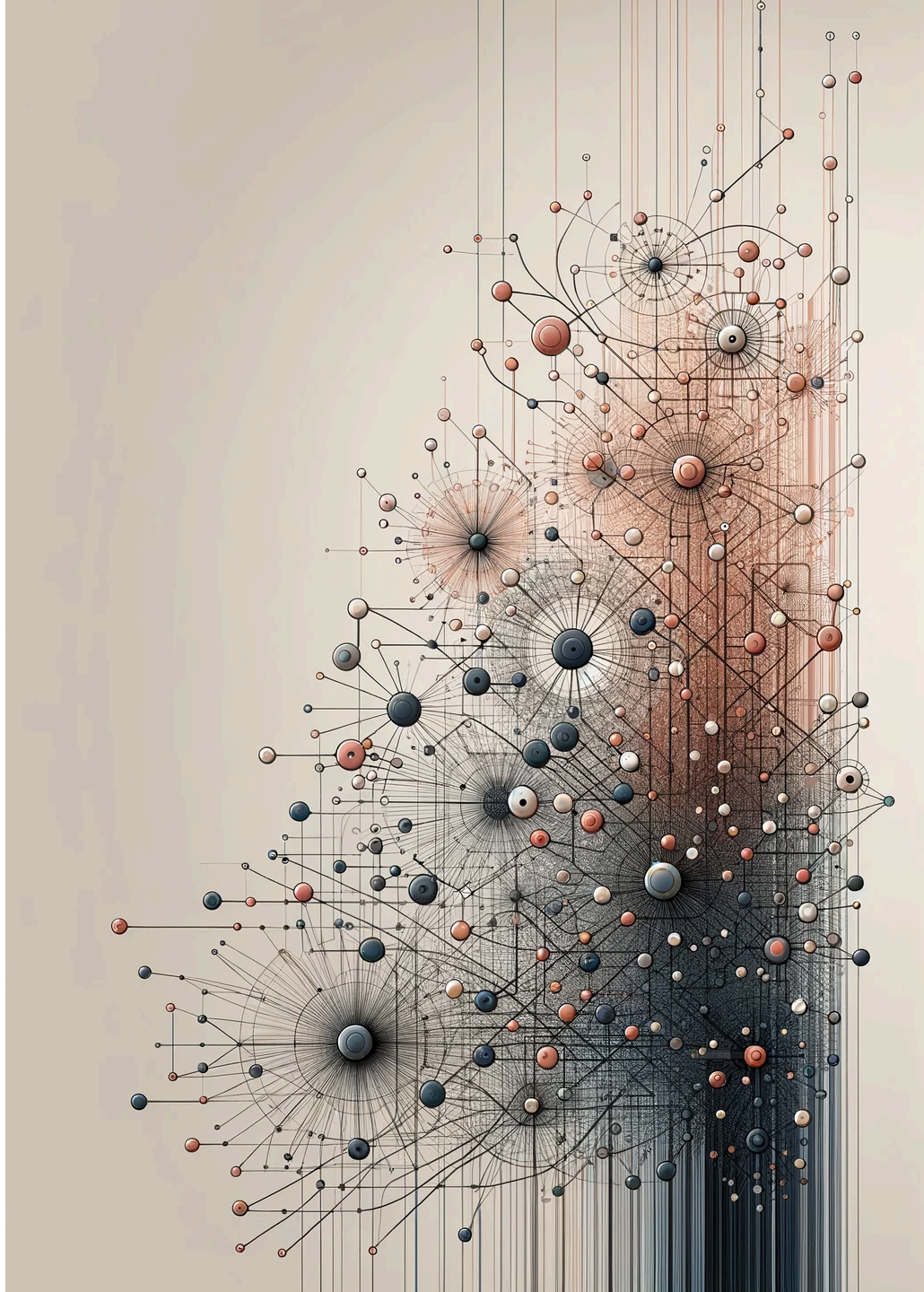
UNIVERSITÀ DI TORINO

Analisi e Visualizzazione delle Reti Complesse

NS02 - Recap on Graphs

Prof. Rossano Schifanella



UNIVERSITÀ DI TORINO

We  graphs

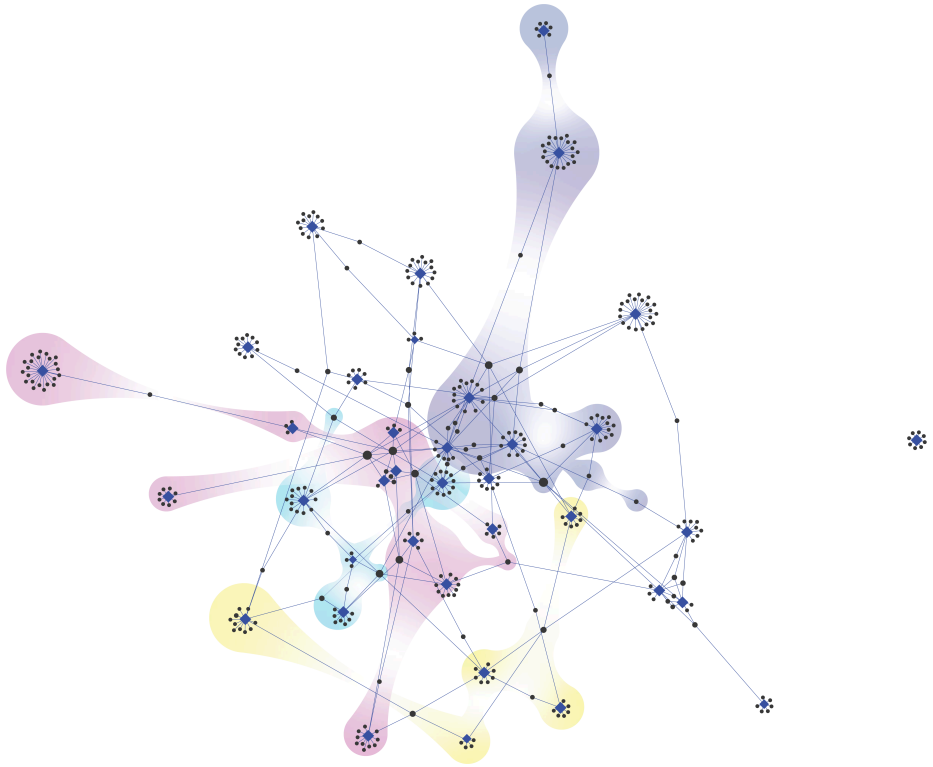
UNIVERSITÀ DI TORINO

Agenda

- Basic definitions of graphs
- Paths and Connectivity
- Distance and BFSearch
- Small world phenomenon (mention)

Why graphs?

- Graphs are a **mathematical model** that helps to represent complex systems in terms of **nodes** and **links**
- We need a **language** to understand the essential elements of a network
- With a language, we will be able to talk appropriately about:
  - **properties** that characterize the structure and behavior of networks
  - **roles** of networks in affecting **processes** occurring on network structures



Basic definitions

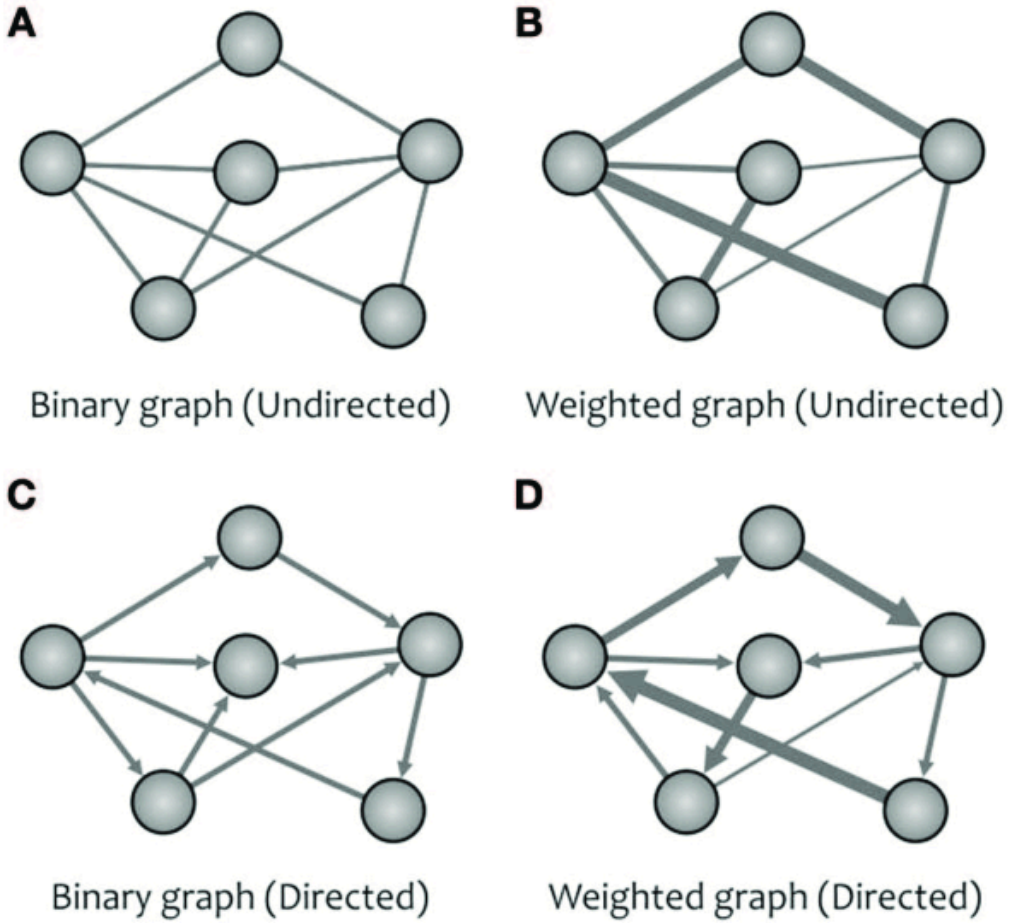
A **graph** is composed of nodes and links $G = (N, L)$

nodes (or vertices) $N = \{n_1, n_2, \ldots, n_l\} = \{1, 2, \ldots, l\}$

links (or edges or arcs) $L = \{(i, j) : i, j \in N\}$

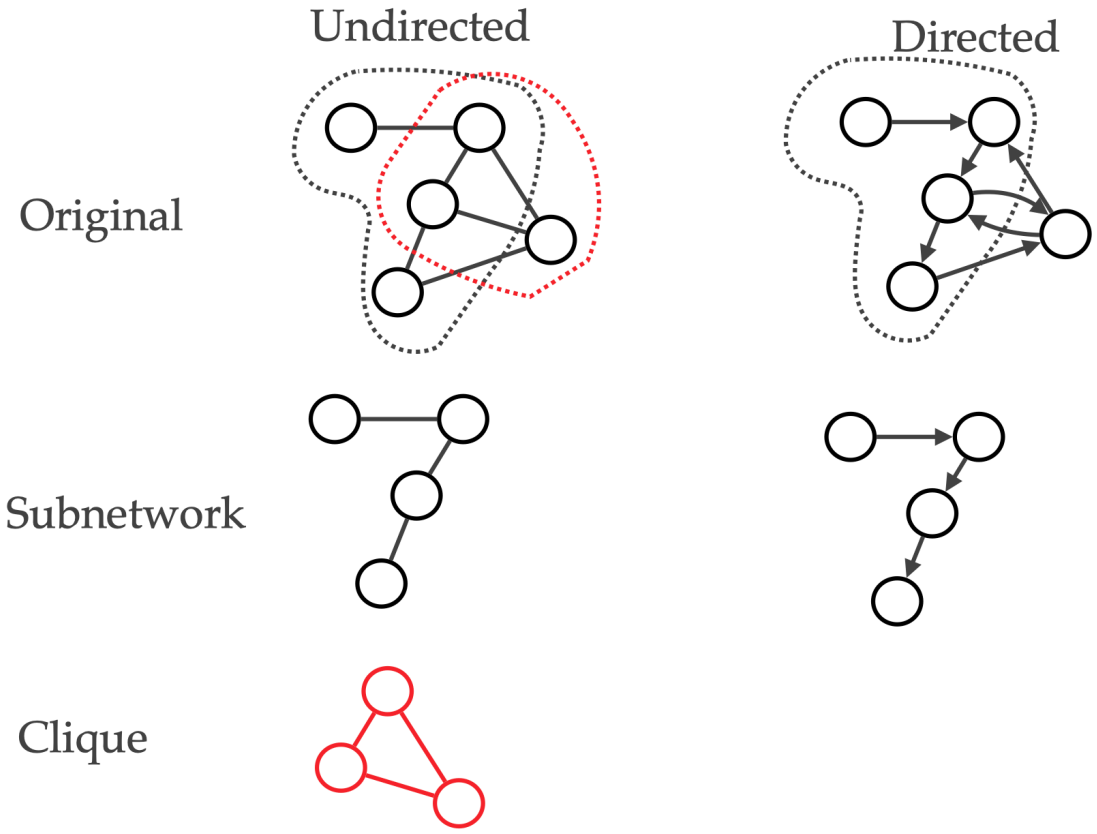
Graphs can be **directed** or **undirected** i, j 

Graphs can be **weighted** or **unweighted** $w_{ij}\ (i, j, w)$


A
B
Binary graph (Undirected)
Weighted graph (Undirected)
C
D
Binary graph (Directed)
Weighted graph (Directed)

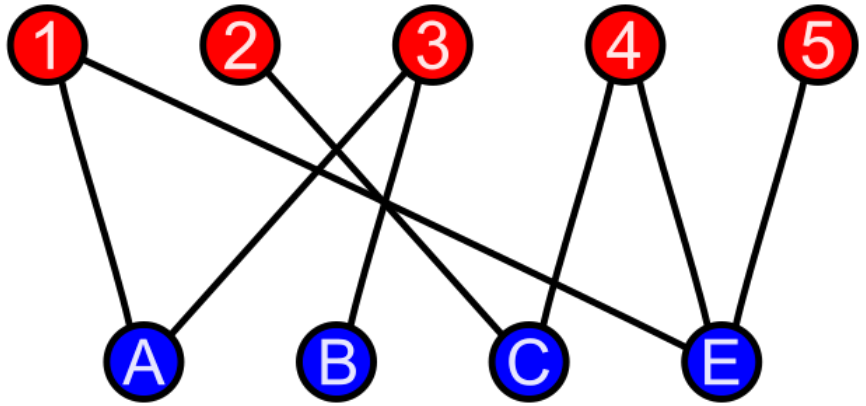
Subnetworks

- A **subnetwork** is a network obtained by selecting a subset of the nodes and all of the links among these nodes
- A **clique** is a complete subnetwork



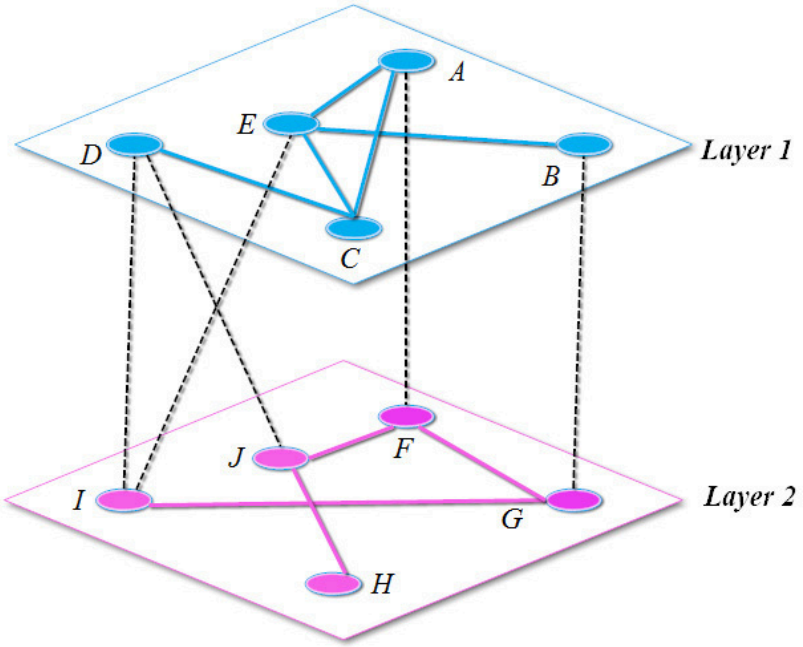
Bipartite graph

- **Two types of nodes**
- Connections only happen between nodes of different type
- Example: actor, movie dataset



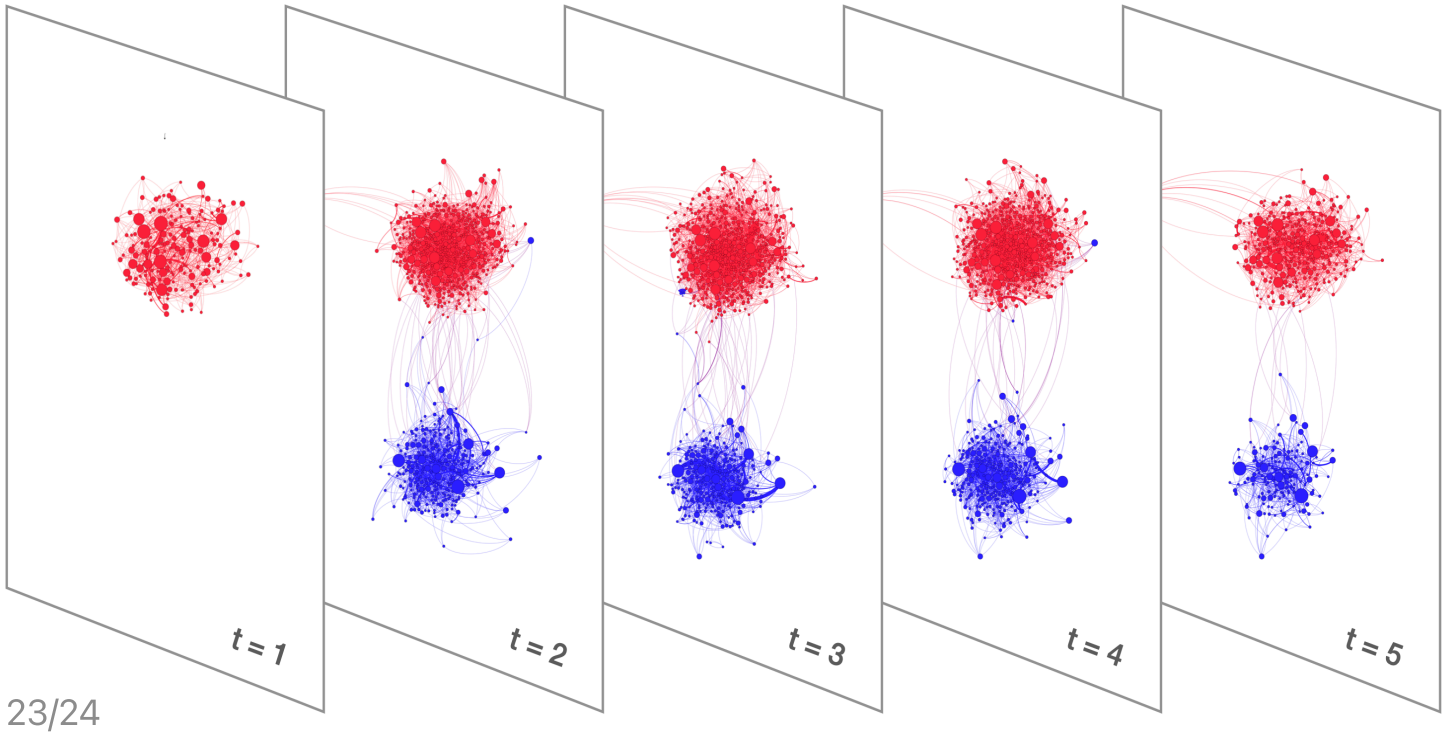
Multilayer networks

- A network can have **multiple layers**, each with its own nodes and edges
    - Example: air transportation networks of distinct airlines, with some but not complete overlap of airport nodes
- **Intralayer links** among nodes in the same layer, **interlayer links** across layers
- If the sets of nodes in the different layers are identical, we call the network a multiplex; interlayer links are **couplings** linking the same node across layers
    - Example: layers to represent different types of relationships in a social network, such as friendship, family ties, coworkers, etc.



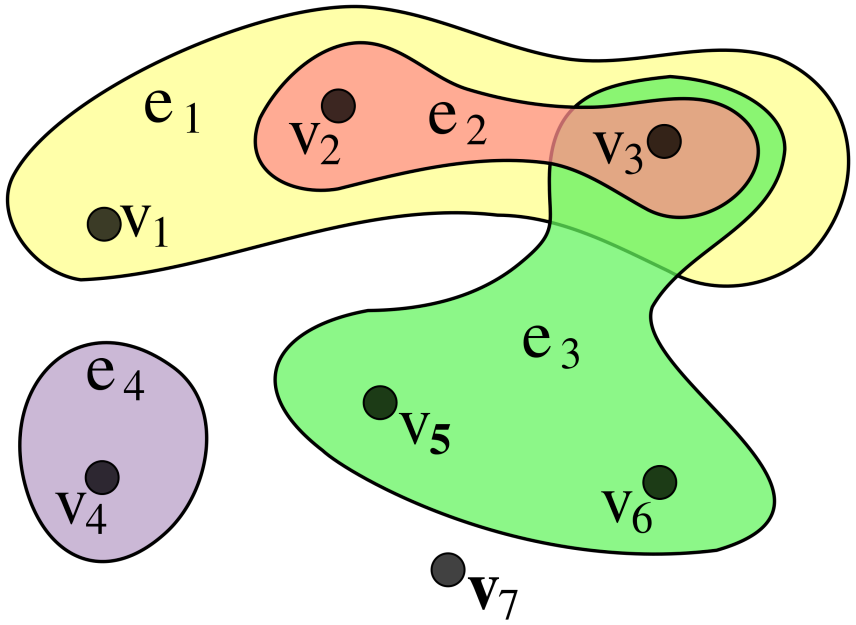
Temporal networks

- A temporal network is a multilayer network in which the layers represent links at different times (temporal snapshots)
  - Example: a Twitter retweet network



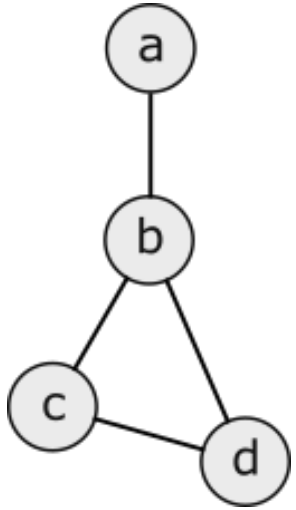
Hypergraphs

- **Interaction may occur between multiple nodes simultaneously**
- The hypergraph is represented by a sequence of **hyperedges** of size greater or equal to 2
  - For example: $E = (v_2, v_3), (v_1, v_2, v_3), (v_4), (v_3, v_5, v_6)$
  - co-authorship network



Neighbors

$$N_a = \{b\}$$
$$N_b = \{a, c, d\}$$
$$N_c = \{b, d\}$$
$$N_d = \{b, c\}$$

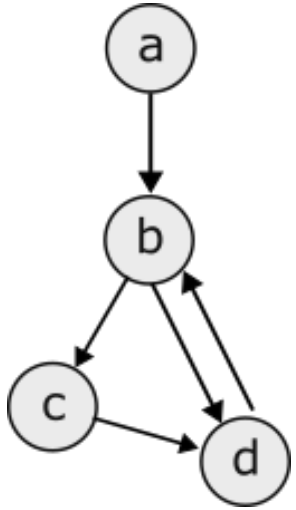


Successors

$S_a = \{b\}$
$S_b = \{c, d\}$
$S_c = \{d\}$
$S_d = \{b\}$

Predecessors

$P_a = \{\}$
$P_b = \{a, d\}$
$P_c = \{b\}$
$P_d = \{b, c\}$



Degree

Number of links (or neighbors)

$i \to N_i \qquad k_i = |N_i|$ **degree**

Singleton: a node whose degree is zero

$N_i = \{\}, k_i = 0$

In directed networks:

$k_i^{in} = |P_i|$ **in-degree**

$k_i^{out} = |S_i|$ **out-degree**

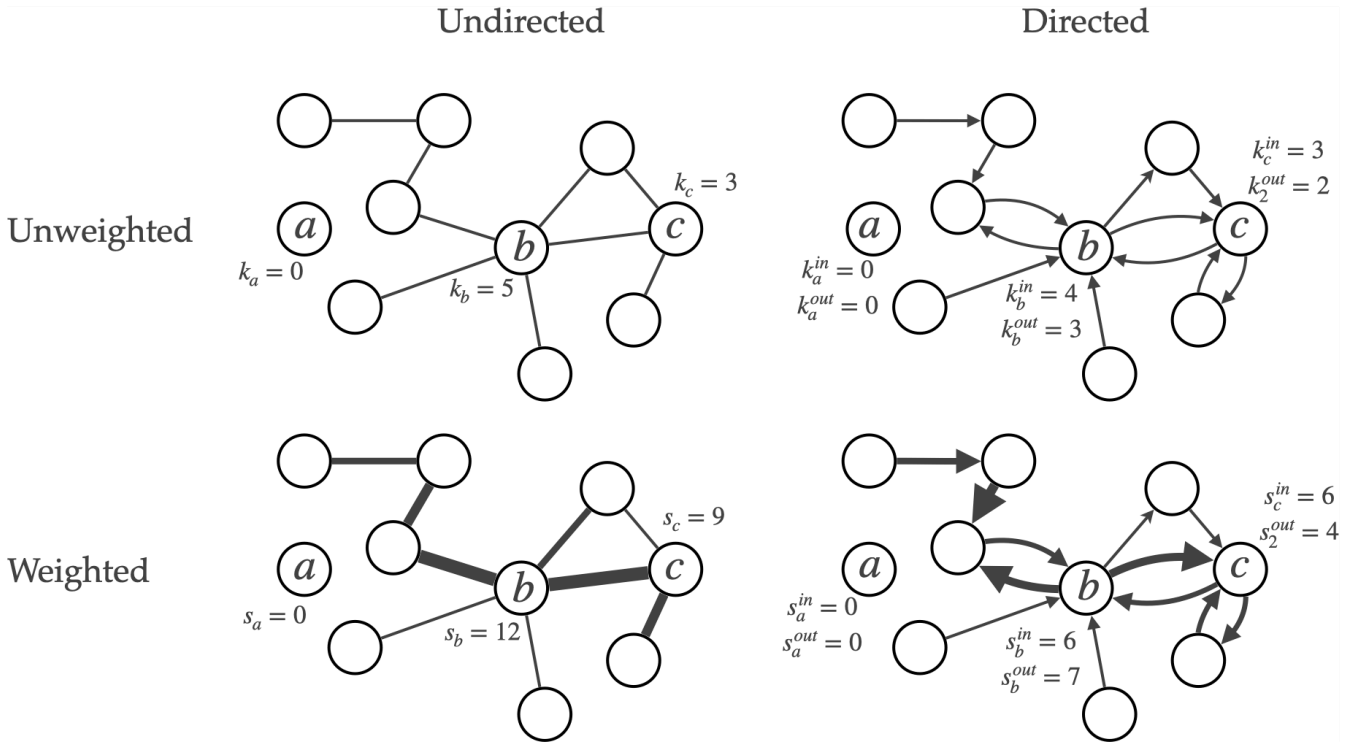
$k_i = k_i^{in} + k_i^{out}$

Strength

strength or **weighted-degree** $s_i = \sum_{j \in N_i} w_{ij}$

in-strength $s_i^{in} = \sum_{j \in P_i} w_{ji}$

out-strength $s_i^{out} = \sum_{j \in S_i} w_{ij}$



Density and sparsity

Network size = number of nodes	$\lvert N \rvert$ or, simply N
Number of links	$\lvert L \rvert$ or, simply L

In an undirected network:

Maximum possible number of links	$L_{max} = \binom{N}{2} = \frac{N(N-1)}{2}$
Density	$d = \frac{L}{L_{max}} = \frac{2L}{N(N-1)}$
Sparsity if $d \ll 1 \Rightarrow sparse$	$\begin{cases} \text{sparse:} & L \approx N \\ \text{dense:} & L \approx N^2 \end{cases}$

In a directed network:

Maximum possible number of links	$L_{max} = N(N-1)$
Density	$d = \frac{L}{L_{max}} = \frac{L}{N(N-1)}$

Average degree

$$\langle k \rangle = \frac{\sum_{i \in N} k_i}{N}$$

In an undirected network:

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

Given

$$d = \frac{2L}{N(N-1)} \Rightarrow L = \frac{dN(N-1)}{2}$$

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

The average degree is also connected to the density d

$$d = \frac{\langle k \rangle}{N-1} = \frac{\langle k \rangle}{k_{max}}$$

UNIVERSITÀ DI TORINO

Example: Facebook

- Rough orders-of-magnitude approximations:
- $N \approx 10^9$
- $L \approx 10^3 * N$
- $d \approx \frac{L}{N^2} = \frac{10^3 N}{N^2} = \frac{10^3}{10^9} = 10^{-6} \ll 1$

Table 1.1 Basic statistics of network examples. Network types can be (D)irected and/or (W)eighted. When there is no label the network is undirected and unweighted. For directed networks, we provide the average in-degree (which coincides with the average out-degree).

Network	Type	Nodes (N)	Links (L)	Density (d)	Average degree ($\langle k \rangle$)
Facebook Northwestern Univ.		10,567	488,337	0.009	92.4
IMDB movies and stars		563,443	921,160	0.000006	3.3
IMDB co-stars	W	252,999	1,015,187	0.00003	8.0
Twitter US politics	DW	18,470	48,365	0.0001	2.6
Enron Email	DW	87,273	321,918	0.00004	3.7
Wikipedia math	D	15,220	194,103	0.0008	12.8
Internet routers		190,914	607,610	0.00003	6.4
US air transportation		546	2,781	0.02	10.2
World air transportation		3,179	18,617	0.004	11.7
Yeast protein interactions		1,870	2,277	0.001	2.4
C. elegans brain	DW	297	2,345	0.03	7.9
Everglades ecological food web	DW	69	916	0.2	13.3

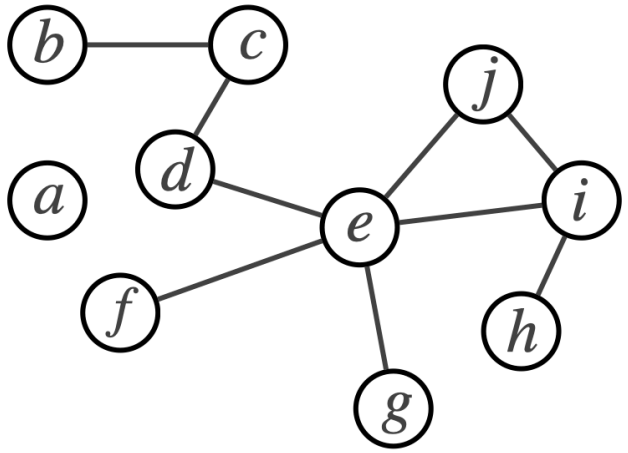
Network representations

Adjacency Matrix

$N \times N$ matrix

$$a_{ij} = \begin{cases} 0 & \text{no edge} \\ 1 & (i,j) \in L \end{cases}$$

In an undirected network: $a_{ij} = a_{ji}$



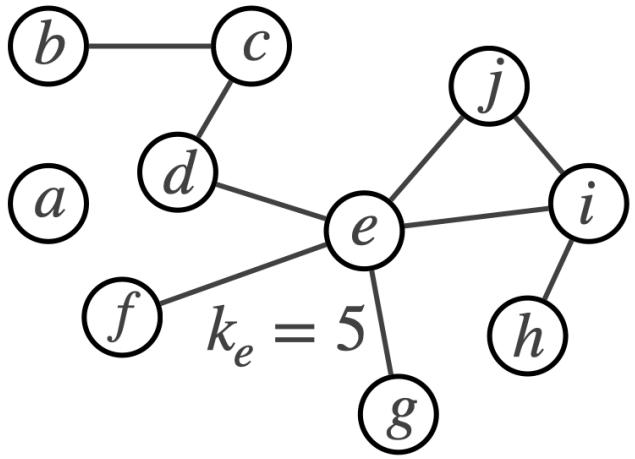
	a	b	c	d	e	f	g	h	i	j
a	0	0	0	0	0	0	0	0	0	0
b	0	0	1	0	0	0	0	0	0	0
c	0	1	0	1	0	0	0	0	0	0
d	0	0	1	0	1	0	0	0	0	0
e	0	0	0	1	0	1	1	0	1	1
f	0	0	0	0	1	0	0	0	0	0
g	0	0	0	0	1	0	0	0	0	0
h	0	0	0	0	0	0	0	0	1	0
i	0	0	0	0	1	0	0	1	0	1
j	0	0	0	0	1	0	0	0	1	0

Network representations

Adjacency Matrix

Degree

$$k_i = \sum_j a_{ij} = \sum_j a_{ji}$$



	a	b	c	d	e	f	g	h	i	j
a	0	0	0	0	0	0	0	0	0	0
b	0	0	1	0	0	0	0	0	0	0
c	0	1	0	1	0	0	0	0	0	0
d	0	0	1	0	1	0	0	0	0	0
e	0	0	0	1	0	1	1	0	1	1
f	0	0	0	0	1	0	0	0	0	0
g	0	0	0	0	1	0	0	0	0	0
h	0	0	0	0	0	0	0	0	1	0
i	0	0	0	0	1	0	0	1	0	1
j	0	0	0	0	1	0	0	0	1	0

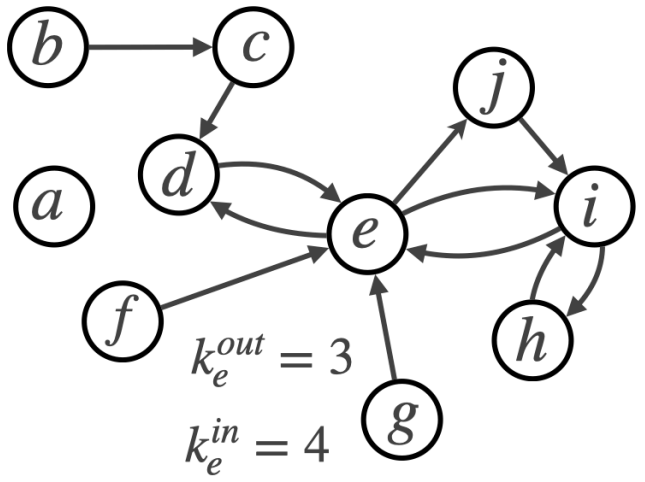
Network representations

Adjacency Matrix

In a **directed** graph the matrix is **not symmetric**

$$k_i^{out} = \sum_j a_{ij}$$

$$k_i^{in} = \sum_j a_{ji}$$



	a	b	c	d	e	f	g	h	i	j
a	0	0	0	0	0	0	0	0	0	0
b	0	0	1	0	0	0	0	0	0	0
c	0	0	0	1	0	0	0	0	0	0
d	0	0	0	0	1	0	0	0	0	0
e	0	0	0	1	0	0	0	0	1	1
f	0	0	0	0	1	0	0	0	0	0
g	0	0	0	0	1	0	0	0	0	0
h	0	0	0	0	0	0	0	0	1	0
i	0	0	0	0	1	0	0	1	0	0
j	0	0	0	0	0	0	0	0	1	0

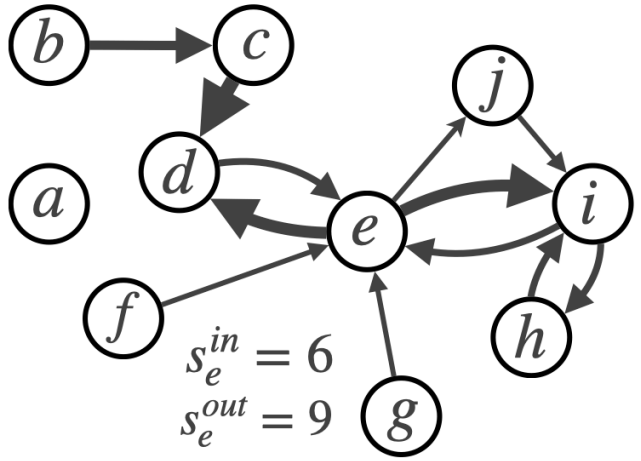
Network representations

Adjacency Matrix

In a **weighted** graph

$$a_{ij} = \begin{cases} 0 & \text{no edge} \\ w_{ij} & (i,j) \in L \end{cases}$$

$$s_i^{out} = \sum_j w_{ij}$$
$$s_i^{in} = \sum_j w_{ji}$$



	a	b	c	d	e	f	g	h	i	j
a	0	0	0	0	0	0	0	0	0	0
b	0	0	3	0	0	0	0	0	0	0
c	0	0	0	4	0	0	0	0	0	0
d	0	0	0	0	2	0	0	0	0	0
e	0	0	0	4	0	0	0	0	4	1
f	0	0	0	0	1	0	0	0	0	0
g	0	0	0	0	1	0	0	0	0	0
h	0	0	0	0	0	0	0	0	2	0
i	0	0	0	0	2	0	0	2	0	0
j	0	0	0	0	0	0	0	0	2	0

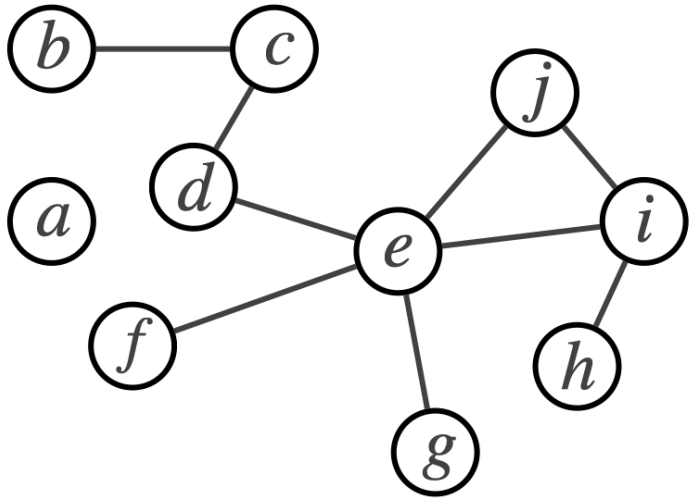
Sparse network representations

- The memory/disk storage needed by an adjacency matrix is proportional to N^2
- In sparse networks (most real-world networks), this is inefficient: most of the space is wasted storing zeros (non-links); for very large networks, adjacency matrices are unfeasible
- It is much more efficient, often necessary, to store only the actual links and assume that if a link is not listed, it means it is not present
- There are two commonly used sparse network representations:
  - **Adjacency list**
  - **Edge list**

Adjacency List

Undirected network:
list each link twice

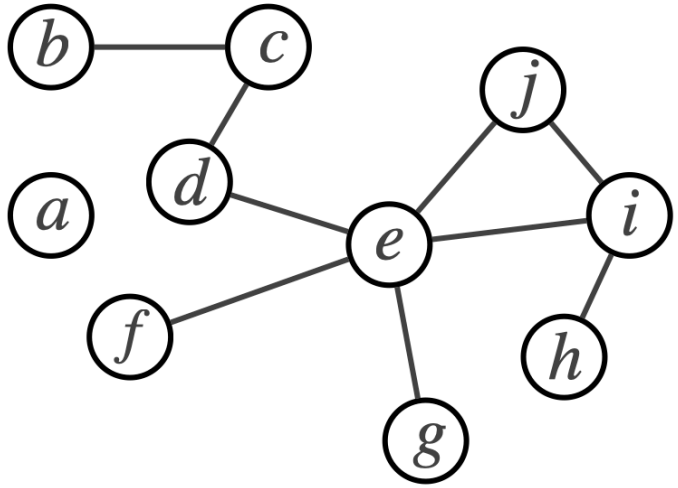
Directed network:
list only existing links




```
a
b    c
c    b d
d    c e
e    d f i j g
f    e
g    e
h    i
i    h j
j    e i
```

Edge list

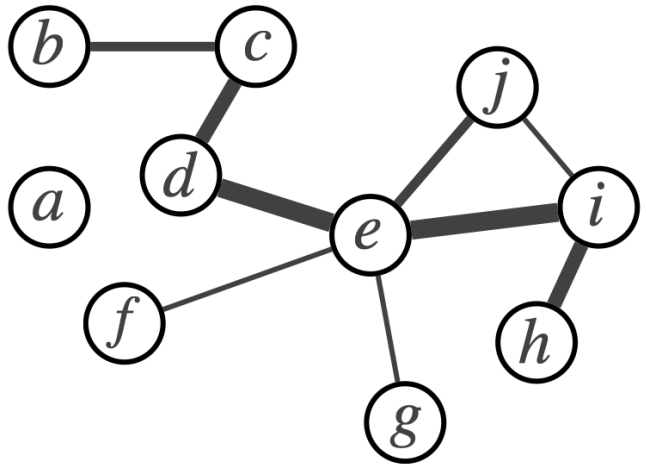
unweigthed



b	c
c	d
d	e
e	f
e	g
e	i
e	j
h	i
i	j

L

weighted



b	c	2
c	d	3
d	e	4
e	f	4
e	g	1
e	i	1
e	j	2
h	i	3
i	j	1

L

Paths and connectivity

Paths and cycles

path: $\{n_1, n_2, \ldots, n_k\}$ where $\forall i : (n_i, n_{i+1}) \in L$

A path has a **length** l

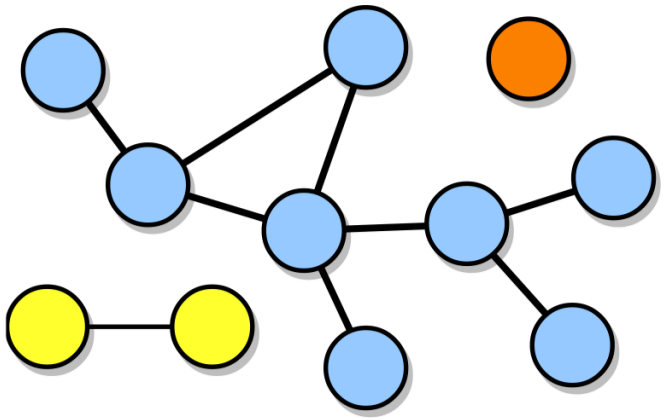
- number of edges in the path
- $l = k - 1$

if we have repeating nodes $\Rightarrow$ **cycles**

no repeating nodes $\Rightarrow$ **simple path**

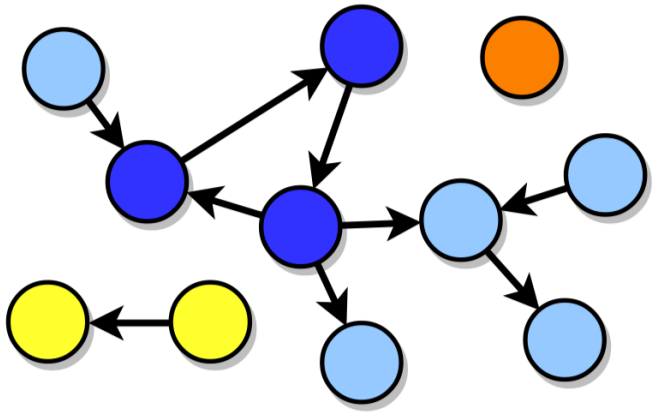
Connectedness and components

- A network is **connected** if there is a path between any two nodes
- If a network is not connected, it is **disconnected** and has multiple connected components
- A connected component is a connected subnetwork
    - The largest one is called **giant component**; it often includes a substantial portion of the network
    - A **singleton** is the smallest-possible connected component



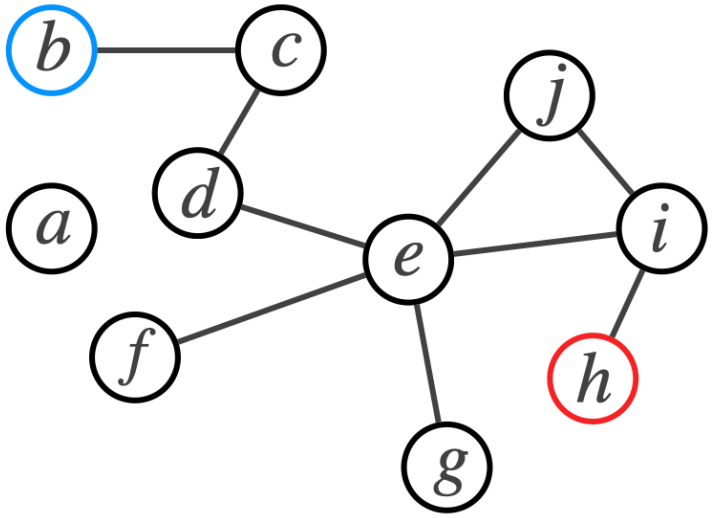
Connectedness and components

- A directed network can be **strongly connected** or **weakly connected** if there is a path between any two nodes, respecting or disregarding the link directions, respectively
- The **in-component** of a strongly connected component S is the set of nodes from which one can reach S, but that cannot be reached from S
- The **out-component** of a strongly connected component S is the set of nodes that can be reached from S, but from which one cannot reach S



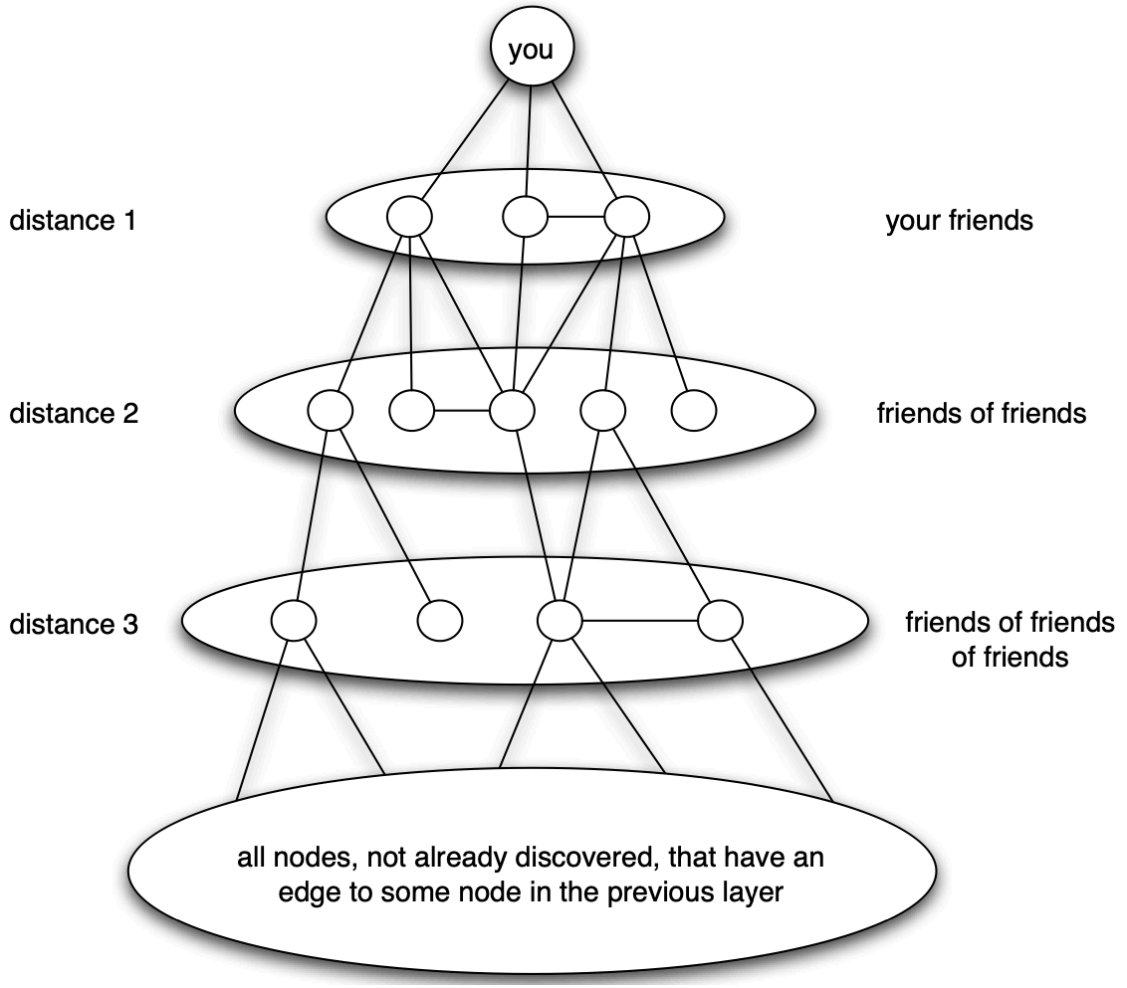
Shortest paths

- Two examples of paths from source b to target h
  - $\{b, c, d, e, i, h\}$
  - $\{b, c, d, e, j, i, h\}$
- **shortest path**: **minimal** path between two nodes
  - in weighted networks: weights can represent distances from two adjacent nodes
- **distance** between two nodes: shortest path length
- a **singleton** is considered at distance ∞ from any other nodes

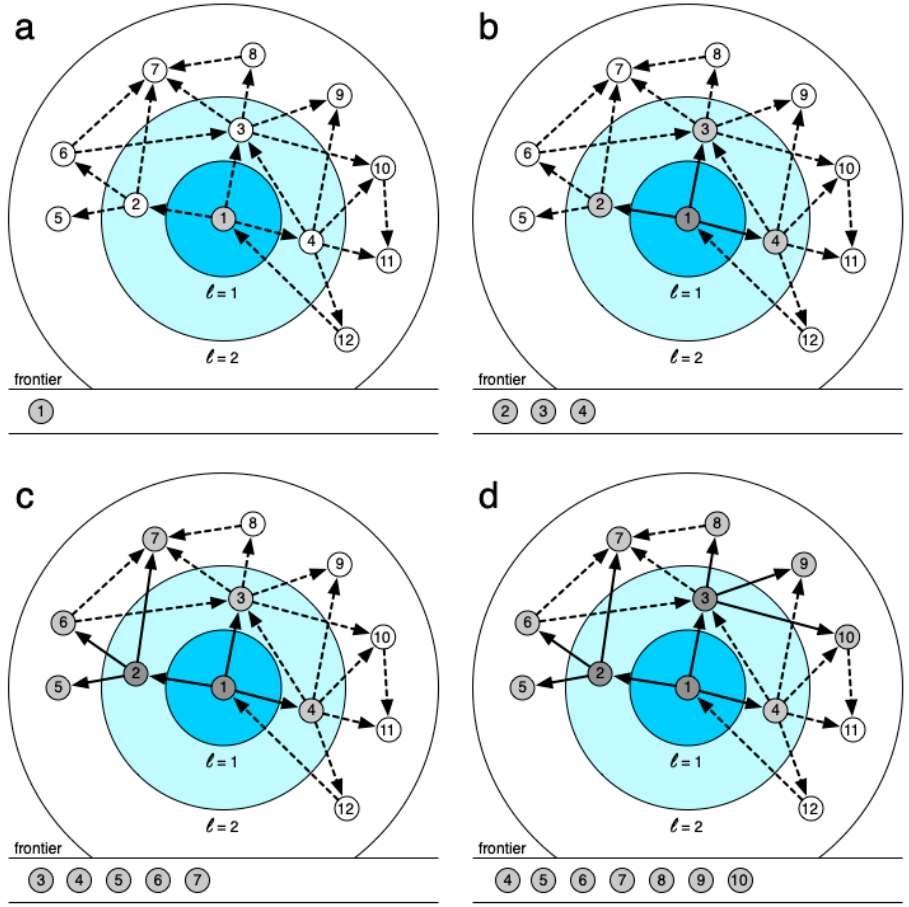


BFS (Breadth First Search)

- One of the most efficient algorithm to find distance
- Start from a **source node** (root)
- Visit the **entire breadth of the network**, within some distance from the source, before moving to a greater depth



- Each node has an attribute storing its distance l from the source
- initially $l(node) = -1$ except for $l(source) = 0$
- A queue (FIFO) holds the frontier, initially contains the source
- A directed shortest path tree, initially contains all the nodes and no links
- Iterate until the frontier is empty:
  - Remove next node i in frontier
  - For each neighbor/successor j of i with $l(j) = -1$:
  - Queue j into frontier



UNIVERSITÀ DI TORINO

Average Path Length and diameter

- **Average Path Length (APL)**
  - undirected network $\langle l \rangle = \frac{2\sum_{ij} l_{ij}}{N(N-1)}$
  - directed netwrok $\langle l \rangle = \frac{\sum_{ij} l_{ij}}{N(N-1)}$
- **Diameter** $l_{max} = max_{ij}(l_{ij})$

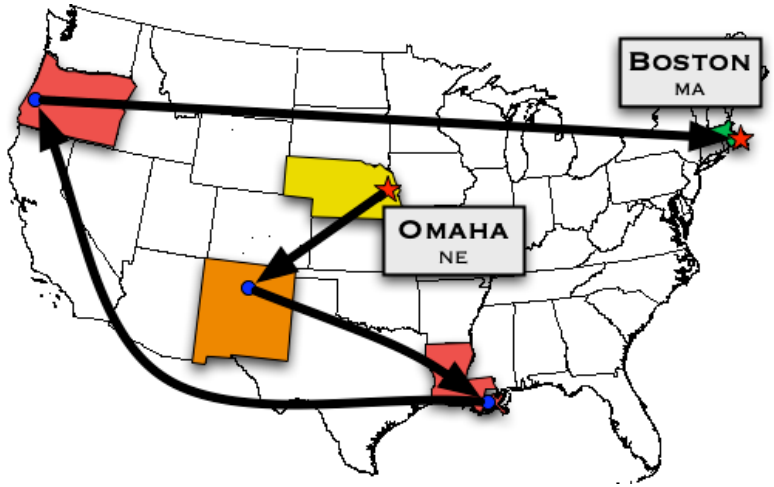
With disconnected components:

$\langle l \rangle = \frac{\sum_{ij} l_{ij}}{N(N-1)} = \infty$ you can rewrite it for undirected network $\langle l \rangle = \left(\frac{\sum_{ij} \frac{1}{l_{ij}}}{\binom{N}{2}} \right)^{-1}$

Small world phenomenon

Milgram's experiment

- Instructions:
  - **send the letter to a personal acquaintance who is more likely to know the target**
- 160 letters to people in Omaha, Nebraska and Wichita, Kansas
- 2 targets in Massachusetts
  - the wife of a student in Sharon and a stockbroker in Boston
- 42 letters made it back (only 26%)
- Average: 6.5 steps (range: 3-12 steps)
- Much lower than most people expected!
- The **small world** effect is still surprising



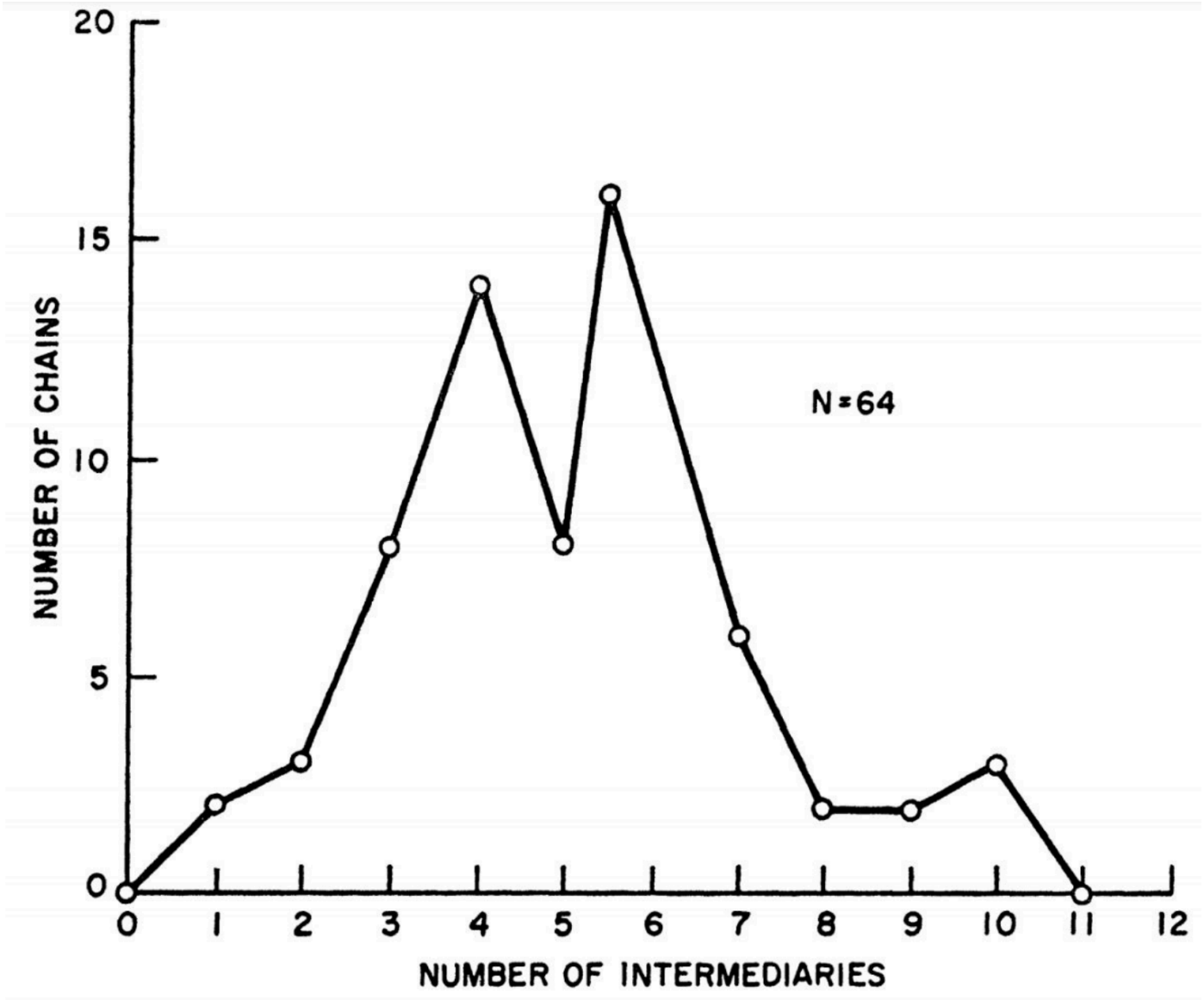


FIGURE 1

Lengths of Completed Chains

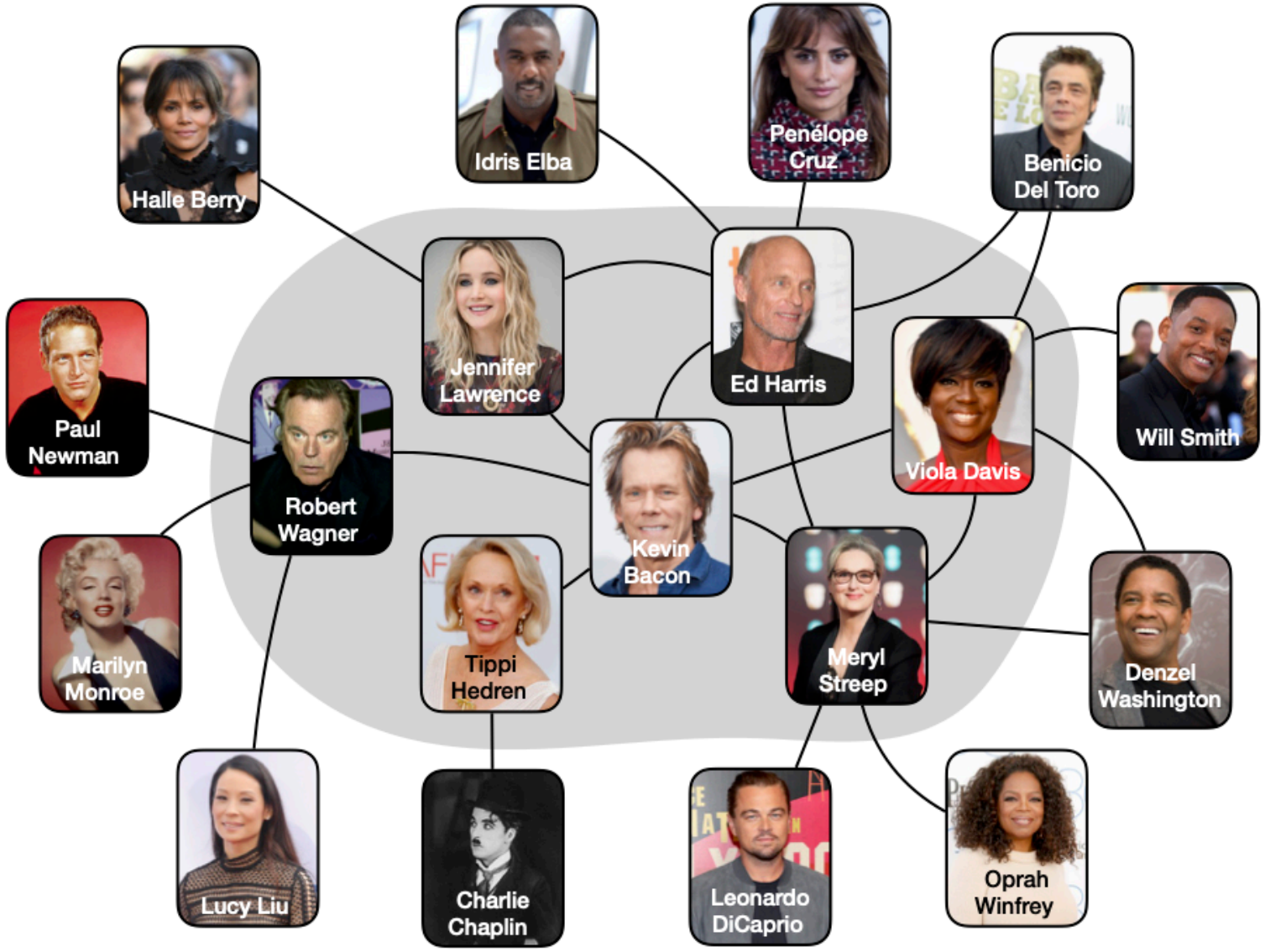
Other small world experiments

- Erdös number
- Oracle of Bacon
- Yahoo! email (Kleinberg, 2003)
- Instant Messaging (Leskovec and Horvitz, 2007)
- Facebook (Boldi, Vigna and others, 2011)

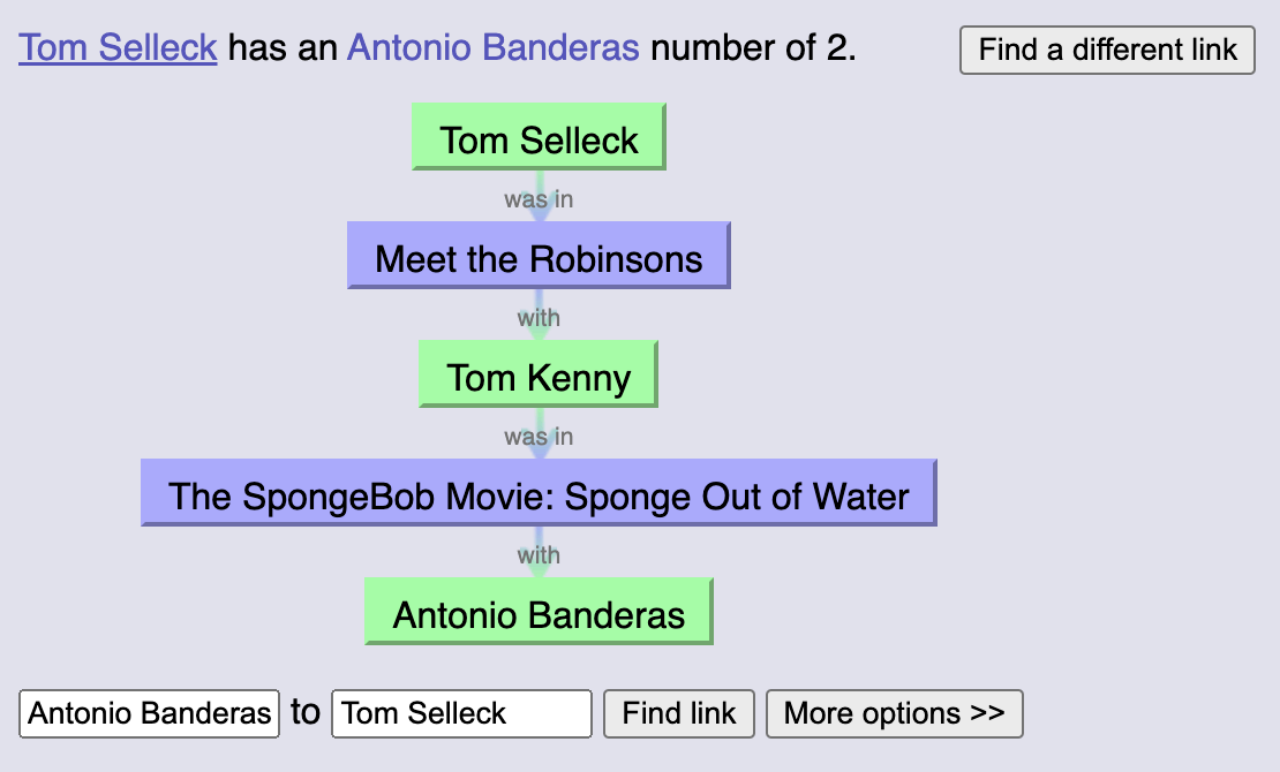
UNIVERSITÀ DI TORINO

Bacon's oracle

- Movie co-star network
- oracleofbacon.org



- Not only Kevin Bacon
- Can you find two stars separated by more than four links?
- Play the game and try!



Instant messaging

- 240 millions of nodes
- median: 7
- average: 6.6

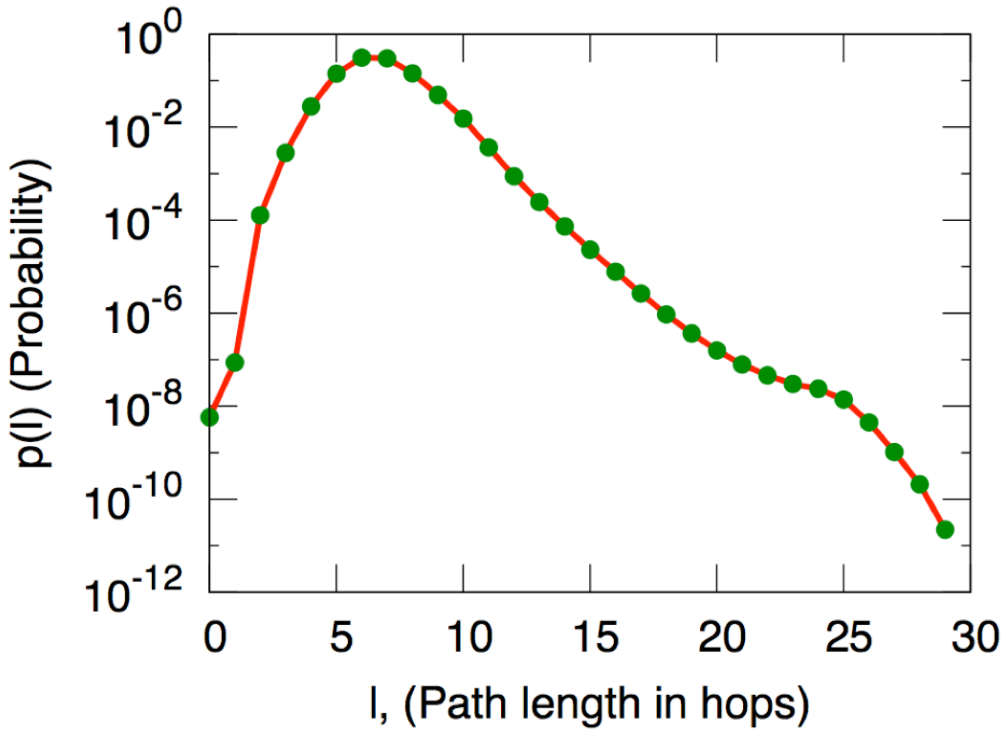


Figure 2.11: The distribution of distances in the graph of all active Microsoft Instant Messenger user accounts, with an edge joining two users if they communicated at least once during a month-long observation period [273].

The New York Times

Separating You and Me? 4.74 Degrees

By John Markoff and Somini Sengupta

Nov. 21, 2011

The world is even smaller than you thought.

Adding a new chapter to the research that cemented the phrase "six degrees of separation" into the language, scientists at Facebook and the University of Milan reported on Monday that the average number of acquaintances separating any two people in the world was not six but 4.74.

The original "six degrees" finding, published in 1967 by the psychologist Stanley Milgram, was drawn from 296 volunteers who were asked to send a message by postcard, through friends and then friends of friends, to a specific person in a Boston suburb.

When a short path is really short?

- It depends on the size of the network!
- Observe the relationship between APL and network size when considering networks (or subnetworks) of different sizes
- We say that the average path length is short when it grows very slowly with the size of the network, say, logarithmically:

$$\langle l \rangle \approx logN$$

UNIVERSITÀ DI TORINO

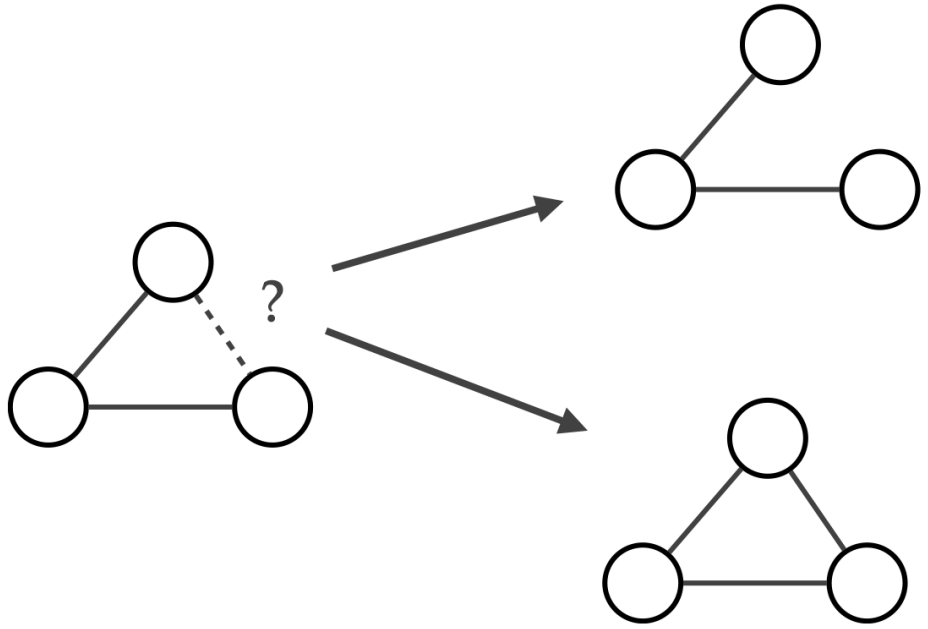
Small world networks

Table 1.1 Basic statistics of network examples. Network types can be (D)irected and/or (W)eighted. When there is no label the network is undirected and unweighted. For directed networks, we provide the average in-degree (which coincides with the average out-degree).

Network	Type	Nodes (N)	Links (L)	Density (d)	Average degree ($\langle k \rangle$)
Facebook Northwestern Univ.		10,567	488,337	0.009	92.4
IMDB movies and stars		563,443	921,160	0.000006	3.3
IMDB co-stars	W	252,999	1,015,187	0.00003	8.0
Twitter US politics	DW	18,470	48,365	0.0001	2.6
Enron Email	DW	87,273	321,918	0.00004	3.7
Wikipedia math	D	15,220	194,103	0.0008	12.8
Internet routers		190,914	607,610	0.00003	6.4
US air transportation		546	2,781	0.02	10.2
World air transportation		3,179	18,617	0.004	11.7
Yeast protein interactions		1,870	2,277	0.001	2.4
C. elegans brain	DW	297	2,345	0.03	7.9
Everglades ecological food web	DW	69	916	0.2	13.3

A friend of a friend

- In social networks, we have **triangles**
- **Very common** phenomenon



Clustering coefficient

The clustering coefficient of a node is the **fraction of pairs of the node's neighbors that are connected to each other**

In alternative: the **ratio between the number of triangles that include the node, and the maximum number of triangles in which the node could participate**

If:

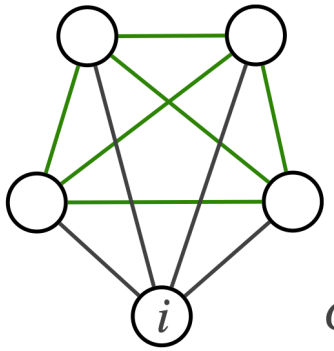
- N_i is the set of the neighbors of i while $k_i = |N_i|$ is the degree of i
- $\tau(i)$ is the number of triangles involving i:

$$C(i) = \frac{\tau(i)}{\tau_{max}(i)} = \frac{\tau(i)}{\binom{k_i}{2}} = \frac{2\tau(i)}{k_i(k_i-1)}$$

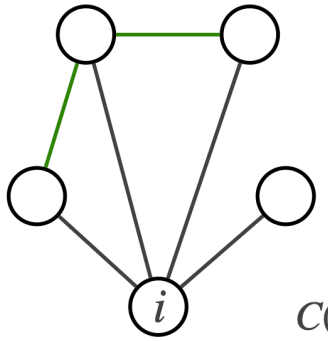
Network clustering coefficient

The clustering coefficient of the network is the **average of the clustering coefficients of the nodes**:

$$C = \frac{\sum_{i:k_i>1} C(i)}{N_{k>1}}$$



$$C(i) = \frac{2 \cdot 6}{4 \cdot 3} = 1$$



$$C(i) = \frac{2 \cdot 2}{4 \cdot 3} = \frac{1}{3}$$

Network clustering coefficient

Some networks, e.g., **social networks**, tend to have **high clustering coefficients because of triadic closure**: we meet through common friends

Other networks, e.g., bipartite and tree-like networks, have low clustering coefficient

Table 1.1 Basic statistics of network examples. Network types can be (D)irected and/or (W)eighted. When there is no label the network is undirected and unweighted. For directed networks, we provide the average in-degree (which coincides with the average out-degree).

Network	Type	Nodes (N)	Links (L)	Density (d)	Average degree ($\langle k \rangle$)
Facebook Northwestern Univ.		10,567	488,337	0.009	92.4
IMDB movies and stars		563,443	921,160	0.000006	3.3
IMDB co-stars	W	252,999	1,015,187	0.00003	8.0
Twitter US politics	DW	18,470	48,365	0.0001	2.6
Enron Email	DW	87,273	321,918	0.00004	3.7
Wikipedia math	D	15,220	194,103	0.0008	12.8
Internet routers		190,914	607,610	0.00003	6.4
US air transportation		546	2,781	0.02	10.2
World air transportation		3,179	18,617	0.004	11.7
Yeast protein interactions		1,870	2,277	0.001	2.4
C. elegans brain	DW	297	2,345	0.03	7.9
Everglades ecological food web	DW	69	916	0.2	13.3

Reading material

[ns1] **Chapter 1** and **Chapter 2** (no homophily/assortativity for now)

[ns2] **Chapter 2**

Q&A