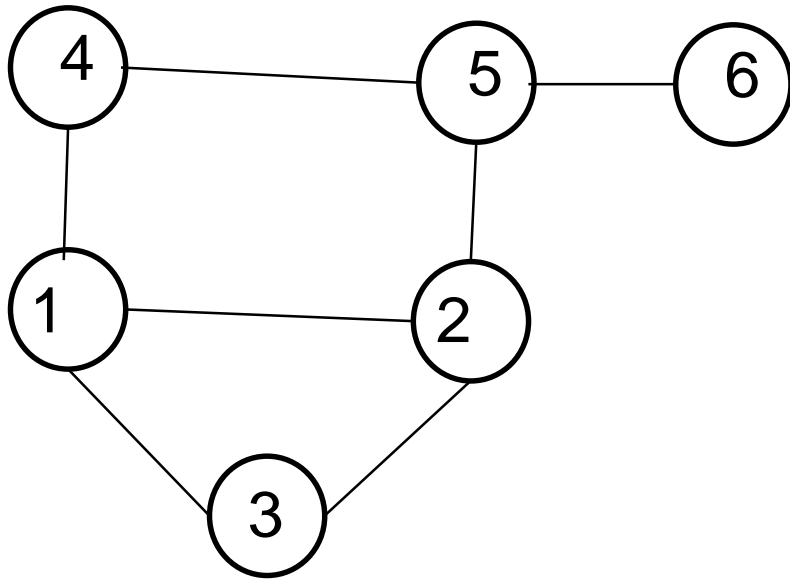
123

132

$$C=3/3=1$$



Example



$C=3/11$

Triplets

213

Closed?

yes

214

no

314

no

123

yes

125

no

325

no

132

yes

145

no

254

no

256

no

456

no



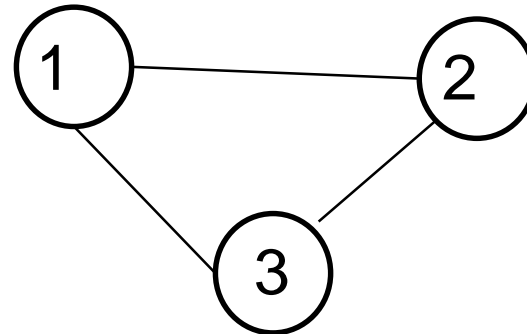
Global Clustering Coefficient Measure based on triangles

A triangle consists of six paths.

213, 312, 123, 321, 132, 231

A triangle consists of three triples, one centered on each of the nodes.

213, 123, 231



Global Clustering Coefficient

Measure based on triangles

The global clustering coefficient is the number of closed paths of length 2 (or 6 x triangles) over the total number of paths of length 2

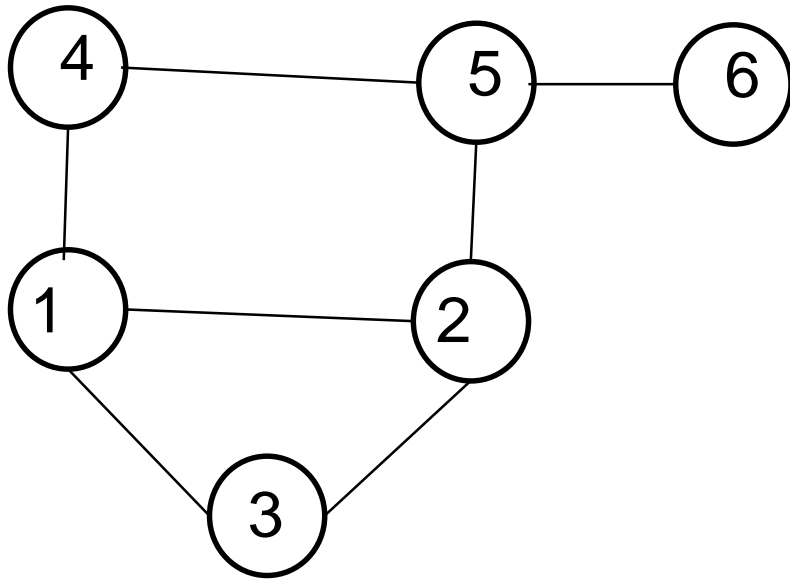
$$\frac{6 * \text{number of triangles}}{\text{number of paths of length 2}}$$

The global clustering coefficient is the number of closed triplets (or 3 x triangles) over the total number of triplets

$$\frac{3 * \text{number of triangles}}{\text{number of triplets}}$$



Example



$$C = 1 * 6 / 22 = 3 / 11$$

Path of length 2

2**1**3 312

214 412

314 413

1**2**3 321

125 521

325 523

1**3**2 231

1**4**5 541

2**5**4 452

256 652

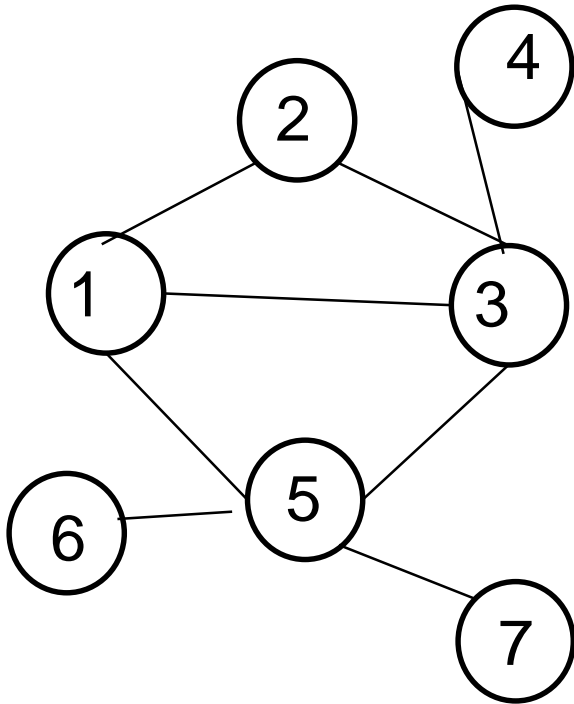
456 654

Triangles

123



Exercise



Local Clustering Coefficient

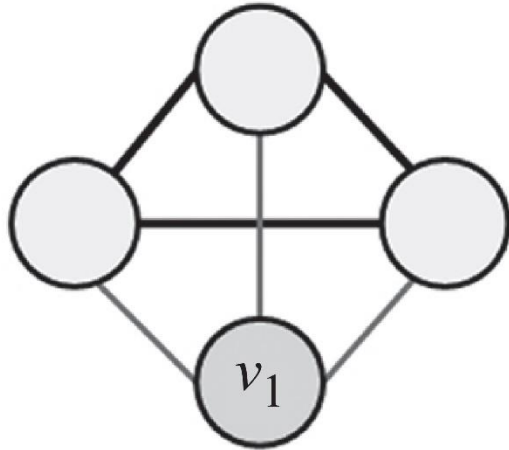
- Local clustering coefficient measures transitivity at the node level
- Commonly employed for undirected graphs, it computes how strongly neighbors of a node v (nodes adjacent to v) are themselves connected

$$C(v_i) = \frac{\text{number of pairs of neighbors of } v_i \text{ that are connected}}{\text{number of pairs of neighbors of } v_i}.$$

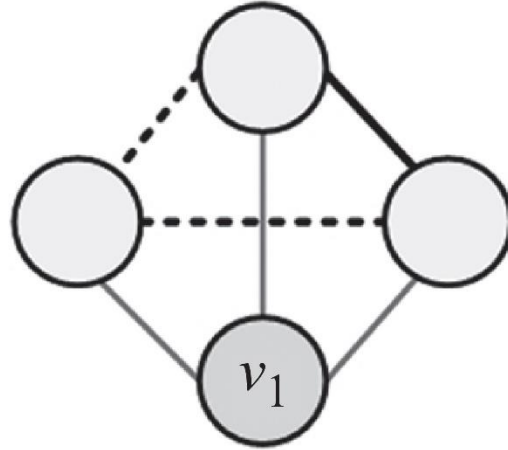
In an undirected graph, the denominator can be rewritten as: $\binom{d_i}{2} = d_i(d_i - 1)/2,$



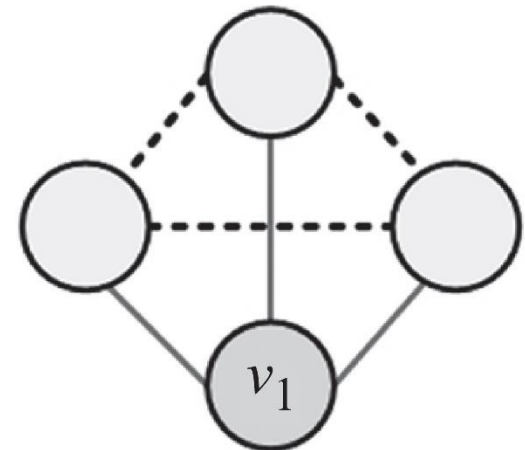
Local Clustering Coefficient:



$$C(v_1) = 1$$



$$C(v_1) = 1/3$$

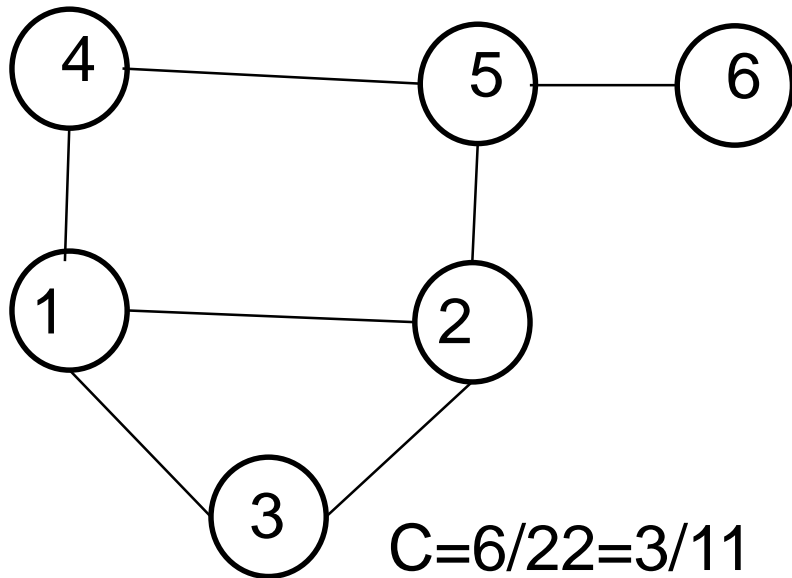


$$C(v_1) = 0$$

- Thin lines depict connections to neighbors
- Dashed lines are the missing connections among neighbors
- Solid lines indicate connected neighbors
 - When none of neighbors are connected $C=0$
 - When all neighbors are connected $C=1$



Example



Node	Pairs of friends	Open or closed triad	Clustering coeff
1	23	closed	$c=1/3$
	24	open	
	34	open	
2	13	closed	$c=1/3$
	15	open	
	35	open	
3	12	closed	$c=1/1$
4	15	open	$c=0$
5	24	open	$c=0/3=0$
	26	open	
	46	open	
6	-	-	-



Example

The clustering coefficient distribution therefore is:

c	Frequency
0	2/5
1/3	2/5
1	1/5

The mean clustering coefficient is: $1/3$

Node	Pairs of friends	Open or closed triad	Clustering coeff
1	23	closed	$c=1/3$
	24	open	
	34	open	
2	13	closed	$c=1/3$
	15	open	
	35	open	
3	12	closed	$c=1/1$
	15	open	$c=0$
	24	open	$c=0/3=0$
4	26	open	
	46	open	
	6	-	-



Average and Global clustering coefficient

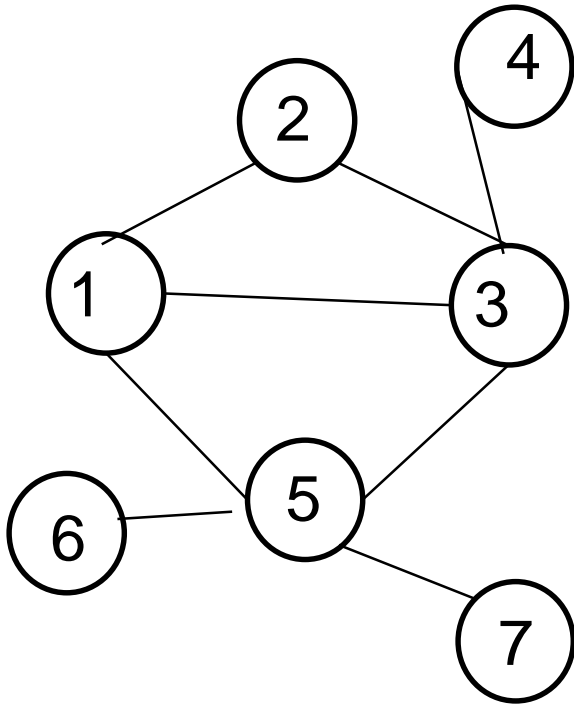
For the previous example, the average clustering is $1/3$ while the global clustering is $3/11$.

These two common measures of clustering can differ. Here the average clustering is higher than the overall clustering, it can also go the other way.

Moreover, it is possible to generate networks where the two measures can produce very different numbers for the same network.

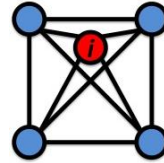


Exercise

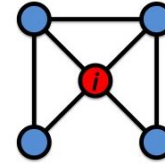


Clustering coefficient in random network

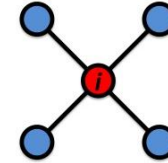
$$C_i = \frac{2 \langle L_i \rangle}{k_i(k_i - 1)}$$



$C_i = 1$



$C_i = 1/2$



$C_i = 0$

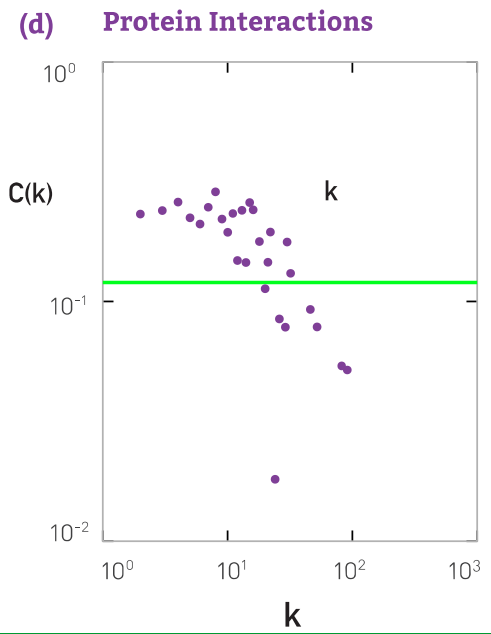
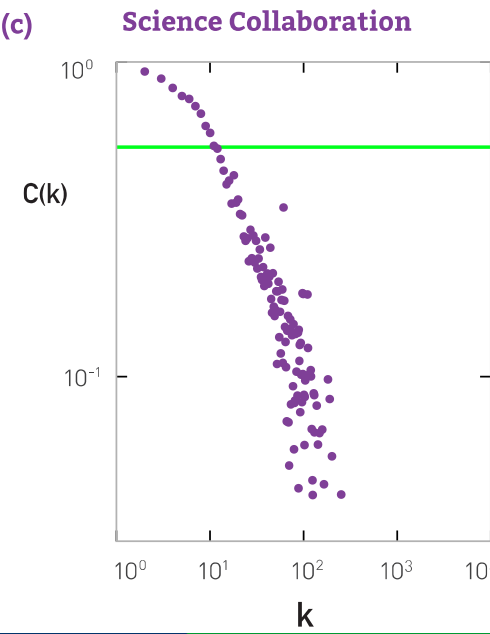
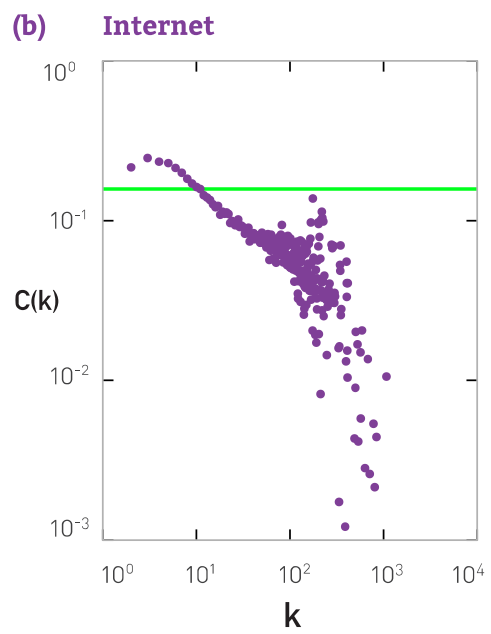
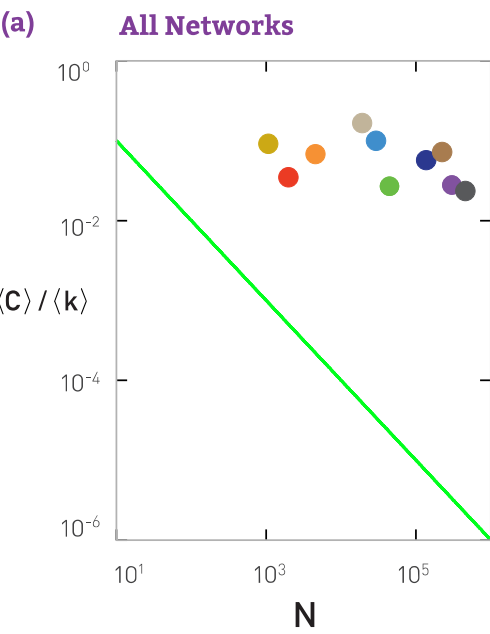
L_i represents the number of links between the k_i neighbors of node i .

Since edges are independent and have the same probability p ,

$$\langle L_i \rangle = p \frac{k_i(k_i - 1)}{2} \quad \Rightarrow \quad C_i = \frac{2 \langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}$$

- The clustering coefficient of random graphs is small.
- C is independent of a node's degree k .





Comparing the average clustering coeff. of real networks with the prediction for random networks

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}$$

C decreases with the system size N .

C is independent of a node's degree k .

The random network model does not capture the clustering of real networks. Instead real networks have a much higher clustering coefficient than expected for a random network of similar N and L .



Clustering coefficient

Random networks

The clustering coefficient of random graphs is small.

For fixed degree C decreases with the system size N .

C is independent of a node's degree k .

$$C_i = \frac{\langle k \rangle}{N}$$

Real networks

A much higher clustering coefficient than expected for a random network of similar N and L .

Independent of N

High-degree nodes tend to have a smaller clustering coefficient than low-degree nodes.



Clustering coefficient

A clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterized by a relatively high density of ties; this likelihood tends to be greater than the average probability of a tie randomly established between two nodes.



Clustering Coefficient

- In real-world networks, friendships are highly transitive, i.e., friends of an individual are often friends with one another
 - These friendships form triads -> high average [local] clustering coefficient
- In May 2011, Facebook had an average clustering coefficient of 0.5 for individuals who had 2 friends.

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
0.081	0.14 (with 100 friends)	0.31	0.33	0.17	0.13



Credits

Reza Zafarani, Mohammad Ali Abbasi, Huan Liu
Social Media Mining: An Introduction
A Textbook by Cambridge University Press
Chapter 3.2.1

Newman, M.E.J.
Networks: An Introduction.
Oxford University Press. 2010.
Chapter 7.9

Albert-László Barabási
Network Science
Chapter 3.9





UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Strong and weak ties

Bridging the local and the global

The strength of weak ties

Granovetter's paper

- Mark Granovetter (born October 20, 1943): an American sociologist and professor at Stanford University.
- 1969: submitted his paper to the American Sociological Review—rejected!
- 1972, submitted a shortened version to the American Journal of Sociology—published in 1973 (Granovetter, 1973).
- According to Current Contents, by 1986, the Weak Ties paper had become a citation classic, being one of the most cited papers in sociology



Granovetter's paper

Bridging the local and the global

“A fundamental weakness of current sociological theory is that it does not relate micro-level interactions to macro-level patterns in any convincing way. Large-scale statistical, as well as qualitative, studies offer a good deal of insight into such macro phenomena as social mobility, community organization, and political structure. At the micro level, a large and increasing body of data and theory offers useful and illuminating ideas about what transpires within the confines of the small group. But how interaction in small groups aggregates to form large-scale patterns eludes us in most cases. I will argue, in this paper, that the analysis of processes in interpersonal networks provides the most fruitful micro-macro bridge. In one way or another, it is through these networks that small-scale interaction becomes translated into large-scale patterns, and that these, in turn, feed back into small groups.»



THE STRENGTH OF TIES

“Most intuitive notions of the "strength" of an interpersonal tie should be satisfied by the following definition:

the strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie.

Each of these is somewhat independent of the other, though the set is obviously highly intracorrelated.

Discussion of operational measures of weights attaching to each of the four elements is postponed to future empirical studies.

It is sufficient for the present purpose if most of us can agree, on a rough intuitive basis, whether a given tie is strong, weak, or absent.»



Granovetter's experiment

- Granovetter interviewed people about how they discovered their jobs
- Most people did so through personal contacts, often described as acquaintances and not close friends

WHY?

- Basic intuition: close friends are part of triad closures and would know what you know and would know others who would know what you know
- "It is the distant acquaintances who are actually to thank for crucial information leading to your new job, rather than your close friends!"



From dyads to small structures

Triadic closure

“The hypothesis which enables us to relate dyadic ties to larger structures is:

The stronger the tie between A and B, the larger the proportion of individuals to whom they will both be tied, that is, connected by a weak or strong tie.

This overlap in their friendship circles is predicted to be least when their tie is absent, most when it is strong, and intermediate when it is weak.”

Motivations: amount of time spent together, similarity



From dyads to small structures

Triadic closure

“The theory of cognitive balance, as formulated by Heider (1958) and especially by Newcomb (1961, pp. 4-23), also predicts this result.

If strong ties A-B and A-C exist, and if B and C are aware of one another, anything short of a positive tie would introduce a "psychological strain" into the situation since C will want his own feelings to be congruent with those of his good friend, A, and similarly, for B and his friend, A.

Where the ties are weak, however, such consistency is psychologically less crucial.”



Bridging local to global

“To derive implications for large networks of relations, it is necessary to frame the basic hypothesis more precisely.

This can be done by investigating the possible triads consisting of strong, weak, or absent ties among A, B, and any arbitrarily chosen friend of either or both”

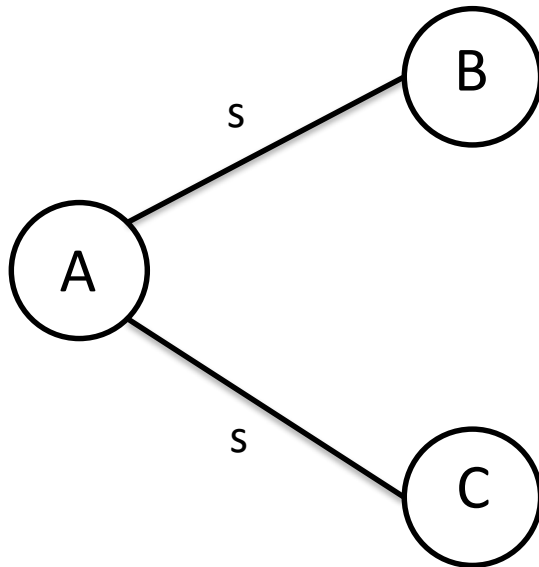


Strong triadic closure

A more extreme version of the triadic closure

Strong Triadic Closure Property (Granovetter):

If a node A has two strong links (to B and C) then a link (strong or weak) must exist between B and C.



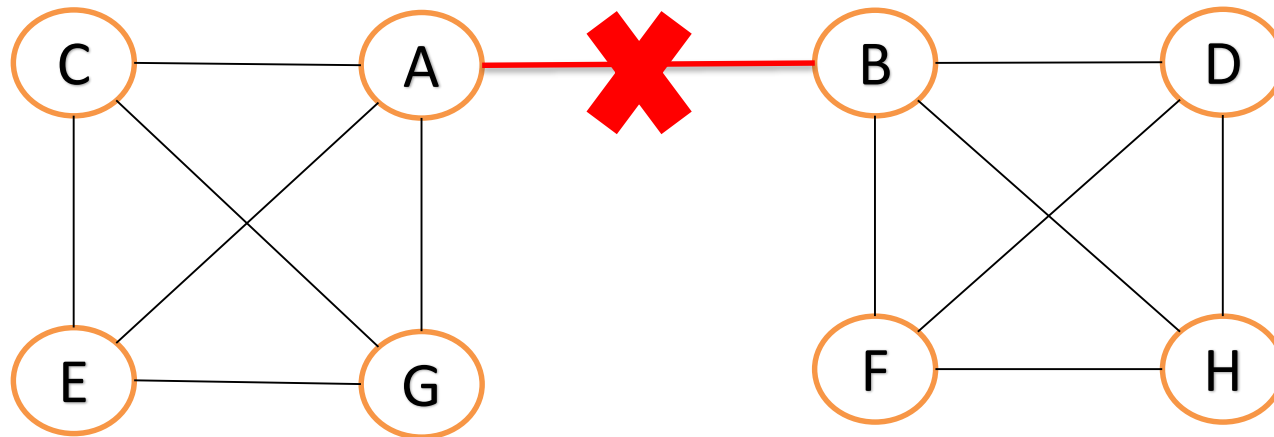
Forbidden triad



Bridge

Let us now introduce another important concept:
bridges

Edge between A and B is a bridge if, when deleted, it would make A and B lie in 2 different components

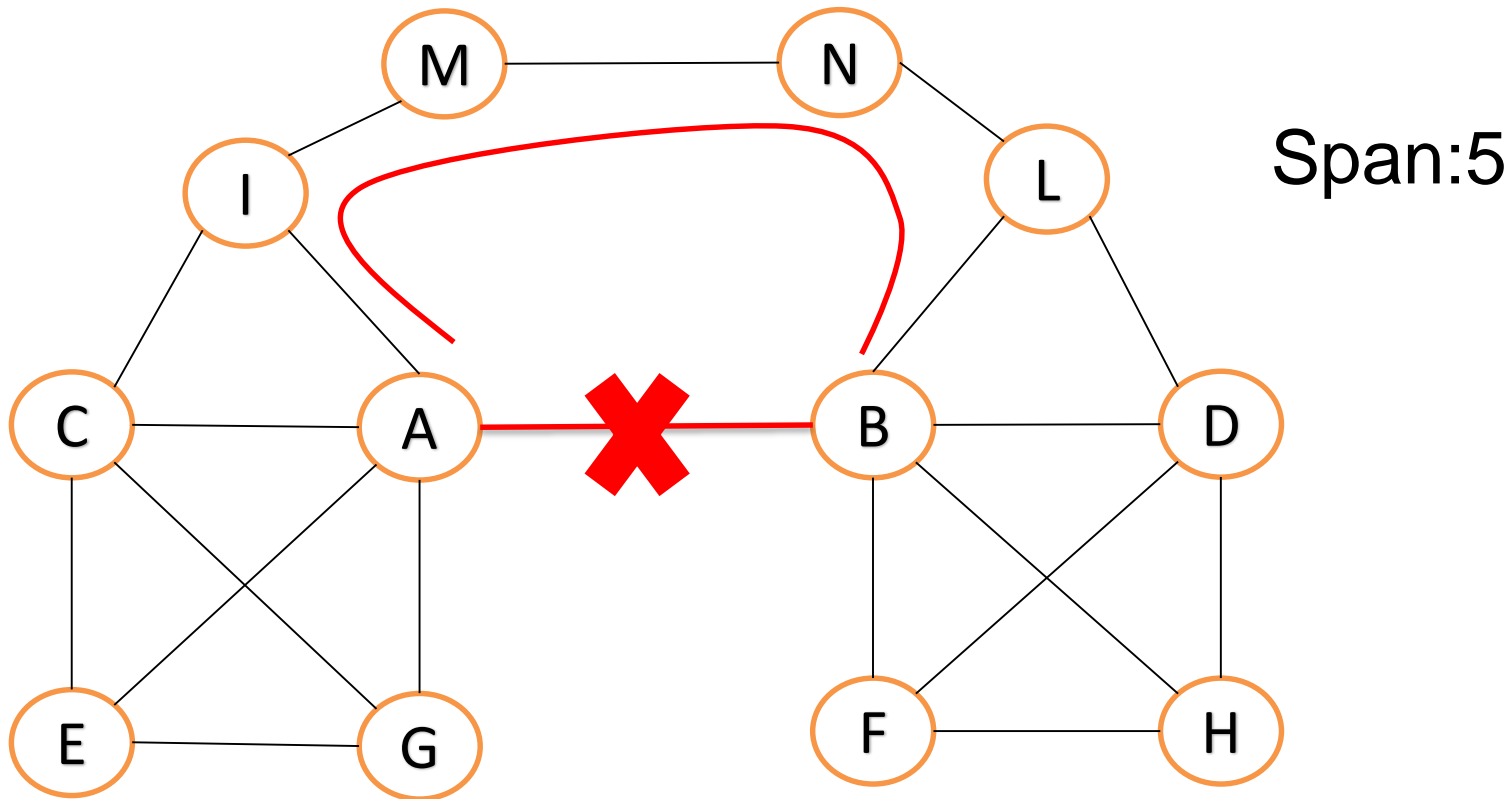


Bridges are presumably extremely rare in real social networks.



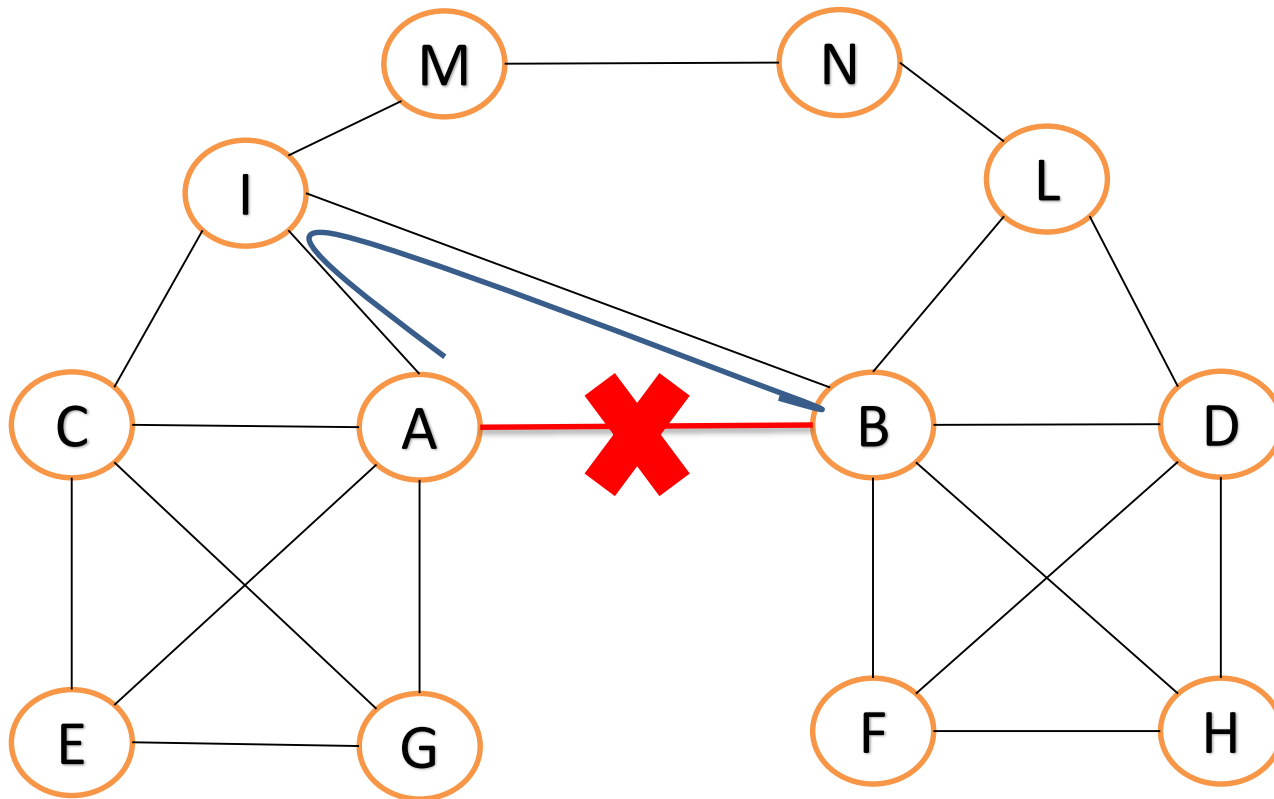
Local bridge

An edge is a local bridge if its endpoints have no friends in common – If deleting the edge would increase the distance of the endpoints to a value more than 2.



Local bridge

An edge is a local bridge if its endpoints have no friends in common – If deleting the edge would increase the distance of the endpoints to a value more than 2.



Triangle
→ not a local
bridge

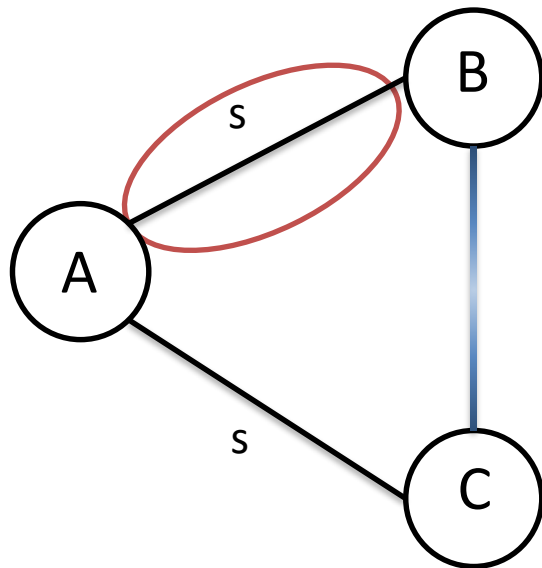


Bridges are weak ties

If node A satisfies the STCP and is involved in at least two strong ties, then any local bridge it is involved in must be a weak tie.

Proof by contradiction (AB is strong and is a bridge)

“Consider the strong tie A-B: if A has another strong tie to C, then forbidding the triad of figure 1 implies that a tie exists between C and B, so that the path A-C-B exists between A and B; hence, A-B is not a bridge.



Weak ties suffer no such restriction, though they are certainly not automatically bridges.

What is important, rather, is that *all bridges are weak ties.*”



The strength of weak ties

Intuitively speaking, this means that whatever is to be diffused can reach a larger number of people, and traverse greater social distance (i.e., path length), when passed through weak ties rather than strong.

If one tells a rumor to all his close friends, and they do likewise, many will hear the rumor a second and third time, since those linked by strong ties tend to share friends.



Almost local bridge

Since a very small fraction of the links in social networks are local bridges, it makes sense to soften this definition

We define the *neighborhood overlap* of an edge connecting A and B to be the ratio:

number of nodes who are neighbors of both A and B

number of nodes who are neighbors of at least one of A or B

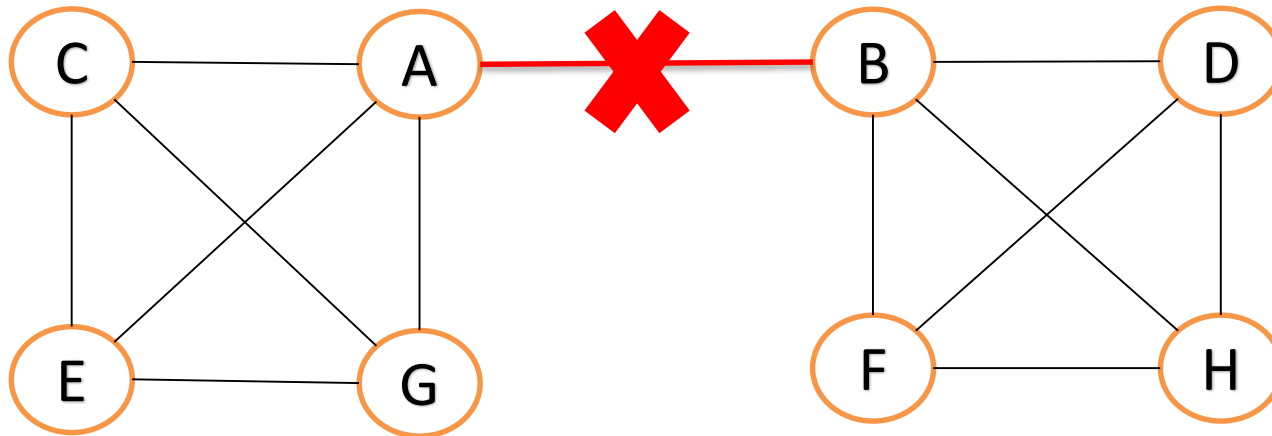
where in the denominator we don't count A or B themselves



Almost local bridge

Neighborhood overlap:

$$O(A, B) = \frac{n(A) \cap n(B)}{n(A) \cup n(B)}$$

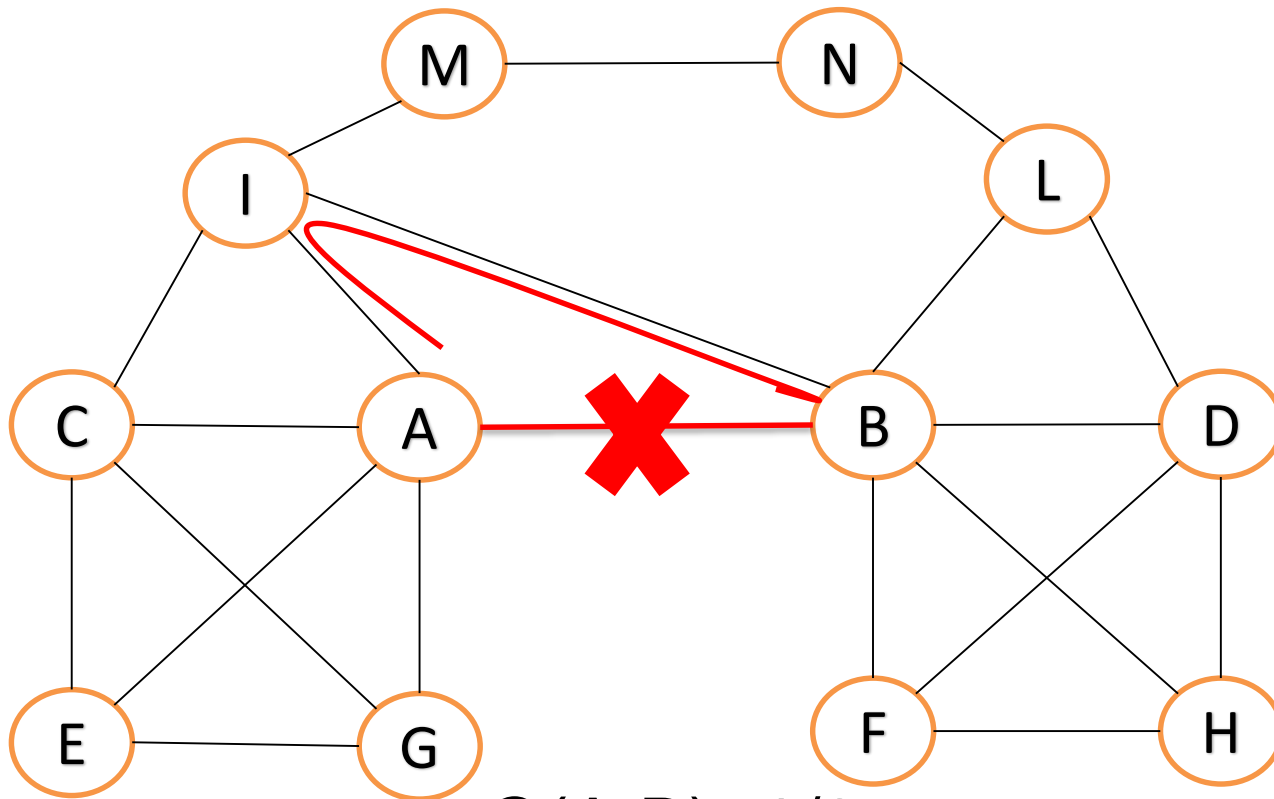


$$O(A, B) = 0$$



Almost local bridge

Neighborhood overlap: $O(A, B) = \frac{n(A) \cap n(B)}{n(A) \cup n(B)}$



edges with very small neighborhood overlap are almost "local bridges."

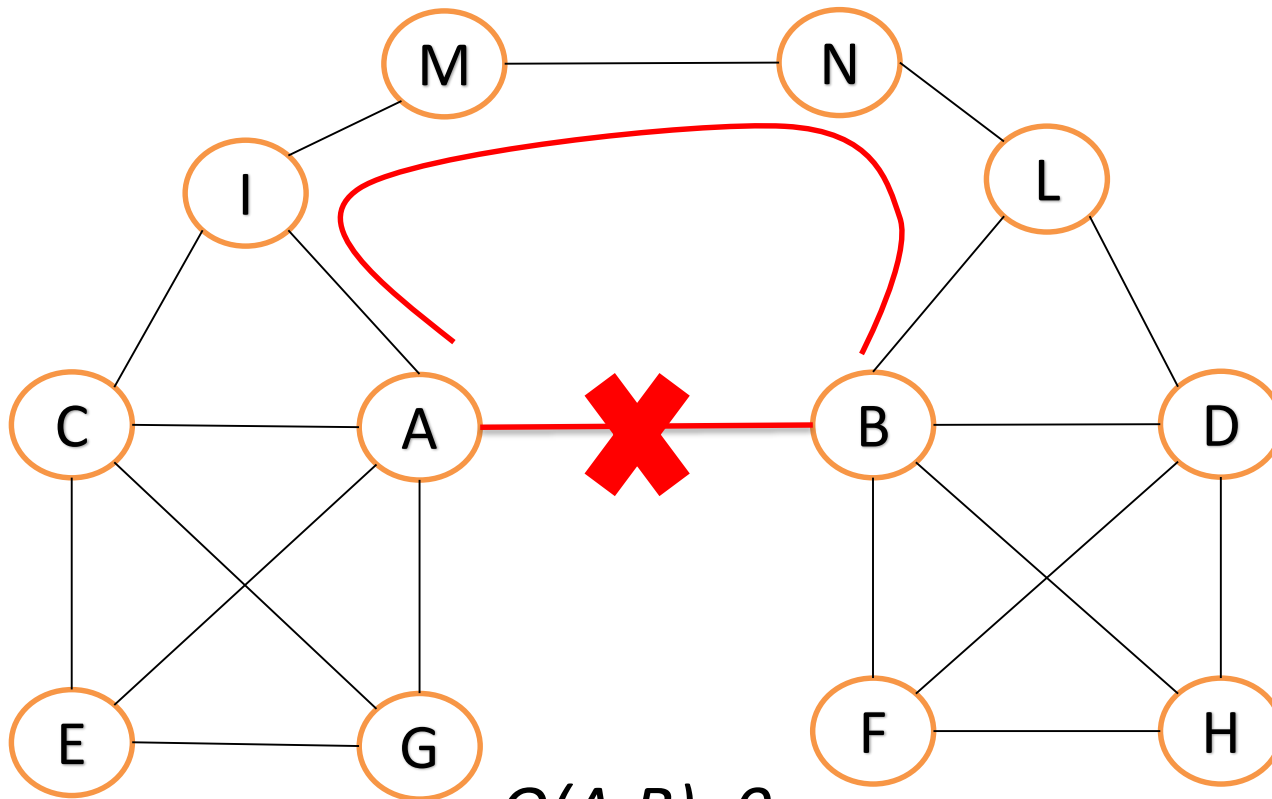
$$O(A, B) = 1/9$$



Almost local bridge

Neighborhood overlap:

$$O(A, B) = \frac{n(A) \cap n(B)}{n(A) \cup n(B)}$$



Local bridge are almost local bridges

$$O(A, B) = 0$$



Case-study: mobile phone networks

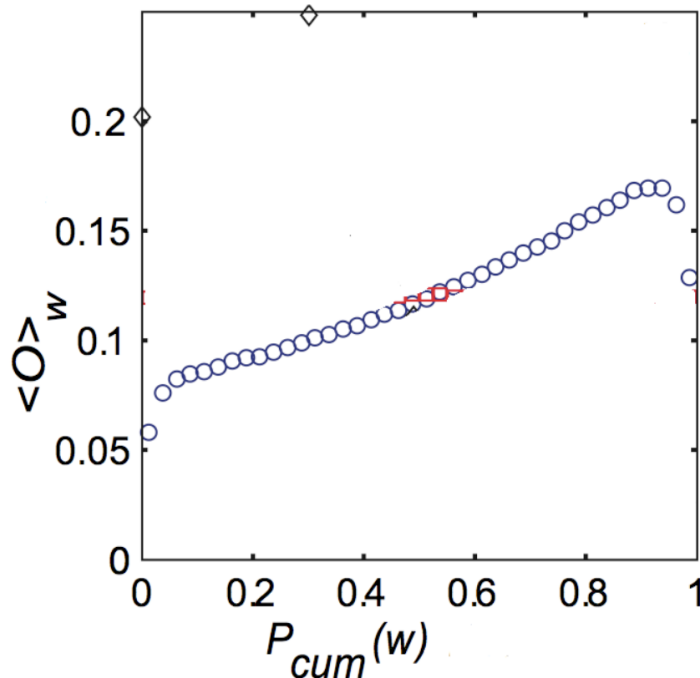


Figure 3.7: A plot of the neighborhood overlap of edges as a function of their percentile in the sorted order of all edges by tie strength. The fact that overlap increases with increasing tie strength is consistent with the theoretical predictions from Section 3.2. (Image from [334].)

Networks, Crowds, and Markets:
Reasoning About a Highly Connected
World

David Easley e Jon Kleinberg
Cambridge University Press, 2010
Chapter 3

Structure and tie strengths in mobile
communication networks

JP Onnela, J Saramäki, J Hyvönen, G Szabó,
D Lazer, K Kaski, J Kertész, A-L Barabási
Proceedings of the National Academy of
Sciences 104 (18), 7332, 2007

Strength:
aggregated duration



Case-study: Facebook

Three categories of links based on usage over a one-month observation period.

- A link represents reciprocal (mutual) communication, if the user both sent messages to the friend at the other end of the link, and also received messages from them during the observation period.
- A link represents one-way communication if the user sent one or more messages to the friend at the other end of the link (whether or not these messages were reciprocated).
- A link represents a maintained relationship if the user followed information about the friend at the other end of the link, whether or not actual communication took place;

Networks, Crowds, and Markets: Reasoning About a Highly Connected World
David Easley e Jon Kleinberg - Cambridge University Press, 2010 - Chapter 3

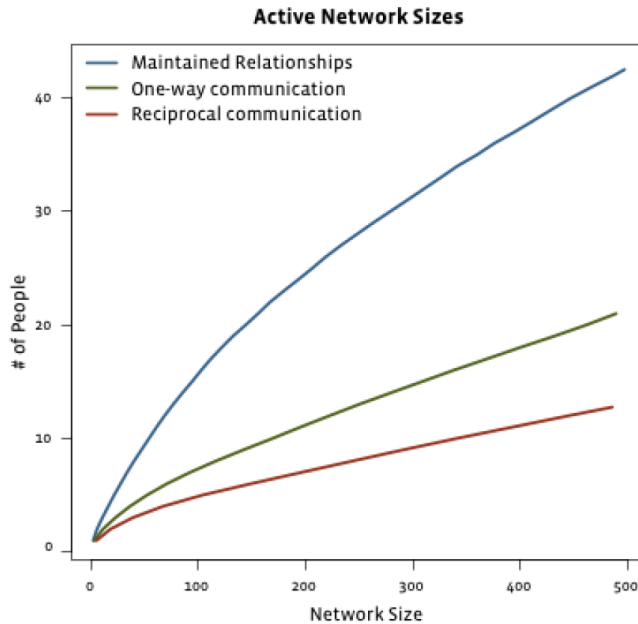
Facebook:

Cameron Marlow

<http://overstated.net/2009/03/09/maintained-relationships-on-facebook>



Case-study: Facebook



Passive engagement: even for users who report very large numbers of friends on their profile pages (on the order of 500), the number with whom they actually communicate is generally between 10 and 20, and the number they follow even passively (e.g. by reading about them) is under 50

Figure 3.9: The number of links corresponding to maintained relationships, one-way communication, and reciprocal communication as a function of the total neighborhood size for users on Facebook. (Image from [286].)

“The stark contrast between reciprocal and passive networks shows the effect of technologies such as News Feed. If these people were required to talk on the phone to each other, we might see something like the reciprocal network, where everyone is connected to a small number of individuals. Moving to an environment where everyone is passively engaged with each other, some event, such as a new baby or engagement can propagate very quickly through this highly connected network.”



Case-study: Facebook

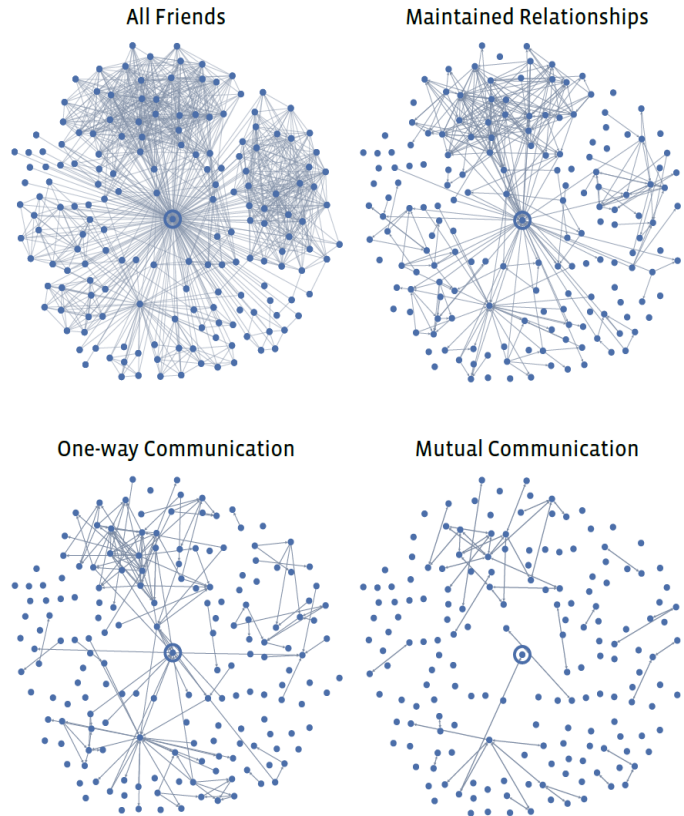
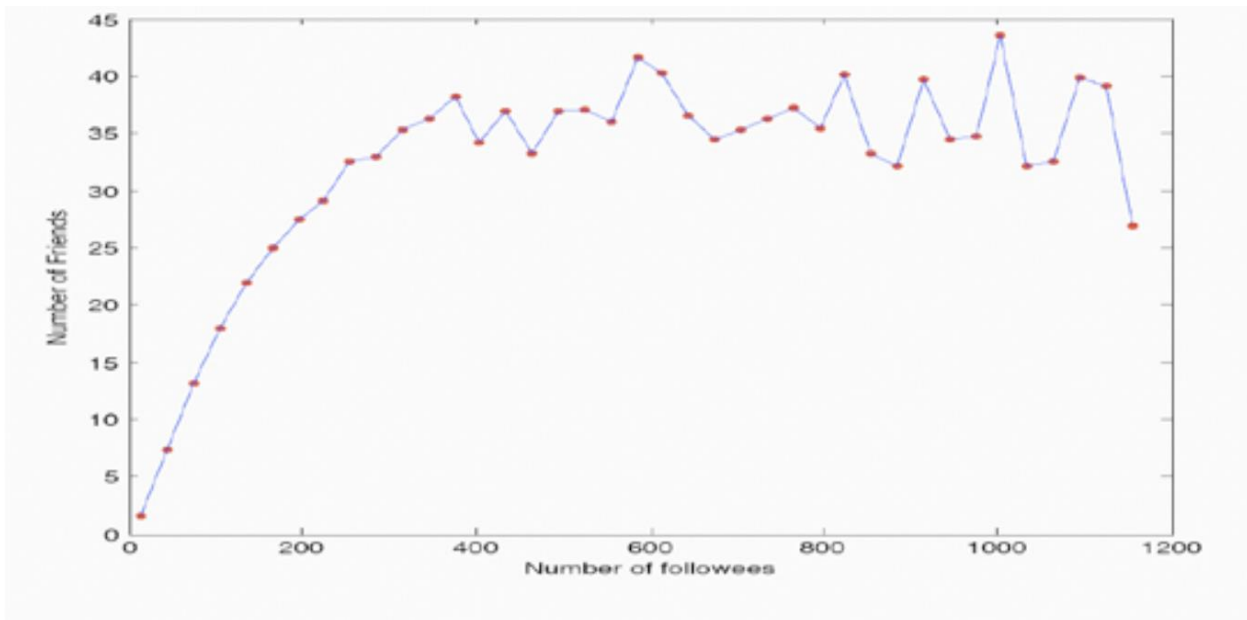


Figure 3.8: Four different views of a Facebook user's network neighborhood, showing the structure of links corresponding respectively to all declared friendships, maintained relationships, one-way communication, and mutual communication. (Image from [286].)



Case-study: Twitter



Even for users who maintain very large numbers of weak ties on-line, the number of strong ties remains relatively modest, in this case stabilizing at a value below 50 even for users with over 1000 followees.

Figure 3.10: The total number of a user's strong ties (defined by multiple directed messages) as a function of the number of followees he or she has on Twitter. (Image from [222].)

Networks, Crowds, and Markets: Reasoning About a Highly Connected World
David Easley e Jon Kleinberg - Cambridge University Press, 2010 - Chapter 3

Huberman, Bernardo A. and Romero, Daniel M. and Wu, Fang, Social Networks that Matter: Twitter Under the Microscope (December 5, 2008).



Strong and weak ties in social networks

Mobile communication networks:

Structure and tie strengths in mobile communication networks

JP Onnela, J Saramäki, J Hyvönen, G Szabó, D Lazer, K Kaski, J Kertész, A-L Barabási

Proceedings of the National Academy of Sciences 104 (18), 7332, 2007

Facebook:

Cameron Marlow

<http://overstated.net/2009/03/09/maintained-relationships-on-facebook>

Twitter:

Huberman, Bernardo A. and Romero, Daniel M. and Wu, Fang, Social Networks that Matter: Twitter Under the Microscope (December 5, 2008).

Available at

SSRN: <https://ssrn.com/abstract=1313405> or <http://dx.doi.org/10.2139/ssrn.1313405>



Sources

The Strength of Weak Ties

Author(s): Mark S. Granovetter

Source: American Journal of Sociology, Vol. 78, No. 6 (May, 1973), pp. 1360-1380

Published by: The University of Chicago Press

Stable URL: <http://www.jstor.org/stable/2776392>

Accessed: 15-06-2015 01:58 UTC

Networks, Crowds, and Markets: Reasoning About a Highly Connected World

David Easley e Jon Kleinberg

Cambridge University Press, 2010

Chapter 3

Reza Zafarani

Social Media Mining

Chapter 2





UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

Six degree of separation

Small-world networks

Small-world model

Six degree of separation

The small-world phenomenon

Stanley Milgram's experiment (1960)



- Random people from Nebraska were asked to send a letter (via intermediaries) to a stock broker in Boston
- S/he could only send to someone with whom they were on a first-name basis

Among the letters that reached the target, the average path length was six.



Stanley Milgram's experiment (1960)

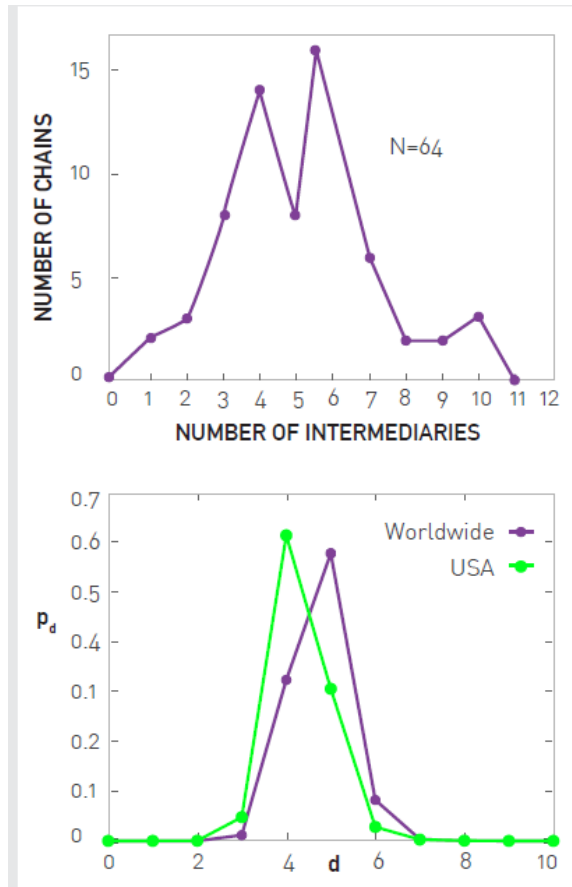


Figure 3.12
Six Degrees? From Milgram to Facebook

Figure 3.12

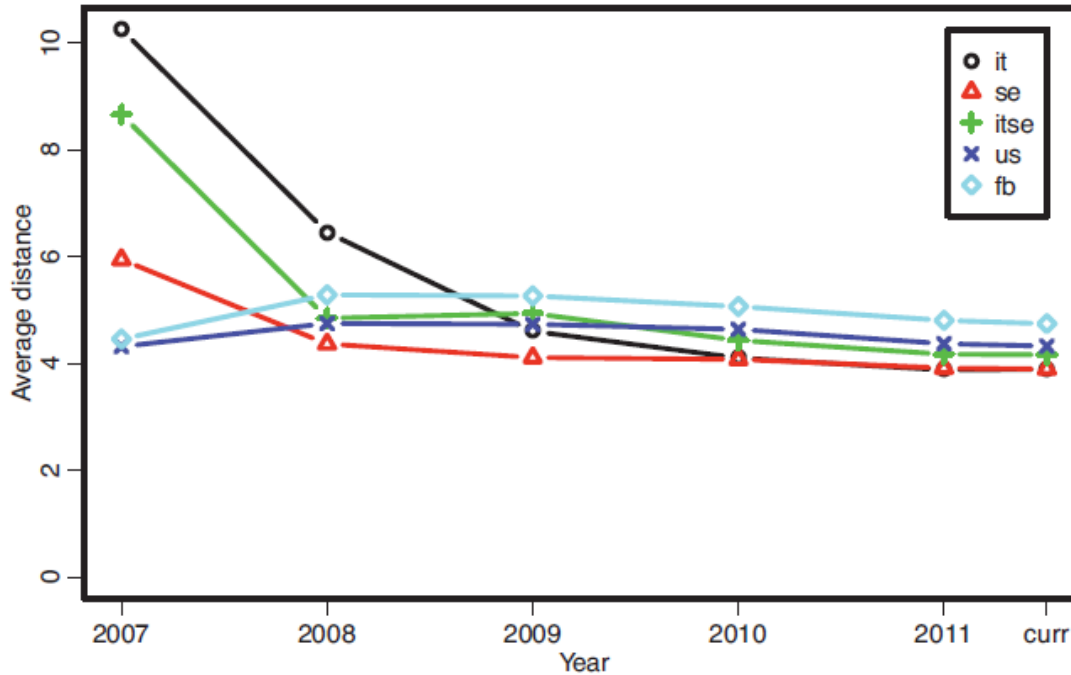
Six Degrees? From Milgram to Facebook

- (a) In Milgram's experiment 64 of the 296 letters made it to the recipient. The figure shows the length distribution of the completed chains, indicating that some letters required only one intermediary, while others required as many as ten. The mean of the distribution was 5.2, indicating that on average six 'handshakes' were required to get a letter to its recipient. The playwright John Guare renamed this 'six degrees of separation' two decades later. After [25].
- (b) The distance distribution, p_d , for all pairs of Facebook users worldwide and within the US only. Using Facebook's N and L (3.19) predicts the average degree to be approximately 3.90, not far from the reported four degrees. After [18].

Source: Barabasi's book



Facebook: four degree «of separation



“We decided to extend our experiments in two directions: regional and temporal. We thus analyse the entire Facebook graph (fb), the SA subgraph (us), the Italian subgraph (it) and the Swedish (se) subgraph. We also analysed a combination of the Italian and Swedish graph (itse) to check whether combining two regional but distant networks could significantly change the average distance, in the same spirit as in the original Milgram’s”

Figure 3. The average distance graph. See also Table 6.

In May 2011, the average path length between individuals in the Facebook graph was 4.7. (4.3 for individuals in the US)

Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. 2012. Four degrees of separation. In Proceedings of the 4th Annual ACM Web Science Conference (WebSci '12). Association for Computing Machinery, New York, NY, USA, 33–42. DOI:<https://doi.org/10.1145/2380718.2380723>



Average distance in social networks

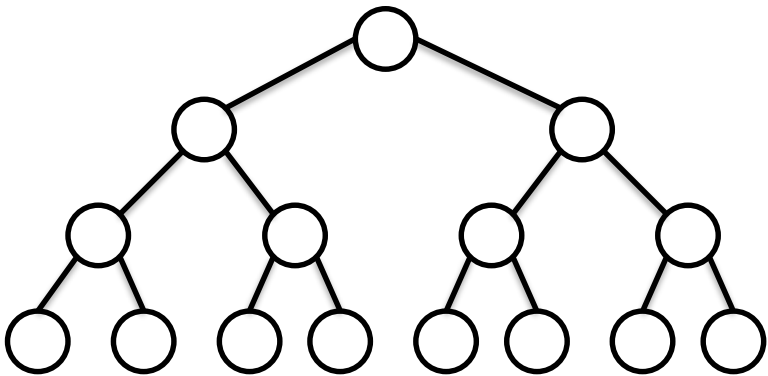
Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
16.12	4.7	5.67	5.88	4.25	5.10



The small-world model

Short paths

Should we be surprised by the fact that the paths between random pairs of people in social networks are so short?



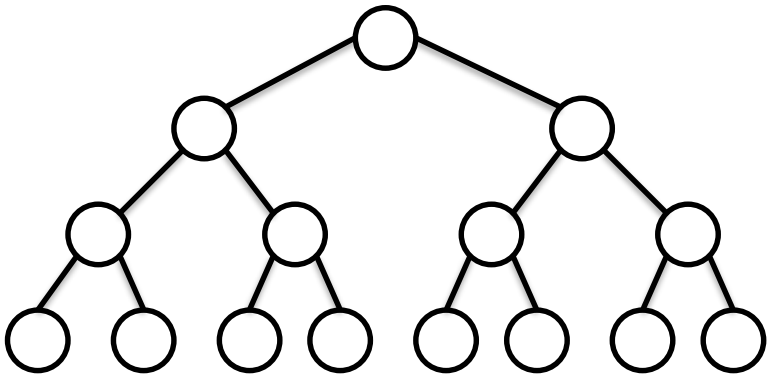
Suppose each of us knows more than 100 other people on a first-name basis (in fact, for most people, the number is significantly larger). Then, taking into account the fact that each of your friends has at least 100 friends other than you, you could in principle be two steps away from over $100 * 100 = 10000$ people. Taking into account the 100 friends of these people brings us to more than $100 * 100 * 100 = 1000000$ people who in principle could be three steps away.

In other words, the numbers are growing by powers of 100 with each step, bringing us to 100 million after four steps, and 10 billion after five steps.



Random graph: diameter

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



$\langle k \rangle$ nodes at distance one ($d=1$).

$\langle k \rangle^2$ nodes at distance two ($d=2$).

$\langle k \rangle^3$ nodes at distance three ($d=3$).

...

$\langle k \rangle^d$ nodes at distance d .

$$N = 1 + \langle k \rangle + \langle k \rangle^2 + \dots + \langle k \rangle^{d_{\max}} = \frac{\langle k \rangle^{d_{\max} + 1} - 1}{\langle k \rangle - 1} \gg \langle k \rangle^{d_{\max}} \Rightarrow$$

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



Small World

In most networks this offers a better approximation to the average distance between two randomly chosen nodes, $\langle d \rangle$, than to d_{\max} .

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

Small world phenomena: the property that the average path length or the diameter depends logarithmically on the system size.

”Small” means that $\langle d \rangle$ is proportional to $\log N$



NETWORK	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\log N}{\log \langle k \rangle}$
Internet	192,244	609,066	6.33	6.98	20	6.58
WWW	223,729	1,497,134	4.60	11.27	92	8.37
Power Grid	2,947	6,094	2.57	16.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.32	39	11.42
Email	107,794	103,731	1.61	6.88	38	18.4
Science Collaboration	22,112	91,239	4.08	6.26	71	4.91
Actor Network	702,388	29,397,908	62.71	3.91	14	3.04
Citation Network	449,673	4,707,958	10.43	11.21	42	9.55
E. Coli Metabolism	1,039	1,802	5.58	2.98	8	4.06
Protein Interactions	2,018	2,939	2.90	6.61	74	7.14

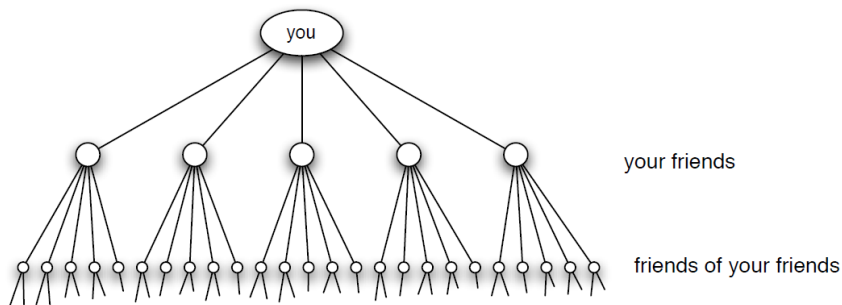
Given the huge differences in scope, size, and average degree, the agreement is excellent.



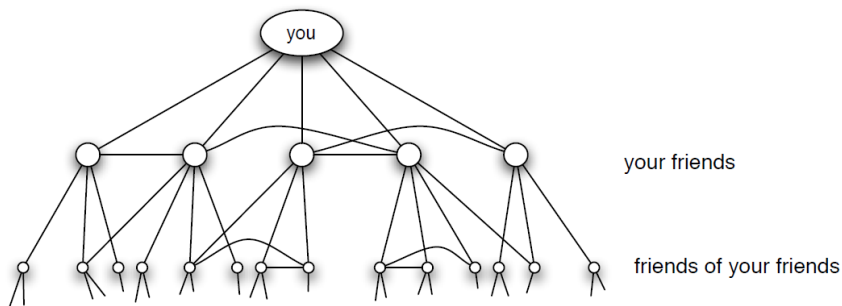
Do social networks deviate from this model?

“The difficulty already manifests itself with the second step, where we conclude that there may be more than 10000 people within two steps of you.

As we've seen, social networks abound in triangles (sets of three people who mutually know each other) and in particular, many of your 100 friends will know each other. As a result, when we think about the nodes you can reach by following edges from your friends, many of these edges go from one friend to another, not to the rest of world. The number 10000 came from assuming that each of your 100 friends was linked to 100 new people; and without this, the number of friends you could reach in two steps could be much smaller.”



(a) Pure exponential growth produces a small world

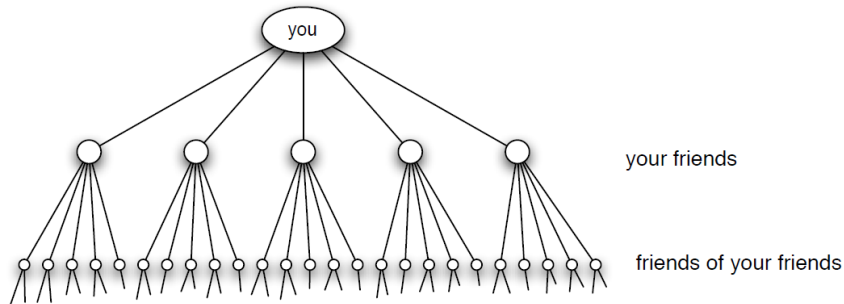


(b) Triadic closure reduces the growth rate

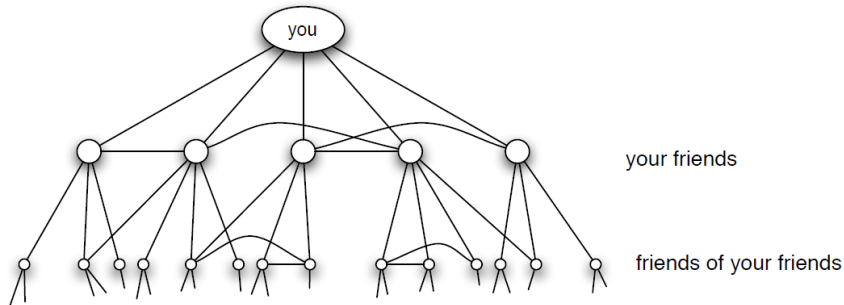
Figure 20.1: Social networks expand to reach many people in only a few steps.



Watts – Strogatz model (1998)



(a) *Pure exponential growth produces a small world*



(b) *Triadic closure reduces the growth rate*

Can we make up a simple model that exhibits both of the features we've been discussing: many closed triads, but also very short paths?

Figure 20.1: Social networks expand to reach many people in only a few steps.

Documentary: <https://www.cornell.edu/video/emergence-of-network-science>

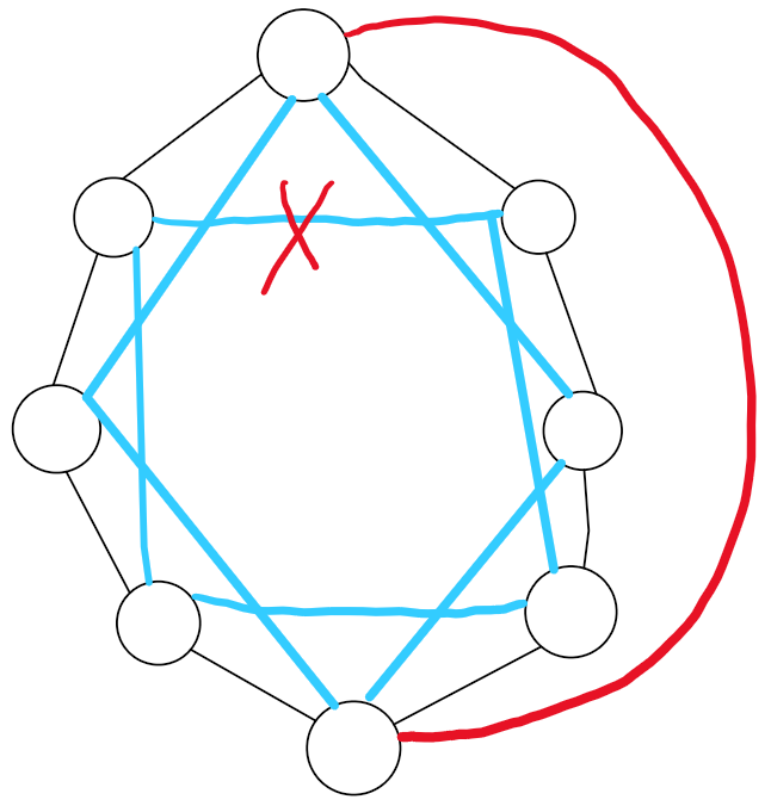
Start: six-degree of separation

Minutes 12-16: small-world model



Watts – Strogatz model (1998)

Regular lattice + rewiring



Triangles
+
Weak ties

High clustering
+
Short paths



Algorithm 4.1 Small-World Generation Algorithm

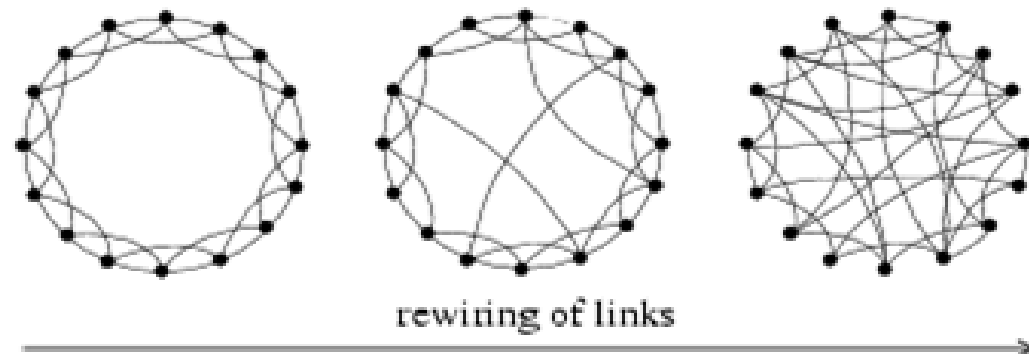
Require: Number of nodes $|V|$, mean degree c , parameter β

- 1: **return** A small-world graph $G(V, E)$
 - 2: $G =$ A regular ring lattice with $|V|$ nodes and degree c
 - 3: **for** node v_i (starting from v_1), and all edges $e(v_i, v_j), i < j$ **do**
 - 4: $v_k =$ Select a node from V uniformly at random.
 - 5: **if** rewiring $e(v_i, v_j)$ to $e(v_i, v_k)$ does not create loops in the graph or multiple edges between v_i and v_k **then**
 - 6: rewire $e(v_i, v_j)$ with probability β : $E = E - \{e(v_i, v_j)\}, E = E \cup \{e(v_i, v_k)\}$;
 - 7: **end if**
 - 8: **end for**
 - 9: **Return** $G(V, E)$
-

regular ring lattice
of degree c :
nodes are
connected to their
previous $c/2$ and
following $c/2$
neighbors.

As in many network generating
algorithms

- Disallow self-edges
- Disallow multiple edges



Case-studies

Network	Original Network				Simulated Graph	
	<i>Size</i>	<i>Average Degree</i>	<i>Average Path Length</i>	<i>C</i>	<i>Average Path Length</i>	<i>C</i>
Film Actors	225,226	61	3.65	0.79	4.2	0.73
Medline Coauthorship	1,520,251	18.1	4.6	0.56	5.1	0.52
E.Coli	282	7.35	2.9	0.32	4.46	0.31
C.Elegans	282	14	2.65	0.28	3.49	0.37



Credits

Reza Zafarani
Social Media Mining
Chapter 2

Barabasi
Network Science
Chapter 3

Networks, Crowds, and Markets: Reasoning About a Highly Connected World
David Easley e Jon Kleinberg
Cambridge University Press, 2010
Chapter 3



Balance and Status

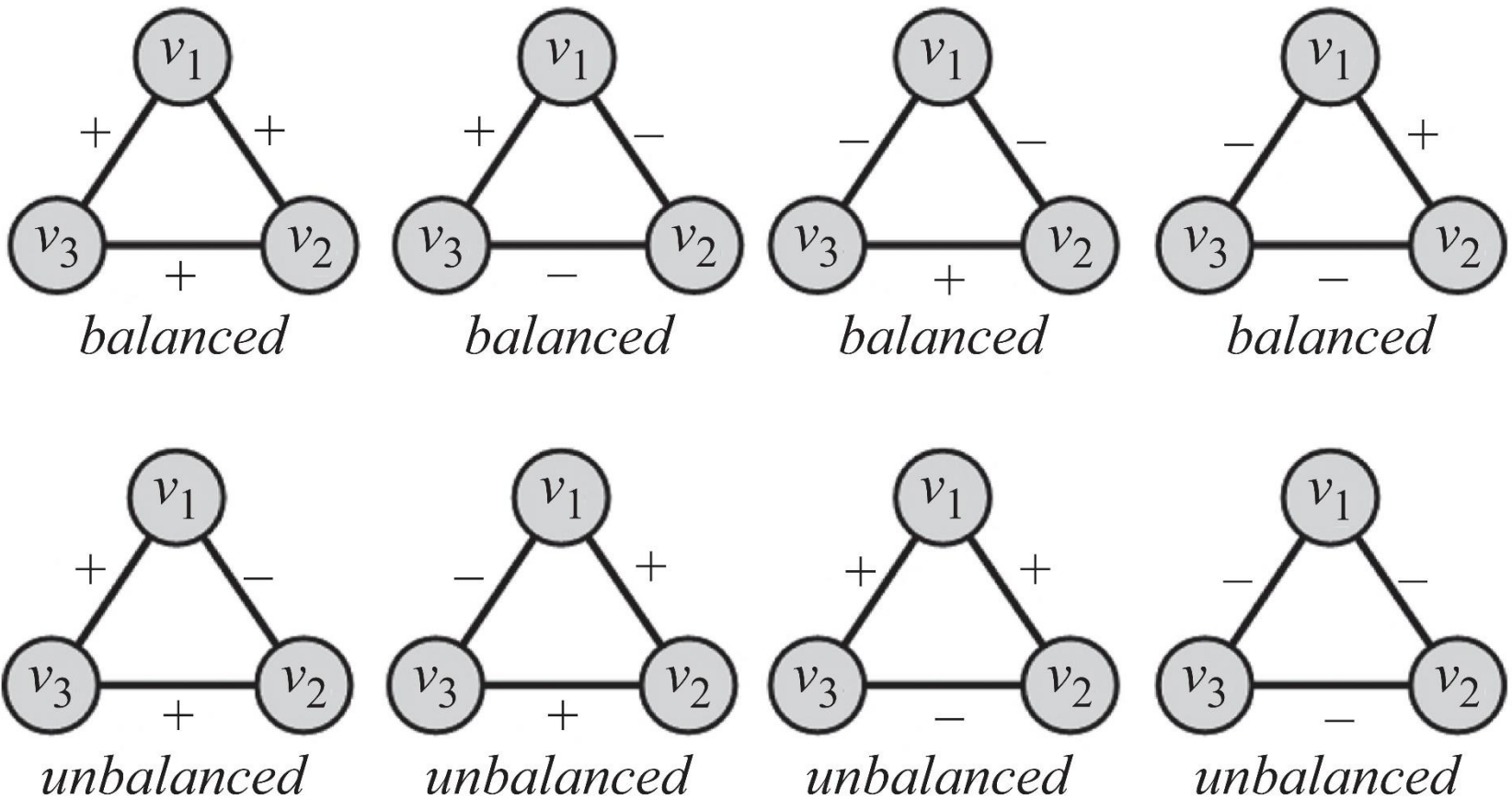
- **Measuring stability based on an observed network**

Social Balance Theory

- Social balance theory discusses consistency in friend/foe relationships among individuals. Informally, social balance theory says friend/foe relationships
 - The friend of my friend is my friend,*
 - The friend of my enemy is my enemy,*
 - The enemy of my enemy is my friend,*
 - The enemy of my friend is my enemy.*
- In the network
 - Positive edges demonstrate friendships ($w_{ij}=1$)
 - Negative edges demonstrate being enemies ($w_{ij}=-1$)
- Triangle of nodes i , j , and k , is balanced, if and only if
 - w_{ij} denotes the value of the edge between nodes i and j

$$w_{ij}w_{jk}w_{ki} \geq 0$$

Social Balance Theory: Possible Combinations



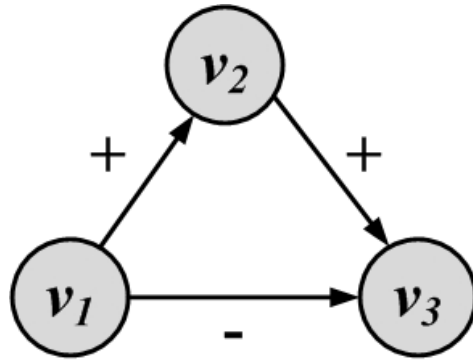
For any cycle if the multiplication of edge values become positive, then the cycle is socially balanced

Social Status Theory

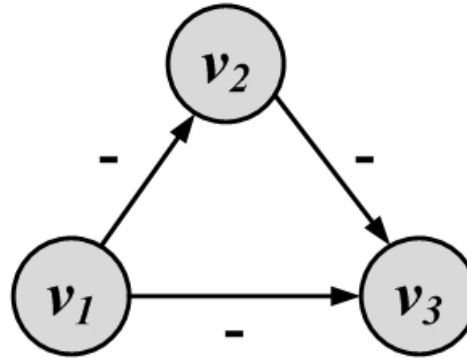
- Status defines how prestigious an individual is ranked within a society
- Social status theory measures how consistent individuals are in assigning status to their neighbors

If X has a higher status than Y and Y has a higher status than Z, then X should have a higher status than Z.

Social Status Theory: Example



Unstable configuration



Stable configuration

- A directed '+' edge from node X to node Y shows that Y has a higher status than X and a '-' one shows vice versa

References

Reza Zafarani
Social Media Mining
Chapter 3

Reciprocity

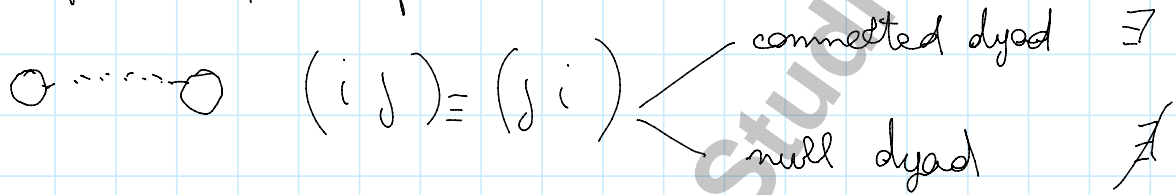
lunedì 2 novembre 2020 12:18

If you'll become my friend, I'll be yours

Definition 1 : dyads

- Undirected network

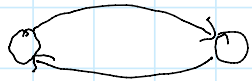
dyad : pair of nodes and links between them



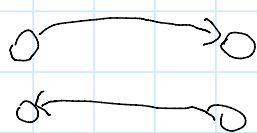
- Directed network



null dyad

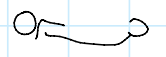
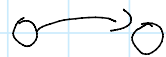
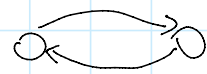


$\exists (i,j) \wedge \exists (j,i)$ symmetric dyad



$\exists (i,j) \wedge \nexists (j,i)$ asymmetric dyad
 $\nexists (i,j) \wedge \exists (j,i)$

$$r = \frac{\# \text{ symmetric dyads in } G(N, L)}{\# \text{ dyads in } G(N, L) \text{ connected}}$$

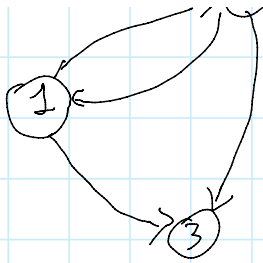


Example
 $G(N, L)$



$G(N, L)$

$$N = \{1, 2, 3\}$$



$$N = \{1, 2, 3\}$$

$$L = \{(1, 2), (1, 3), (2, 1), (2, 3)\}$$

Dyads

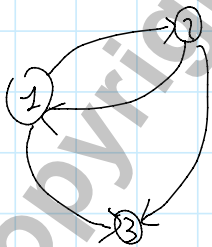
$\{1, 2\}$	$(1, 2) \wedge (2, 1)$	symmetric
$\{1, 3\}$	$(1, 3) \nexists (3, 1)$	asymmetric
$\{2, 3\}$	$(2, 3) \nexists (3, 2)$	asymmetric

$$r = \frac{1}{3}$$

• Definition 2

link (ij) is reciprocated if $\exists (ji)$

$$r = \frac{\# \text{ reciprocated link } G(N, L)}{|L|}$$



link
 (ij)

Reciprocated?
 $\exists (ji)?$

$(1, 2)$

$(2, 1) \times$

$(1, 3)$

\nexists

$(2, 1)$

$(1, 2) \times$

$$r = \frac{2}{4} = \frac{1}{2}$$

(2 1)

(1 2) x

4

z

(2 3)

∅

A_{ij} of $G(N, L)$

$$\exists (i, j) \quad A_{ji} = 1$$

$$\exists (j, i) \quad A_{ij} = 1$$

Not reciprocated link: $\exists (i, j) \quad \nexists (j, i)$

$$A_{ji} = 1$$

$$A_{ij} = 0$$

A link to be reciprocated

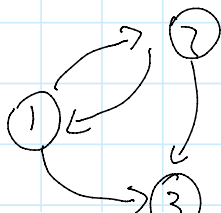
$$A_{ij} \cdot A_{ji} = 1 \quad \leftarrow$$

A link not be reciprocated

$$A_{ij} \cdot A_{ji} = 0$$

$$2 = \frac{\sum_{i,j=1}^{|L|} A_{ij} \cdot A_{ji}}{\sum_{i,j=1}^{|L|} A_{ij}} = \frac{\text{Trace } A^2}{|L|} \quad \uparrow$$

Example

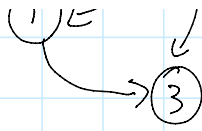


$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

7 7

7 7

7



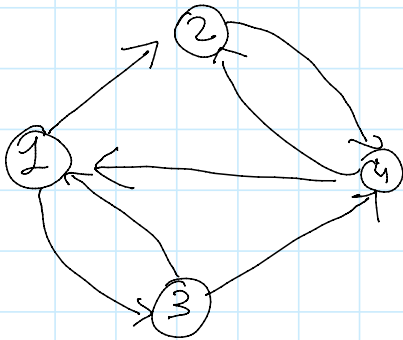
$$L \neq 0 \checkmark$$

$$A^2 = A \cdot A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

$$r = \frac{2}{4}$$



Exercise



Definition 1

Dyads

- {1 2}
- {1 3}
- {2 4}
- {3 4}
- {1 4}

Symmetric?

- n
- y
- y
- 3
- 3

$$r = \frac{2}{5}$$

Definition 2

Links

- (1 2)
- (1 3)
- (2 4)
- (3 1)
- (3 4)
- (4 2)
- (4 1)

Reciprocated

- 3
- y
- y
- 3
- 3
- y

$$r = \frac{4}{7}$$

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 1 \\ 2 & 0 & 0 & 1 \end{bmatrix}$$

$$r = \frac{4}{7}$$



Exercise

1. Consider an undirected network with 10 nodes and 5 links. Model it with an Erdos-Renyi random network $G(N, p)$.
2. Consider a directed network with 10 nodes and 5 links. Model it with an Erdos-Renyi random network $G(N, p)$.
3. Consider an undirected network with 10 nodes and an average degree equal to 1. Model it with an Erdos-Renyi random network $G(N, p)$.
4. Consider a directed network with 10 nodes and mean in-degree and out-degree equal to 1. Model it with an Erdos-Renyi random network $G(N, p)$.
5. Consider an undirected network with 10 nodes and density equal to 0.1. Model it with an Erdos-Renyi random network $G(N, p)$.
6. Consider a directed network with 10 nodes and density equal to 0.1. Model it with an Erdos-Renyi random network $G(N, p)$.

Social Media Mining

Influence and Homophily Assortativity

Reza Zafarani

Mohammad Ali Abbasi

Huan Liu

Social Forces

- Social forces connect individuals in different ways
- Among connected individuals, one often observes high social similarity or assortativity
 - This similarity is exhibited by similar behavior, similar interests, similar activities, and shared attributes such as language, among others.
 - In networks with assortativity, similar nodes are connected to one another more often than dissimilar nodes.
 - In social networks, a high similarity between friends is observed
- Friendship networks are examples of assortative networks

Why connected people are similar?

- **Influence**

- Influence is the process by which an individual (the influential) affects another individual such that the influenced individual becomes more similar to the influential figure.

- If most of one's friends switch to a mobile company, he might be influenced by his friends and switch to the company as well.

- **Homophily**

- It is realized when similar individuals become friends due to their high similarity.

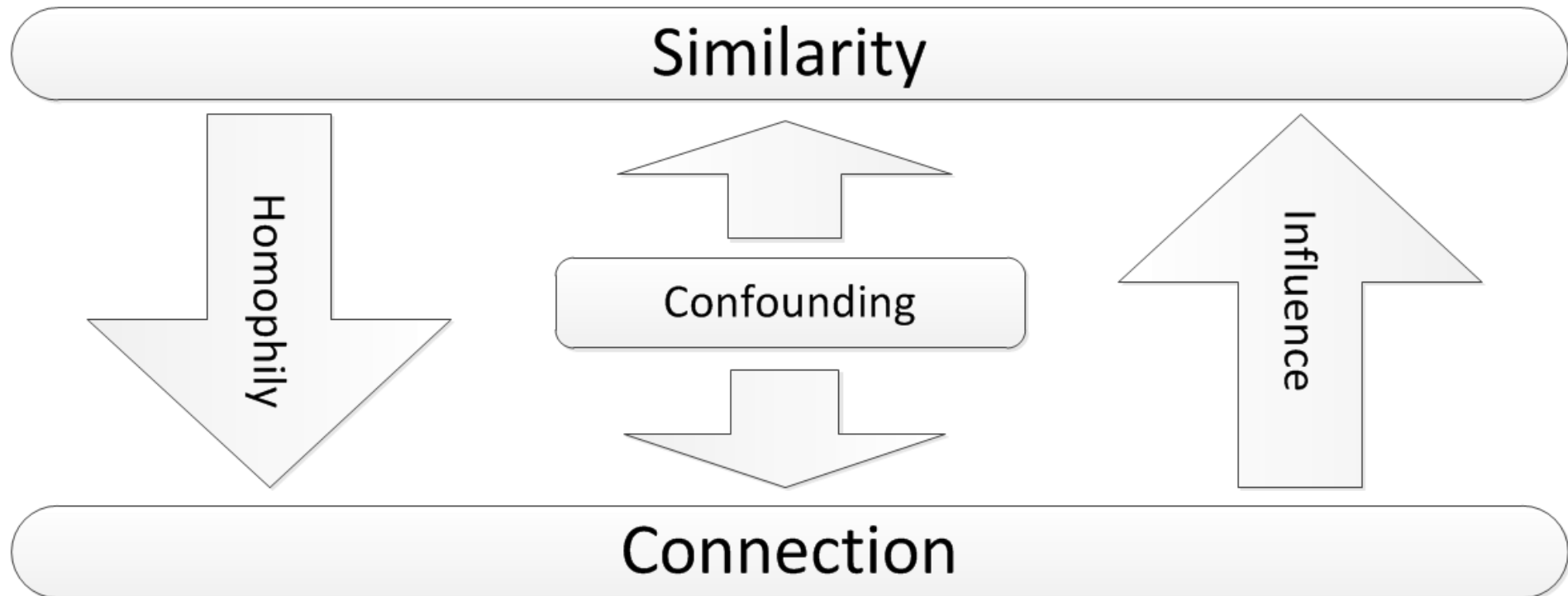
- Two musicians are more likely to become friends.

- **Confounding**

- Confounding is environment's effect on making individuals similar

- Two individuals living in the same city are more likely to become friends than two random individuals

Influence, Homophily, and Confounding



Homophily

Similar individuals
become friends

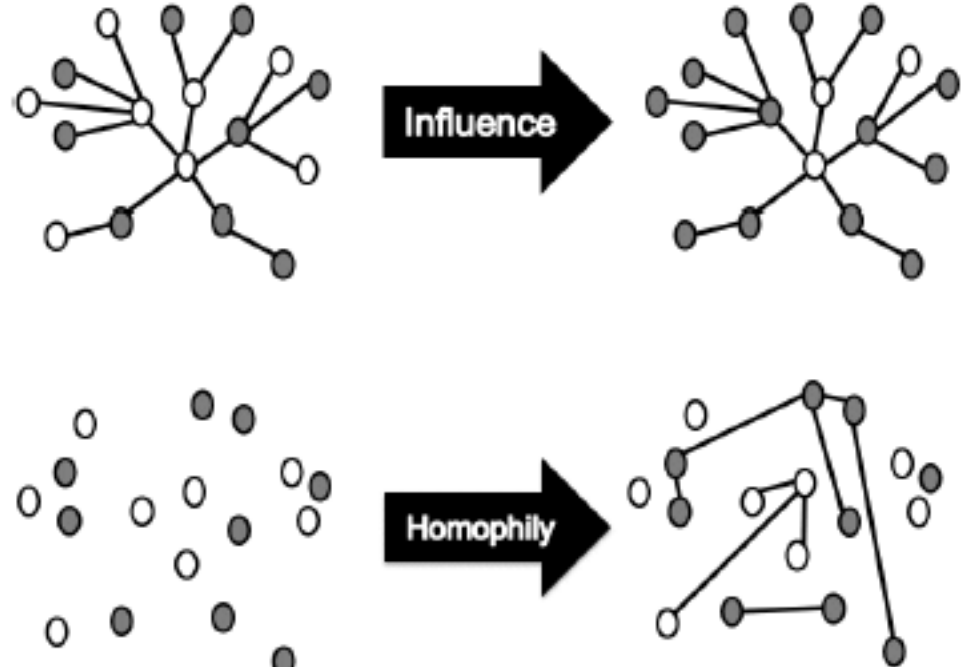
Influence

Friends become
similar

Source of Assortativity in Networks

Both influence and homophily generate similarity in social networks but in different ways

- **Homophily** selects similar nodes and links them together
- **Influence** makes the connected nodes similar to each other



Assortativity: An Example

The city's draft tobacco control strategy says more than 60% of under-16s in Plymouth smoke regularly

BBC News Sport Weather Travel TV Radio More... Search

DEVON BBC RADIO DEVON Listen Live Listen Again

Page last updated at 14:58 GMT, Monday, 14 June 2010 15:58 UK

E-mail this to a friend Printable version

Patches for Plymouth's young smokers

By Jo Irving
BBC Devon website



More than 60% of Plymouth's under-16s smoke

BBC Local
Devon
Things to do
People & Places
Nature & Outdoors
History
Religion & Ethics
Arts & Culture
BBC Introducing
TV & Radio
Local BBC Sites
News
Sport
Weather
Travel
Neighbouring Sites
Cornwall
Dorset
Somerset
Related BBC Sites
England

MORE FROM DEVON
NEWS
SPORT
WEATHER
TRAVEL

ELSEWHERE ON THE WEB
Plymouth NHS Trust Stop Smoking Service

Smoking Behavior In a Group of Friends: why is happening?

- Smoker friends influence their non-smoker friends

Influence

- Smokers become friends

Homophily

- There are lots of places that people can smoke

Confounding

Our goal in this chapter?

- How can we measure assortativity?
- How can we measure influence or homophily?
- [
- How can we model influence or homophily? **NO**
- How can we distinguish the two? **NO**

]

Measuring Assortativity

Measuring Assortativity

Nominal attributes

Measuring Assortativity for Nominal Attributes

- Where nominal attributes are assigned to nodes (language), we can use edges that are between nodes of the same type (i.e., attribute value) to measure assortativity of the network
 - Node attributes could be nationality, race, sex, etc.

$$\frac{1}{m} \sum_{(v_i, v_j) \in E} \delta(t(v_i), t(v_j)) = \frac{1}{2m} \sum_{ij} A_{ij} \delta(t(v_i), t(v_j))$$

$t(v_i)$ denotes type of vertex v_i

$$\delta(x, y) = \begin{cases} 0, & \text{if } x \neq y \\ 1, & \text{if } x = y \end{cases}$$

Kronecker delta function

Assortativity Significance

- Assortativity significance measures the difference between the measured assortativity and its expected assortativity
 - The higher this value, the more significant the assortativity observed
- **Example**
 - Consider a school where half the population is white and half the population is Hispanic. It is expected for 50% of the connections to be between members of different races. If all connections in this school were between members of different races, then we have a significant finding

Assortativity Significance: Measuring

Assortativity



The expected assortativity in the whole graph



$$\begin{aligned} Q &= \frac{1}{2m} \sum_{ij} A_{ij} \delta(t(v_i), t(v_j)) - \frac{1}{2m} \sum_{ij} \frac{d_i d_j}{2m} \delta(t(v_i), t(v_j)) \\ &= \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \delta(t(v_i), t(v_j)). \end{aligned}$$

This measure is called modularity

Normalized Modularity

The maximum happens when all vertices of the same type are connected to one another

$$\begin{aligned} Q_{\text{normalized}} &= \frac{Q}{Q_{\text{max}}}, \\ &= \frac{\frac{1}{2m} \sum_{ij} (A_{ij} - \frac{d_i d_j}{2m}) \delta(t(v_i), t(v_j))}{\max \left(\frac{1}{2m} \sum_{ij} A_{ij} \delta(t(v_i), t(v_j)) - \frac{1}{2m} \sum_{ij} \frac{d_i d_j}{2m} \delta(t(v_i), t(v_j)) \right)} \\ &= \frac{\frac{1}{2m} \sum_{ij} (A_{ij} - \frac{d_i d_j}{2m}) \delta(t(v_i), t(v_j))}{\frac{1}{2m} 2m - \frac{1}{2m} \sum_{ij} \frac{d_i d_j}{2m} \delta(t(v_i), t(v_j))}, \\ &= \frac{\sum_{ij} (A_{ij} - \frac{d_i d_j}{2m}) \delta(t(v_i), t(v_j))}{2m - \sum_{ij} \frac{d_i d_j}{2m} \delta(t(v_i), t(v_j))}. \end{aligned}$$

Modularity: Matrix Form

- Let $\Delta \in \mathbb{R}^{n \times k}$ denote the indicator matrix and let k denote the number of types

$$\Delta_{x,k} = \begin{cases} 1, & \text{if } t(x) = k; \\ 0, & \text{if } t(x) \neq k \end{cases}$$

- The Kronecker delta function can be reformulated using the indicator matrix

$$\delta(t(v_i), t(v_j)) = \sum_k \Delta_{v_i,k} \Delta_{v_j,k}$$

- Therefore, $(\Delta \Delta^T)_{i,j} = \delta(t(v_i), t(v_j))$

Normalized Modularity: Matrix Form

Let Modularity matrix be:

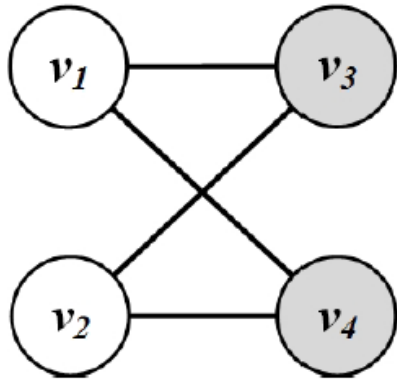
$$B = A - dd^T / 2m \quad \text{where } d \in \mathbb{R}^{n \times 1}$$

Is the degree vector

Then, modularity can be reformulated as

$$Q = \frac{1}{2m} \sum_{ij} \underbrace{\left(A_{ij} - \frac{d_i d_j}{2m} \right)}_{B_{ij}} \underbrace{\delta(t(v_i), t(v_j))}_{(\Delta \Delta^T)_{i,j}} = \frac{1}{2m} \text{Tr}(B \Delta \Delta^T) = \frac{1}{2m} \text{Tr}(\Delta^T B \Delta).$$

Modularity Example



$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \Delta = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{d} = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix}, m = 4.$$

$$B = A - \mathbf{d}\mathbf{d}^T/2m = \begin{bmatrix} -0.5 & -0.5 & 0.5 & 0.5 \\ -0.5 & -0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \end{bmatrix}.$$

$$\frac{1}{2m} \text{Tr}(\Delta^T B \Delta) = -0.5.$$

the number of edges between nodes of the **same color** is less than the expected number of edges between them

Measuring Assortativity

Ordinal attributes

Measuring Assortativity for Ordinal Attributes

- A common measure for analyzing the relationship between ordinal values is covariance.
- It describes how two variables change together.
- In our case we are interested in how values of nodes that are connected via edges are correlated.

Covariance Variables

- We construct two variables X_L and X_R , where for any edge $(v_i; v_j)$ we assume that x_i is observed from variable X_L and x_j is observed from variable X_R .
- In other words, X_L represents the ordinal values associated with the left node of the edges and X_R represents the values associated with the right node of the edges
- Our problem is therefore reduced to computing the covariance between variables X_L and X_R

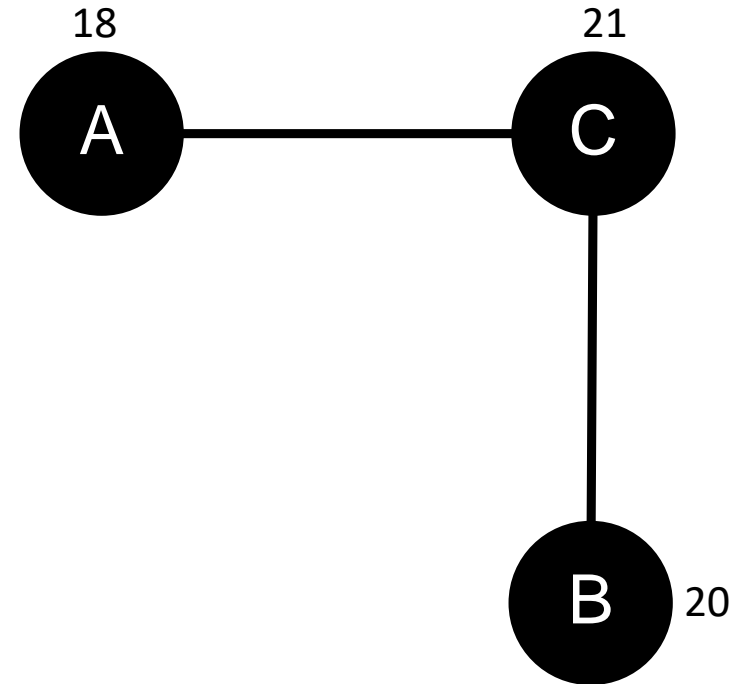
Covariance Variables: Example

List of edges:
((A, C),
(C, A),
(C, B),
(B, C))

- $X_L : (18, 21, 21, 20)$
- $X_R : (21, 18, 20, 21)$



$$E(X_L) = E(X_R),$$
$$\sigma(X_L) = \sigma(X_R).$$



Covariance

For two given column variables X_L and X_R the covariance is

$$\begin{aligned}\sigma(X_L, X_R) &= \mathbf{E}[(X_L - \mathbf{E}[X_L])(X_R - \mathbf{E}[X_R])] \\ &= \mathbf{E}[X_L X_R - X_L \mathbf{E}[X_R] - \mathbf{E}[X_L] X_R + \mathbf{E}[X_L] \mathbf{E}[X_R]] \\ &= \mathbf{E}[X_L X_R] - \mathbf{E}[X_L] \mathbf{E}[X_R] - \mathbf{E}[X_L] \mathbf{E}[X_R] + \mathbf{E}[X_L] \mathbf{E}[X_R] \\ &= \mathbf{E}[X_L X_R] - \mathbf{E}[X_L] \mathbf{E}[X_R].\end{aligned}$$

$\mathbf{E}(X_L)$ is the mean of the variable and $\mathbf{E}(X_L X_R)$ is the mean of the multiplication

$$\begin{aligned}E(X_L) &= E(X_R) = \frac{\sum_i (X_L)_i}{2m} = \frac{\sum_i d_i x_i}{2m} \\ E(X_L X_R) &= \frac{1}{2m} \sum_i (X_L)_i (X_R)_i = \frac{\sum_{ij} A_{ij} x_i x_j}{2m}.\end{aligned}$$

Covariance

$$\begin{aligned}\sigma(X_L, X_R) &= \mathbf{E}[X_L X_R] - \mathbf{E}[X_L]\mathbf{E}[X_R] \\ &= \frac{\sum_{ij} A_{ij} x_i x_j}{2m} - \frac{\sum_{ij} d_i d_j x_i x_j}{(2m)^2} \\ &= \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2m} \right) x_i x_j.\end{aligned}$$

Normalizing Covariance

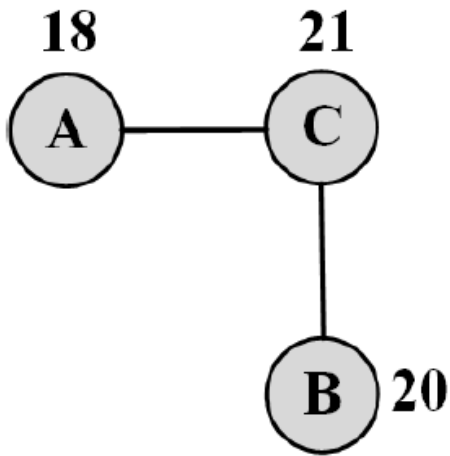
Pearson correlation $\rho(X,Y)$ is the normalized version of covariance

$$\rho(X_L, X_R) = \frac{\sigma(X_L, X_R)}{\sigma(X_L)\sigma(X_R)}.$$

In our case: $\sigma(X_L) = \sigma(X_R)$

$$\begin{aligned}\rho(X_L, X_R) &= \frac{\sigma(X_L, X_R)}{\sigma(X_L)^2}, \\ &= \frac{\frac{1}{2m} \sum_{ij} (A_{ij} - \frac{d_i d_j}{2m}) x_i x_j}{\mathbf{E}[(X_L)^2] - (\mathbf{E}[X_L])^2} \\ &= \frac{\frac{1}{2m} \sum_{ij} (A_{ij} - \frac{d_i d_j}{2m}) x_i x_j}{\frac{1}{2m} \sum_{ij} A_{ij} x_i^2 - \frac{1}{2m} \sum_{ij} \frac{d_i d_j}{2m} x_i x_j}.\end{aligned}$$

Correlation Example



$$X_L = \begin{bmatrix} 18 \\ 21 \\ 21 \\ 20 \end{bmatrix}, X_R = \begin{bmatrix} 21 \\ 18 \\ 20 \\ 21 \end{bmatrix}$$

$$\rho(X_L, X_R) = -0.67.$$

Social Influence

- **Measuring Influence**
- **Modeling Influence**

Social Influence: Definition

- the act or power of producing an effect without apparent exertion of force or direct exercise of command

Measuring the Influence

Measuring Influence

- Measuring influence is assigning a number to each node that represents the influential power of that node
- The influence can be measured either based on prediction or observation

Prediction-based Measurement

- In prediction-based measurement, we assume that an individual's attribute or the way she is situated in the network predicts how influential she will be.
- For instance, we can assume that the gregariousness (e.g., number of friends) of an individual is correlated with how influential she will be. Therefore, it is natural to use any of the centrality measures discussed in Chapter 3 for prediction-based influence measurements.
- An example:
 - On Twitter, in-degree (number of followers) is a benchmark for measuring influence commonly used

Observation-based Measurement

- In observation-based we quantify influence of an individual by measuring the amount of influence attributed to the individual
 - When an individual is the role model
 - Influence measure: size of the audience that has been influenced
 - When an individual spreads information:
 - Influence measure: the size of the cascade, the population affected, the rate at which the population gets influenced
 - When an individual increases values:
 - Influence measure: the increase (or rate of increase) in the value of an item or action

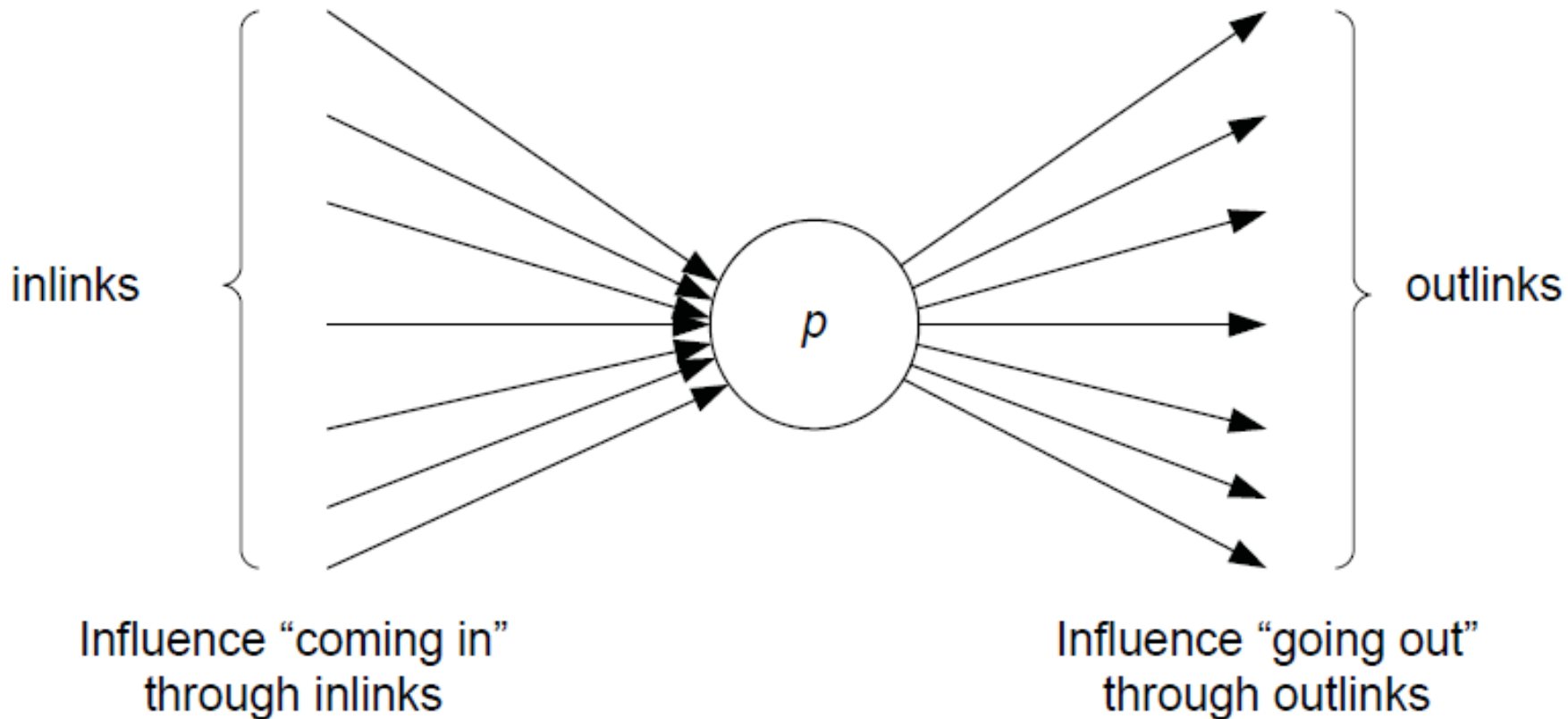
Case Studies for Measuring Influence in Social Media

- **Measuring Social Influence on Blogosphere**
- **Measuring Social Influence on Twitter**

Measuring Social Influence on Blogosphere

- The goal of measuring influence in blogosphere is to figure out most influential bloggers on the blogosphere
- Due to limited time an individual has, following the influentials is often a good heuristic of filtering what's uninteresting
- One common measure for quantifying influence of bloggers is to use indegree centrality
- Due to the sparsity of in-links, more detailed analysis is required to measure influence in blogosphere

iFinder: A System to measure influence on blogosphere



- **Recognition**
 - Recognition for a blogpost is the number of the links that point to the blogpost (in-links).
 - Let I_p denotes the set of in-links that point to blogpost p .
- **Activity Generation**
 - Activity generated by a blogpost is the number of comments that p receives.
 - c_p denotes the number of comments that blogpost p receives.
- **Novelty**
 - The blogpost's novelty is inversely correlated with the number of references a blogpost employs. In particular the more citations a blogpost has it is considered less novel.
 - O_p denotes the set of out-links for blogpost p .
- **Eloquence**
 - Eloquence is estimated by the length of the blogpost. Given the unformal nature of blogs and the bloggers tendency to write short blogposts, longer blogposts are believed to be more eloquent. So the length of a blogpost l_p can be employed as a measure of eloquence

Measuring Social Influence on Twitter

- In Twitter, users have an option of following individuals, which allows users to receive tweets from the person being followed
- Intuitively, one can think of the number of followers as a measure of influence (in-degree centrality)

Measuring Social Influence on Twitter: Measures

- **Indegree**
 - The number of users following a person on Twitter
 - Indegree denotes the “audience size” of an individual.
- **Number of Mentions**
 - The number of times an individual is mentioned in a tweet, by including @username in a tweet.
 - The number of mentions suggests the “ability in engaging others in conversation”
- **Number of Retweets:**
 - Tweeter users have the opportunity to forward tweets to a broader audience via the retweet capability.
 - The number of retweets indicates individual’s ability in generating content that is worth being passed on.

Measuring Social Influence on Twitter: Measures

- Each one of these measures by itself can be used to identify influential users in Twitter.
- This can be performed by utilizing the measure for each individual and then ranking individuals based on their measured influence value.
- Contrary to public belief, number of followers is considered an inaccurate measure compared to the other two.
- We can rank individuals on twitter independently based on these three measures.
- To see if they are correlated or redundant, we can compare ranks of an individuals across three measures using rank correlation measures.

- Spearman's rank correlation is the Pearson's correlation coefficient for ordinal variables that represent ranks (i.e., takes values between 1 . . . n); hence, the value is in range [-1,1].
- Popular users (users with high in-degree) do not necessarily have high ranks in terms of number of retweets or mentions.

Measures	Correlation Value
Indegree vs Retweets	0.122
Indegree vs Mentions	0.286
Retweets vs Mentions	0.638

Social Media Mining

Network Measures

Reza Zafarani

Mohammad Ali Abbasi

Huan Liu

Network Similarity

How similar are two nodes in a network?

- **Neighbourhood**
- **Attributes**
- **Contents**

Structural Equivalence

- In structural equivalence, we look at the neighborhood shared by two nodes; the size of this neighborhood defines how similar two nodes are.

For instance, two brothers have in common sisters, mother, father, grandparents, etc. This shows that they are similar, whereas two random male or female individuals do not have much in common and are not similar.

Structural Equivalence: Definitions

Vertex similarity

$$\sigma(v_i, v_j) = |N(v_i) \cap N(v_j)|.$$

Jaccard Similarity:

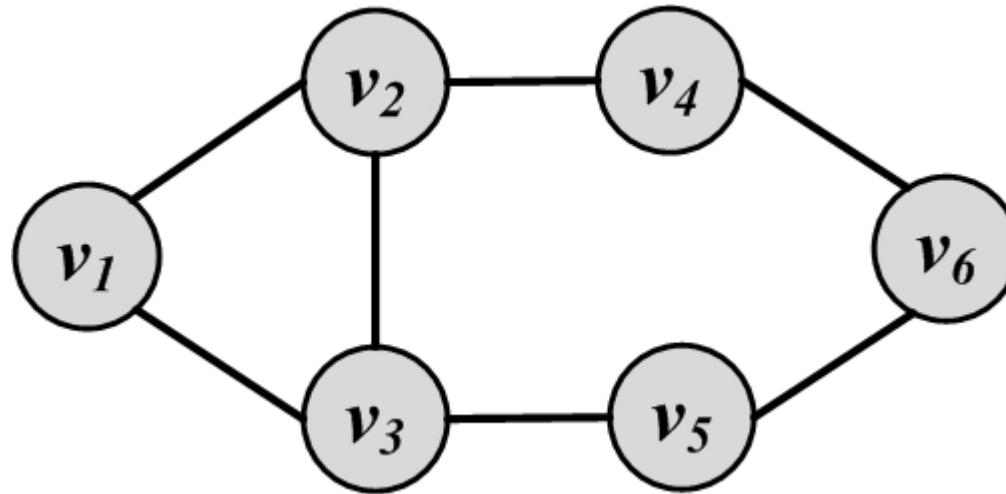
$$\sigma_{Jaccard}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|}$$

Cosine Similarity:

$$\sigma_{Cosine}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{\sqrt{|N(v_i)||N(v_j)|}}$$

- Range: [0,1]
- In general, the definition of neighborhood $N(v)$ excludes the node itself v .
 - Nodes that are connected and do not share a neighbor will be assigned zero similarity
 - This can be rectified by assuming nodes to be included in their neighborhoods

Similarity: Example



$$\sigma_{Jaccard}(v_2, v_5) = \frac{|\{v_1, v_3, v_4\} \cap \{v_3, v_6\}|}{|\{v_1, v_3, v_4, v_6\}|} = 0.25$$

$$\sigma_{Cosine}(v_2, v_5) = \frac{|\{v_1, v_3, v_4\} \cap \{v_3, v_6\}|}{\sqrt{|\{v_1, v_3, v_4\}| |\{v_3, v_6\}|}} = 0.40$$

Regular Equivalence

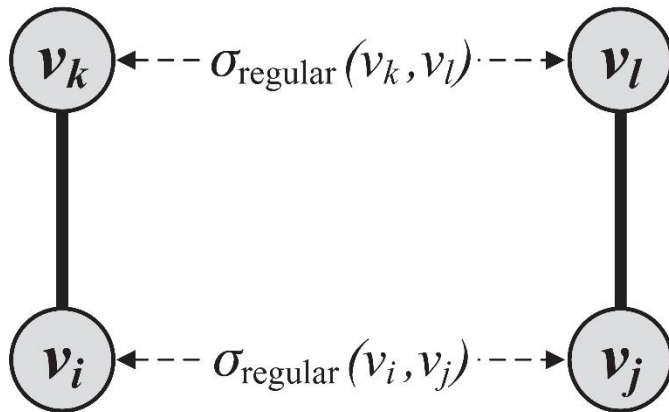
- In regular equivalence, we do not look at neighborhoods shared between individuals, but how neighborhoods themselves are similar

For instance, athletes are similar not because they know each other in person, but since they know similar individuals, such as coaches, trainers, other players, etc.

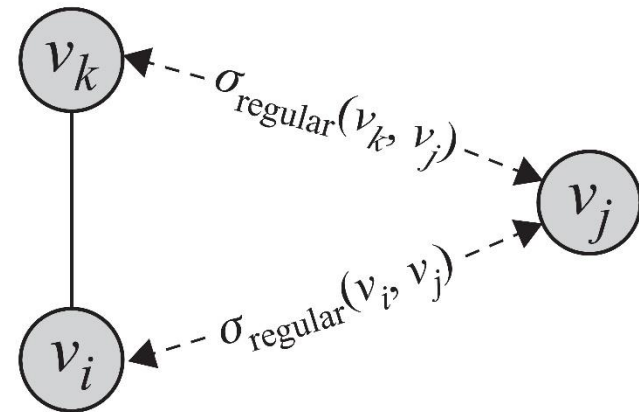
Regular Equivalence

- v_i, v_j are similar when their neighbors v_k and v_l are similar

$$\sigma_{regular}(v_i, v_j) = \alpha \sum_{k,l} A_{i,k} A_{j,l} \sigma_{Regular}(v_k, v_l).$$



(a) Original Formulation

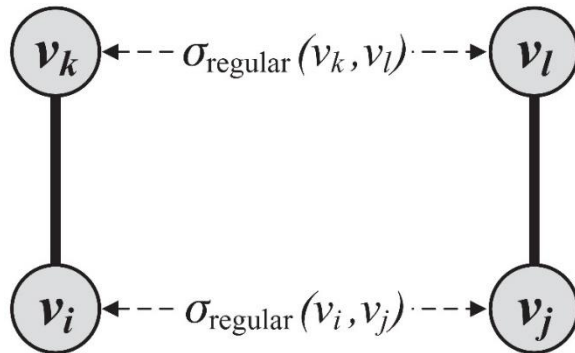


(b) Relaxed Formulation

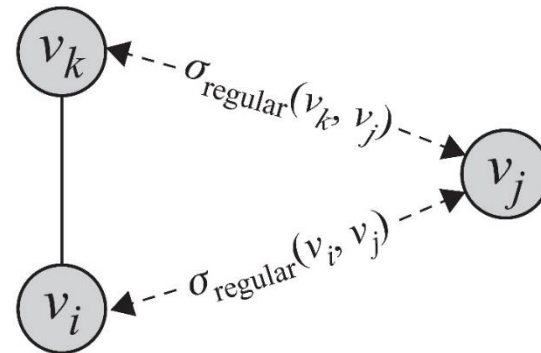
- The equation (left figure) is hard to solve since it is self referential so we relax our definition using the right figure.
- v_i is similar to v_j when v_j is similar to v_i 's neighbors v_k

Regular Equivalence

- v_i is similar to v_j is similar when v_j is similar to v_i 's neighbors v_k



(a) Original Formulation



(b) Relaxed Formulation

$$\sigma_{regular}(v_i, v_j) = \alpha \sum_k A_{i,k} \sigma_{Regular}(v_k, v_j)$$



$$\sigma_{regular} = \alpha A \sigma_{Regular}$$


- In vector format

Regular Equivalence


- v_i is similar to v_j is similar when v_j is similar to v_i 's neighbors v_k


$$\sigma_{regular}(v_i, v_j) = \alpha \sum_k A_{i,k} \sigma_{Regular}(v_k, v_j)$$

- In vector format

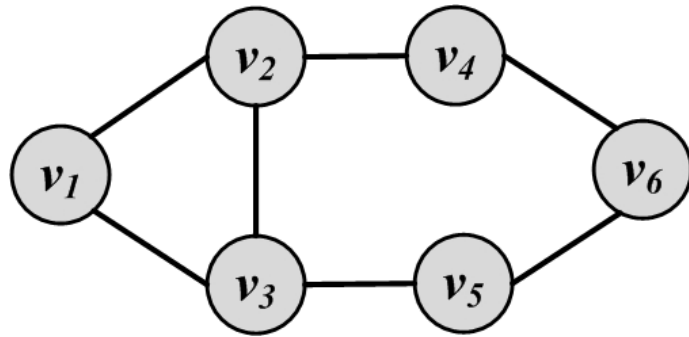

$$\sigma_{regular} = \alpha A \sigma_{Regular}$$

A vertex is highly similar to itself, we guarantee this by adding an identity matrix to the equation


$$\sigma_{regular} = \alpha A \sigma_{Regular} + \mathbf{I}$$


$$\sigma_{regular} = (\mathbf{I} - \alpha A)^{-1}$$

Regular Equivalence: Example



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

The largest eigenvalue of A is 2.43

Set $\alpha = 0.4 < 1/2.43$

$$\sigma_{regular} = (I - 0.4A)^{-1} = \begin{bmatrix} 1.43 & 0.73 & 0.73 & 0.26 & 0.26 & 0.16 \\ 0.73 & 1.63 & 0.80 & 0.56 & 0.32 & 0.26 \\ 0.73 & 0.80 & 1.63 & 0.32 & 0.56 & 0.26 \\ 0.26 & 0.56 & 0.32 & 1.31 & 0.23 & 0.46 \\ 0.26 & 0.32 & 0.56 & 0.23 & 1.31 & 0.46 \\ 0.16 & 0.26 & 0.26 & 0.46 & 0.46 & 1.27 \end{bmatrix}$$

- Any row/column of this matrix shows the similarity to other vertices
- We can see that vertex 1 is most similar (other than itself) to vertices 2 and 3
- Nodes 2 and 3 have the highest similarity

Crawling

Cheick Tidiane Ba

Copyright Università degli Studi di Milano



What is a crawler

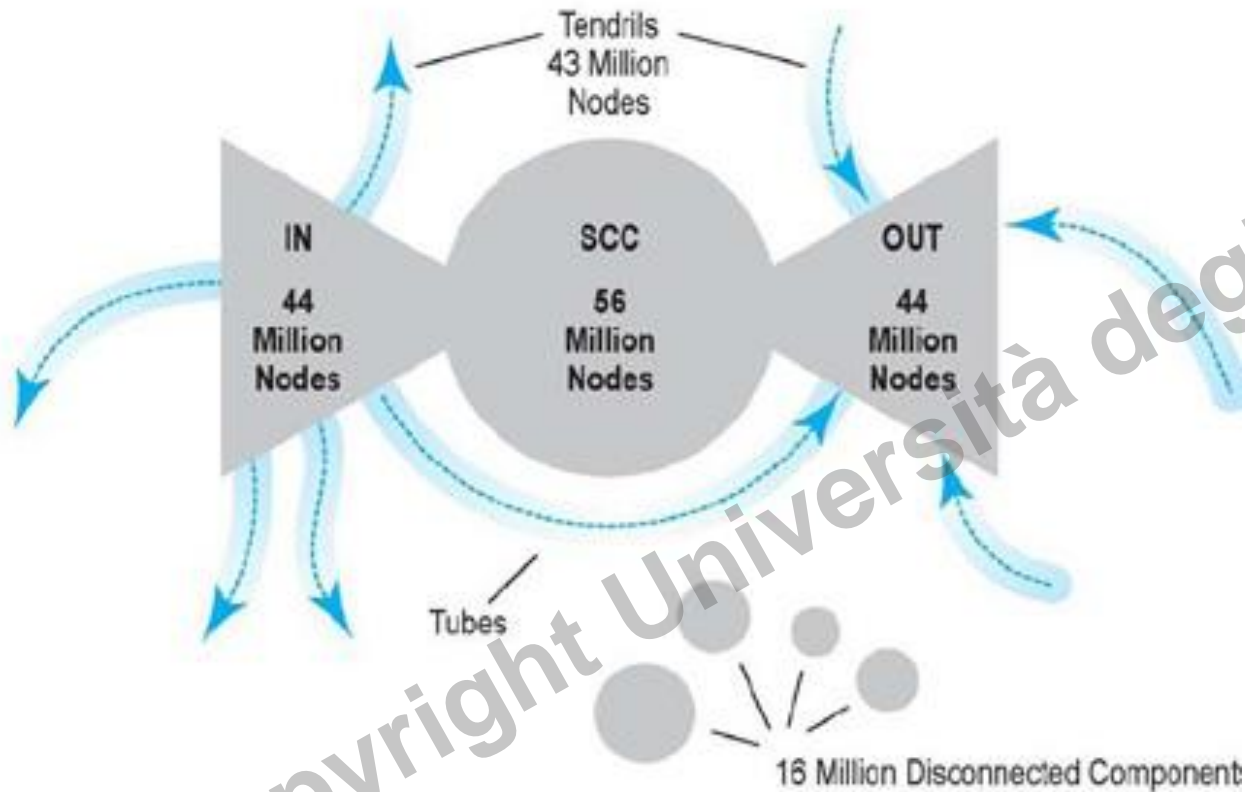
- A Web crawler, sometimes called a spider or spiderbot and often shortened to crawler, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing (web spidering).
- We make the distinction between
 - Crawling: the activity of download of web pages, while visiting the web
 - Web scraping: extracting data from websites.

Why we care

- Data is key for machine learning and business decisions
- It is important to understand the issues behind data retrieval,
- As data scientists we may need to address those issues and configure scraping tools to obtain data
 - Understanding concepts helps us deal with those tools

Basic organization of a large-scale, distributed web crawler

The web is enormous



- The web is a network
- Large scale
- Reconstruction of the structure depends on where we start decides the result

Source: K. Laudon & C. Trever, E-Commerce
2009 (5th Edition), Prentice Hall.

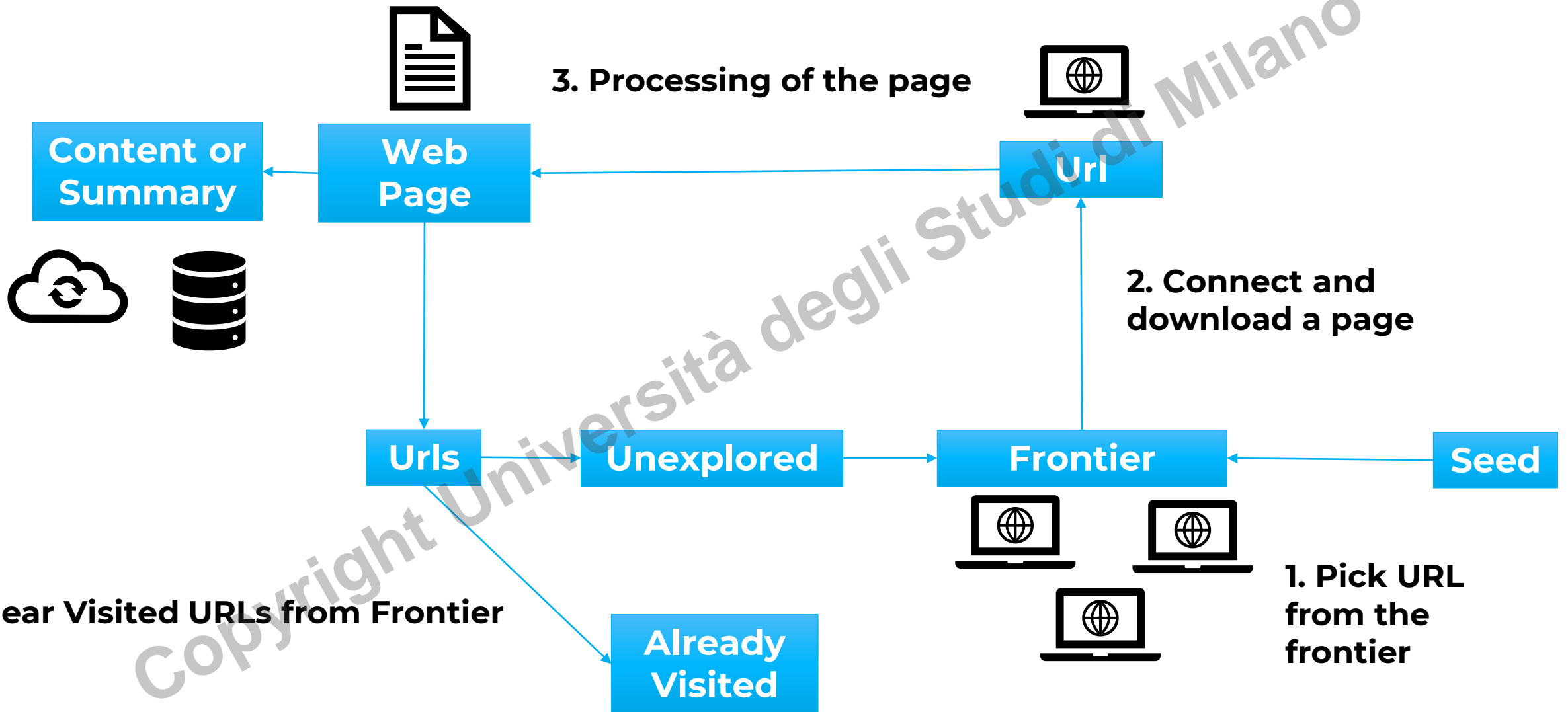
URLs classification

- **Seed:** set of urls
- **Frontier:** URLs available but that have not been visited yet.
 - From the seed set
 - Found in pages we have already visited
- **Visited URLs:** downloaded pages that have been analyzed and processed.
- **Unknown URLs:** everything else

What does the crawler do

- The crawler loads the seed set in the frontier set.
- While there are URLs left:
 1. Pick URL in the frontier
 2. Connect and download a page
 3. Processing of the page (URLs extraction, summary)
 4. Move the URL from the frontier to the set of Visited URLs
 5. Filter extracted URLs:
 - remove the already visited links
 - new URLs are added to the frontier

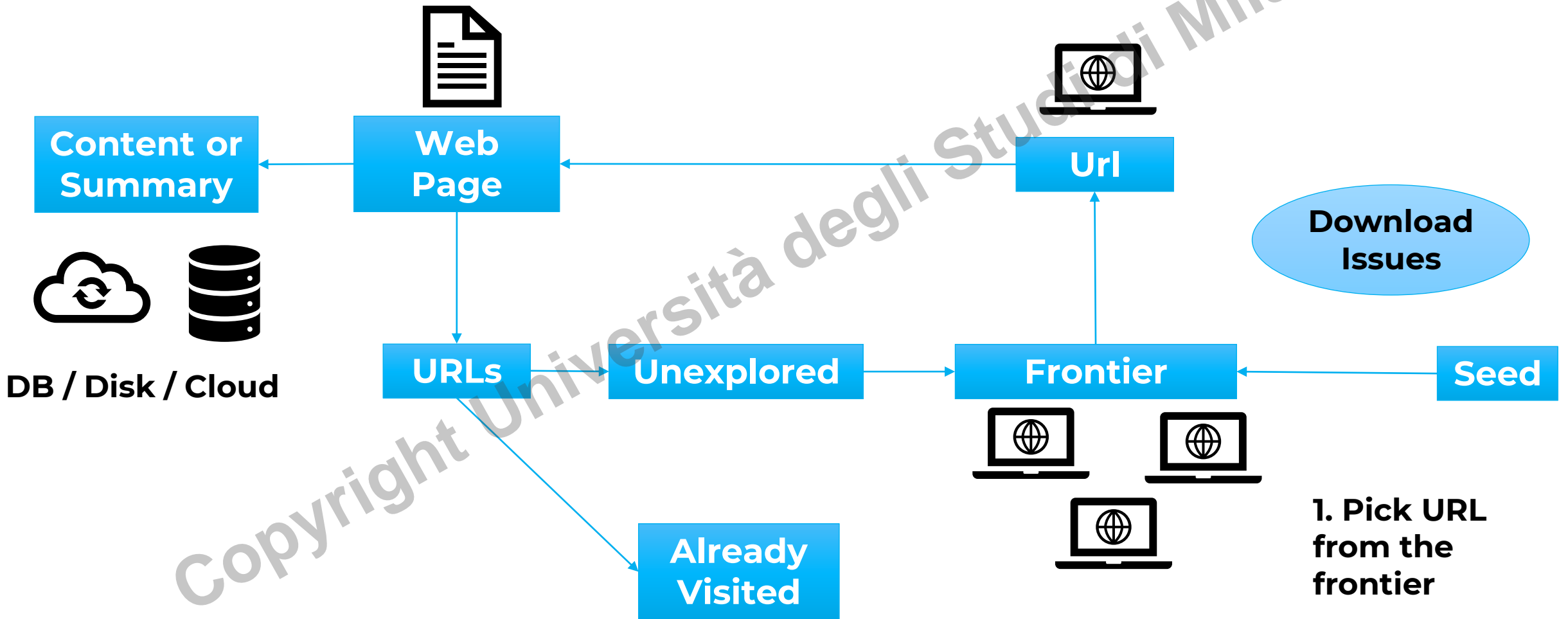
The system



Download Issues

Copyright Università degli Studi di Milano

The system



Download - Algorithmic issues

- We are dealing with a graph
- While there are a lot of visiting algorithms for graphs they can't be applied
- The issue is that they require to know a priori how many nodes are available and which ones.
- In the crawling process we have important choices to make.

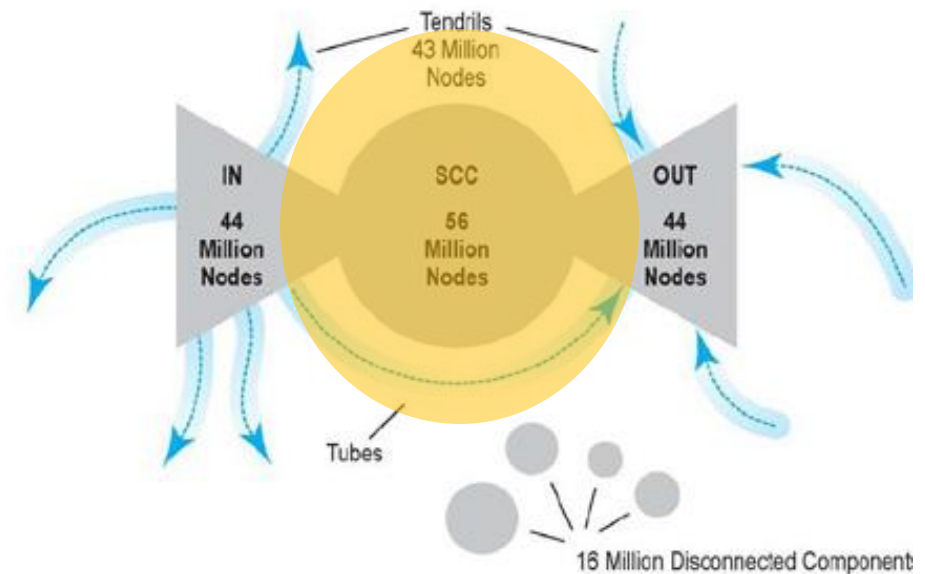
Copyright Università degli Studi di Milano

Download - Algorithmic issues

- **First choice: policy**
 - the way we choose an url in frontier.
- **The policy influences the crawling order and the obtained structure**
- **The policy can prioritize different aspects and could be changed with time.**
- **Examples of policy:**
 - **Content-based**
 - The basis of Scraping processes for specific content
 - **Priority to most frequent urls**
 - **Priority to long urls**
 - We want to visit the sub pages first (sort of a depth first visit)

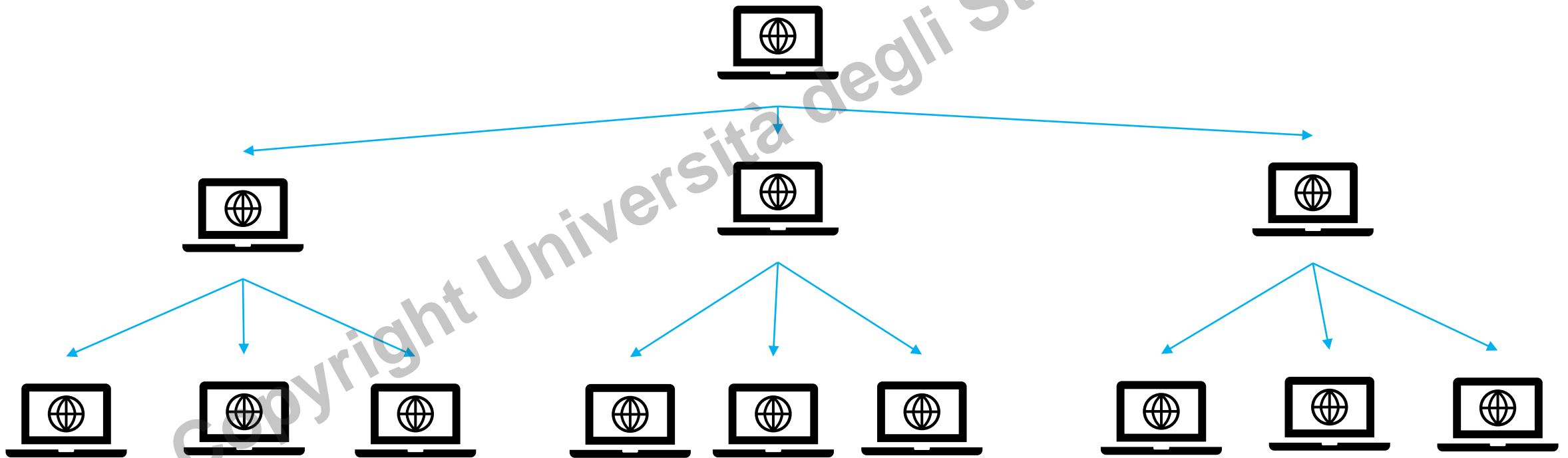
Download - Algorithmic issues

- Where we start is just as important as how we choose the next page
- The seed set must be chosen carefully
- What we want for the seed
 - Limited set
 - In the main component
 - Content driven selection
 - Theme focus
 - General purpose needs more variety



Download - Resource issues

- The Frontier grows rapidly, much more quickly than the number of visited websites
- The growth could be exponential



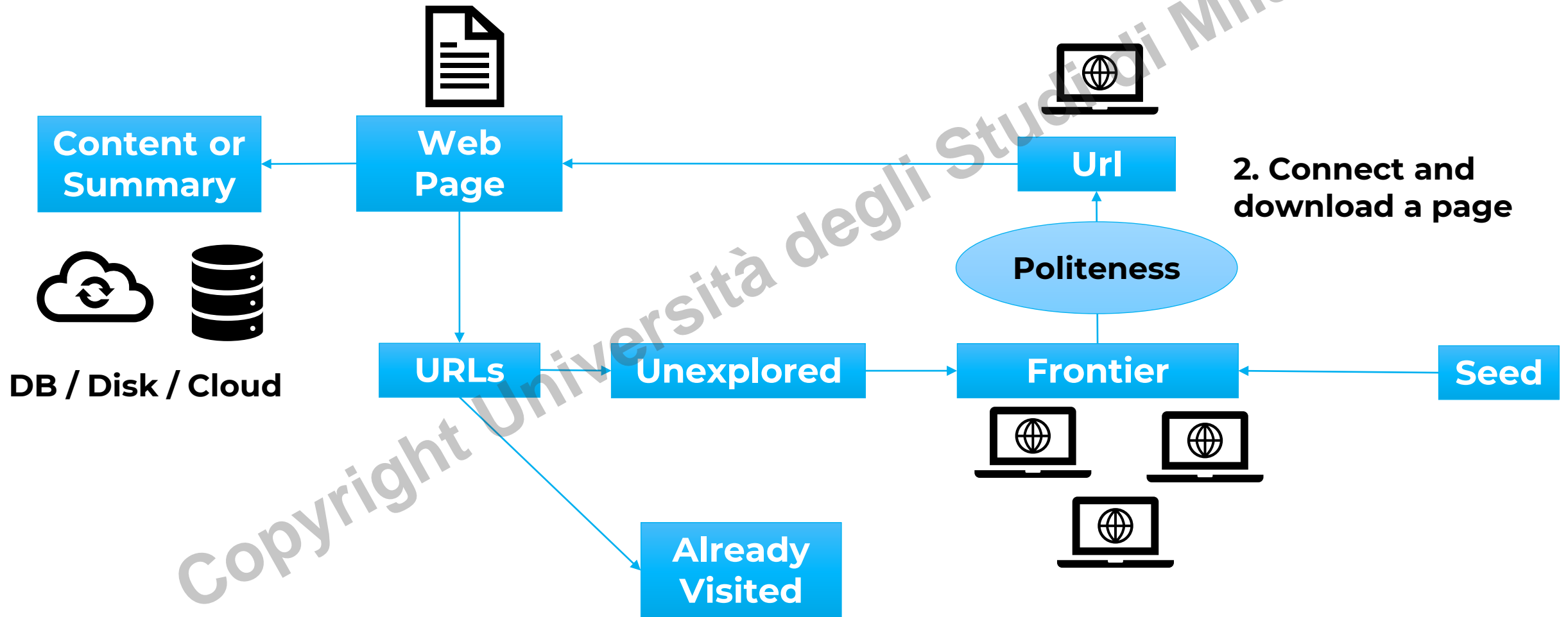
Download - Resource issues

- **We need Resources**
 - Ram
 - Storage (Disk or Cloud)
- **The resources available may not be enough**
- **The crawler may need to stop downloading: the programmer should include a procedure to perform a Graceful Degradation**
 - When RAM is full, rely on Storage
 - When Disk is about to be full, stop and don't compromise the operative system

Politeness

Copyright Università degli Studi di Milano

The system



Politeness

- We should not exceed with the amount of time dedicated to downloads from a single site or server.
- While we may have a lot resources, shall i use them all
- The issue is that we can create issues for those we handle the websites or servers.

Copyright Università degli Studi di Milano

Why do i care about Politeness

- **Some hosting sites make clients pay based on the transmitted data or bytes**
 - **Unexpected costs**
- **Some hosting services may have a limit on traffic**
 - **We may be slowing down other legit users or blocking them out completely**
- **Some pages require a lot CPU work to be loaded.**
 - **Again some hosts may have high costs for this kind of resource.**

Why do i care about Politeness

- **Crawlers are automatic and unsupervised. They visit sites without approval**
- **Story time: you may accidentally destroy a database**
 - A testing link to an http delete command was hidden in a crawled page
 - Visiting that URL deleted an entire database
 - Only protection is an accurate log system

What are the consequences

- **The consequences**
 - **Potentially Ban by IP**
 - **Legal issues**
- **Mechanisms so that sites can signal that they wish for crawler to limit to certain sections or URLs or that they may not be crawled at all**
- **One of them is the “robots.txt” files**

Robots.txt

- File used by websites to indicate which parts of the website can be visited by a crawler, if any.
- It doesn't have legal value, it can be ignored
- But it avoids us issues with website managers
- The file must be downloaded the first time we access the website. Then checked periodically (e.g. 6 hours), looking for updates.
- Example:
 - <https://www.nytimes.com/robots.txt>
 - <https://www.facebook.com/robots.txt>
 - <https://corriere.it/robots.txt>
 - <https://www.reddit.com/robots.txt>

Example: Facebook

- We can notice that except for all the big of IT that have some specified rules, for everyone else all is forbidden

```
# Notice: Collection of data on Facebook through automated means is
# prohibited unless you have express written permission from Facebook
# and may only be conducted for the limited purpose contained in said
# permission.
# See: http://www.facebook.com/apps/site\_scraping\_tos\_terms.php
```

```
User-agent: Applebot
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /dialog/
Disallow: /fbml/ajax/dialog/
Disallow: /feeds/
Disallow: /file_download.php
```

```
User-agent: *
Disallow: /
```

Example: Corriere.it

- **User-agent: ***
 - Rules applied to everyone, even browsers.
- **Notes:**
 - More disorganized compared to big websites
 - We can find sitemaps
 - Urls that specify the structure of the website
 - Some pages are not reachable following links in the pages
 - Usually are links generated through javascript.

```
Disallow: /corrierepedia
Disallow: /firme/pierluigi-battista_80
Disallow: /cook/ricette/ricerca
Allow: /sitemap*.xml
Allow: /notizie-ultima-ora/rss_col.xml
Allow: /rss/*.xml
Allow: /salute/sitemap-dizionario.xml
Allow: /rss/ultimora.xml
Allow: /rss/homepage.xml
Allow: /notizie-ultima-ora/sitemap-news.xml
#richiesta da Ruggiero BG27112011
Allow: /editorspicks/

Sitemap: https://www.corriere.it/rss/homepage.xml
Sitemap: https://www.corriere.it/sitemap/sitemap_100.xml
Sitemap: https://www.corriere.it/salute/sitemap-dizionario.xml
Sitemap: https://www.corriere.it/sitemap_100_english.xml
Sitemap: https://www.corriere.it/sitemap_100_chinese.xml
Sitemap: https://www.corriere.it/rss/ultimora.xml
```

Crawl delay

- An option that specifies a waiting time is crawl-delay
- This options specifies an amout of time in seconds that the crawler should wait before making consecutives connections
- E.g.
 - crawl-delay:2

Copyright Università degli Studi di Milano

How a crawler can prevent issues

- Option 1: limit the time of a request between each request.
- Even if it not specified by a robots.txt file
- In practical terms a sleep()

Copyright Università degli Studi di Milano

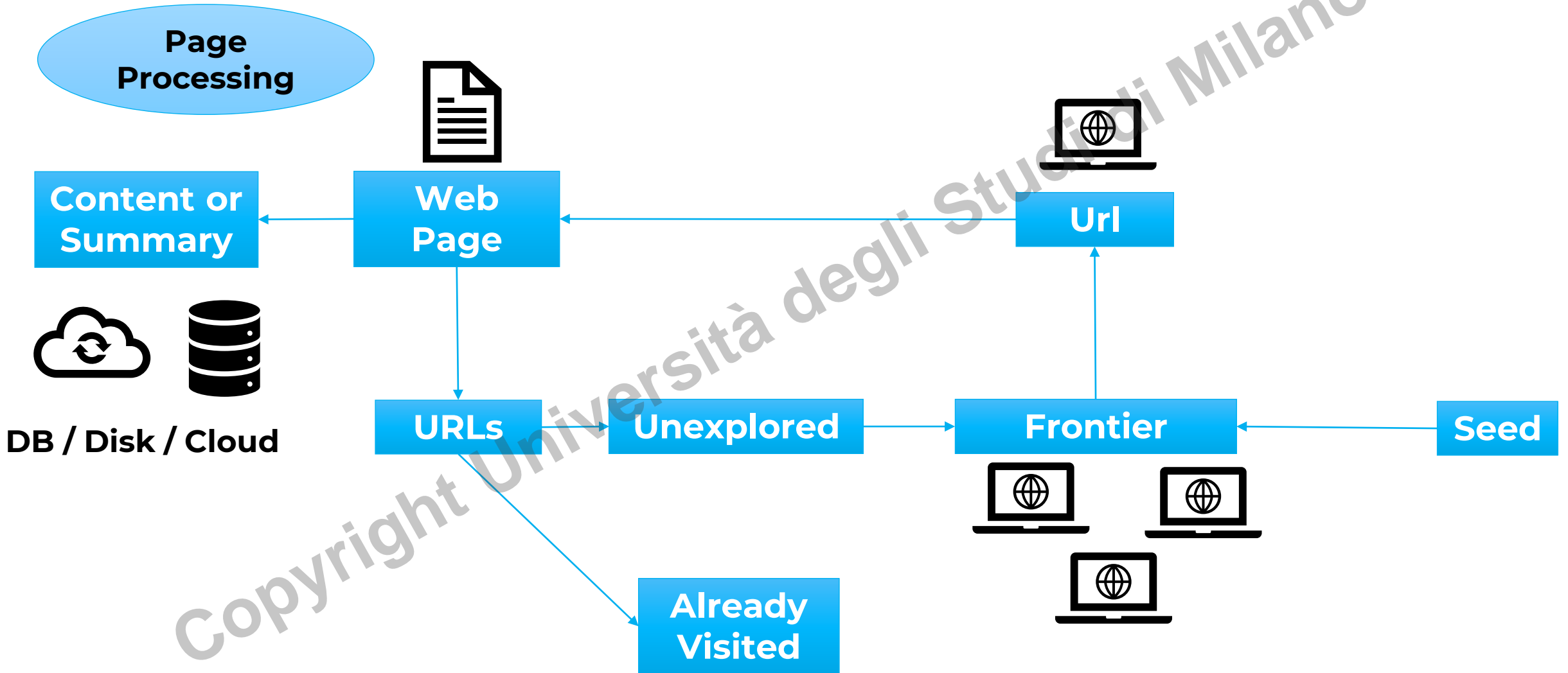
How a crawler can prevent issues

- **Option 2: limit the fraction of time of download compared to the time passed not downloading.**
- **Given:**
 - a fraction of time p
 - maximum download time s e.g (1s)
- **We want that the proportion of download time and non-download time to be equal to p**
- **Caveats:**
 - We need to monitor s
 - Slow resources could require a bigger s time, so we need to adjust.

Page Processing and Storage

Copyright Università degli Studi di Milano

The system



Content processing

- Given a new page, we may want to:
 - 1) extract new URLs
 - 2) Save the content
 - Apply some processing functions
 - Usually compressed before saving in a DB.
 - For this large-scale data, usually a distributed database, e.g. Mongo or Google's Bigtable

Copyright Università degli Studi di Milano

Content processing

- **Potential processing steps include**
 - Remove dates
 - Delete markup, html tags or attributes
 - Remove links
 - Remove headers
 - Remove or execute javascript code (if we want to save the dynamic content of a page, we need to be careful about code with infinite loops or broken code (sometimes we have traps))
- **We may not be interested in the entire page**
 - Extract the interesting values and bundle them up, in a file or json structure

Duplicate pages

- Delete duplicate pages to save storage space
- Some pages are exact duplicates non memorizzare pagine che non cambiano di molto
 - google.com, google.it
- Some pages are quasi-duplicates, as they update small parts like dates or random ids
 - nytimes.com, nyt.com
- Some websites are crawler traps
 - They generate random links to trap crawlers

The screenshot shows the top navigation bar of The New York Times website. On the left, there is a menu icon and a search icon. The date is displayed as "Wednesday, November 18, 2020" with "Today's Paper" below it. The main title "The New York Times" is centered in a large, black, serif font. To the right of the title, there are buttons for "SUBSCRIBE NOW" and "LOG IN". Below the title, there is a row of navigation links: World, U.S., Politics, N.Y., Business, Opinion, Tech, Science, Health, Sports, Arts, Books, Style, Food, Travel, Magazine, T Magazine, Real Estate, and Video. In the top right corner, there is a weather widget showing "6°C" and "9° 4°", and a market data widget showing "S&P 500 -0.93% ↓".

When can we check for duplicates

- We can check at two different moments in time
 1. After storage
 - Save all the pages, clear duplicates after the crawling process
 2. Before storage
 - We need to identify potential duplicates before saving them

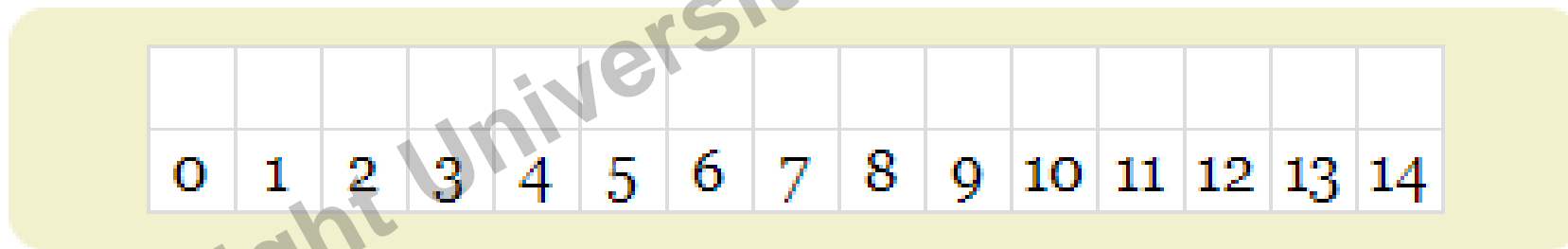
Copyright Università degli Studi di Milano

Bloom filters

- Data structure for the detection of quasi duplicates
- Compact representation of big sets of elements
- Characterized by
 - Rapid answer
 - Memory efficient
- Probabilistic data structure
 - The price for efficiency
- What we can do:
 - Add elements to the list of seen elements
 - And ask if we have seen a certain element already

Bloom filters

- Bit Vector
- Compact space
- We can keep it in RAM for efficient checks



Bloom filters

- It tells us that the element either definitely is not in the set or may be in the set.
- Asking if an element is contained, can yield 2 results:
 - false: definitely is not in the set
 - true: may be in the set
- The accuracy/ probability of error of the answer depends on the amount on the number of elements we plan to save (insert)

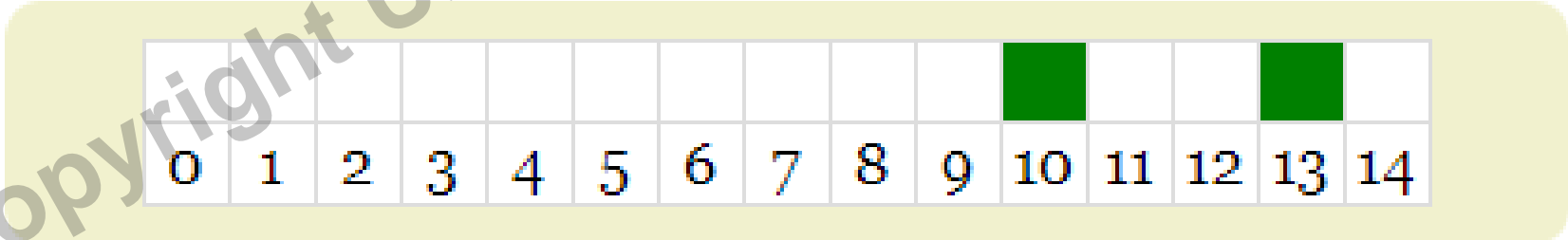
Bloom Filter example

- Given a Bloom filter of 15 bits
- A set of URLs X
- Two hash function h_1, h_2
- I want to add the following web page summary: «the fox is on the table»

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Bloom Filter example

- I want to add to our set the summary x e.g. «the fox is on the table»
- Apply each hash function
 - E.g. $h_1(x) = 10$; $h_2(x) = 13$
- Set the bits in position 10 and 13 to 1

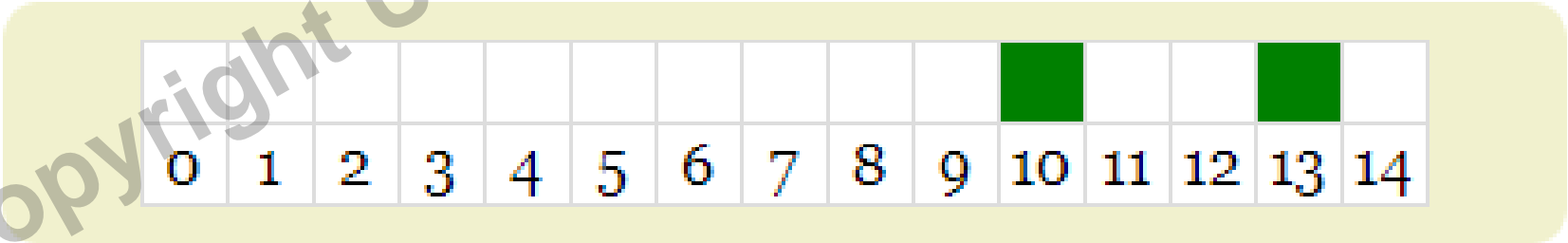


A diagram of a Bloom filter array consisting of 15 bits, indexed from 0 to 14. The bits at positions 10 and 13 are highlighted in green, indicating they are set to 1. All other bits are white, indicating they are 0.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
---	---	---	---	---	---	---	---	---	---	----	----	----	----	----

Bloom Filter example

- I want to check if the summary x e.g. «the fox is on the table»
- Apply each hash function
 - E.g. $h1(x) = 10$; $h2(x) = 13$
- Check the bit values bit values in positions 10 and 13
 - The AND combination of the bit values can be 1 (true) or 0 (false)



A diagram of a Bloom filter array consisting of 15 bits, indexed from 0 to 14. The bits at positions 10 and 13 are highlighted in green, indicating they are set to 1. All other bits are white, indicating they are 0.

										1			1	
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Bloom Filter example

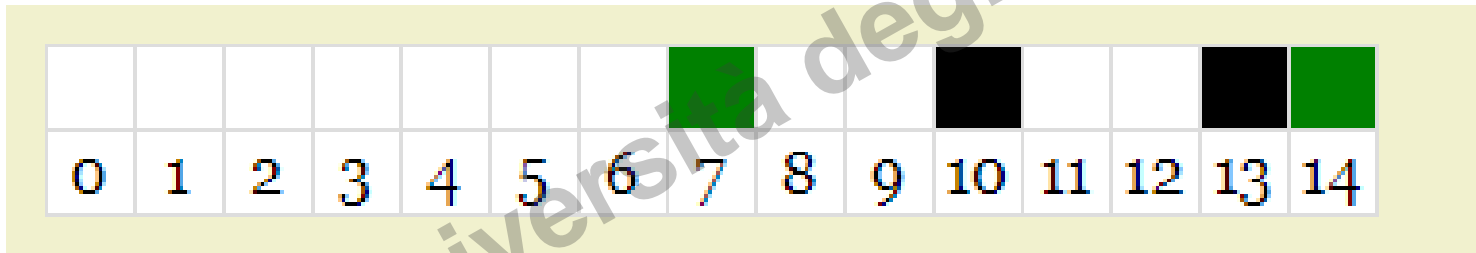
- **Reminder: asking if an element is contained, can yield 2 results:**
 - **False (0) : definitely is not in the set**
 - **True (1): may be in the set**
- **The AND combination of the bits can be 1 or 0**
- **If all those values are set to 1 in the bit vector, it might be because another element or some combination of other elements could have set the same bits**
 - **So I say that may be in the set, we don't know for sure**
- **if at least 1 of those values are set to 0, you know that the element isn't in the set**
 - **We know for sure**

Why bloom filters

- **The filter is useful as it limits the access to Storage.**
 - Avoid access to slower storage like and HDD
 - Particulary effective if we have enough bits to obtain often negative answers
 - Allows us to deal with large sets, without checking them directly
 - Important for big elements like URLs
- **This data structure offers us graceful degradation (when full we always check the disk)**
- **Key property: we can influence the probability of error by knowing how many elements i may have to save**

Graceful degradation

- The filter fills up as we add new elements
- Adding «the cat is on the table»
 - $h1(x) = 14$, $h2(x) = 7$



- With time we have more false positives
 - We obtain 1 but we didn't actually see the element before
- As it fills up, we'll check the disk more often

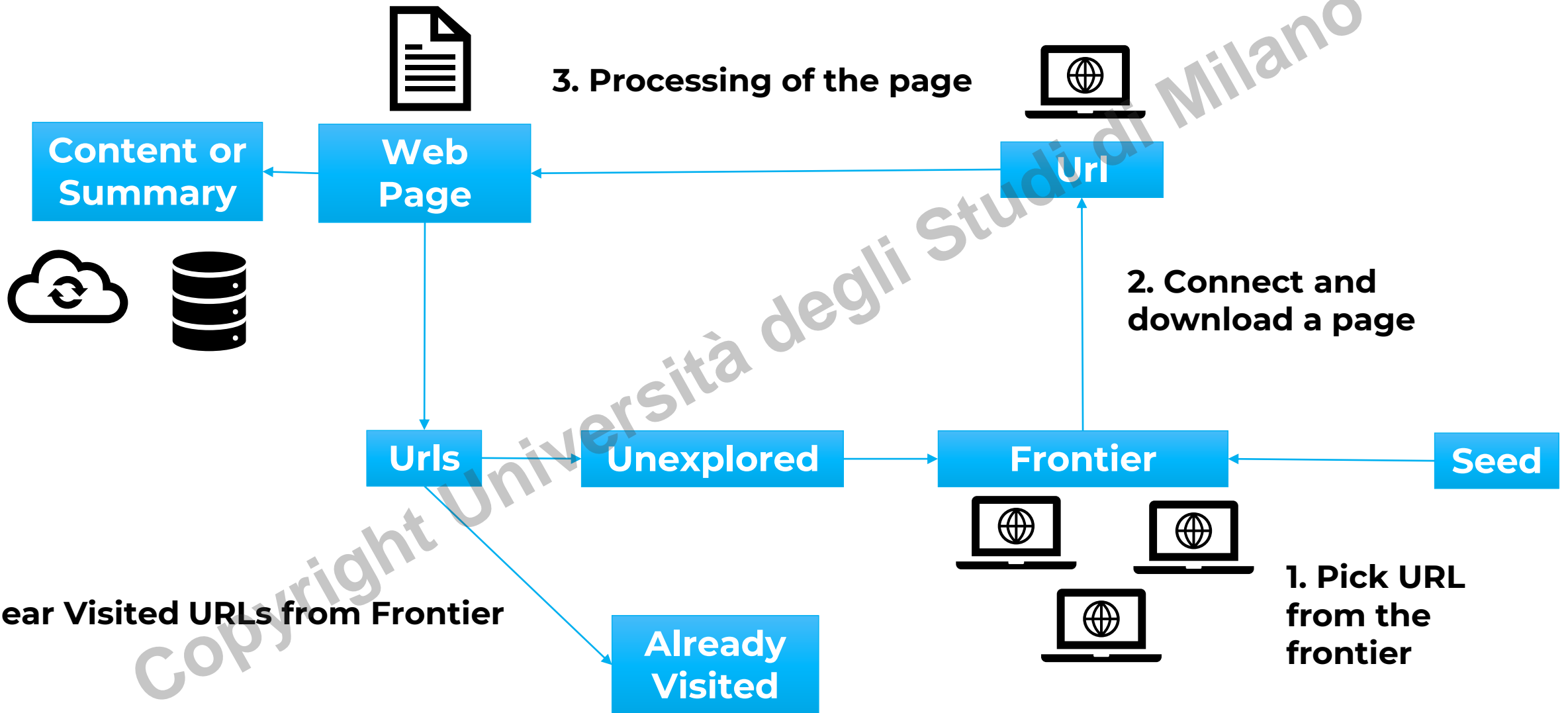
Probability of error

- A mathematical analysis show us that we can choose a proper number of hash functions
- The choice depends on the number of bits m , and the probability of error we aim to obtain
- So we can decide the right amount of bits required
- The analysis gives us the probability to observe a positive answer, (false or true) after n inserts
- This probability is a majoration of the probability of a false positive.

Copyright Università degli Studi di Milano

Recap

The system



Next lesson

- Quick recap
- Bloom Analysis
- Frontier data structure
- Load management for distributed crawling systems

Copyright Università degli Studi di Milano

Copyright Università degli Studi di Milano

References

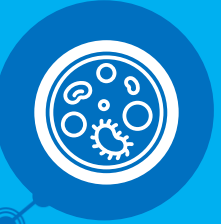
References

- <http://vigna.di.unimi.it/algoweb/>
- <https://lilimlib.github.io/bloomfilter-tutorial/> (bloom filter demo)
- Burton H. Bloom. Space-time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970
- <https://www.cs.princeton.edu/courses/archive/spring02/cs493/lec6.pdf> (Bloom filter analysis, section 3.1)

Thank you for the attention!

For any question send an email at
cheick.ba@unimi.it

Copyright Università degli Studi di Milano



COMMUNITY DETECTION

SOURCES

Books:

Barabasi, Network Science, Chapter 9

Zafarani, Social Media Mining, Chapter 6 (particularly the introductory part)

Newman, Networks

Papers:

Santo Fortunato, Darko Hric, Community detection in networks: A user guide, Physics Reports, Volume 659, 11 November 2016, Pages 1-44, ISSN 0370-1573, <https://doi.org/10.1016/j.physrep.2016.09.002>.

(<http://www.sciencedirect.com/science/article/pii/S0370157316302964>)

Santo Fortunato, Community detection in graphs, Physics Reports, Volume 486, Issues 3–5, February 2010, Pages 75-174, ISSN 0370-1573, <https://doi.org/10.1016/j.physrep.2009.11.002>.

(<http://www.sciencedirect.com/science/article/pii/S0370157309002841>)

Why to study communities?

- individuals often form groups based on their interests and we are interested in identifying these groups. Consider the importance of finding groups with similar reading tastes by an online book seller for recommendation purposes.
- groups provide a clear global view of user interactions, whereas a local-view of individual behavior is often noisy and ad hoc (mesoscale).
- some behaviors are only observable in a group setting and not on an individual level. This is because the individual's behavior can fluctuate, but group collective behavior is more robust to change.

Communities

- Two types of communities:
 - **Explicit Groups**: formed by user subscriptions
 - **Implicit Groups**: implicitly formed by social interactions
 - (individuals calling Canada from the United States need not be friends) -> the phone operator considers them one community for marketing purposes
- We may see *group*, *cluster*, *cohesive subgroup*, or *module* in different contexts instead of “community”

Examples of explicit social media community

- **Facebook**
 - Facebook has groups and communities. In both, users can post messages and images, can comment on other messages, can like posts, and can view activities of other users
- **Google+**
 - Circles in Google+ represent communities
- **Twitter**
 - Communities form as lists. Users join lists to receive information in the form of tweets
- **LinkedIn**
 - LinkedIn provides *Groups* and *Associations*. Users can join professional groups where they can post and share information related to the group

COMMUNITY DETECTION

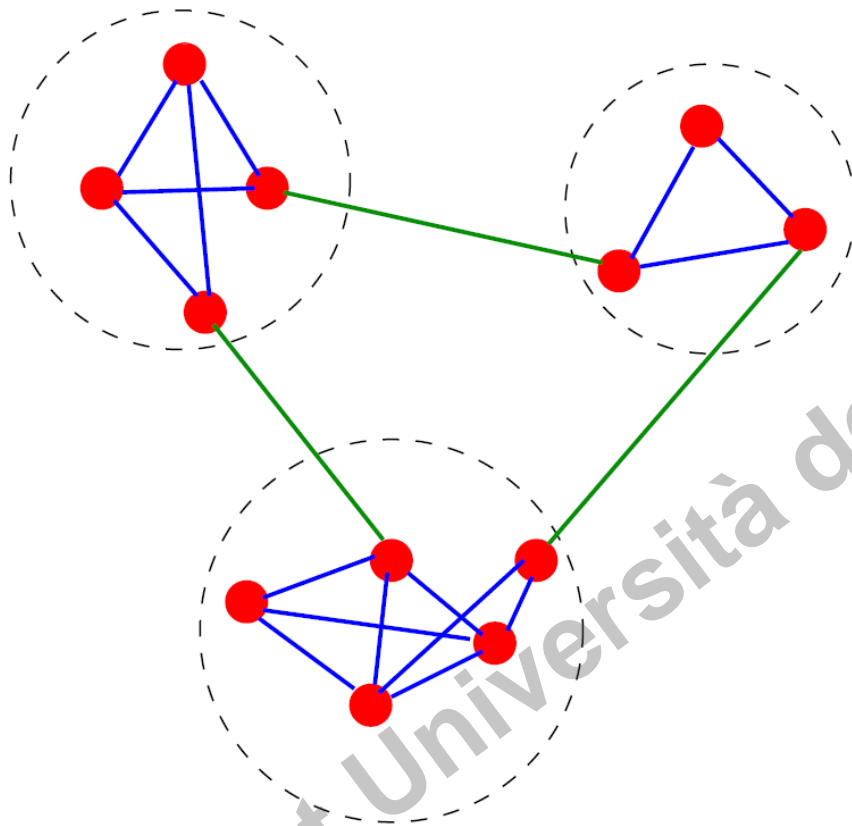
DISCOVERING IMPLICIT COMMUNITIES

COMPUTE SETS OF NODES BASED
ON THEIR CONNECTIVITY

Hypothesis:

The network community structure is
encoded in its wiring diagram

Real networks have community structure



Real networks are not random.
Weak ties seem to bridge groups of tightly coupled nodes (communities)

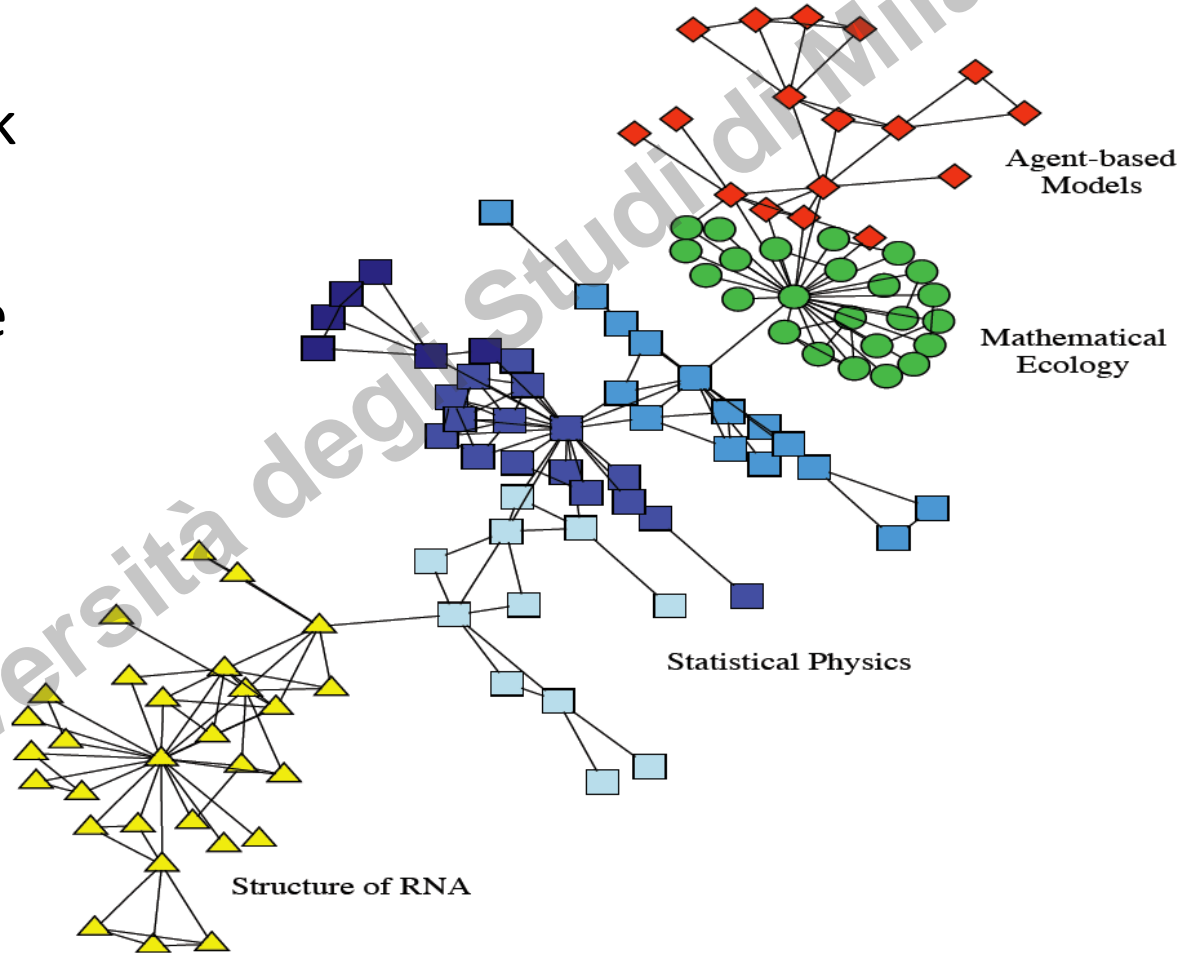
A simple graph with three communities, enclosed by the dashed circles

Source: S. Fortunato / Physics Reports 486 (2010) 75–174

Example: scientist collaboration network

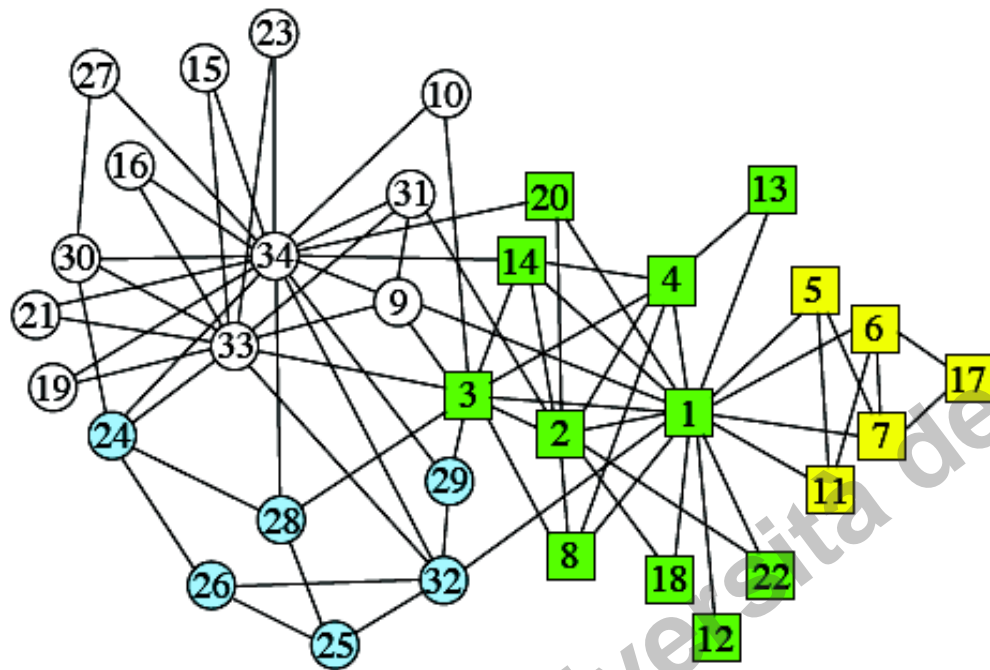
- Collaboration network between scientists working at the Santa Fe Institute. Edges are placed between scientists that have published at least one paper together.

The colors indicate high level communities and correspond to research divisions of the institute



Source: S. Fortunato / Physics Reports 486 (2010) 75–174

Example: Zachary's Karate Club



Zachary observed 34 members of a karate club over two years. Edges connect individuals who were observed to interact outside the activities of the club.

During the course of observation, the club members split into two groups because of the disagreement between the administrator of the club and the club's instructor (nodes: 1 and 34), and the members of one group left to start their own club

Disjoint communities

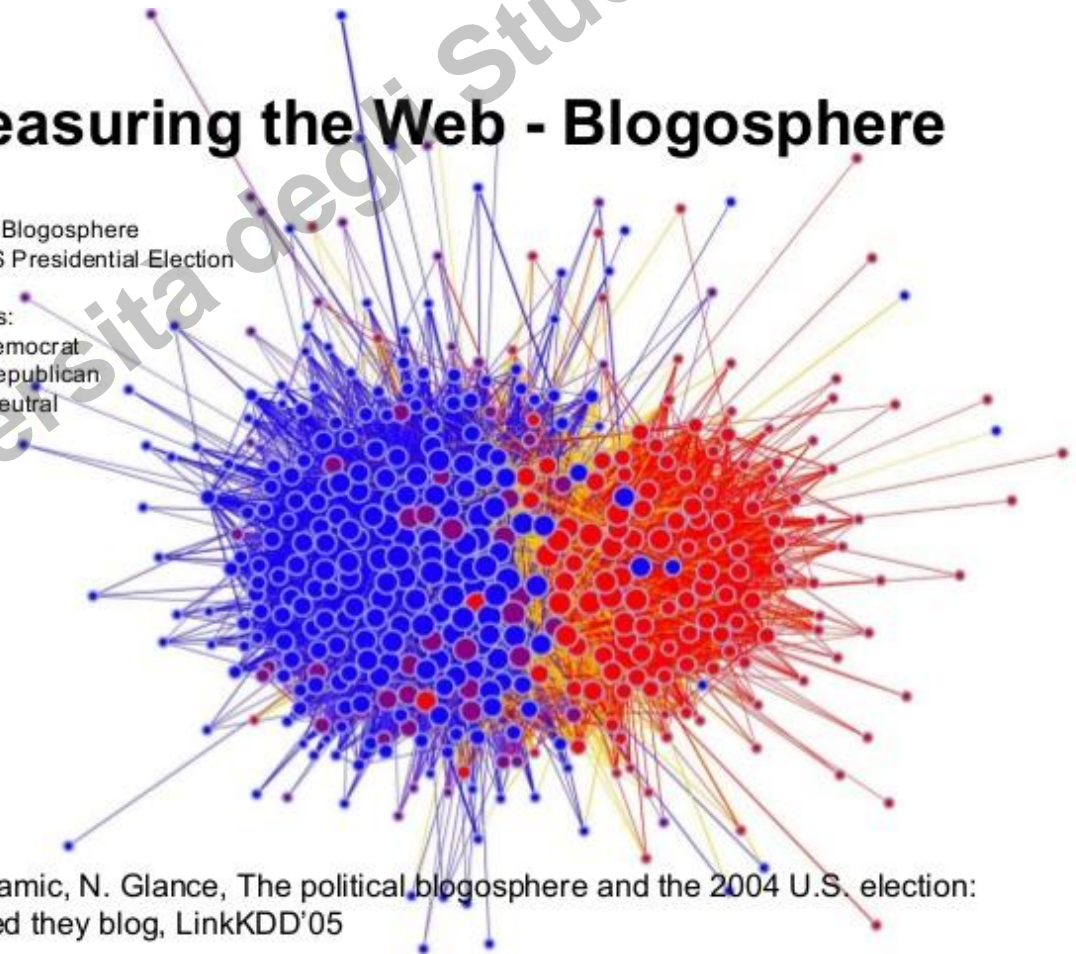
Separating networks into disjoint subsets seems to make sense when communities are somehow “adversarial”

Measuring the Web - Blogosphere

Political Blogosphere
2004 US Presidential Election

Bloggers:
Blue - Democrat
Red - Republican
Pink - Neutral

L. Adamic, N. Glance, The political blogosphere and the 2004 U.S. election: divided they blog, LinkKDD'05



Communities

Disjoint communities (i.e., groups of friends who don't know each other) e.g. my American friends and my Australian friends

Overlapping communities (i.e., groups with some intersection) e.g. my friends and my girlfriend's friends

Defining communities

There is no unique definition of community

Intuition:

There are more links inside a community than links connected with the rest of the network

Hypothesis:

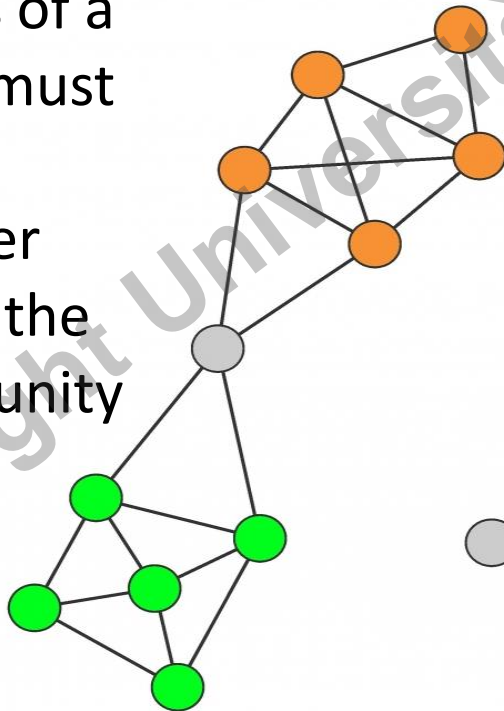
a community is a locally dense connected subgraph in a network

Hypotheses

Connectedness Hypothesis:

A community corresponds to a connected subgraph.

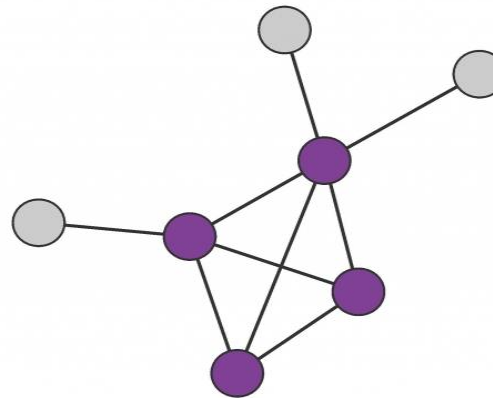
All members of a community must be reached through other members of the same community



Density Hypothesis:

Communities correspond to locally dense neighborhoods of a network.

Nodes of the same community has higher probability of linking to other members of the same community than to nodes outside it



Local definition: maximum cliques

One of the first paper on community defined a community as a group of individuals whose members all know each other

- It is a connected subgraph with maximal link density
- Triangles are frequent; larger cliques are rare.
- Finding the cliques of a network is computationally rather demanding, being a so-called NP-complete problem.
- Too restrictive: communities do not necessarily correspond to complete subgraphs, as many of their nodes do not link directly to each other.
- Relaxing cliques
 - *n-clique, n-clan, n-club, k-plex*
 - *k-core*: maximal subgraph that each vertex is adjacent to at least *k* other vertices in the subgraph

Almost loca definitions

Graph G , a connected subgraph C and node i

The *internal degree* k_i^{int} of node i is the number of links that connect i to other nodes in C .

The *external degree* k_i^{ext} is the number of links that connect i to the rest of the network.

If $k_i^{ext}=0$, each neighbor of i is within C , hence C is a good community for node i . If $k_i^{int}=0$, then node i should be assigned to a different community

k_i^{int}, k_i^{ext} : internal and external degrees of $i \in C$

k_{int}^C, k_{ext}^C : internal and external degrees of C

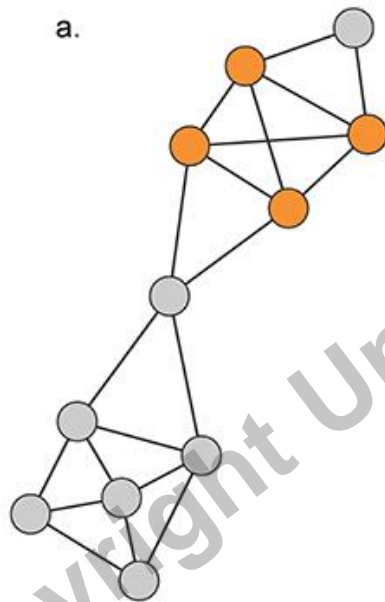
Sum of the Internal and external degrees of all $v \in C$

Almost local definitions

strong community:

each node has more links within the community than with the rest of the graph.

$$k_i^{\text{int}}(C) > k_i^{\text{ext}}(C)$$

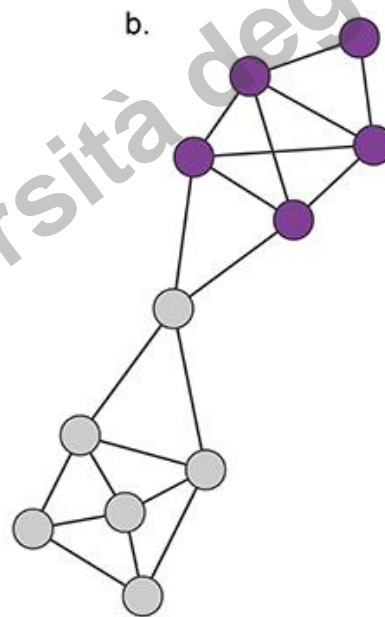


Clique

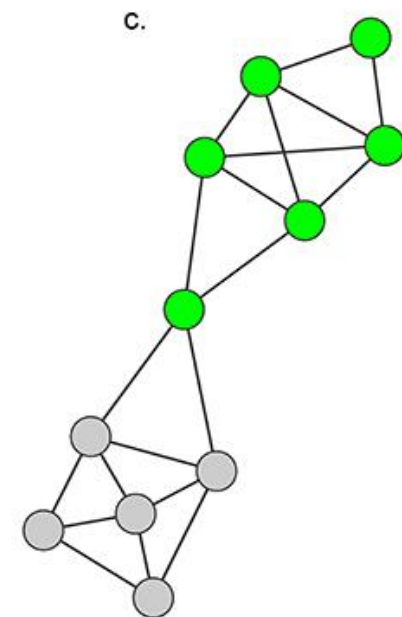
weak community:

the total internal degree of the subgraph exceeds its total external degree,

$$\sum_{i \in C} k_i^{\text{in}}(C) > \sum_{i \in C} k_i^{\text{out}}(C)$$



Strong community

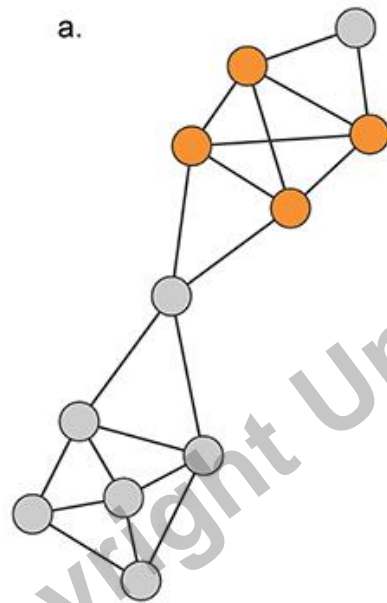


Weak community

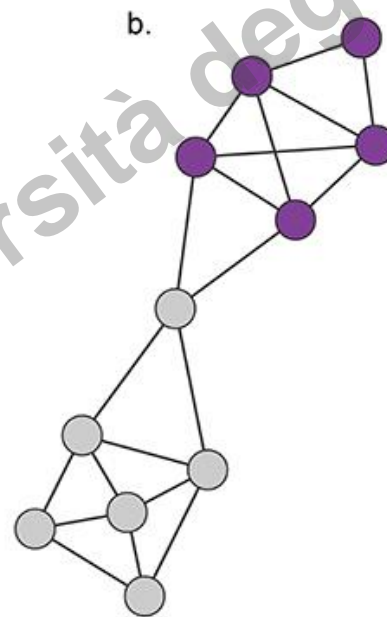
Almost local definitions

Clique \rightarrow *strong community* \rightarrow *weak*

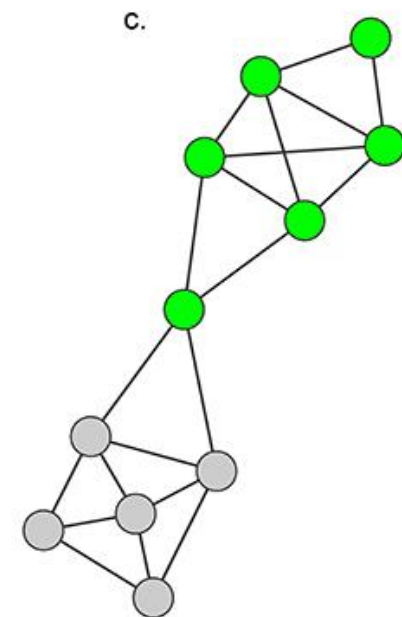
Is the converse true? No



Clique



Strong community



Weak community

Partitions

Definition:

We call a partition a division of a network into an arbitrary number of groups, such that each node belongs to one and only one group.

Community detection:

the number and size of the communities are unknown at the beginning.

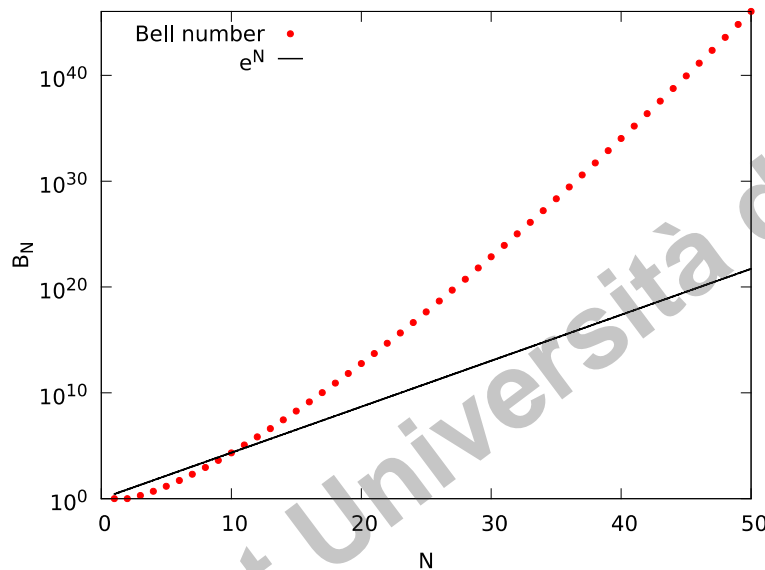
Partition detection:

division of a network into groups of nodes, so that each node belongs to one group.

the number and size of the communities are known at the beginning.

Partitions

How many ways can we partition a network into communities?



The number of possible partitions is given by the Bell number and grows faster than exponentially

Brute-force approaches that aim to identify communities by inspecting all possible partitions are computationally infeasible

Global definition: modularity

Randomly wired networks lack an inherent community structure

By comparing the link density of a community with the link density obtained for the same group of nodes for a randomly rewired network, we could decide if the original community corresponds to a dense subgraph, or its connectivity pattern emerged by chance.

Global definition: modularity

Systematic deviations from a random configuration allow us to define a quantity called *modularity*

It measures the quality of each partition.

It allows us to decide if a particular community partition is better than some other one.

Modularity optimization offers a novel approach to community detection.

Global definition: with respect to the whole graph

- Null model: A random graph where some structure properties are matched with the original graph
- Intuition: a subgraph is a community if the number of internal links exceeds the expectation over all realizations of the null model

Modularity measures the difference between the network's real wiring diagram (A_{ij}) and the expected number of links between i and j if the network is randomly wired (p_{ij})

Definition:
$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - p_{ij}) \delta(C_i, C_j)$$

- p_{ij} : expected number of links between i and j in the null model
- random graph: $p_{ij} = p, \forall i, j$

Modularity is the fraction of the links that fall within the given groups minus the expected such fraction if links were distributed at random

Higher Modularity Implies Better Partition

The higher is M for a partition, the better is the corresponding community structure

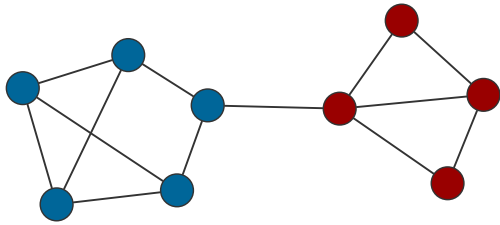
Modularity

- Range: $[-\frac{1}{2}, 1)$
- if we treat the whole graph as one community $Q = 0$
- if each vertex is one community $Q < 0$

Maximal Modularity Hypothesis

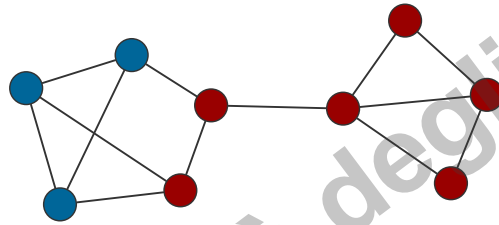
(a)

Optimal Partition
 $M = 0.4$



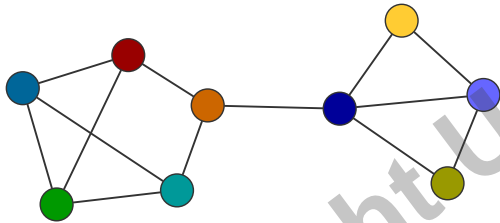
(b)

Suboptimal Partition
 $M = 0.22$



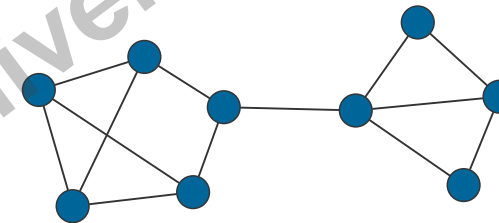
(c)

Negative Modularity
 $M = -0.12$



(d)

Single Community
 $M = 0$



- *Optimal partition*: maximizes the modularity.
- *Sub-optimal* but positive modularity.
- *Negative Modularity*: if we assign each node to a different community.
- *Zero modularity*: Assigning all nodes to the same community, we obtain , independent of the network structure.
- *Modularity is size dependent*.

Modularity-based community detection methods

- Modularity maximization

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

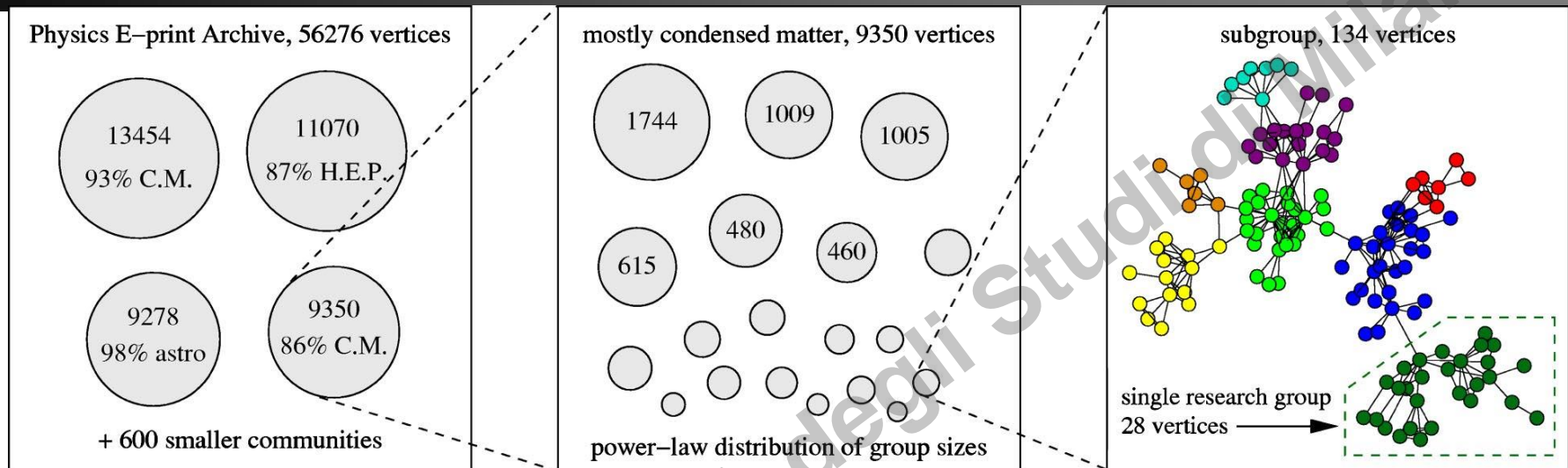
- Finding the best value for Q is NP hard
- Hence we use heuristics

Modularity maximization: greedy algorithm

Greedy techniques [Newman], iteratively joins nodes if the move increases the partition's modularity.

- 1. Start with all nodes as isolated that is assign each node to a community of its own, e.g. start with “communities”.
- 2. Inspect each pair of communities connected by at least one link and compute the modularity variation (on the full network) obtained if we merge these two communities.
- 3. Identify the community pair for which ΔM is the largest and merge them.
- 4. Repeat Step 2 and 3 until all nodes are merged into a single community.
- 5. Record M for each step and select the partition for which the modularity is maximal.
- Issues: limit resolution and modularity maxima

Modularity maximization: resolution limit



The community structure of the collaboration network of physicists. The greedy algorithm predicts four large communities, each composed primarily of physicists of similar interest. These four large communities (together containing 77% of all nodes) coexist with 600 smaller communities, resulting in an overall modularity $M=0.713$.

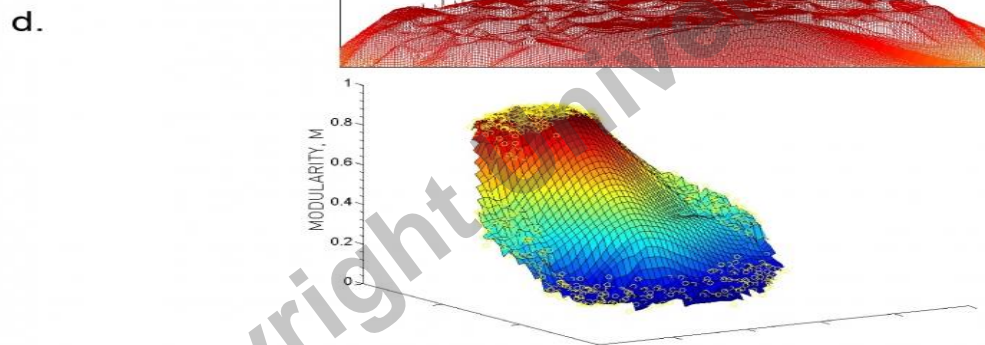
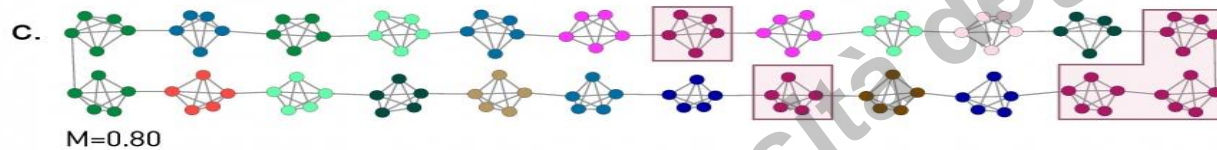
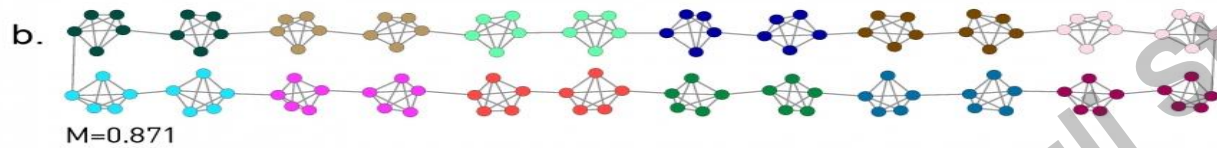
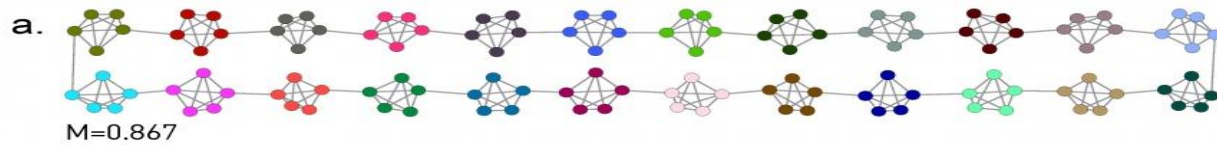
Identifying Subcommunities

We can identify subcommunities by applying the greedy algorithm to each community, treating them as separate networks. This procedure splits the condensed matter community into many smaller subcommunities, increasing the modularity of the partition to $M=0.807$.

Research Groups

One of these smaller communities is further partitioned, revealing individual researchers and the research groups they belong to.

Modularity maxima



All algorithms based on maximal modularity rely on the assumption that a network with a clear community structure has an optimal partition with a maximal M .

In practice we hope that M_{max} is easy to find and that the communities predicted by all other partitions are distinguishable from those corresponding to M_{max} .

Yet, this optimal partition is difficult to identify among a large number of close to optimal partitions.

Modularity maximization: fast modularity

The greedy algorithm is neither particularly fast nor particularly successful at maximizing M .

Scalability: Due to the sparsity of the adjacency matrix, the update of the matrix involves a large number of useless operations.

The use of data structures for sparse matrices can decrease the complexity of the computational algorithm to $O(N \log^2 N)$

See:

Clauset, Aaron, "Fast Modularity" Community Structure Inference Algorithm.

<http://www.cs.unm.edu/~aaron/research/fastmodularity.htm> (2012).

Modularity maximization: Louvain algorithm

Louvain method: Finding communities in large networks

The modularity optimization algorithm achieves a computational complexity of $O(L)$.

Hence it allows us to identify communities in networks with millions of nodes.

*Fast unfolding of communities in large networks,
Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre,
Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)*

Louvain algorithm: weighted network of N nodes

“Our algorithm is divided in two phases that are repeated iteratively.

Phase 1

- First, we assign a different community to each node of the network. So, in this initial partition there are as many communities as there are nodes.
- Then, for each node i we consider the neighbors j of i and we evaluate the gain of modularity that would take place by removing i from its community and by placing it in the community of j . The node i is then placed in the community for which this gain is maximum (in case of a tie we use a breaking rule), but only if this gain is positive. If no positive gain is possible, i stays in its original community.
- This process is applied repeatedly and sequentially for all nodes until no further improvement can be achieved and the first phase is then complete. Let us insist on the fact that a node may be, and often is, considered several times.
- This first phase stops when a local maxima of the modularity is attained, i.e., when no individual move can improve the modularity.

One should also note that the output of the algorithm depends on the order in which the nodes are considered. Preliminary results on several test cases seem to indicate that the ordering of the nodes does not have a significant influence on the modularity that is obtained.”

Louvain algorithm: weighted network of N nodes

Our algorithm is divided in two phases that are repeated iteratively.

Phase 2

We construct a new network whose nodes are the communities identified during phase 1.

The weight of the link between two nodes is the sum of the weight of the links between the nodes in the corresponding communities. Links between nodes of the same community lead to weighted self-loops.

Once phase 2 is completed, we repeat phases 1 - 2, calling their combination a *pass*.

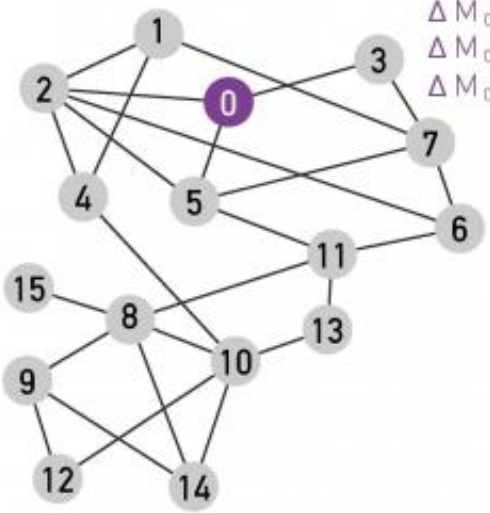
The number of communities decreases with each pass.

The passes are repeated until there are no more changes and maximum modularity is attained.

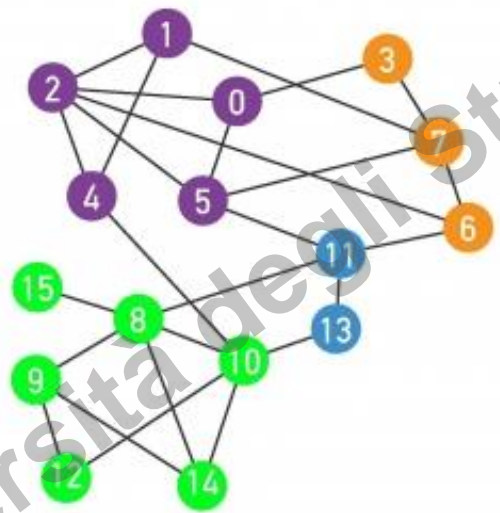
Louvain algorithm

1ST PASS

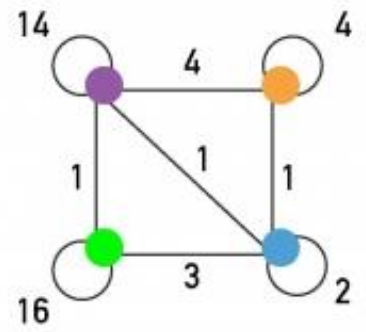
$\Delta M_{0,2} = 0.023$
 $\Delta M_{0,3} = 0.032$
 $\Delta M_{0,4} = 0.026$
 $\Delta M_{0,5} = 0.026$



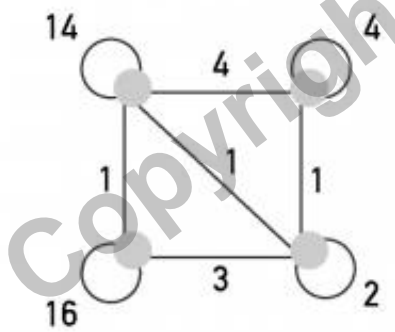
STEP I



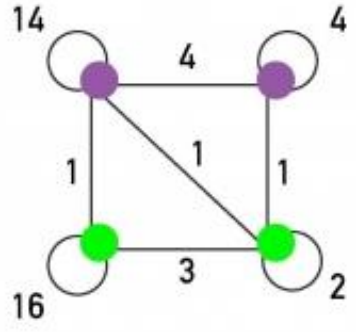
STEP II



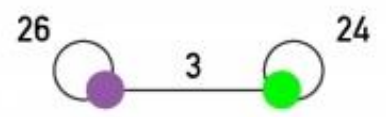
2ND PASS



STEP I



STEP II



Copyright Università degli Studi di Milano

Some studies that use the Louvain method

Twitter social network (2.4M nodes 38M links, Twitter)
Divide and Conquer: Partitioning Online Social Networks
Josep M. Pujol, Vijay Erramilli, Pablo Rodriguez
arXiv 0905.4918, 2010

LinkedIn social network (21M nodes, LinkedIn)
Mapping search relevance to social networks
Jonathan Haynes, Igor Perisic
Proceedings of the 3rd Workshop on Social Network Mining and Analysis, 2010

Audio sharing networks (Freesound)
Community structure in audio clip sharing
Gerard Roma, Perfecto Herrera
International Conference on Intelligent Networking and Collaborative Systems, INCoS 2010

Mobile phone networks (4M nodes, 100M links)
Tracking the Evolution of Communities in Dynamic Social Networks
Greene, D.; Doyle, D.; Cunningham, P.;
International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2010

Flickr 1.8M/22M, LiveJournal 5.3M/77M, YouTube 1.1M/4.5M
Real World Routing Using Virtual World Information
Pan Hui, Sastry N.
International Conference on Computational Science and Engineering, 2009

Citation network (6M nodes, ISI database)
Subject clustering analysis based on ISI category classification
Lin Zhang, Xinhai Liu, Frizo Janssens, Liming Liang and Wolfgang Glänzel
Journal of Informetrics, Volume 4, Issue 2, April 2010

Retail transaction data network
Market basket analysis with networks
Troy Raeder, Nitesh V. Chawla
Social Network Analysis and Mining, 2010

Human brain functional networks
Hierarchical Modularity in Human Brain Functional Networks
David Meunier, Renaud Lambiotte, Alex Fornito, Karen D. Ersche and Edward T. Bullmore
Neuroinformatics, 3: 37, 2009

Overlapping communities

Clique finder

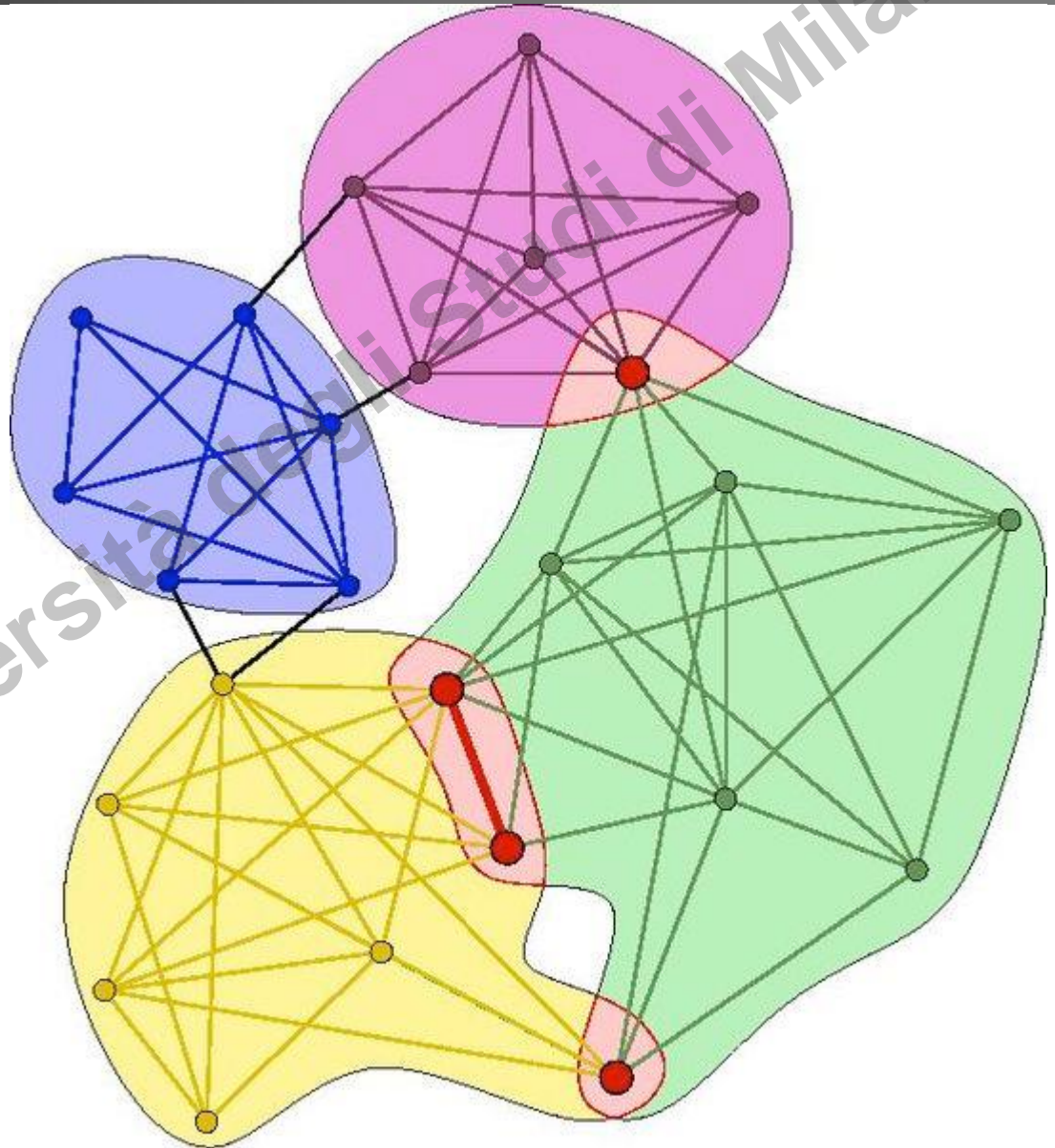
<http://cfinder.org>

Uncovering the overlapping community structure of complex networks in nature and society

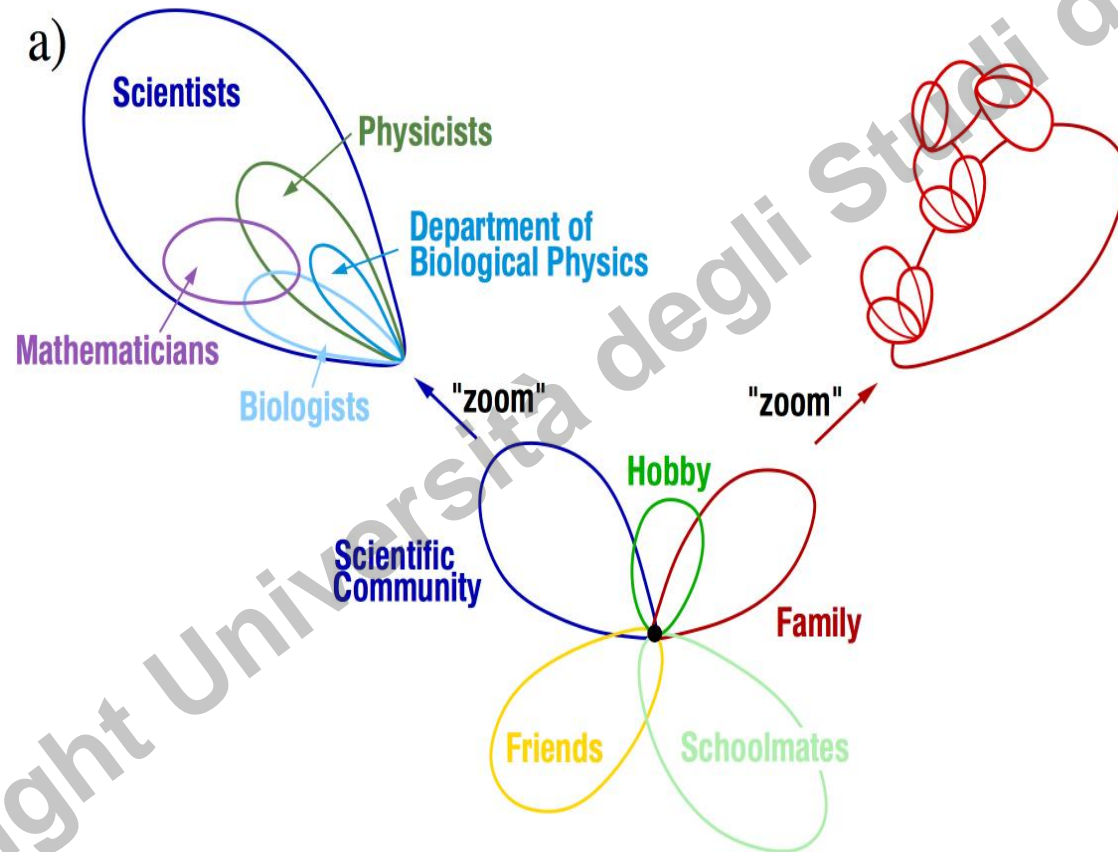
G. Palla, I. Derényi,

I. Farkas, and T. Vicsek:

Nature 435, 814–818 (2005)

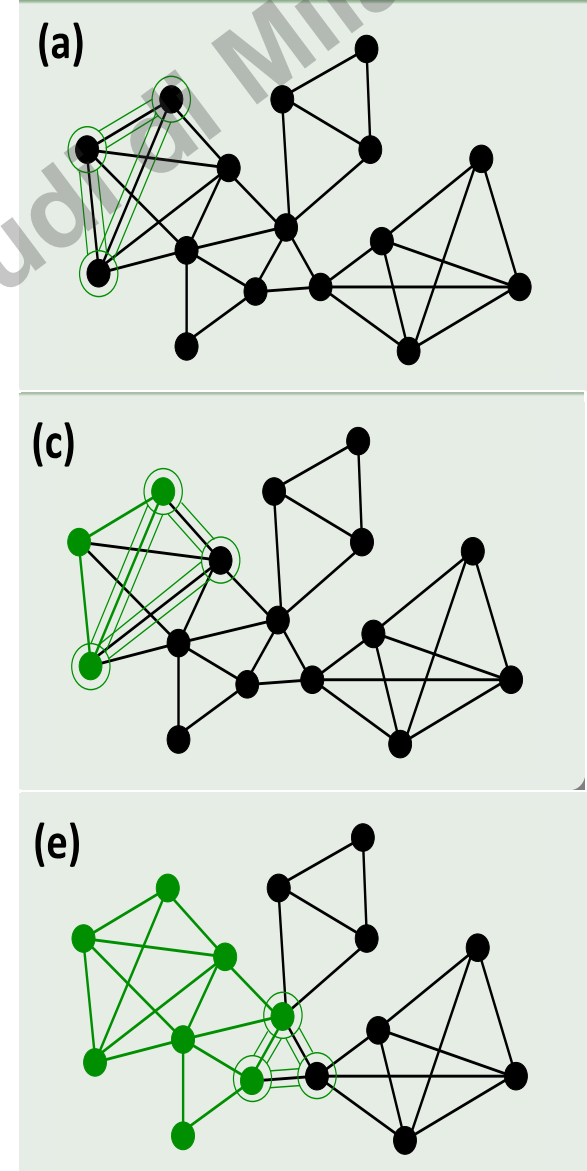


Overlapping communities

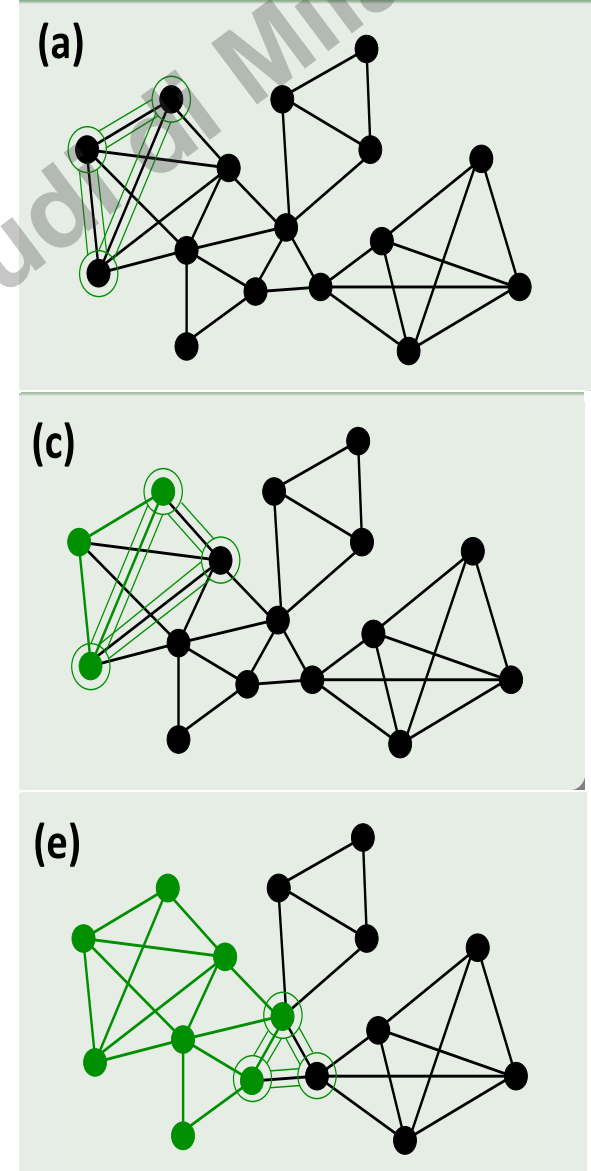
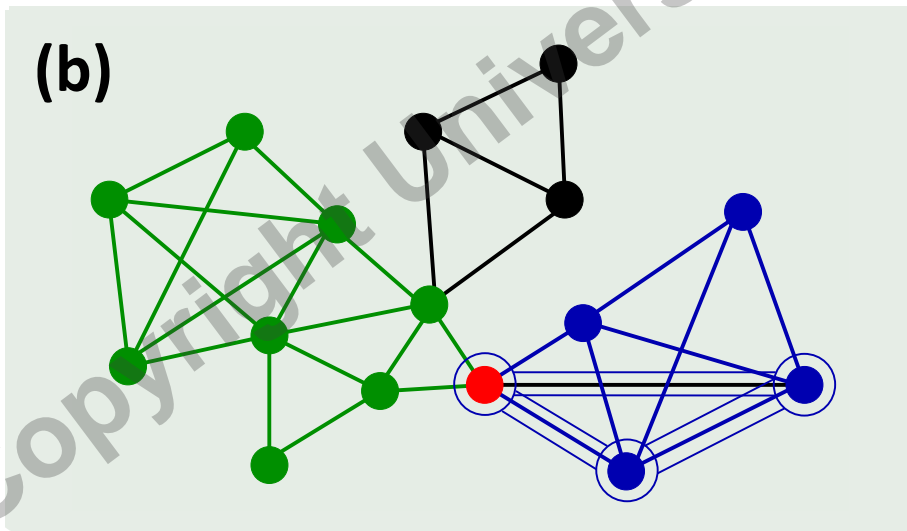


Overlapping communities: clique percolation

- Two k -cliques (complete subgraphs of k nodes) 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
- Other k -cliques that can not be reached from a particular k -clique correspond to other k -clique-communities



- Other k -cliques that can not be reached from a particular k -clique correspond to other k -clique-communities

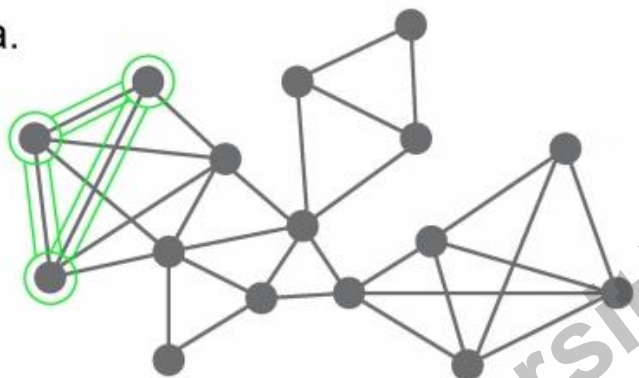


CPM: 4-clique

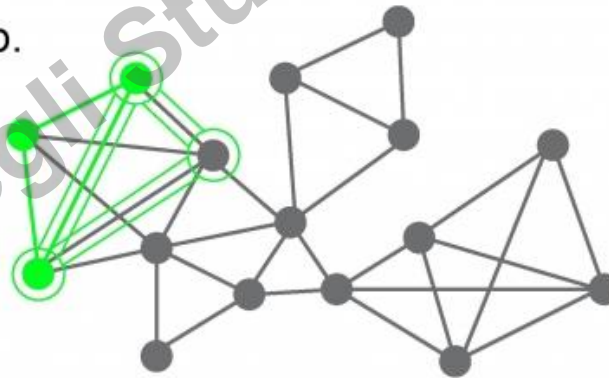
d.

$k=4$ community structure of a small network, consisting of complete four node subgraphs that share at least three nodes. Orange nodes belong to multiple communities.

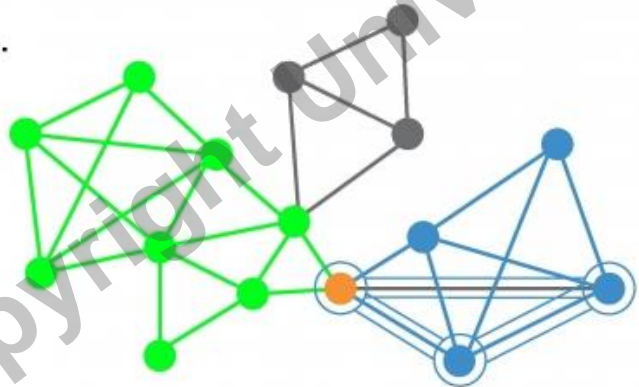
a.



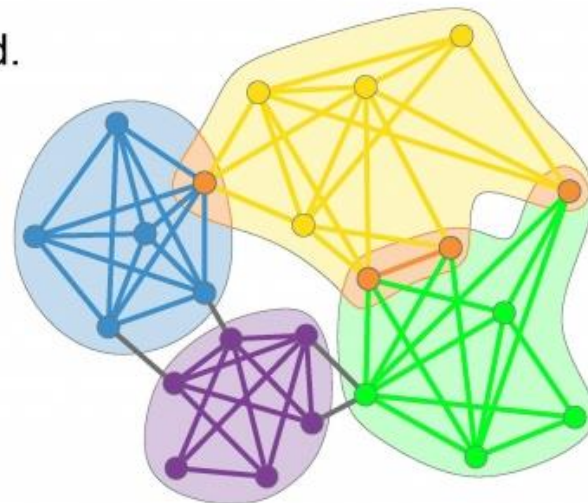
b.

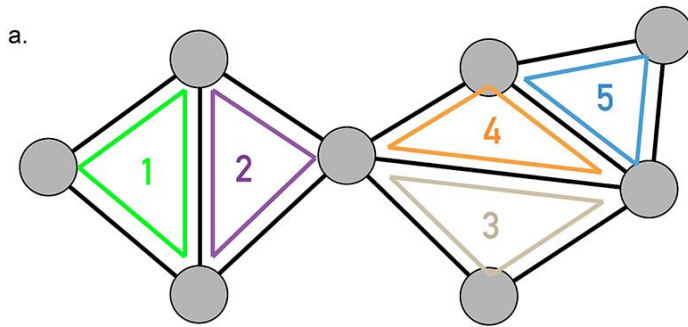


c.



d.

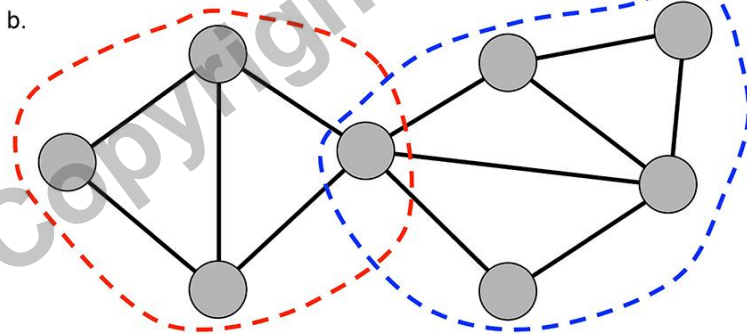
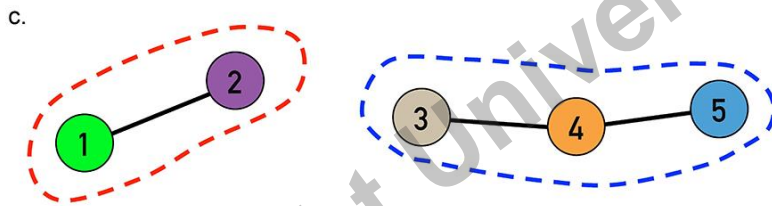




b.

$O =$

	1	2	3	4	5
1	0	1	0	0	0
2	1	0	0	0	0
3	0	0	0	1	0
4	0	0	1	0	1
5	0	0	0	1	0



CFinder algorithm

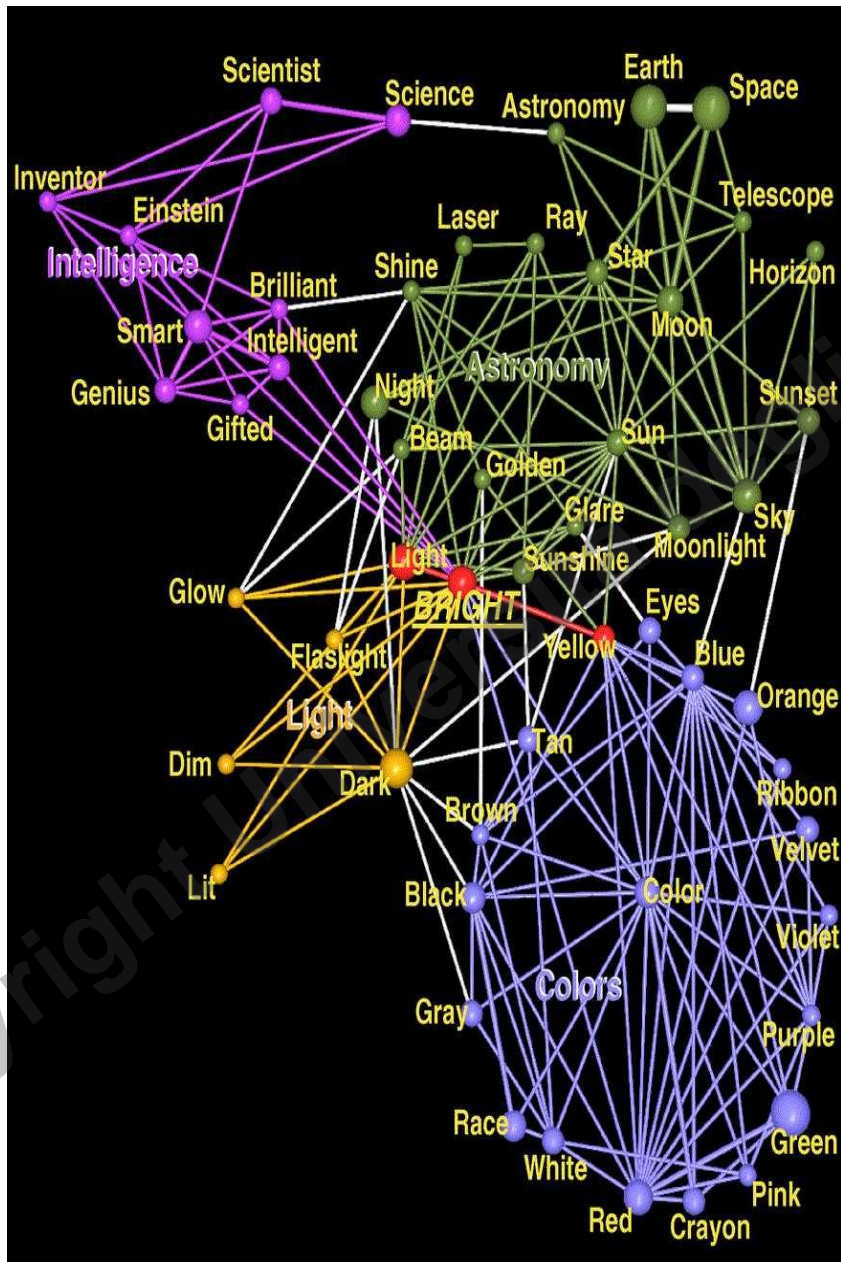
The main steps of the CFinder algorithm. Starting from the network shown in the figure, our goal is to identify all cliques. All five $k=3$ cliques present in the network are highlighted. The overlap matrix O of the $k=3$ cliques. This matrix is viewed as an adjacency matrix of a network whose nodes are the cliques of the original network.

The matrix indicates that we have two connected components, one consisting of cliques (1,2) and the other of cliques (3, 4, 5). The connected components of this network map into the communities of the original network.

The two clique communities predicted by the adjacency matrix.

The two clique communities shown in (c), mapped on the original network.

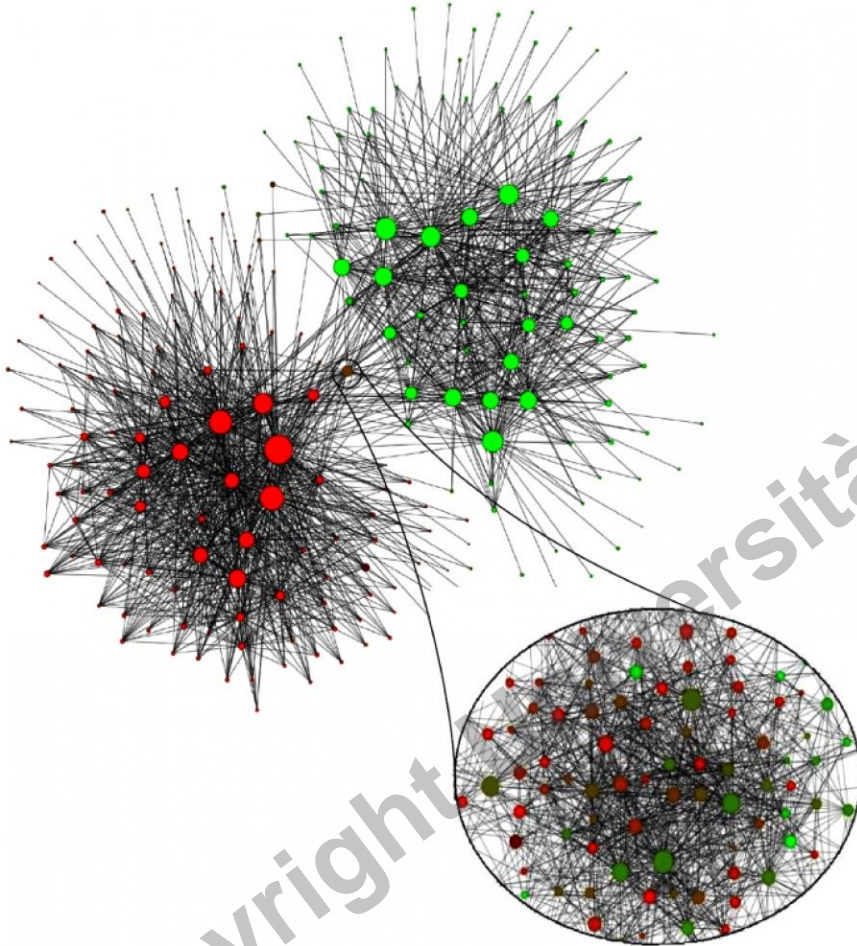
Example



bright:

- community containing light-related, *glow* or *dark*;
- community capturing different colors (*yellow*, *brown*)
- community consisting of astronomical terms (*sun*, *ray*).
- community linked to intelligence (*gifted*, *brilliant*).

Example: mobile call

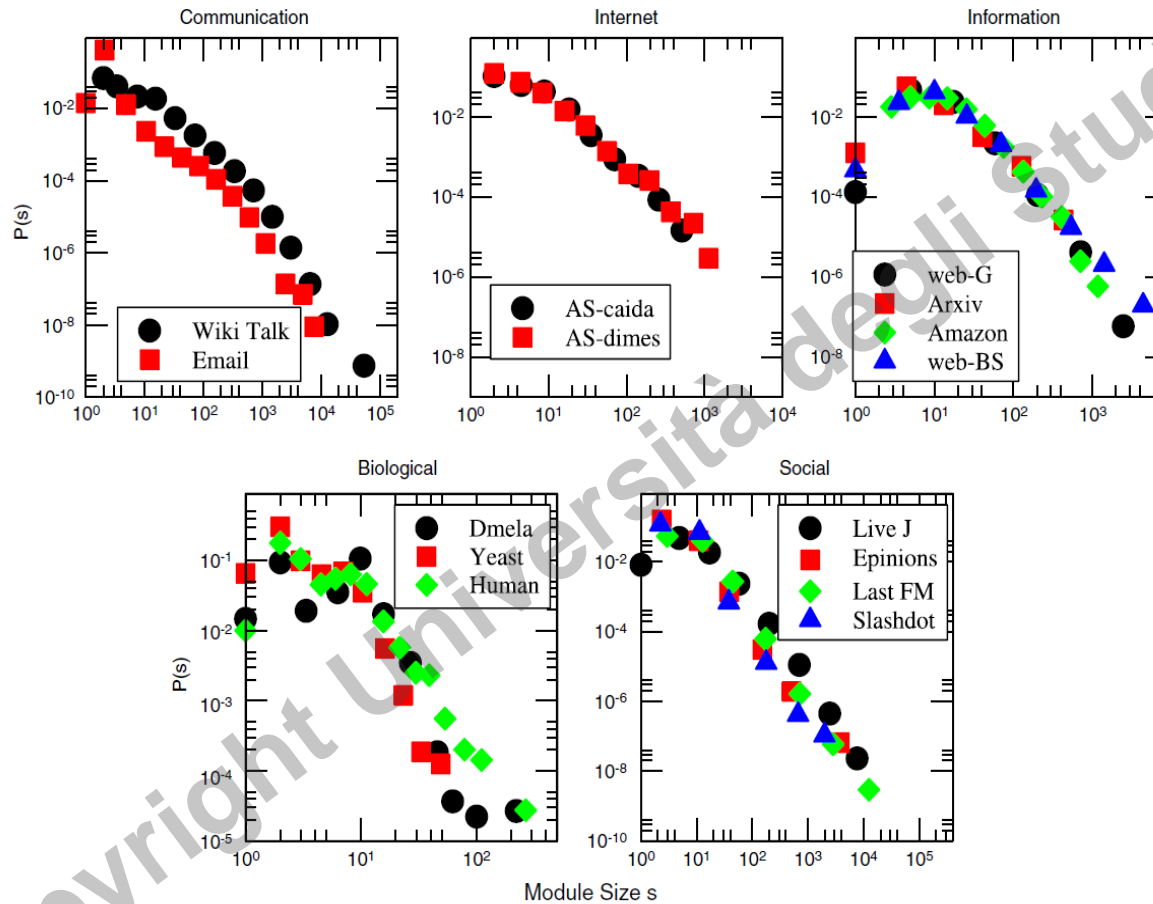


Communities extracted from the call pattern of the consumers of the largest Belgian mobile phone company. The network has about two million mobile phone users. The nodes correspond to communities, the size of each node being proportional to the number of individuals in the corresponding community. The color of each community on a red–green scale represents the language spoken in the particular community, red for French and green for Dutch. Only communities of more than 100 individuals are shown. The community that connects the two main clusters consists of several smaller communities with less obvious language separation, capturing the culturally mixed Brussels, the country’s capital.

Communities: size

S. Fortunato, D. Hric / Physics Reports 659 (2016) 1–44

21



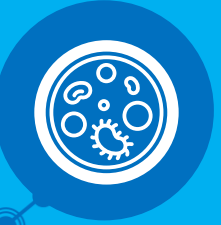
Numerous small communities coexist with a few very large ones.

Fig. 18. Distribution of community sizes in real networks. The clusters were detected with Infomap [42], but other methods yield qualitatively similar results. Various classes of networks are considered. All distributions are broad, spanning several orders of magnitude. Source: Reprinted figure with permission from [58].

Crawling – Part 2

Cheick Tidiane Ba

Copyright Università degli Studi di Milano



What is a crawler

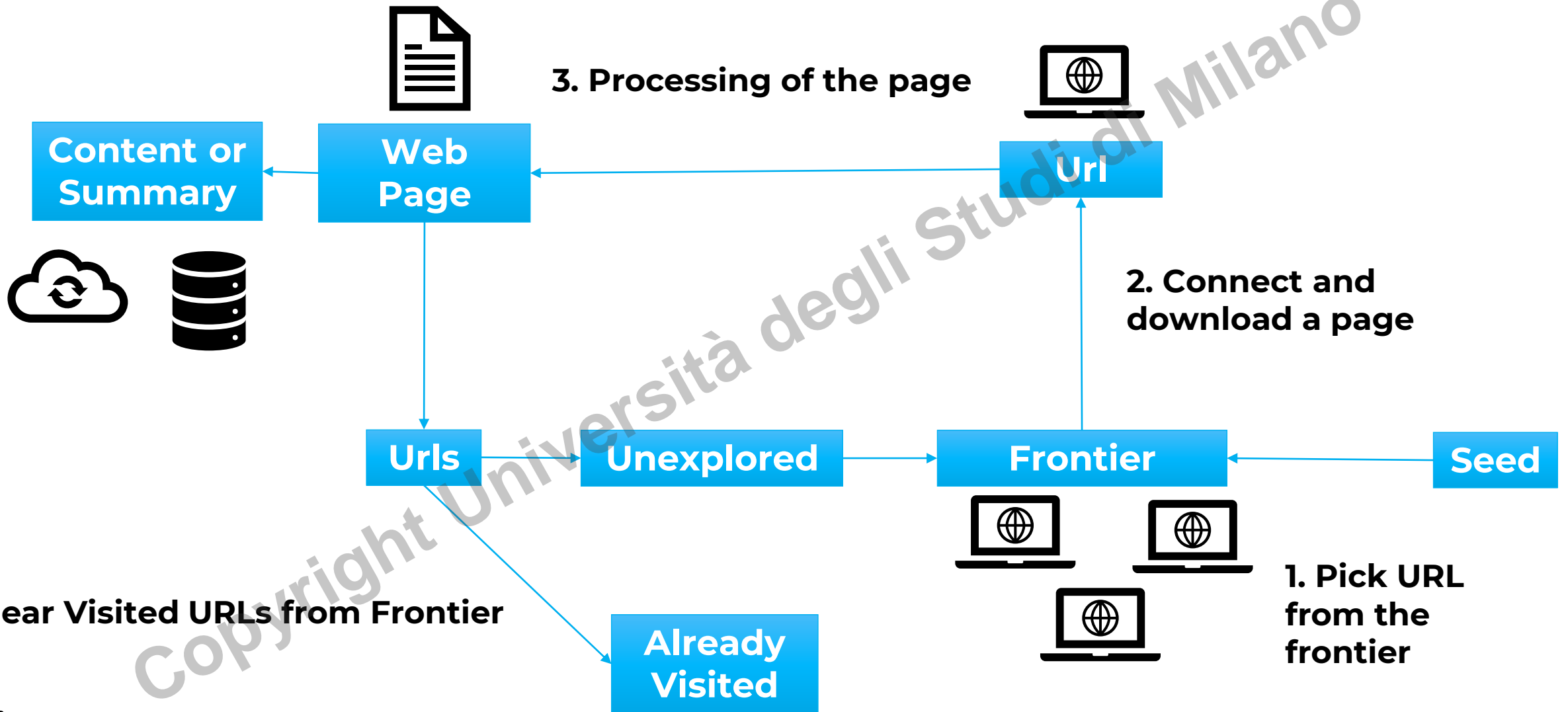
- A Web crawler, sometimes called a spider or spiderbot and often shortened to crawler, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing (web spidering).
- We make the distinction between
 - Crawling: the activity of download of web pages, while visiting the web
 - Web scraping: extracting data from websites.

Why we care

- Data is key for machine learning and business decisions
- It is important to understand the issues behind data retrieval,
- As data scientists we may need to address those issues and configure scraping tools to obtain data
 - Understanding concepts helps us deal with those tools

Basic organization of a large-scale, distributed web crawler

Recap



Recap

- **Policy:** the way we choose an url in frontier
- **Seed set must be chosen carefully** Where we start is just as important as how we choose the next page
- **Resource issues:** Frontier grows rapidly (exponential) Ram or Storage (Disk or Cloud)
- **Graceful Degradation:** property that describes the ability to deal with this issues

Recap

- **Politeness: Issues downloads from a single site or server.**
 - robots.txt
 - limit the time of a request between each request.
 - limit the time of a request between each request.

Copyright Università degli Studi di Milano

Recap

- **Content processing**
 - 1) extract new URLs
 - 2) Save content
- **Avoid duplicate pages to save storage space**
- **Bloom filters Probabilistic data structure**
 - **Characterized by**
 - Rapid answer
 - Memory efficient
- **Track seen elements**
 - Add elements to the list of seen elements
 - And ask if we have seen a certain element already

Copyright Università degli Studi di Milano

Bloom Filters - Analysis

Bloom Filter example

- Given a Bloom filter of 15 bits
- A set of URLs X
- Two hash function h_1, h_2
- I want to add the following web page summary: «the fox is on the table»

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Bloom Filter example

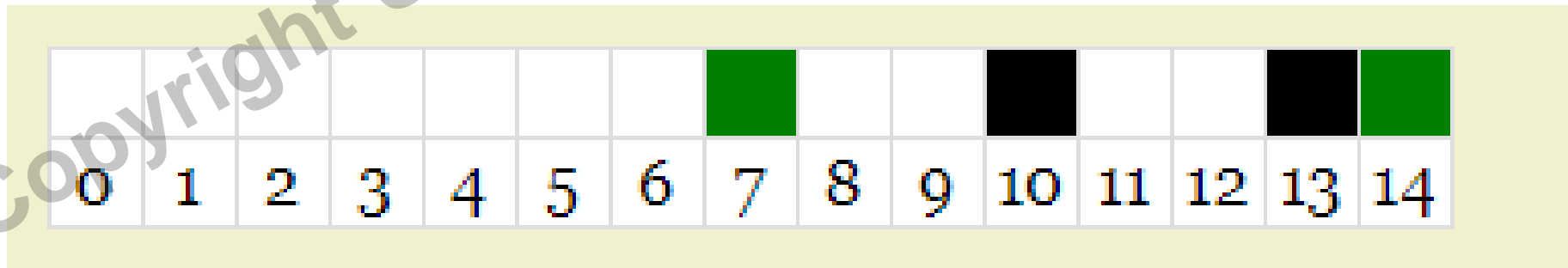
- I want to add to our set the summary x e.g. «the fox is on the table»
- Apply each hash function
 - E.g. $h1(x) = 10$; $h2(x) = 13$
- Set the bits in position 10 and 13 to 1
- Adding «the cat is on the table»
 - $h1(x) = 14$, $h2(x) = 7$

A diagram of a Bloom filter array with 15 bits, indexed from 0 to 14. The bits are represented by colored squares: white for 0, green for 1, and black for 0. The bits at positions 7, 10, 13, and 14 are set to 1 (green), while all other bits are 0 (white).

							1			1			1	1
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Bloom Filter example

- I want to check if the summary x e.g. «the fox is on the table»
- Apply each hash function
 - E.g. $h_1(x) = 10$; $h_2(x) = 13$
- Check the bit values bit values in positions 10 and 13
 - The AND combination of the bit values can be 1 (true) or 0 false



Analysis

- **Given:**
 - m = number of bits in bit vector
 - d = number of hash function
 - n = insert operations
- Find the probability of a (false or true) positive after n insert operations
- Probability that a certain bit b is set to 1 after one insert operations is: $\frac{1}{m}$
- Probability that a certain bit b is set to 0 after one insert operations is the complementary event:
 - $1 - \frac{1}{m}$

Analysis

- Probability that a certain bit b is set to 0 after one insert operations is:
 - $1 - \frac{1}{m}$
- Probability that a certain bit b is set to 0 after n insert operations is
 - $(1 - \frac{1}{m})^n$
- Probability that a certain bit b is set to 0 after n insert operations is
 - $(1 - \frac{1}{m})^{dn}$

Analysis

- **Positive: we check d bits and find all 1s**

- $(1 - (1 - \frac{1}{m})^{dn})^d$

- **Property:**

- $(1 + \frac{\alpha}{n})^n \rightarrow e^\alpha$ for $n \rightarrow \infty$

- **In our case we have** $(1 - (1 - \frac{1}{m})^{dn})^d$

- $(1 - (1 - \frac{1}{m})^{dn})^d \approx (1 - e^{-\frac{dn}{m}})^d$

Analysis

- If we consider $p = e^{-nd/m}$ we can express $d = -\binom{m}{n} \ln p$
- We need to minimize:
 - $1 - p^{-\binom{m}{n} \ln p} = e^{-\binom{m}{n} \ln p \ln(1-p)}$
- To minimize we need the first derivative:
 - $-\binom{m}{n} e^{-\binom{m}{n} \ln p \ln(1-p)} \left(\frac{\ln(1-p)}{p} - \frac{\ln p}{1-p} \right)$
- The first derivative is zero when:
 - $(1-p) \ln(1-p) = p \ln p$
- A solution is for sure when $1-p = p$
 - So, $p = \frac{1}{2}$
 - It can be proven it's the only one (See References)

Analysis

- Given $p = \frac{1}{2}$, and the previous equation $d = -\binom{m}{n} \ln p$
- The probability of (false) positives is minimized for $d \approx m \ln 2/n$

Copyright Università degli Studi di Milano

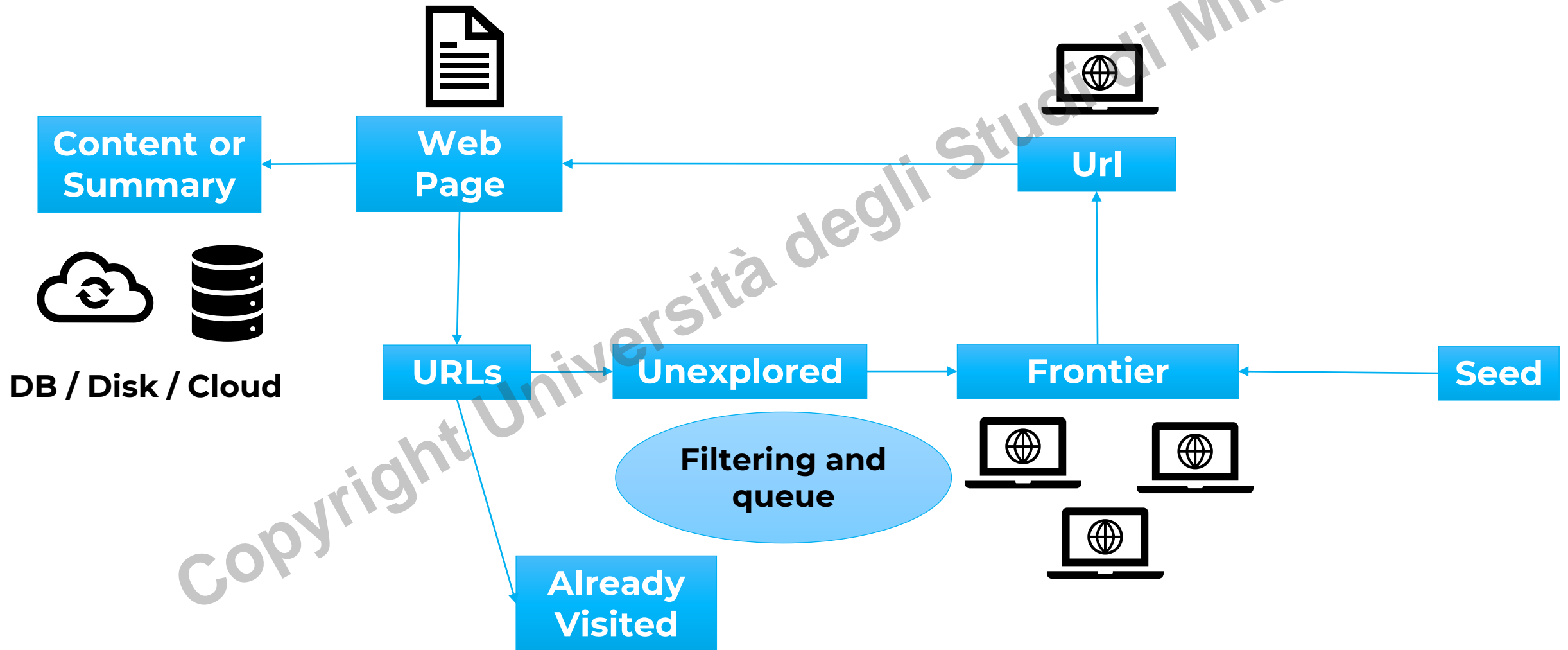
Analysis

- The conclusion of the analysis is that given:
 - m = number of bits in bit vector
 - d = number of hash function
 - n = insert operations
- The probability of (false) positives is minimized for $d \approx m \ln 2/n$
- In this case the probability of a (false) positive is 2^{-d}
- This means that we can exponentially improve the probability of error by increasing the number of hash functions d or working on the number of bit m
- The proportion is given by $m \approx dn/\ln 2 \approx 1,44 dn$

Copyright Università degli Studi di Milano

Frontier

The system



What we need

- **We want to :**
 - **Filter URLs**
 - **Clear Visited URLs from Frontier**
 - **Add new URLs to the Frontier**
 - **Implement a Policy**
- **When it comes to duplicates, we could use the Bloom Filter**
- **While good in theory, we have other data structures that are more suited for the frontier**
- **Not just duplicates but also act as a queue for the Frontier**

Sieve

- The Sieve is a data structure that accepts URLs that may need a visit as input. Then it emits, sometimes as blocks, URLs ready for the visit.
- Each URL that is inserted in the Sieve can be emitted only once, no matter how many times is inserted.
- The Sieve has the properties of a dictionary, a priority queue.
- It represents at the same time the frontier, the visited set and the queue of urls to be visited. Combined in one structure, performance is better.

Sieve in Mercator [HN99]

- Data structure that returns URLs
- Characteristics:
 - Return urls in FIFO (first in first out) order. It's the equivalent of a policy for breadth first visit
 - Costant Memory Space: remind that the frontier grows exponentially, our data structure can't risk
 - Probabilistic: we are not going to work with URLs; instead we'll work with Signatures, output of hash functions applied to the URLs

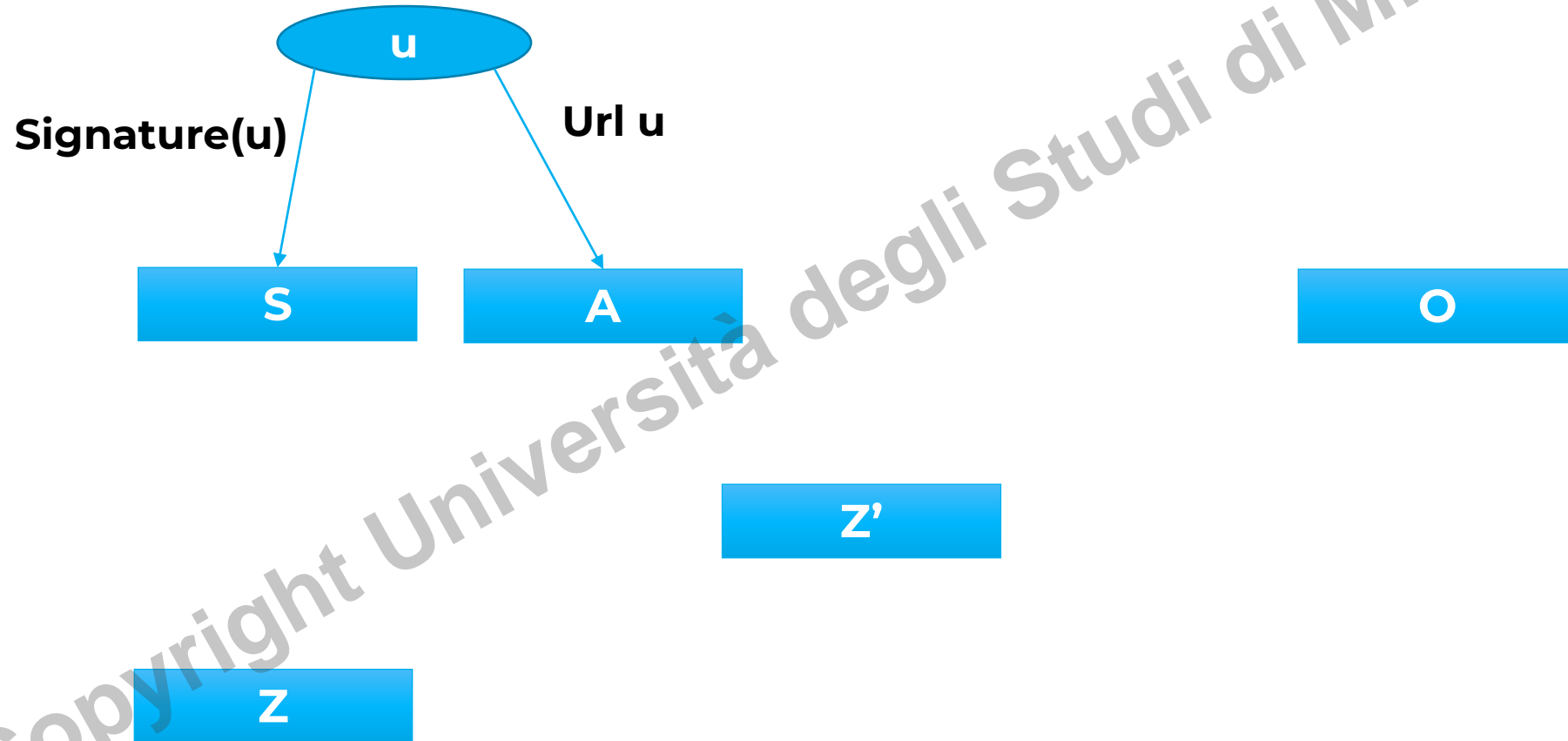
Working with Signatures

- Crawler work with Signatures
- Value obtained by applying an hash function to a URL, obtaining a value of with k bits.
- We expect collisions. They are dependent on the number of bits k .
- If we indicate U as the set of unknown nodes with n the number of bits, a collision estimate is $n^2/2U$.
- The first collision will happen after $O(\sqrt{n})$
- With an hash function that generates signatures of 64 bits, we'll have 100 collision every billion: acceptable.

Sieve

- The Sieve is formed by
- **S** vector in central memory that contains the signatures. Ex: 64 bits. Fixed dimension n , starts empty and gets filled
- **Z** file on disk that will contain all the signatures we meet over
- A auxiliary file, on disk, empty at beginning.
- **O** output file, in which we store the new urls

Adding URLs

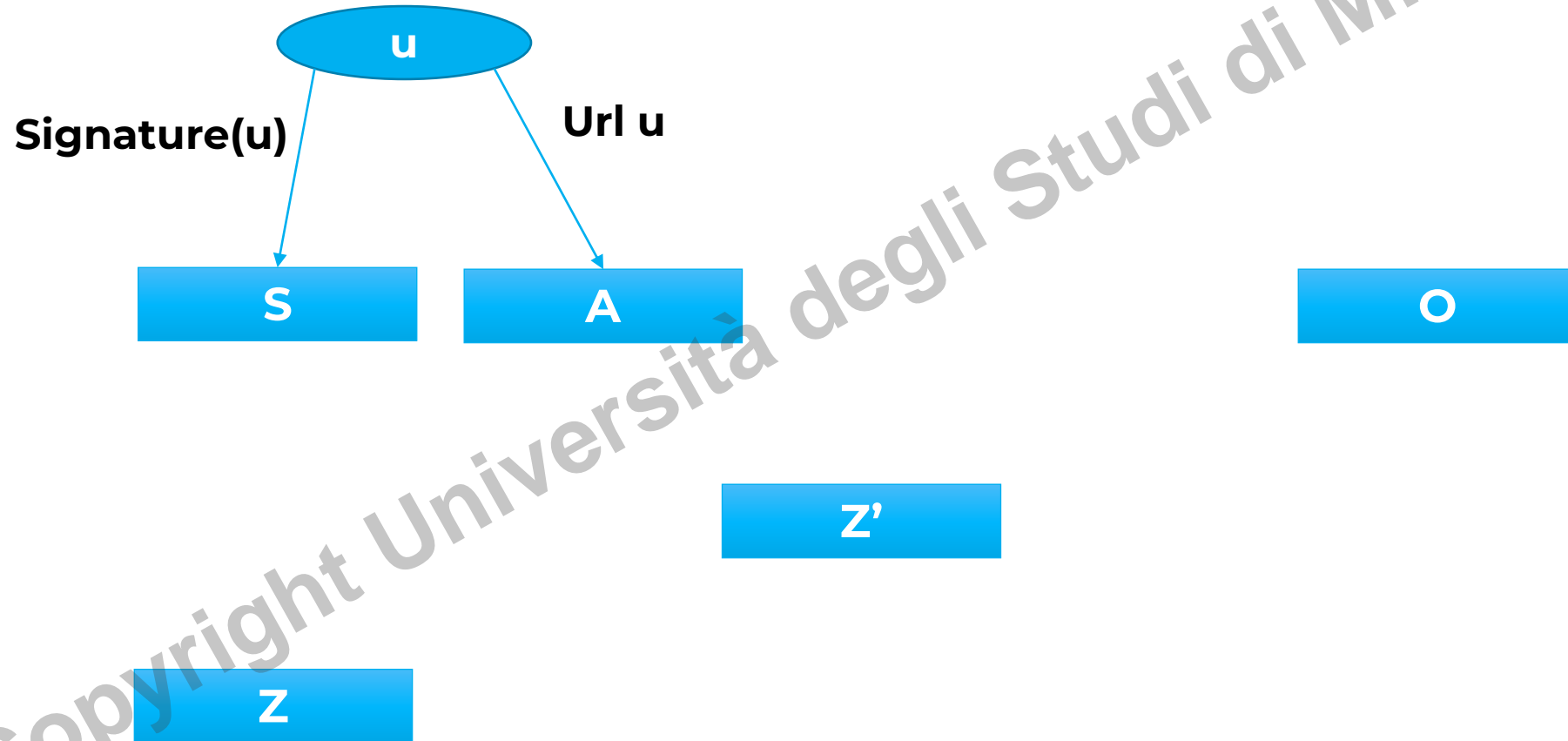


Adding URLs

- Given an URL u , we compute its signature/hash $h(u)$.
 - The URL u is added to A
 - The signature is stored in S
- S will fill up as we add more and more URLs.
- We need to do a Flush

S	0	1	1	2	1
A	A	B	B	C	B

Adding URLs



Flush - Sorting

- **1a) Sorting of S.**

- **Indirect sorting:** sorting with an external auxiliary array, so that we can maintain both the original and the sorted sequence
- **Stable sorting:** Sorting that respects the insert order of duplicates.

- **1b) Remove duplicates signatures in S**

- We mark them as useless, we consider only the first one. We can because the sorting is stable

Indirect sorting

- **Sorting with an external auxiliary array, so that we can maintain both the original and the sorted sequence**
- **Example:**
 - **S: b a c e d g -> abcdeg**
 - **V: 0 1 2 3 4 5 -> 1 0 2 4 3 5**

Copyright Università degli Studi di Milano

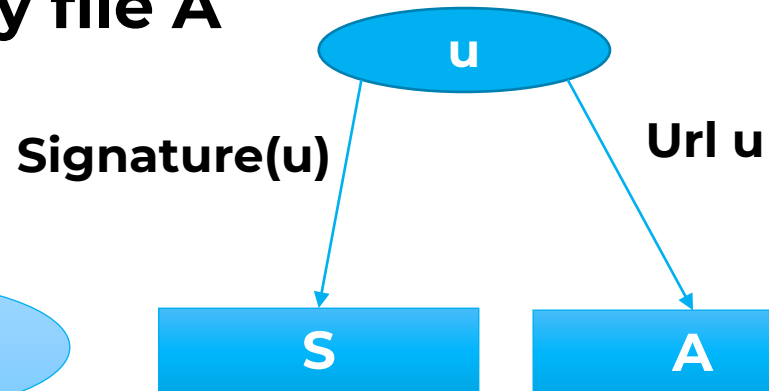
Stable sorting

- **Sorting that respects the insert order of duplicates.**
- **Example of not stable:**
 - **b a c e d e -> a b c d e e**
 - **0 1 2 3 4 5 -> 1 0 2 4 5 3**
 - **Not stable because the first appearance of e comes after.**

Copyright Università degli Studi di Milano

Flush - Sorting

In S we have signatures; the corresponding URLs are in the auxiliary file A



1. Sorting

S

A

O

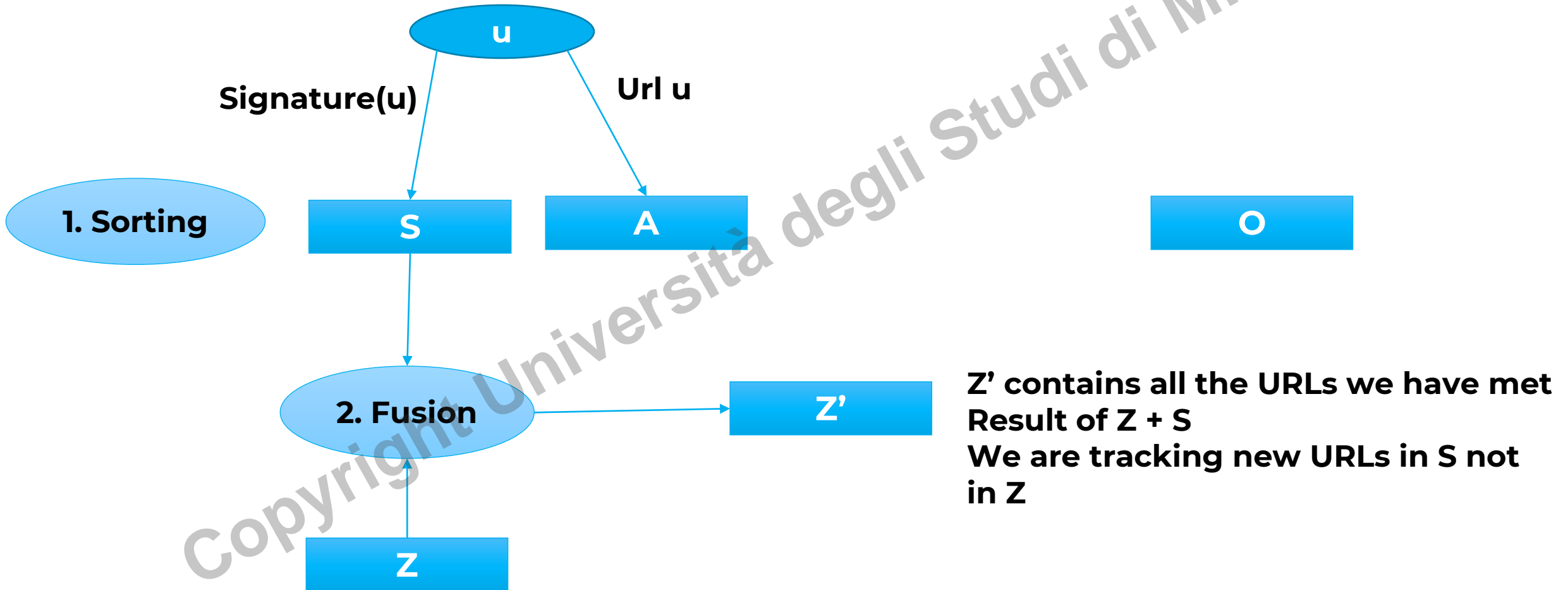
Z'

Z

Flush - Fusion

- We need to generate:
 - O : List of URLs to be visited
 - We need to know the signatures that are in S but they are not yet in Z ($==$ Visited set).
- 2) Fusion of Z with the signatures in S
 - Save the temporary result in a new file Z' ,
 - During the fusion we keep track of the signatures in S that that are not in Z .
 - Efficient (linear time) as Z and S are sorted.

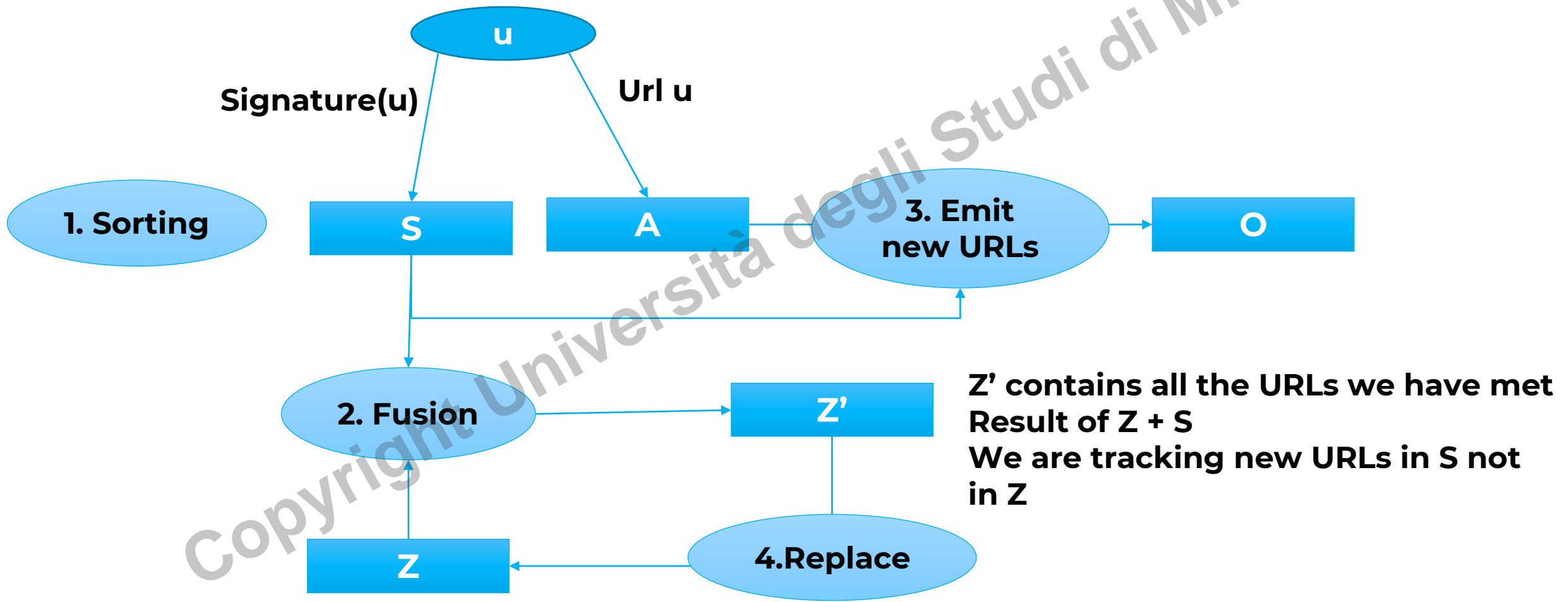
Flush - Fusion



Flush – new URLs

- To produce the urls to be visited, for output O
- 3) Scan A and S in parallel, and for every signature marked as new in S: emit the URL in output file O.
 - The scan of A is sequential, since S is sorted like A that's why we used and indirect sorting.
- 4) Replace Z with Z'; clear S and A.

Flush – new URLs



Adding URLs

S	0	1	1	2	1
A	A	B	B	C	B

S					
A					

Z	
----------	--

Z'	
-----------	--

O	
----------	--

Copyright Università degli Studi di Milano

Flush - Sorting

S	0	1	1	2	1
A	A	B	B	C	B

S	0	1	1	1	2
A	A	B	B	B	C

Z _____

Z' _____

O _____

Copyright Università degli Studi di Milano

Flush - Fusion

S	0	1	1	2	1
A	A	B	B	C	B

S	0	1	1	1	2
A	A	B	B	B	C

Z

Z' 0,1,2

O

Flush - Emit New URLs

S	0	1	1	2	1
A	A	B	B	C	B

S	0	1	1	1	2
A	A	B	B	B	C

Z

Z' 0,1,2

O A,B,C

Flush – Replace Z

S	0	1	1	2	1
A	A	B	B	C	B

S	0	1	1	1	2
A	A	B	B	B	C



O	A,B,C
---	-------

Flush - Clear

S					
A					

S					
A					

Z	0,1,2
---	-------

Z'	
----	--

O	A,B,C
---	-------

Copyright Università degli Studi di Milano

Adding more URLs

S	1	1	0	3	5
A	B	B	A	D	F

S				
A				

Z 0,1,2

Z'

O

Flush - Sorting

S	1	1	0	3	5
A	B	B	A	D	F

S	0	1	1	3	5
A	A	B	B	D	F

Z	0,1,2
----------	-------

Z'	
-----------	--

O	
----------	--

Flush – Emit New URLs

S	1	1	0	3	5
A	B	B	A	D	F

S	0	1	1	3	5
A	A	B	B	D	F

Z 0,1,2

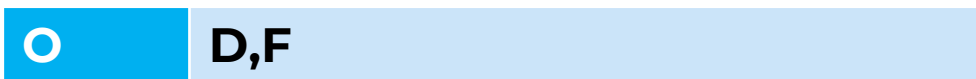
Z' 0,1,2,3,5

O D,F

Flush – Replace Z

S	1	1	0	3	5
A	B	B	A	D	F

S	0	1	1	3	5
A	A	B	B	D	F



Flush – Clear

S					
A					

S				
A				

Z 0,1,2,3,5

Z'

O D,F

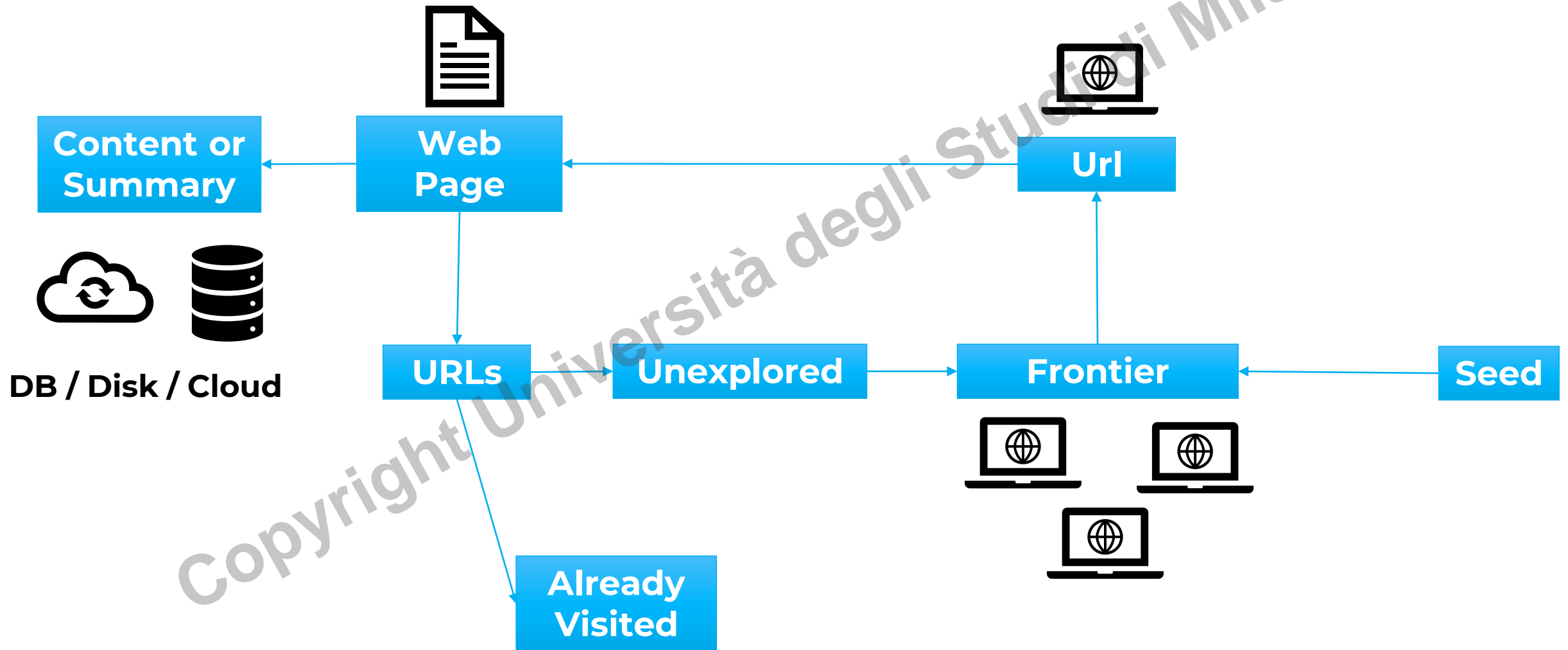
Observations

- **Z** after the Flush contains again all the signatures of URLs we never visited.
- In **O** we have all the URLs which signature was not in **Z** and (barring collisions) all the unvisited urls.
- The Urls in **O** are emitted in FIFO order
- The marking procedure and the fusion can be done at the same time

Conclusion

Copyright Università degli Studi di Milano

The system



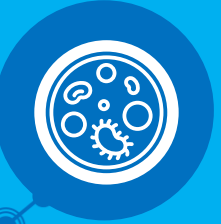
References

- <http://vigna.di.unimi.it/algoweb/>
- <https://lmlib.github.io/bloomfilter-tutorial/> (bloom filter demo)
- Burton H. Bloom. Space-time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970
- <https://www.cs.princeton.edu/courses/archive/spring02/cs493/lec6.pdf> (Bloom filter analysis, section 3.1)
- <https://courses.cs.washington.edu/courses/cse454/15wi/papers/mercator.pdf> [HN99]
- <http://vigna.di.unimi.it/ftp/papers/BUbiNG.pdf> (Sieve implemented)

Thank you for the attention!

For any question send an email at
cheick.ba@unimi.it

Copyright Università degli Studi di Milano



Social Media Mining

Information Diffusion in Social Media

Reza Zafarani, Mohammad Ali Abbasi, Huan Liu

Social Media Mining: An Introduction

Cambridge University Press, 2014

Definition

- In February 2013, during the third quarter of Super Bowl XLVII, a power outage stopped the game for 34 minutes.
- Oreo, a sandwich cookie company, tweeted during the outage: “Power out? No Problem, You can still dunk it in the dark”.
- The tweet caught on almost immediately, reaching nearly 15,000 retweets and 20,000 likes on Facebook in less than 2 days.
- A simple tweet diffused into a large population of individuals.
- It helped the company gain fame with minimum budget in an environment where companies spent as much as 4 million dollars to run a 30 second ad during the super bowl.
- This is an example of Information Diffusion.

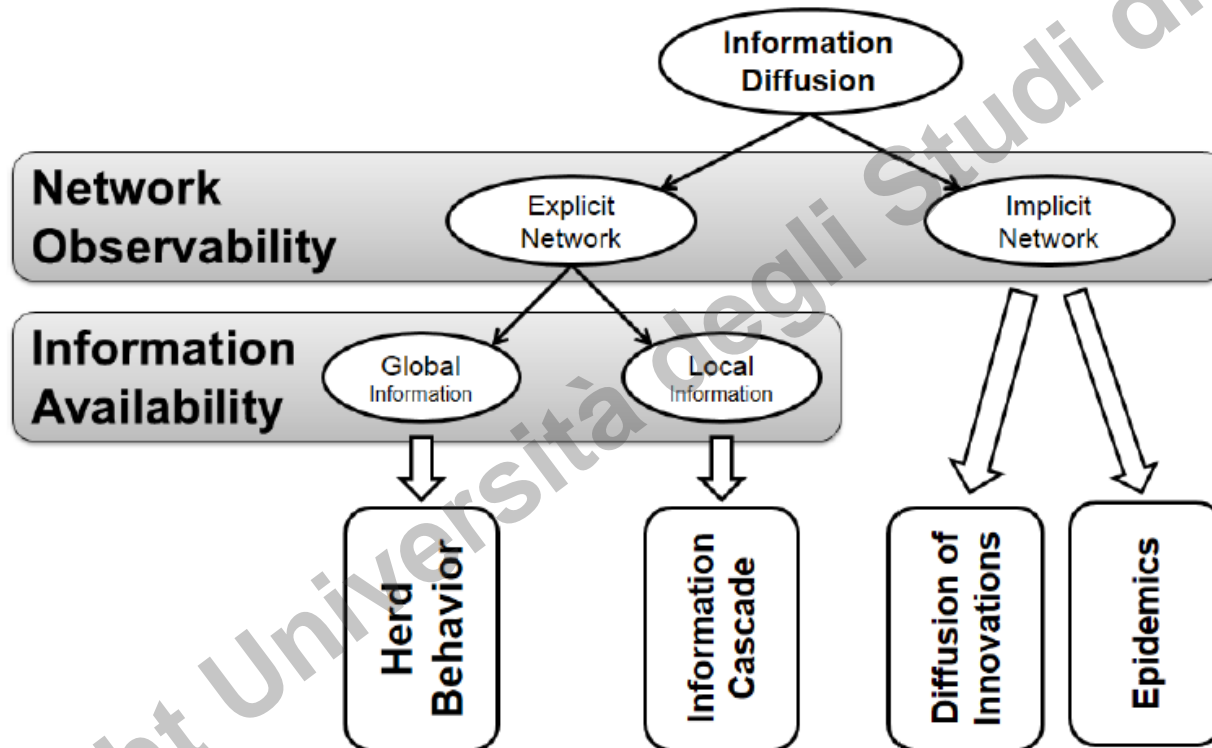
- Information diffusion is studied in a plethora of sciences.
- Our focus is on techniques that can model information diffusion.
- **Information diffusion:** process by which a piece of information (knowledge) is spread and reaches individuals through interactions.

Information Diffusion

A diffusion process involves three elements:

- **Sender(s).** A sender or a small set of senders that initiate the information diffusion process;
- **Receiver(s).** A receiver or a set of receivers that receive diffused information. Commonly, the set of receivers is much larger than the set of senders and can overlap with the set of senders;
- **Medium.** This is the medium through which the diffusion takes place. For example, when a rumor is spreading, the medium can be the personal communication between individuals

Information Diffusion Types



We define the process of interfering with information diffusion by expediting, delaying, or even stopping diffusion as **Intervention**

Information diffusion types

Herd behavior takes place when individuals observe the actions of all others and act in an aligned form with them.

Information cascade describes the process of diffusion when individuals merely observe their immediate neighbors.

In information cascades and herd behavior, the *network* of individuals is *observable*;

In *herding*, individuals decide based on global information (*global dependence*);

In *information cascades* decisions are made based on knowledge of immediate neighbors (*local dependence*)

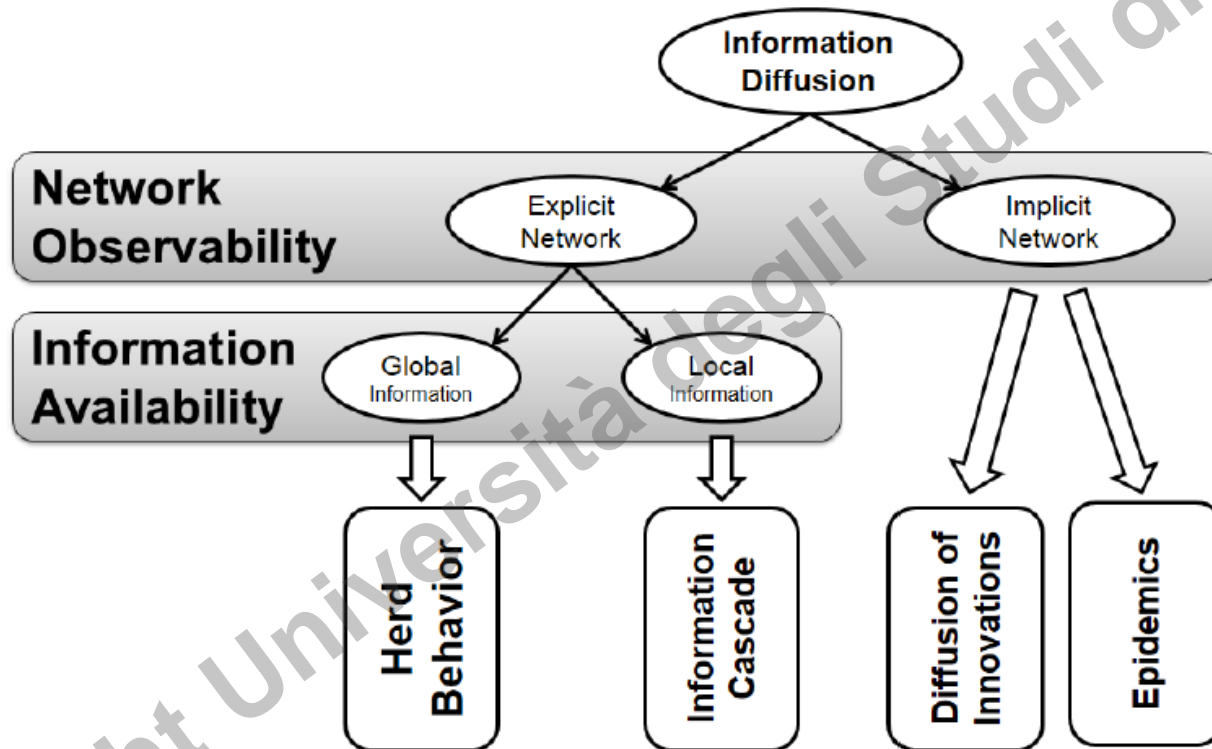
Information diffusion types

Diffusion of innovations provides a bird's-eye view of how an innovation (e.g., a product, music video, or fad) spreads through a population. It assumes that *interactions* among individuals are *unobservable* and that the *sole available information* is *the rate* at which products are being adopted throughout a certain period of time. This information is particularly interesting for companies performing market research, where the sole available information is the rate at which their products are being bought. These companies have no access to interactions among individuals.

Epidemic models are similar to diffusion of innovations models, with the difference that the innovation's analog is a pathogen and adoption is replaced by infection.

Another difference is that in epidemic models, *individuals do not decide* whether to become infected or not and infection is considered a random natural process, as long as the individual is exposed to the pathogen.

Information Diffusion Types



We define the process of interfering with information diffusion by expediting, delaying, or even stopping diffusion as *Intervention*

Herd Behavior

- **Network is observable**
- **Only public information is available**

Herd Behavior

Herd behavior describes when a group of individuals performs actions that are highly correlated without any plans

Main Components of Herd Behavior

- A method to transfer behavior among individuals or to observe their behavior
- A connection between individuals

Examples of Herd Behavior

- Flocks, herds of animals, and humans during sporting events, demonstrations, and religious gatherings

Herd Behavior Example

- Consider people participating in an online auction.
- In this case, individuals can observe the behavior of others by monitoring the bids that are being placed on different items.
- Individuals are connected via the auction's site where they can not only observe the bidding behaviors of others, but can also often view profiles of others to get a feel for their reputation and expertise.
- In these online auctions, it is common to observe individuals participating actively in auctions, where the item being sold might otherwise be considered unpopular.
- This is due to individuals trusting others and assuming that the high number of bids that the item has received is a strong signal of its value. In this case, Herd Behavior has taken place.

Herd Behavior: Popular Restaurant Experiment

- Assume you are on a trip in a metropolitan area that you are less familiar with.
- Planning for dinner, you find restaurant **A** with excellent reviews online and decide to go there.
- When arriving at **A**, you see **A** is almost empty and restaurant **B**, which is next door and serves the same cuisine, almost full.
- Deciding to go to **B**, based on the belief that other diners have also had the chance of going to **A**, is an example of herd behavior

- In this example, when **B** is getting more and more crowded, herding is taking place.
- Herding happens because we consider crowd intelligence trustworthy.
- We assume that there must be private information not known to us, but known to the crowd, that resulted in the crowd preferring restaurant B over A.
- In other words, we assume that, given this private information, we would have also chosen B over A.

Herd behavior

In general, when designing a herding experiment, the following four conditions need to be satisfied:

1. There needs to be a decision made.

In this example, the decision involves going to a restaurant.

2. Decisions need to be in sequential order.

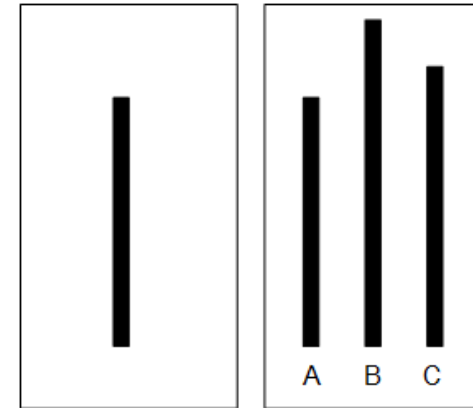
3. Decisions are not mindless, and people have private information that helps them decide.

4. No message passing is possible.

Individuals do not know the private information of others, but can infer what others know from what they observe from their behavior.

Conformity pressure: Solomon Asch's Experiment

- In one experiment, he asked groups of students to participate in a vision test where they were shown two cards, one with a single line segment and one with 3 lines, and the participants were required to match line segments with the same length.
- Each participant was put into a group where all other group members were collaborators with Asch. These collaborators were introduced as participants to the subject.
 - Asch found that in control groups with no pressure to conform, only 3% of the subjects provided an incorrect answer.
 - However, when participants were surrounded by individuals providing an incorrect answer, up to 32% of the responses were incorrect.



Herd Behavior: Milgram's Experiment

- Stanley Milgram asked one person to stand still on a busy street corner in New York City and stare straight up at the sky.
 - About 4% of all passersby stopped to look up.
- When 5 people stand on the sidewalk and look straight up at the sky, 20% of all passersby stopped to look up.
- Finally, when a group of 18 people look up simultaneously, almost 50% of all passersby stopped to look up.

Herding: Elevator Example



<http://www.youtube.com/watch?v=zNNzoyzHcw>

Network Observability in Herd Behavior

In herd behavior, individuals make decisions by observing all other individuals' decisions

- In general, herd behavior's network is close to a complete graph where nodes can observe at least most other nodes and they can observe public information
 - For example, they can see the crowd

Designing a Herd Behavior Experiment (NO)

- There needs to be a decision made.
 - In our example, it is going to a restaurant
- Decisions need to be in sequential order;
- Decisions are not mindless and people have private information that helps them decide; and
- No message passing is possible. Individuals don't know the private information of others, but can infer what others know from what is observed from their behavior.

Herding: Urn Experiment (NO)

- There is an urn in a large class with three marbles in it



- During the experiment, each student comes to the urn, picks one marble, and checks its color in private.
- The student predicts majority blue or red, writes her prediction on the blackboard, and puts the marble back in the urn.
- Students can't see the color of the marble taken out and can only see the predictions made by different students regarding the majority color on the board

Urn Experiment: First and Second Student (NO)

- First Student:

- *Board: -*

- Observed: B \rightarrow Guess: B

- or-

- Observed: R \rightarrow Guess: R

- Second Student:

- *Board: B*

- Observed: B \rightarrow Guess: B

- or-

- Observed: R \rightarrow Guess: R/B (flip a coin)

Urn Experiment: Third Student (NO)

- *If board: B, R*
 - Observed: B \rightarrow Guess: B, or
 - Observed: **R** \rightarrow Guess: **R**
- *If board: B, B*
 - Observed: B \rightarrow Guess: B, or
 - **Observed: R \rightarrow Guess: B (Herding Behavior)**

The forth student and onward

- Board: B,B,B
- **Observed: B/R \rightarrow Guess: B**

Bayes's Rule in the Herding Experiment (NO)

Each student tries to estimate the conditional probability that the urn is majority-blue or majority-red, given what she has seen or heard

- She would guess majority-blue if:

$$\Pr[\text{majority-blue} \mid \text{what she has seen or heard}] > 1/2$$

- From the setup of the experiment we know:

$$\Pr[\text{majority-blue}] = \Pr[\text{majority-red}] = 1/2$$

$$\Pr[\text{blue} \mid \text{majority-blue}] = \Pr[\text{red} \mid \text{majority-red}] = 2/3$$

Bayes's Rule in the Herding Experiment (NO)

$$\Pr[\text{majority-blue}|\text{blue}] = \Pr[\text{blue}|\text{majority-blue}] * \Pr[\text{majority-blue}] / \Pr[\text{blue}]$$

$$\begin{aligned}\Pr[\text{blue}] &= \Pr[\text{blue}|\text{majority-blue}] * \Pr[\text{majority-blue}] \\ &+ \Pr[\text{blue}|\text{majority-red}] * \Pr[\text{majority-red}] \\ &= 2/3 * 1/2 + 1/3 * 1/2 = 1/2\end{aligned}$$

$$\Pr[\text{majority-blue}|\text{blue}] = (2/3 * 1/2) / (1/2)$$

- So the first student should guess “**blue**” when she sees “**blue**”
- The same calculation holds for the second student

Bayes's Rule in the Herding Experiment: Third Student (NO)

$$\Pr[\text{majority-blue} | \text{blue, blue, red}] = \frac{\Pr[\text{blue, blue, red} | \text{majority-blue}] * \Pr[\text{majority-blue}]}{\Pr[\text{blue, blue, red}]}$$

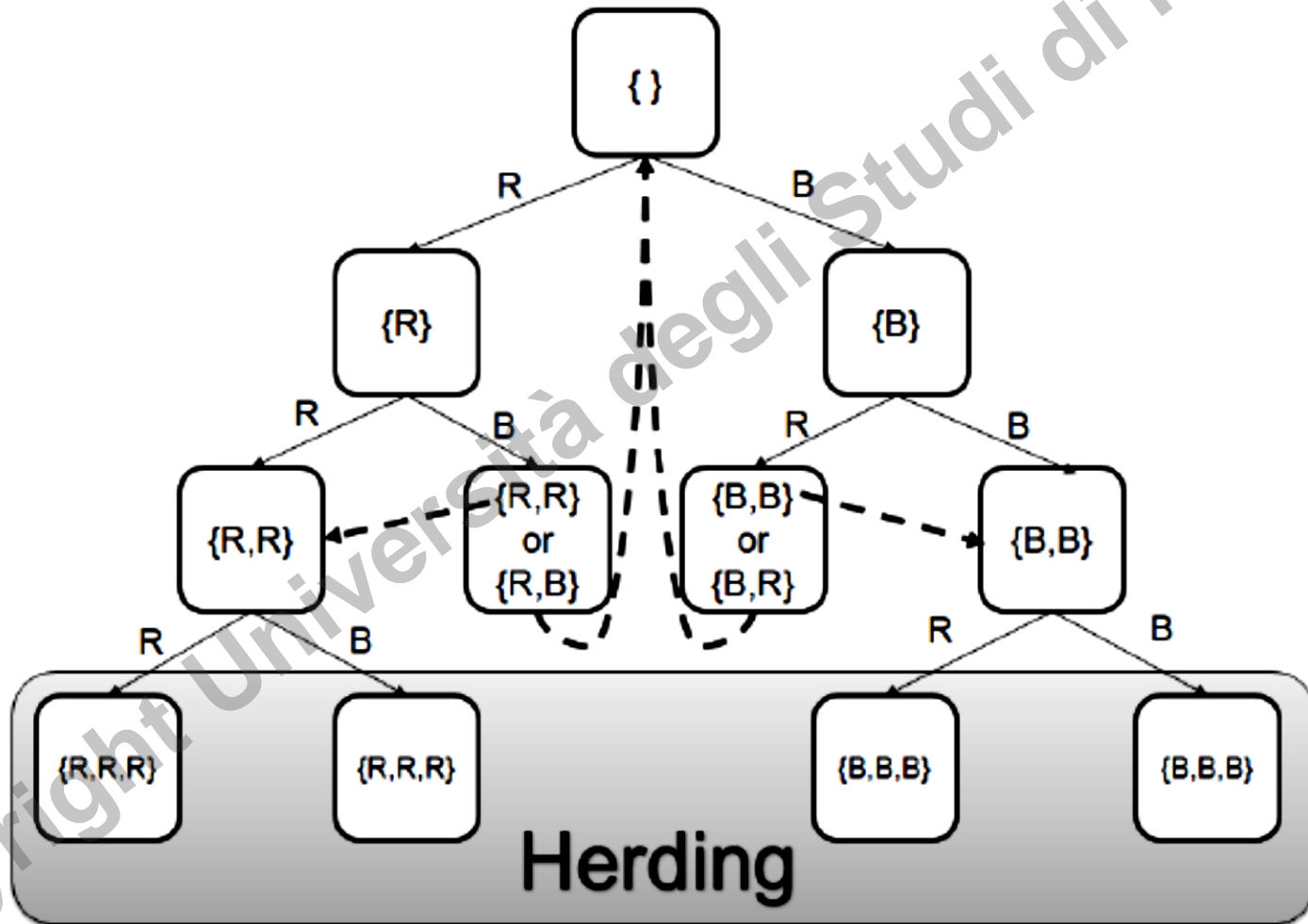
$$\Pr[\text{blue, blue, red} | \text{majority-blue}] = 2/3 * 2/3 * 1/3 = 4/27$$

$$\begin{aligned} \Pr[\text{blue, blue, red}] &= \Pr[\text{blue, blue, red} | \text{majority-blue}] * \Pr[\text{majority-blue}] \\ &+ \Pr[\text{blue, blue, red} | \text{majority-red}] * \Pr[\text{majority-red}] \\ &= (2/3 * 2/3 * 1/3) * 1/2 + (1/3 * 1/3 * 2/3) * 1/2 = 1/9 \end{aligned}$$

$$\Pr[\text{majority-blue} | \text{blue, blue, red}] = (4/27 * 1/2) / (1/9) = 2/3$$

- So the third student should guess “blue” even when she sees “red”
- All future students will have the same information as the third student

Urn Experiment (NO)



Herding Intervention (NO)

In herding, the society only has access to public information.

Herding may be intervened by **releasing private information** which was not accessible before

The little boy in “The Emperor’s New Clothes” story intervenes the herd by shouting “he’s got no clothes on”

Herding Intervention (NO)

Milgram Experiment: To intervene the herding effect, we need one person to tell the herd that there is nothing in the sky

How Does Intervention Work? (NO)

- When a new piece of private information releases, the herd reevaluate their guesses and this may create completely new results
- The Emperor's New Clothes
 - When the boy gives his private observation, other people compare it with their observation and confirm it
 - This piece of information may change others guess and ends the herding effect
- In general, intervention is possible by providing private information to individuals not previously available. Consider an urn experiment where individuals decide on majority red over time. Either
 - 1) a private message to individuals informing them that the urn is majority blue or
 - 2) writing the observations next to predictions on the board stops the herding and changes decisions.

Information Cascade

- **In the presence of a network**
- **Only local information is available**

Information Cascade

- In social media, individuals commonly repost content posted by others in the network. This content is often received via immediate neighbors (friends).
- An Information Cascade occurs as information propagates through friends
- An information cascade is defined as a piece of information or decision being cascaded among a set of individuals, where
 - 1) individuals are connected by a network and
 - 2) individuals are only observing decisions of their immediate neighbors (friends).
- Therefore, cascade users have less information available to them compared to herding users, where almost all information about decisions are available.

In cascading, local information is available to the users, but in herding the information about the population is available.

Underlying Assumptions for Cascade Models

- The network is represented using a directed graph. Nodes are actors and links depict the communication channels between them. A node can only influence nodes that it is connected to;
- Decisions are binary - nodes can be either active or inactive. An active nodes means that the node decided to adopt the behavior, innovation, or decision;
- A node, once activated, can activate its neighboring nodes;
- Activation is a progressive process, where nodes change from inactive to active, but not vice versa 1.

Independent Cascade Model (ICM)

- **Independent Cascade Model** is a sender centric model of cascade
 - In this model each node has one chance to activate its neighbors
- Considering nodes that are active as senders and nodes that are being activated as receivers,
 - The *linear threshold model* concentrates on the receiver (to be discussed later).
 - The independent cascade model concentrates on the sender

Independent Cascade Model (ICM)

- In *Independent Cascade Model*, the node that is activated at time t , has one chance, at time step $t + 1$, to activate its neighbors
- Let v be an active node at time t , for any neighbor w of it, there's a probability p_{vw} that node w gets activated at time $t + 1$.
- A node v activated at time t has a single chance of activating its neighbors and that activation can only happen at $t + 1$
- We start with a set of active nodes and we continue until no further activation is possible.

ICM Algorithm

Algorithm 7.1 Independent Cascade Model (ICM)

Require: Diffusion graph $G(V, E)$, set of initial activated nodes A_0 , activation probabilities $p_{v,w}$

```
1: return Final set of activated nodes  $A_\infty$ 
2:  $i = 0$ ;
3: while  $A_i \neq \{\}$  do
4:
5:    $i = i + 1$ ;
6:    $A_i = \{\}$ ;
7:   for all  $v \in A_{i-1}$  do
8:     for all  $w$  neighbor of  $v, w \notin \cup_{j=0}^i A_j$  do
9:       rand = generate a random number in  $[0,1]$ ;
10:      if rand  $< p_{v,w}$  then
11:        activate  $w$ ;
12:         $A_i = A_i \cup \{w\}$ ;
13:      end if
14:    end for
15:  end for
16: end while
17:  $A_\infty = \cup_{j=0}^i A_j$ ;
18: Return  $A_\infty$ ;
```

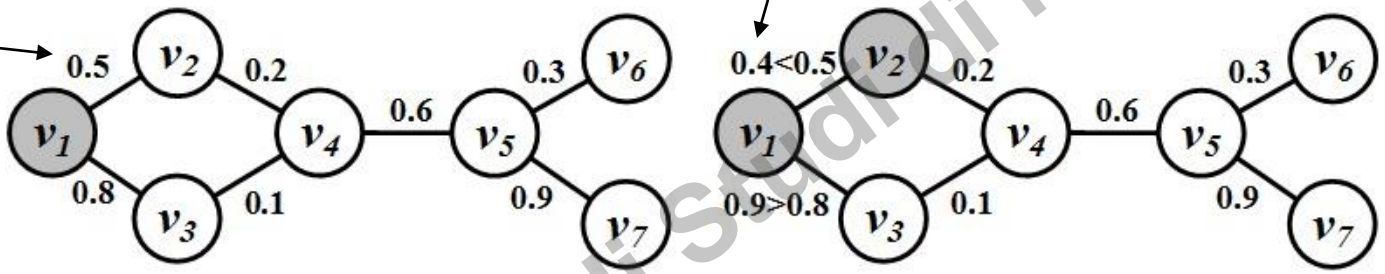
Node activation in ICM is a probabilistic process.

Thus, we might get different results for different runs.

Independent Cascade Model: An Example

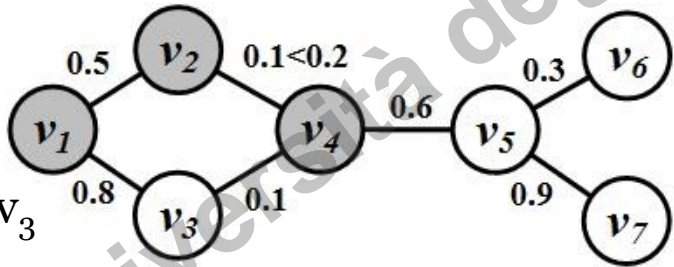
p_{vw} probability that node w gets activated

Random number generated at time $t+1$



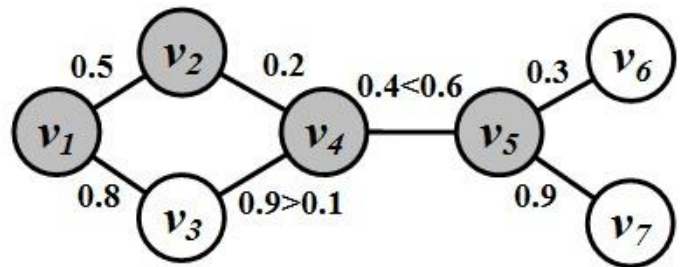
Step 1

Step 2

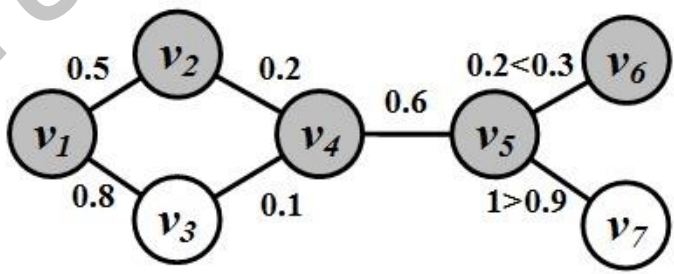


Step 3: v_1 can't activate v_3 as it was activated at step 1

Step 3



Step 4



Step 5

After five steps, five nodes get activated and the ICM procedure converges.

Maximizing the Spread of Cascades

Copyright Università degli Studi di Milano

Spread maximization

Consider a network of users and a company that is marketing a product.

The company is trying to advertise its product in the network.

The company has a limited budget; therefore, not all users can be targeted.

However, when users find the product interesting, they can talk with their friends (immediate neighbors) and market the product.

Their neighbors, in turn, will talk about it with their neighbors, and as this process progresses, the news about the product is spread to a population of nodes in the network.

The company plans on *selecting a set of initial users* such that the *size of the final population talking about the product is maximized*.

Maximizing the spread of cascades

- **Maximizing the Spread of Cascades** is the problem of finding a small set of nodes in a social network such that their aggregated spread in the network is maximized
- Applications
 - Product marketing
 - Influence

Problem Setting

- Given
 - A limited budget for initial advertising (e.g., give away free samples of product)
 - Estimating spread between individuals
- Goal
 - To trigger a large spread (e.g., further adoptions of a product)
- Question
 - Which set of individuals should be targeted at the very beginning?

Problem Statement

- Spread of node set S : $f(S)$
 - An expected number of active nodes, if set S is the initial active set
- Problem:
 - Given a parameter k (budget), find a k -node set S to maximize $f(S)$
 - A constrained optimization problem with $f(S)$ as the objective function

f(S): Properties

- Non-negative (obviously)
- Monotone: $f(S + v) \geq f(S)$
- Submodular:
 - Let N be a finite set
 - A set function is submodular iff

$$f: 2^N \mapsto \mathfrak{R}$$

$$\forall S \subset T \subset N, \forall v \in N \setminus T,$$

$$f(S + v) - f(S) \geq f(T + v) - f(T)$$

Some Facts Regarding this Problem

- **Bad News**
 - For a submodular function monotone non-negative f , finding a k -element set S for which $f(S)$ is maximized is an NP-hard optimization problem
 - It is NP-hard to determine the optimum for influence maximization for both independent cascade model and linear threshold model (to be introduced in next chapter).
- **Good News**
 - We can use Greedy Algorithm
 - Start with an empty set S
 - For k iterations:
 - Add node v to S that maximizes $f(S + v) - f(S)$.
 - How good (or bad) it is?
 - Theorem: The greedy algorithm is a $(1 - 1/e)$ approximation.
 - The resulting set S activates at least $(1 - 1/e) > 63\%$ of the number of nodes that any k set S could activate (optimum).

Cascade Maximization: A Greedy approach

Maximizing the cascade is a NP-hard problem but it is proved that the greedy approaches gives a solution that is at least 63 % of the optimal.

Given a network and a parameter k , *which k nodes should be selected to be in the activation set B in order to maximize the cascade in terms of the total number of active nodes?*

- Let $\sigma(B)$ denote the expected number of nodes that can be activated by B , the optimization problem can be formulated as follows:

$$\max_{B \subseteq V} \sigma(B) \text{ s.t. } |B| \leq k$$

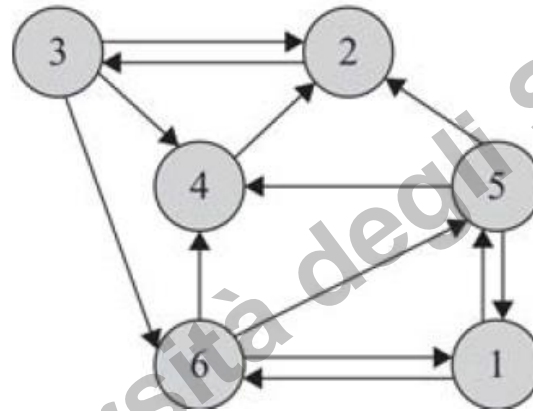
The Algorithm

- Start with $B = \emptyset$
 - Evaluate $\sigma(v)$ for each node, and pick the node with maximum σ as the first node v_1 to form $B = \{v_1\}$
 - Select a node which will increase $\sigma(B)$ most if the node is included in B .
- *Essentially, we greedily find a node $v \in V \setminus B$ such that*

$$v = \arg \max_{v \in V \setminus B} \sigma(B \cup \{v\})$$

Example

Example 7.3. For the following graph, assume that node i activates node j when $|i - j| \equiv 2 \pmod{3}$. Solve cascade maximization for $k = 2$.



To find the first node v , we compute $f(\{v\})$ for all v . We start with node 1. At time 0, node 1 can only activate node 6, because

$$|1 - 6| \equiv 2 \pmod{3}, \quad (7.11)$$

$$1 - 6 = 5$$

$$5 / 3 = 1 \text{ with remainder } 2$$

Example

$$6 - 4 = 2$$

$$2 / 3 = 0 \text{ with remainder } 2$$

$$6 - 5 = 1$$

$$1 / 3 = 0 \text{ with remainder } 1$$

At time 1, node 1 can no longer activate others, but node 6 is active and can activate others. Node 6 has outgoing edges to nodes 4 and 5. From 4 and 5, node 6 can only activate 4:

$$|6 - 4| \equiv 2 \pmod{3} \quad (7.13)$$

$$|6 - 5| \not\equiv 2 \pmod{3}. \quad (7.14)$$

At time 2, node 4 is activated. It has a single out-link to node 2 and since $|4 - 2| \equiv 2 \pmod{3}$, 2 is activated. Node 2 cannot activate other nodes; therefore, $f(\{1\}) = 4$. Similarly, we find that $f(\{2\}) = 1$, $f(\{3\}) = 1$, $f(\{4\}) = 2$, $f(\{5\}) = 1$, and $f(\{6\}) = 4$. So, 1 or 6 can be chosen for our first node. Let us choose 6. If 6 is initially activated, nodes 1, 2, 4, and 6 will become activated at the end. Now, from the set $\{1, 2, 3, 4, 5, 6\} \setminus \{1, 2, 4, 6\} = \{3, 5\}$, we need to select one more node. This is because in the setting for this example, $f(\{6, 1\}) = f(\{6, 2\}) = f(\{6, 4\}) = f(\{6\}) = 4$. In general, one needs to compute $f(S \cup \{v\})$ for all $v \in V \setminus S$ (see Algorithm 7.2, line 5). We have $f(\{6, 3\}) = f(\{6, 5\}) = 5$, so we can select one node randomly. We choose 3. So, $S = \{6, 3\}$ and $f(S) = 5$.

