



UNIVERSITÀ DEGLI STUDI DI PADOVA

Network Science

A.Y. 26/27

ICT for Internet & multimedia, Data science, Physics of data

Course overview

Network science 26/27



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Lecturer

tomaso.erseghe@unipd.it
room 217, DEI/A



lectures: **thu** 16:30-18:00 & **fri** 16:30-18:00

office hours: contact me by email



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

In this course you'll also meet





Basic requirements (that you already satisfy)



Calculus and linear algebra
Familiarity with a programming language
(**Python**, R, MatLab, C, Java, etc.)
Probability theory / Statistics

Other useful knowledge

Networking processes in economics,
telecommunications, semantics, etc ...
Optimization, machine learning,
deep learning, etc ...





Which programming language?

☐ Python



very good at scraping data (e.g., via Twitter APIs), polishing, plotting graphs, implementing algorithms

☐ R



very good for memory storage, plotting graphs, implementing algorithms

☐ MatLab



MATLAB®

An alternative for algorithms and graph plotting

University license available

<https://www.ict.unipd.it/servizi/servizi-utenti-istituzionali/contratti-software-e-licenze/matlab>



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

What about you?



Why did you pick the course?



Which is your background?
Who knows about deep learning?



Do you know Python?
and CoLab?



What do you expect
from this course?



Do you have a laptop?



Are you interested in an
interdisciplinary work?



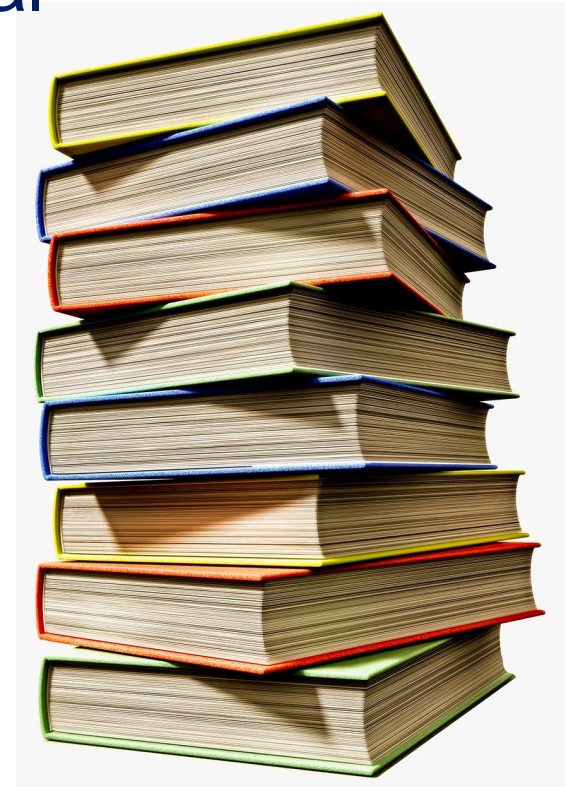
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Textbooks

No textbook! 😊

Slides/videos & additional material
available

@ stem.elearning.unipd.it





- ❑ A.L. Barabási, «Network science»
<http://barabasi.com/networksciencebook>
(these slides = Ch.1 “Introduction”)
- ❑ J. Leskovec, «Machine learning with graphs»
<http://web.stanford.edu/class/cs224w>
- ❑ M. Newman, «Networks: an introduction»
Oxford University Press, 2010
- ❑ R. van der Hofstad, «Random graphs and complex networks»
<http://www.win.tue.nl/~rhofstad/NotesRGCN.html>



Project based exam

☐ Written exam

multiple choice questions (30 min)
2 open questions (30+30 min)



☐ Project

extract network analytics using your
preferred programming **language(s)**
oral presentation: **slides + code**
10 min presentation (slides)
5 min for questions



Final grade: 40% written exam, 60% project



- ☐ Jan 29, 2027 (Fri) - 10:00 (written) 11:30 (oral), Le
- ☐ IP day Feb 12, 2027 (Fri) - 9:00 (oral), tbd
- ☐ Feb 19, 2027 (Fri) - 10:00 (written) 11:30 (oral), Le
- ☐ July 2, 2027 (Fri) - 10:00 (written) 11:30 (oral), Le
- ☐ Sep 10, 2027 (Fri) - 10:00 (written) 11:30 (oral), Le

PS: You will be asked to enrol in
www.uniweb.unipd.it

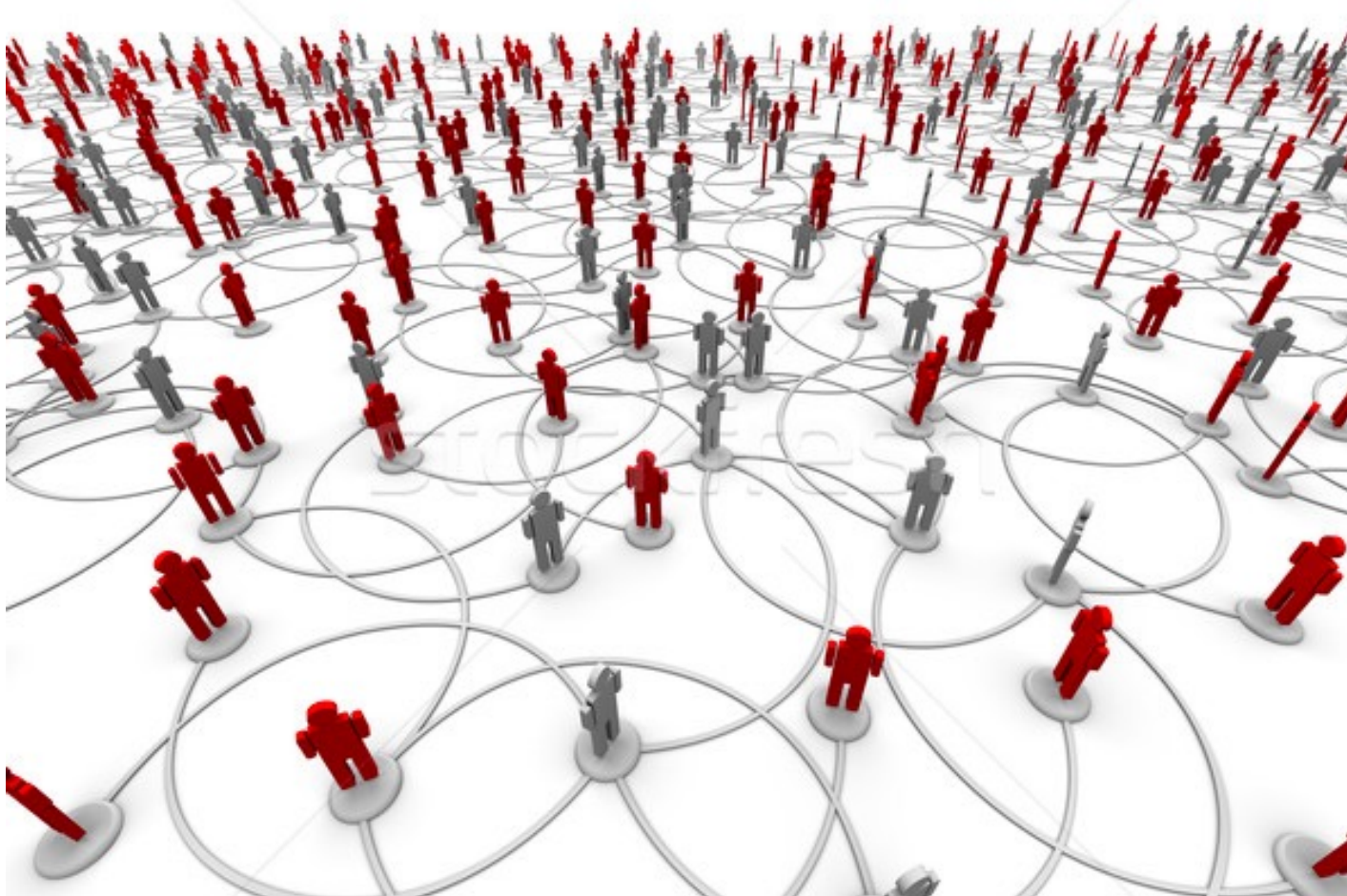
Contents

a brief overview



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

This course is about networks



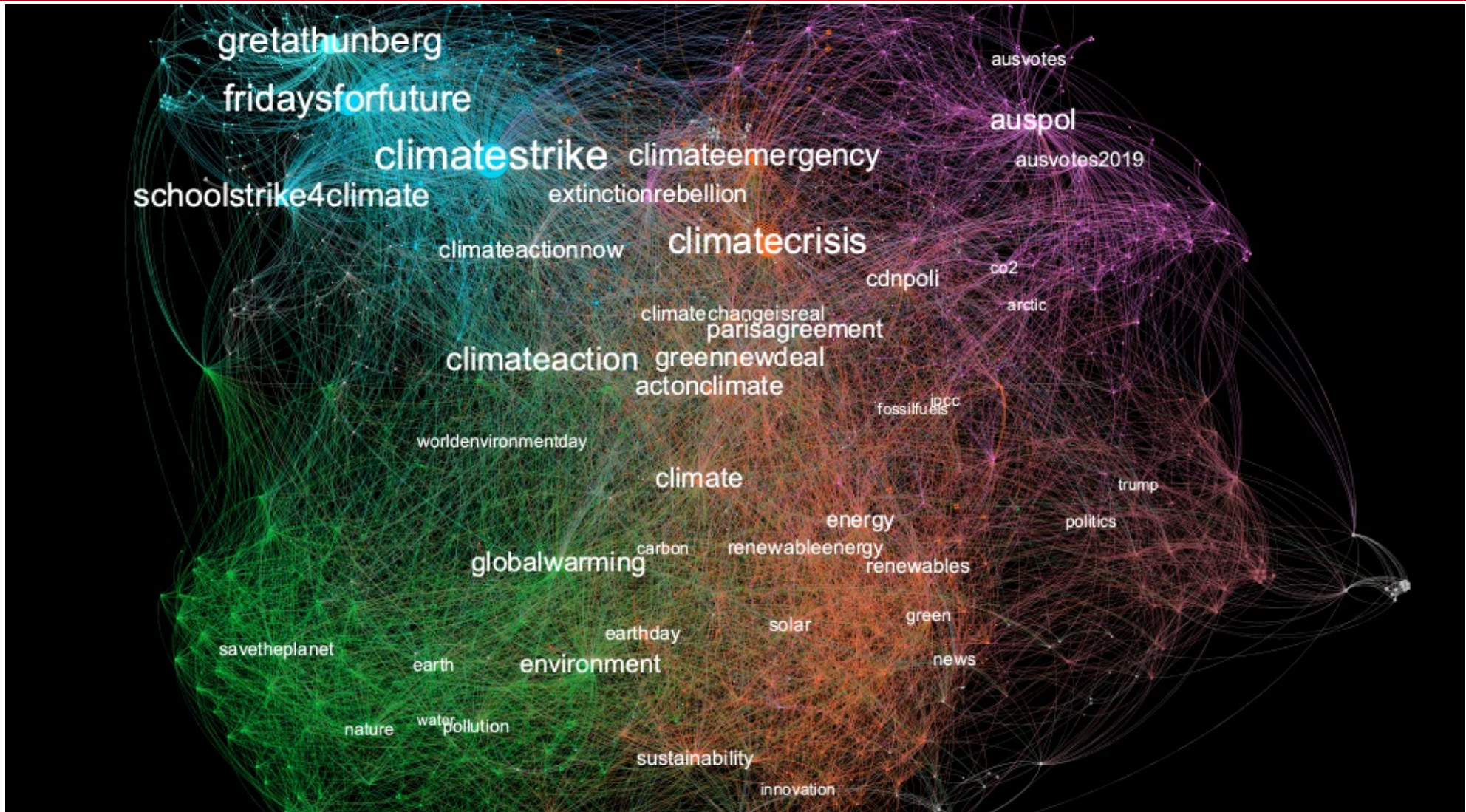
Network = anything that interconnects
e.g., people sharing friendship in a social network platform



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Network example

2019 hashtag network related to #climatechange
from Twitter, after #gretathunberg

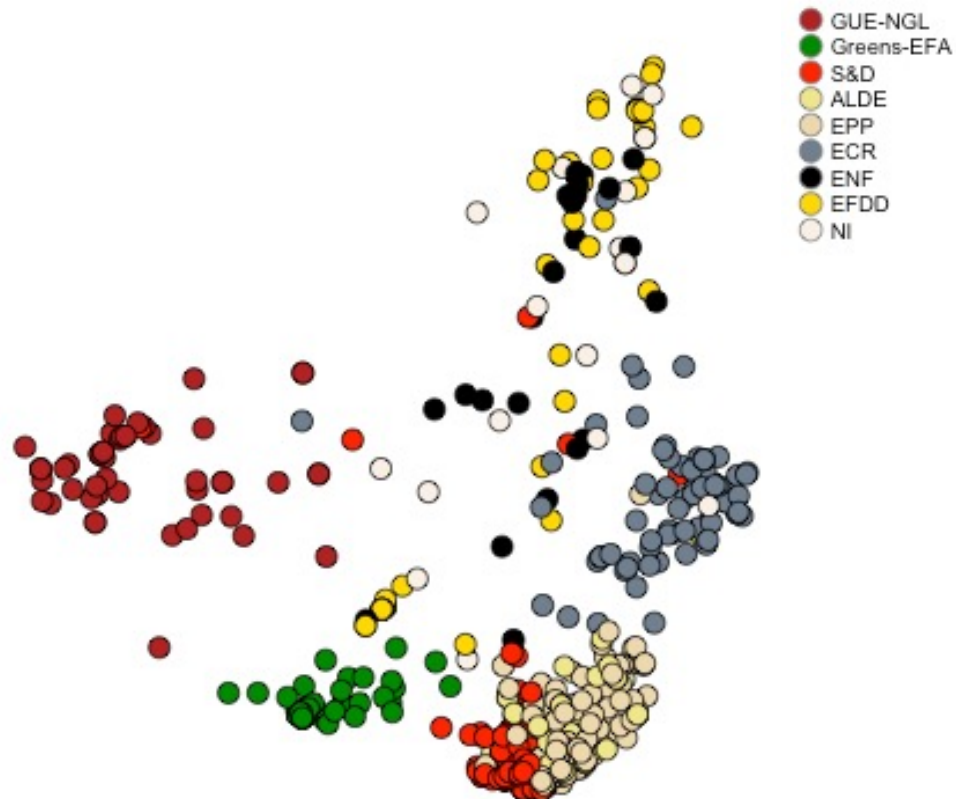




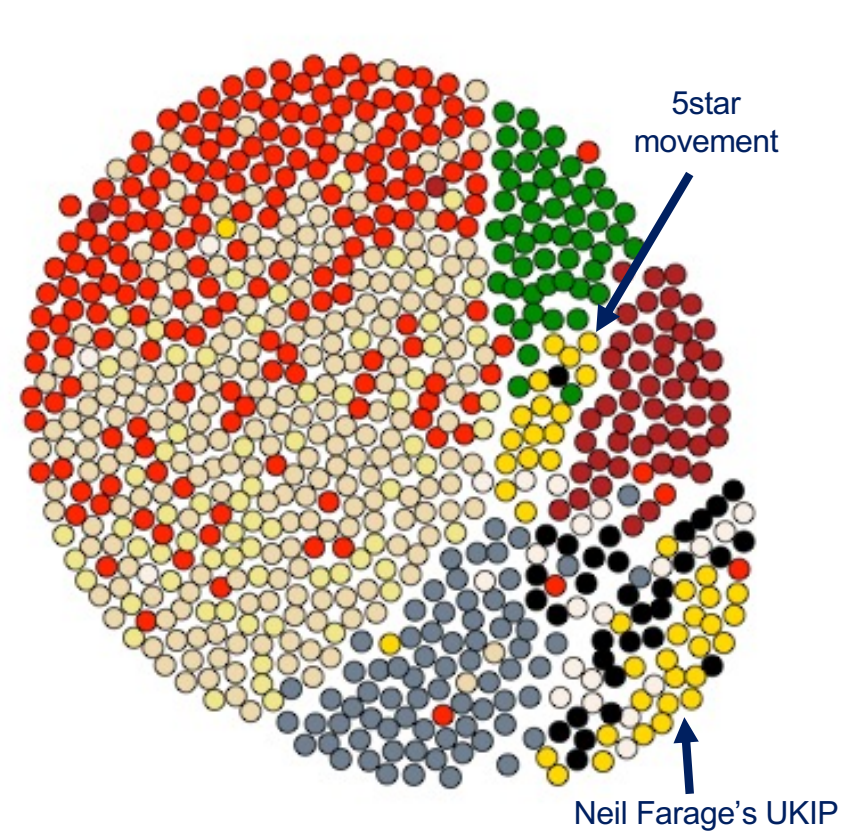
Network examples (cont'd)

April-May 2016 political network (votes at the EU parliament)

spectral clustering layout



SimRank force directed layout



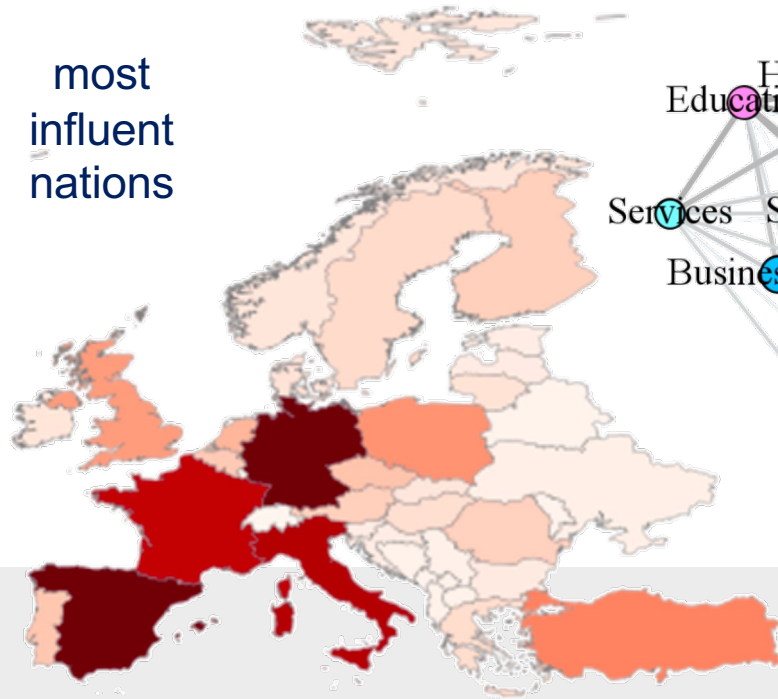


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

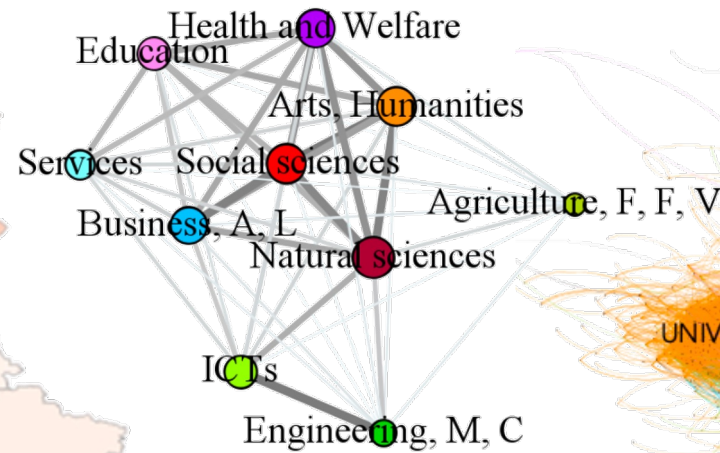
Network examples (cont'd)

Erasmus exchanges network 2019

most
influent
nations



fields of study



most influent
communities
and their top
universities

Which are the more connected
Italian institutions?

01



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



POLITECNICO
DI MILANO



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



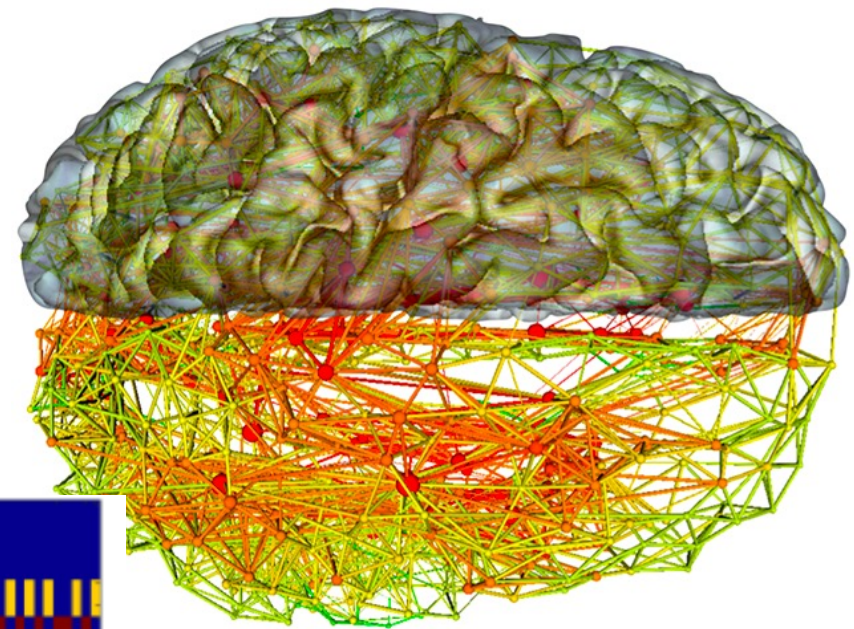
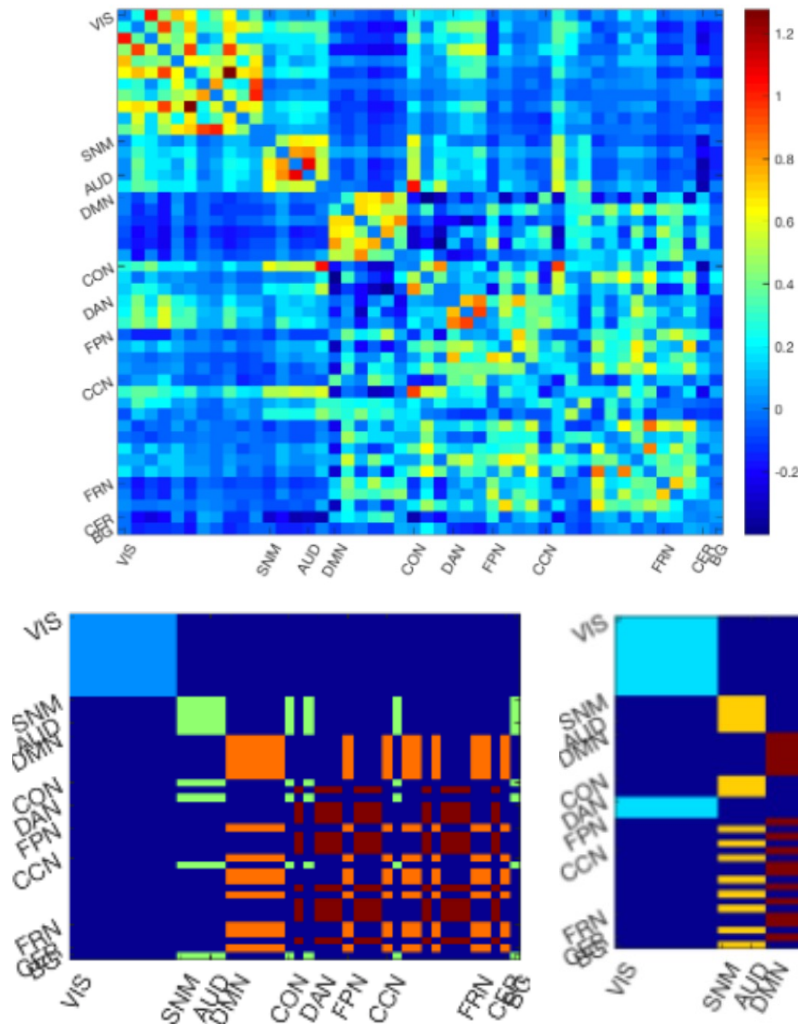
SAPIENZA
UNIVERSITÀ DI ROMA



UNIVERSITÀ
DEGLI STUDI
FIRENZE

Network examples (cont'd)

the brain network – functional connectivity





What is then network science?

Network science

From Wikipedia, the free encyclopedia

For other uses, see [Network \(disambiguation\)](#).

Network science is an academic field which studies [complex networks](#) such as [telecommunication networks](#), [computer networks](#), [biological networks](#), cognitive and [semantic networks](#), and [social networks](#), considering distinct elements or actors represented by *nodes* (or *vertices*) and the connections between the elements or actors as *links* (or *edges*). The field draws on theories and methods including [graph theory](#) from mathematics, [statistical mechanics](#) from physics, [data mining](#) and [information visualization](#) from computer science, [inferential modeling](#) from statistics, and [social structure](#) from sociology. The [United States National Research Council](#) defines network science as "the study of network representations of physical, biological, and social phenomena leading to [predictive models](#) of these phenomena."^[1]





Social network analysis

From Wikipedia, the free encyclopedia

Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory.^[1]

It characterizes networked structures in terms of *nodes* (individual actors, people, or things within the network) and the *ties*, *edges*, or *links* (relationships or interactions) that connect them. Examples of social structures commonly visualized through social network analysis include social media networks,^{[4][3]} memes spread,^[4] information circulation,^[5] friendship and acquaintance networks, business networks, knowledge networks,^{[6][7]} difficult working relationships,^[8] social networks,

Social network analysis has emerged as a key technique in modern sociology. It has also gained a significant following in anthropology, biology,^[12] demography, communication studies,^{[3][13]} economics, geography, history, information science, organizational studies,^{[6][8]} political science, public health,^{[14][7]} social psychology, development studies, sociolinguistics, and computer science^[15] and is now commonly available as a consumer tool (see the list of SNA software).^{[16][17][18][19]}





And how do we study networks?

With a **holistic** character

(the whole is greater than the sum of its parts)

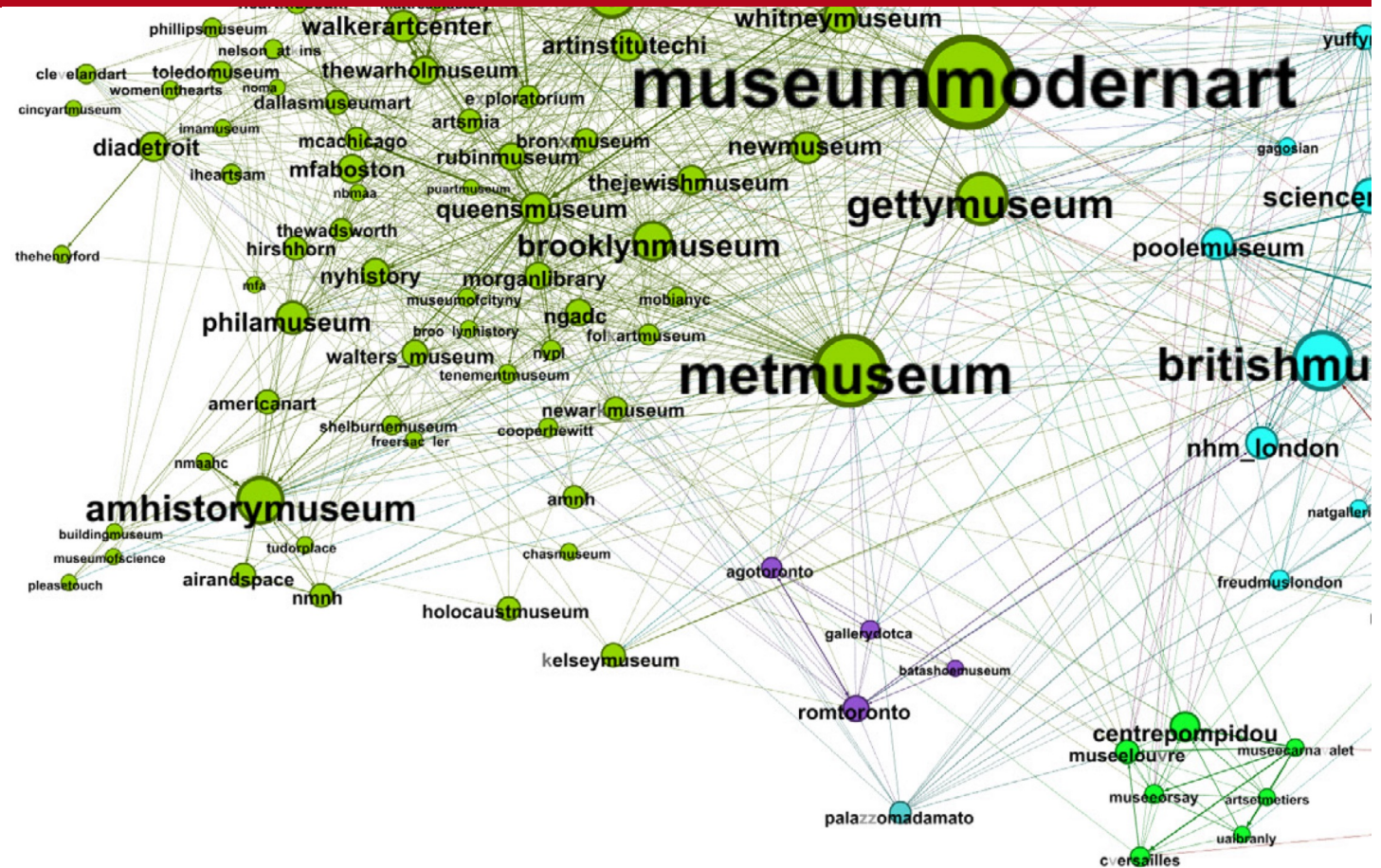
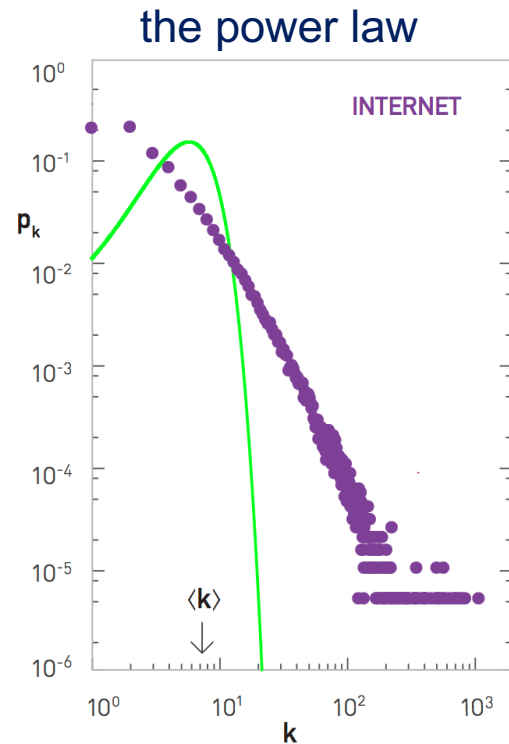
With **mathematical** rigour

The approach is

empirical (driven by concrete data),
precise (requires a proper formalism),
interdisciplinary (can be applied to several fields), and
challenging (in data size and in objectives)

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

And what do we study?

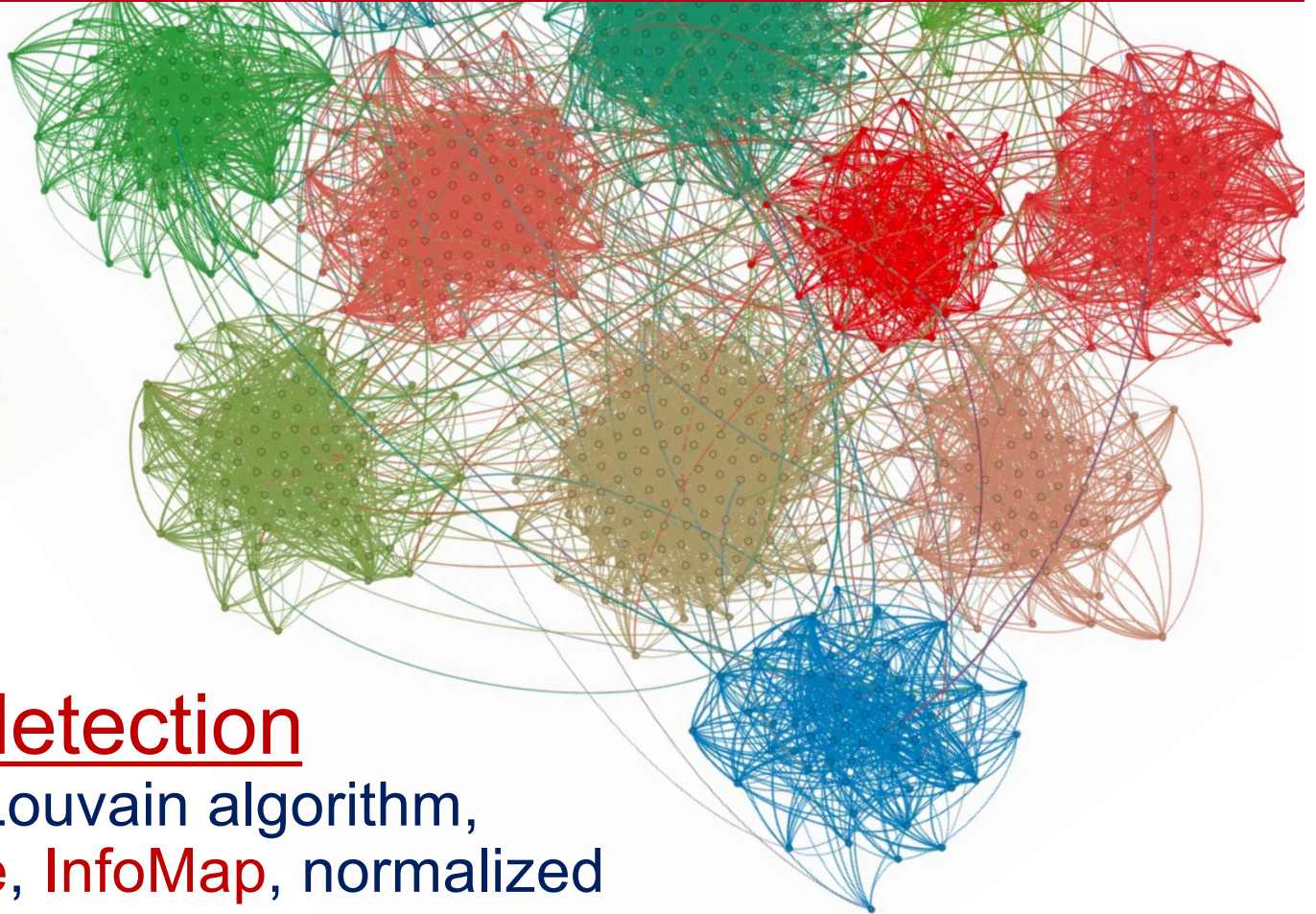
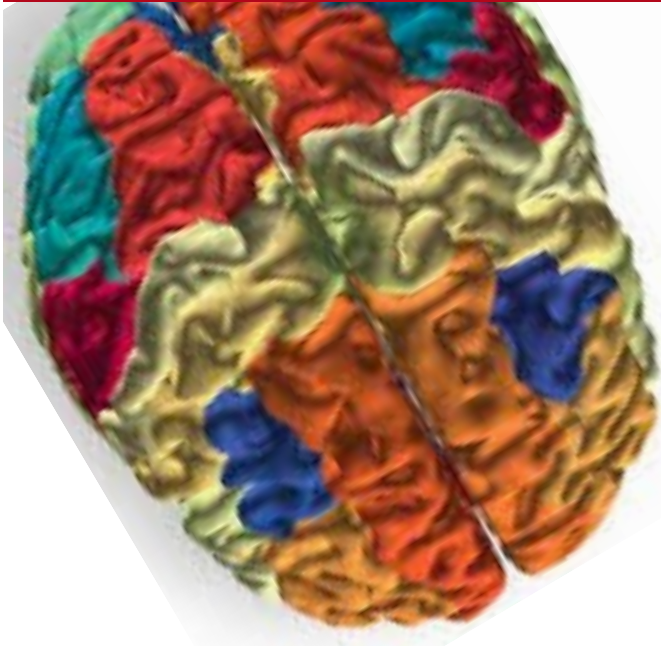


many network analytics, e.g., centrality

degree, PageRank, HITS, betweenness, etc.



And what do we study? (cont'd)



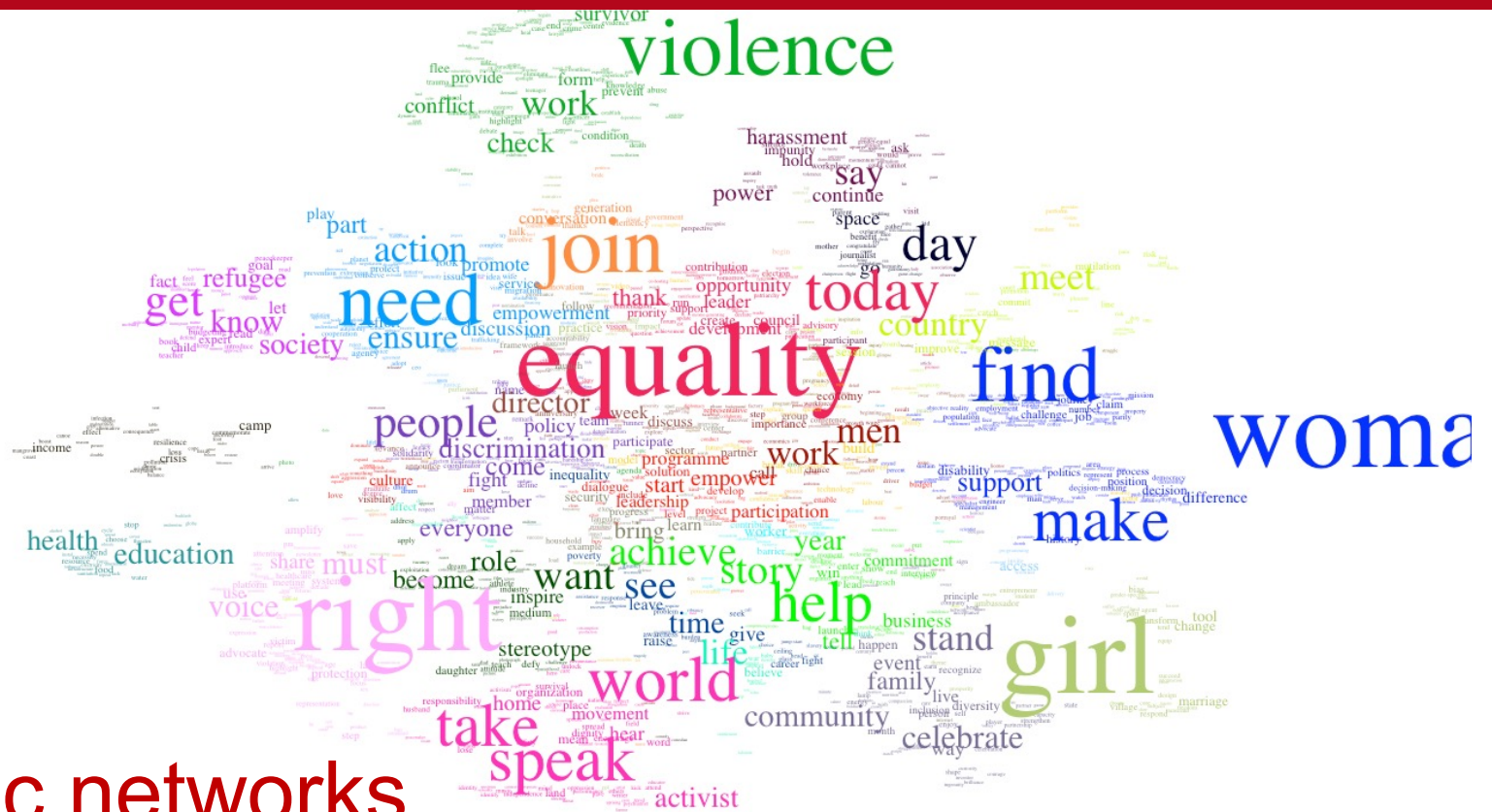
community detection

- ❑ modularity, Louvain algorithm, conductance, InfoMap, normalized mutual information, overlapping communities, BigCLAM, stochastic block models



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

And what do we study? (cont'd)



semantic networks

- network layout, data collection, sentiment analysis, BERTAgent, topic detection, latent Dirichlet allocation, variational autoencoders, BERTopic and the Transformer architecture

Project

a brief overview



What about the project?

create your own group (1 to 3 people)

choose your **dataset** (possibly **create**
your own dataset)

apply the ideas learned during the course

show that you can do **clever** things

try extracting meaningful measures/analytics
that describe an interesting aspect of your network

write good code

each **contributor** to the group should focus on
a different aspect (no everything together)

present the project in a clear and convincing way,
using clear and convincing plots

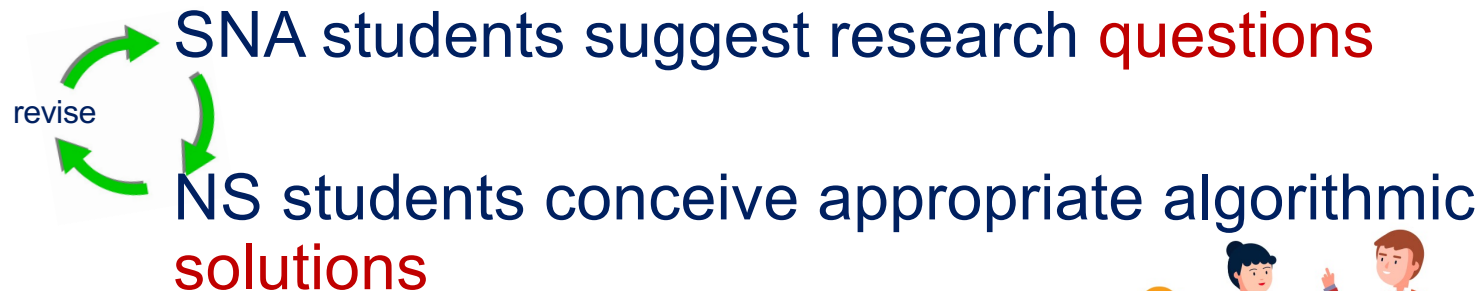




What about interdisciplinary projects?

mainly related to semantic networks

in collaboration with the twin course of
Social Network Analysis @ Communication
Strategies



in brainstorming sessions
the instructor will help/give feedback 😊





UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Interdisciplinary projects 26/27

MIME
Master's degree ICT Internet Multimedia Engineering

DI
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DSL
DIPARTIMENTO DI STUDI
LINGUISTICI E LETTERARI

Dipartimento di
Psicologia dello Sviluppo
e della Socializzazione

INTERDISCIPLINARY PROJECTS PRESENTATION

Network Science &
Social Networks Analysis

AULA MAGNA LEPSCHY

DEI - VIA GRADENIGO 6 - PADOVA

Fri Feb 12, 2027, 9:00





IP examples from past years

on Twitter

- ❑ 2019 – Evolution of Climate Change Perception on Twitter - Focusing on Greta Thunberg Impact
- ❑ 2019 – UN Women Twitter profile's reaction to the MeToo movement
- ❑ 2020 – NBA and Premier League players around #blacklivesmatter and the racial issue on Twitter
- ❑ 2020 – Republicans vs Democrats on Twitter
- ❑ 2020 – Haters gonna (make you) hate - Semantic analysis of hate during 2019 European elections
- ❑ 2021 – Sports brands and eco-sustainability
- ❑ 2022 – Sexism in Politics
- ❑ 2022 – What is the perception around the world in terms of Menstruation Stigma in 2021?
- ❑ 2022 – Cancel culture on social media - Social network analysis on famous cases of cancellation

on TikTok

- ❑ 2022 – PoliTok: How do Italian politicians use TikTok as tool to promote their political ideas and influence the young generation during the 2022 elections?

other

- ❑ 2019 – Noodles and Spaghetti - How people make pasta in eastern countries
- ❑ 2021 – Erasmus+ Programme: a social network analysis study of the 2014-2019 exchanges
- ❑ 2021 – Nationality vs. movie prestige: from the Oscars to International Film Festivals

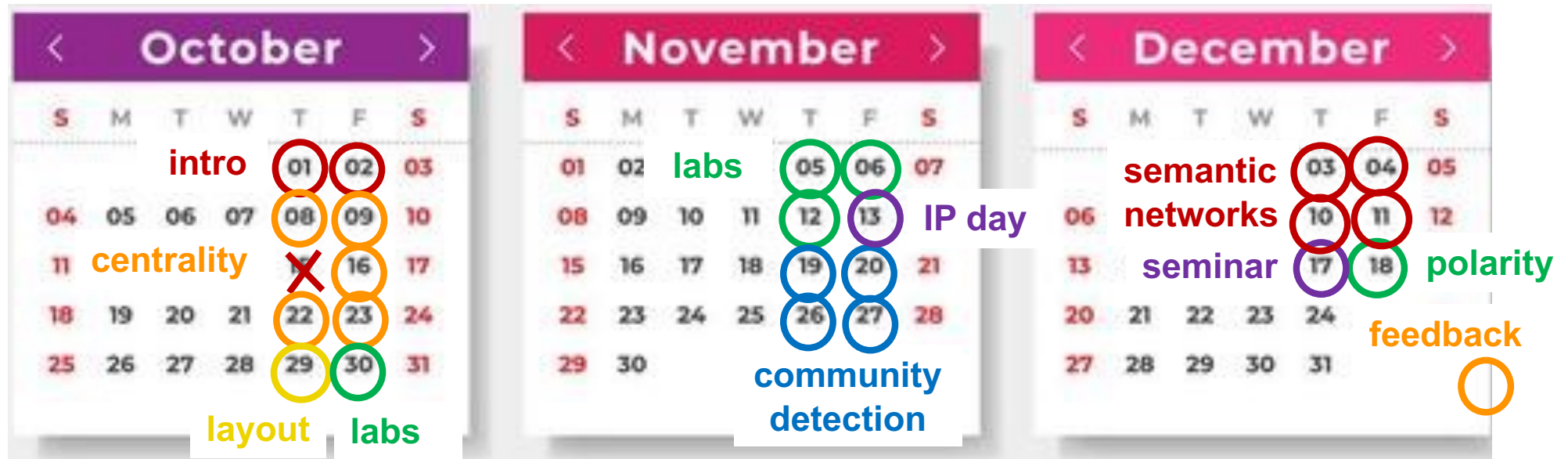
Calendar

tentative



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Calendar





❑ Misc (4 lectures)

introduction; graphs; graph layout: ForceAtlas2, Gephi, **UMAP**; robustness; homophily

❑ Centrality (5 lectures)

degree centrality, power law, preferential attachment, fitness, Bianconi-Barabasi model, scale-free regime, PageRank, convergence properties, Local PageRank, Approximate and signed PageRank, **Row-normalized PageRank**, HITS, closeness, betweenness, clustering coefficient

❑ **Community detection** (4 lectures)

modularity, Louvain algorithm, consensus clustering, **Modularity for directed and signed networks and overlapping communities**, Minimum cut criterion, spectral clustering, **InfoMap**, **Normalized mutual information**, **F1 score**, **Dice correlation**, **BigCLAM**, **stochastic block models**, Dendrograms, Girvan-Newman, **HDBSCAN**

❑ **Semantic networks** (3 lectures)

Reddit, cleaning steps: **spaCy**, **LIWC**, **BERTAgent**, **semantic networks**, **TF-IDF**, **modularity**, **latent Dirichlet analysis**, **variational autoencoders**, **BERTopic**, **performance comparison**

❑ Python labs (4 lectures)

❑ IP projects (1+1 lectures)



- ☐ Enrol @ stem.elearning.unipd.it 😊
- ☐ Have a laptop available
- ☐ Ensure you know **Python**'s basics
- ☐ Activate a **Google account** (with the @unipd.it email)
 - Google **Drive**
 - Google **CoLab**
- ☐ Activate a **Reddit account** (using Google's account)
 - Reddit **apps** <https://www.reddit.com/prefs/apps>
- ☐ Install **Gephi** on your laptop <https://gephi.org/>
- ☐ Review everything you know about **deep learning** and/or **optimization**
- ☐ Organize yourselves into **working groups** (max 3 people)



Graphs

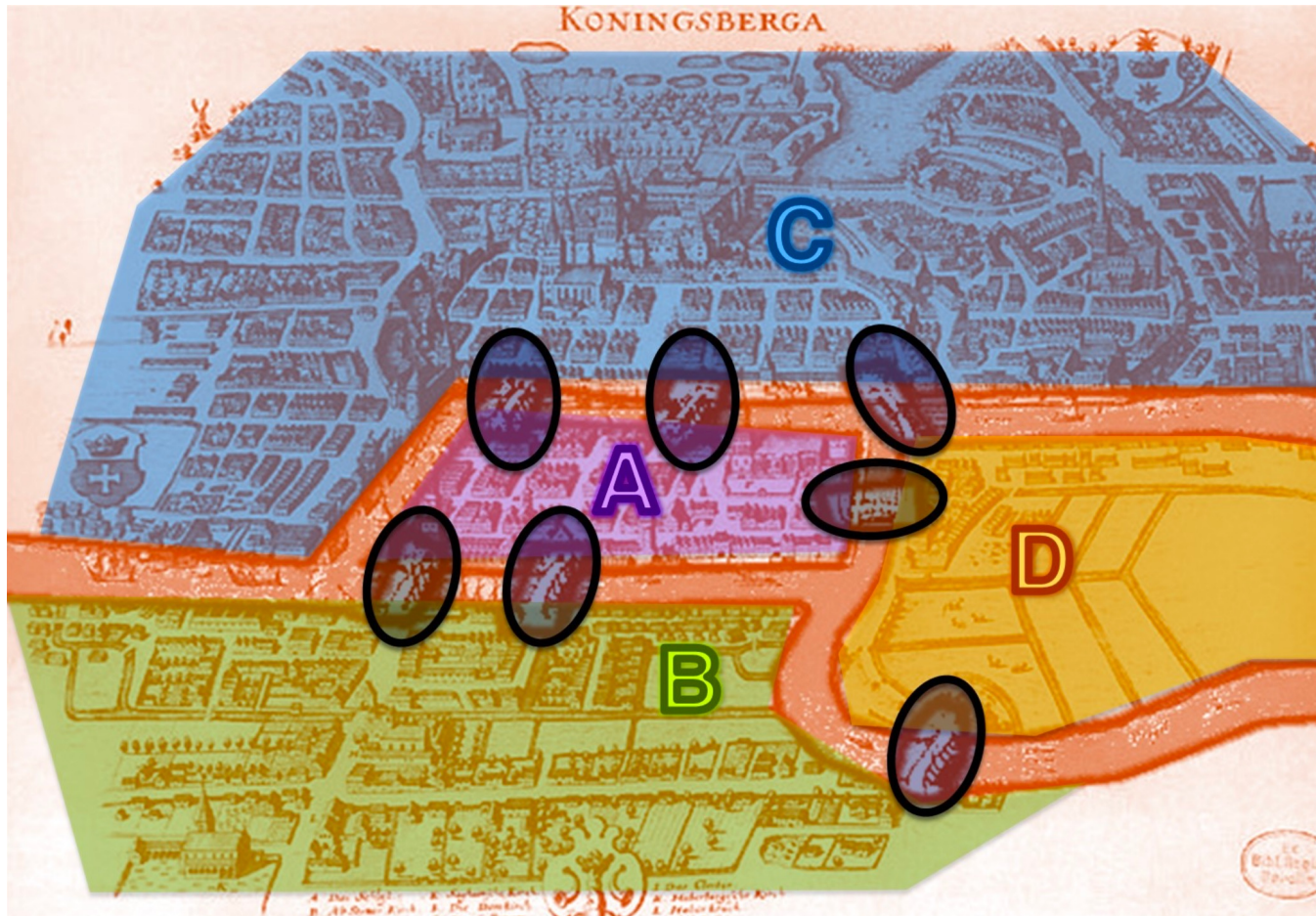
an introduction



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

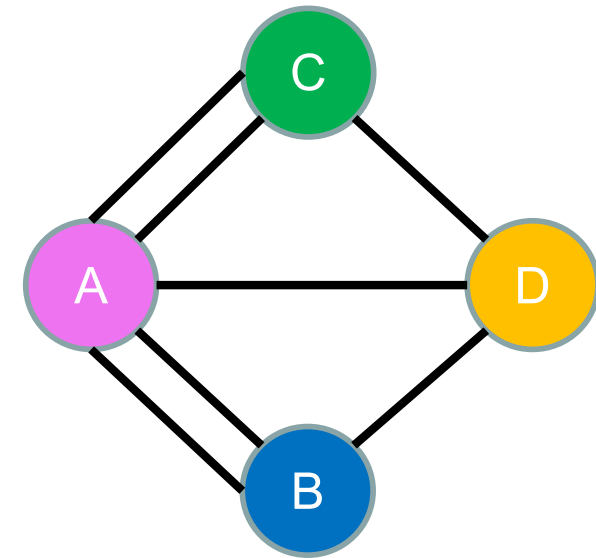
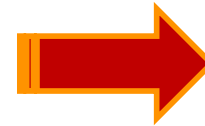
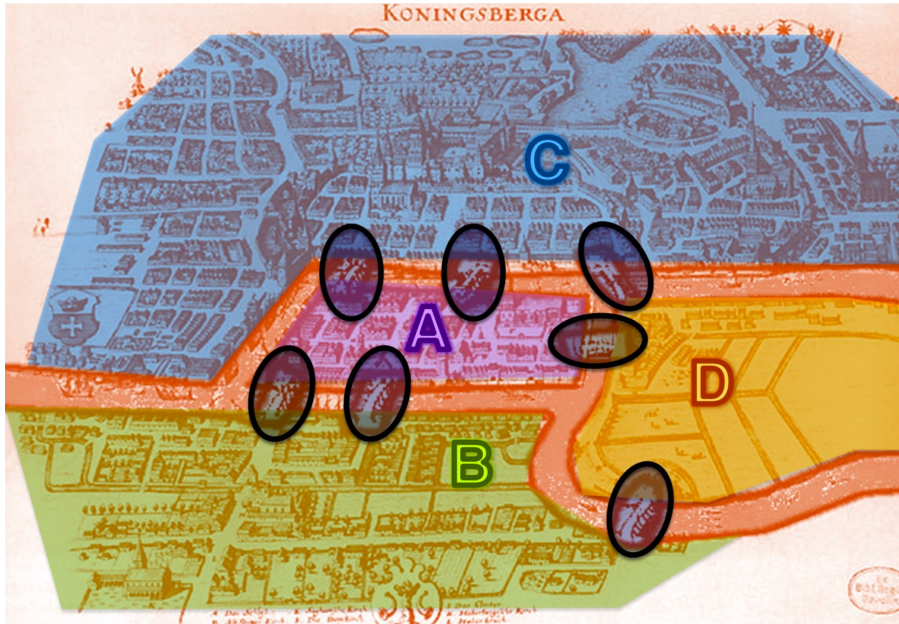
Euler and the 7 bridges of Königsberg

(Prussia, 1736) today Kaliningrad



How to walk through the city by crossing each bridge only once?

Networks as graphs



Graph $G(\mathcal{V}, \mathcal{E})$: network

□ Vertices (set $\mathcal{V}$): nodes, people, concepts

□ Edges (set $\mathcal{E}$): links, relations, associations

↑
mathematics

↑
technology

↑
social
psychology

↑
social
cognition



Directed versus undirected

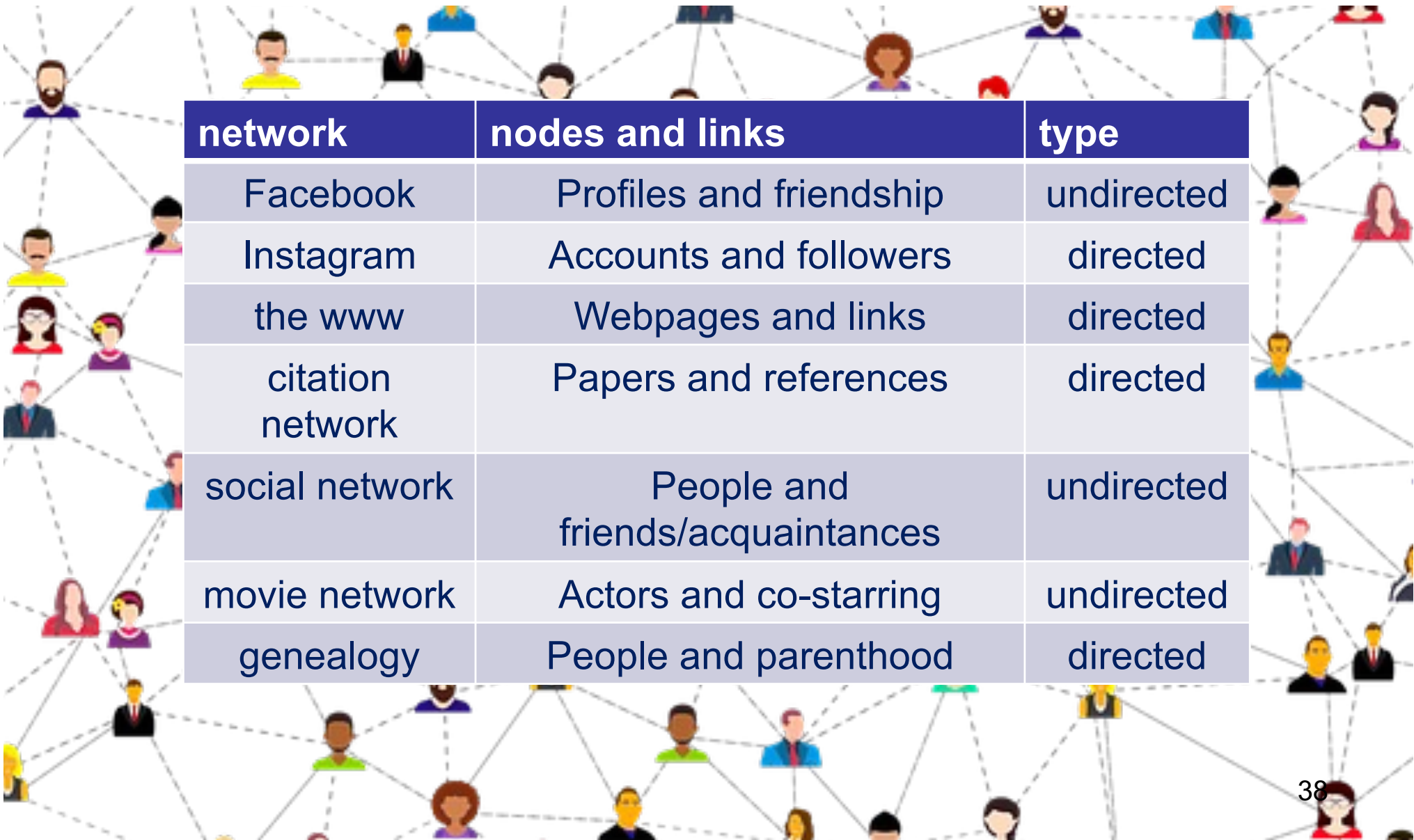
- ❑ A connection relationship can have a privileged direction or can be mutual
- ❑ Either a **directed** or an **undirected** link



- ❑ If the network has only (un)directed links, it is also called itself (un)directed network
- ❑ Certain networks can have both types



Some examples

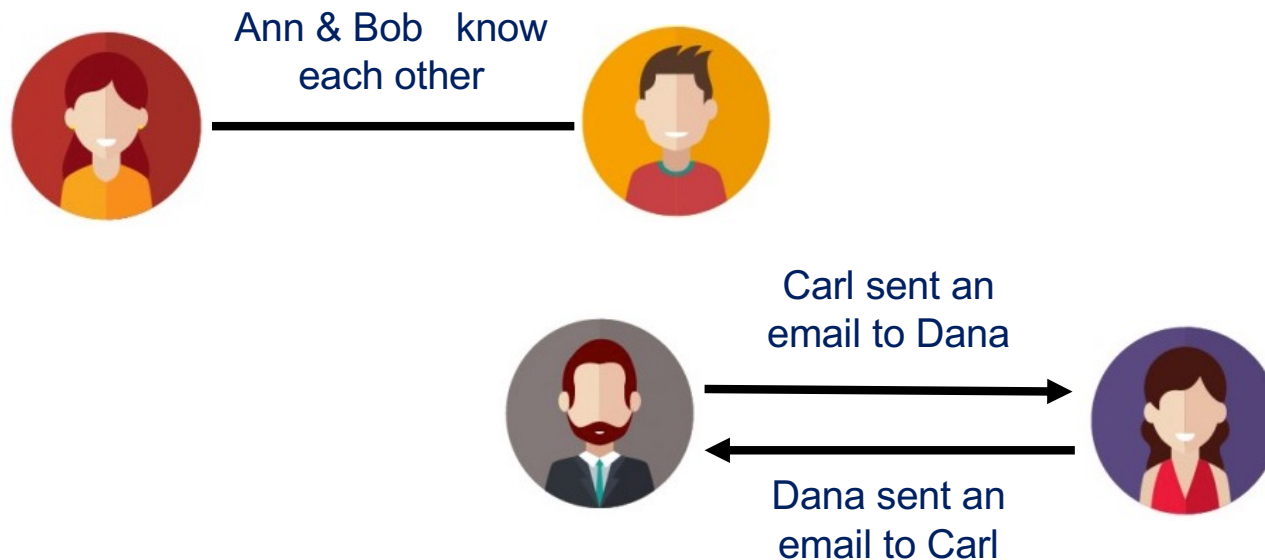


network	nodes and links	type
Facebook	Profiles and friendship	undirected
Instagram	Accounts and followers	directed
the www	Webpages and links	directed
citation network	Papers and references	directed
social network	People and friends/acquaintances	undirected
movie network	Actors and co-starring	undirected
genealogy	People and parenthood	directed



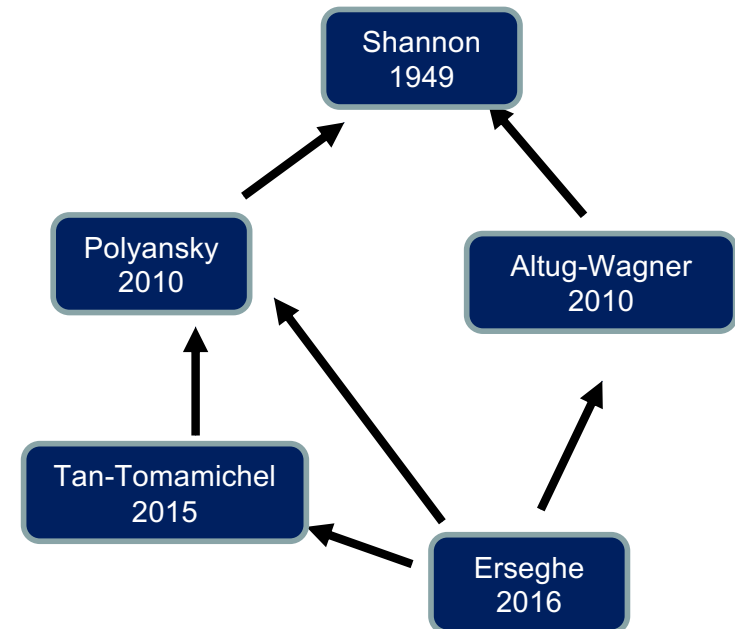
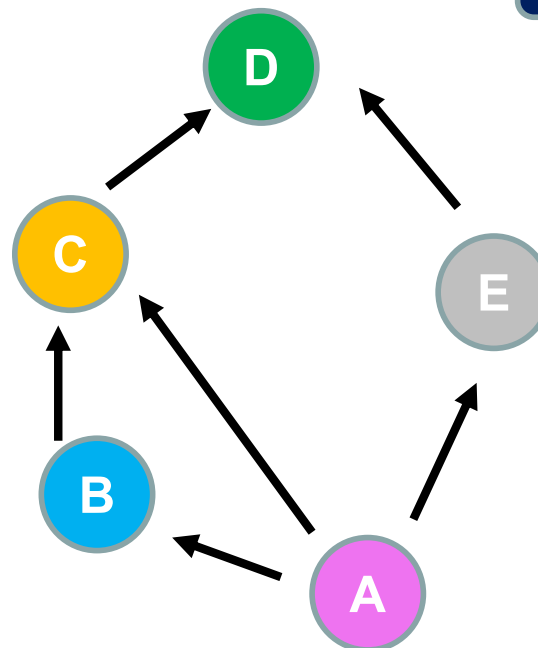
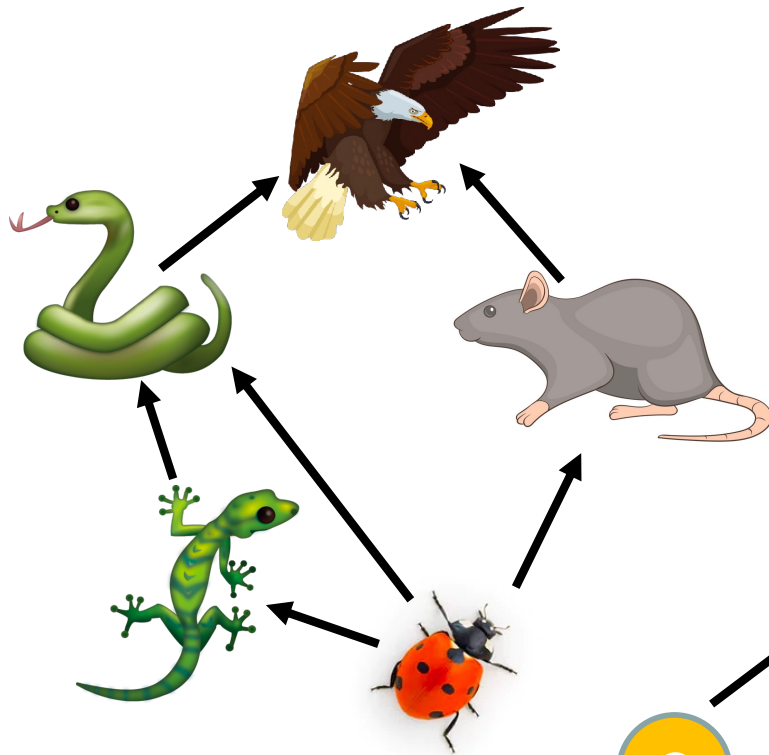
Directed versus undirected

- ❑ At first glance **undirected** → **directed** by duplicating links, but not necessarily quite the same though





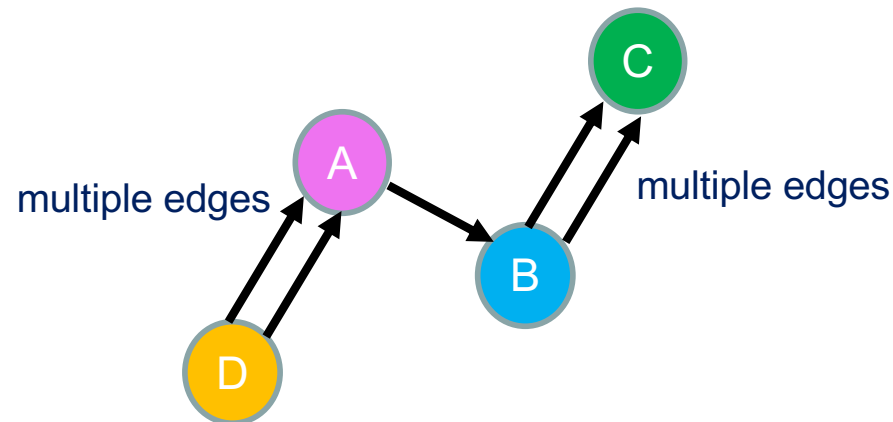
Generality of representation



Weighted graphs

and adjacency matrix

- ❑ Multi-graphs (or pseudo-graphs)
Some network representations require **multiple** links (e.g., number of citations from one author to another)

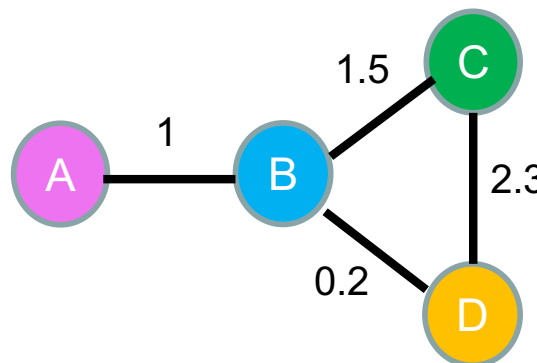




□ Weighted graph

Usually a **weight** w_{ij} is associated to a link $(i,j) \in \mathcal{E}$, e.g., to underline that the links are not identical (strong/weak relationships)

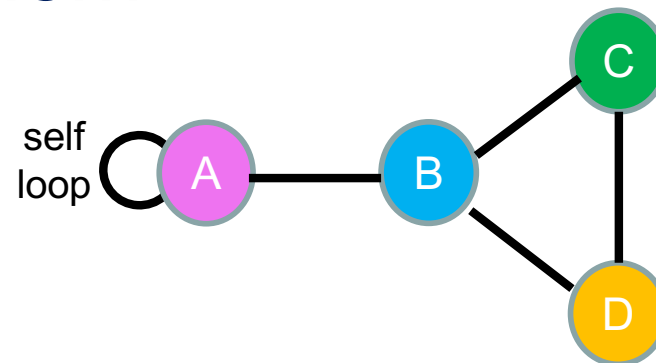
Can be seen as a generalization of multi-graphs (weight = # of links)



- ❑ In many networks nodes do not interact with themselves

if $i \in \mathcal{V}$ then $(i,i) \notin \mathcal{E}$

- ❑ To account for self-interactions, we add **loops** to represent them



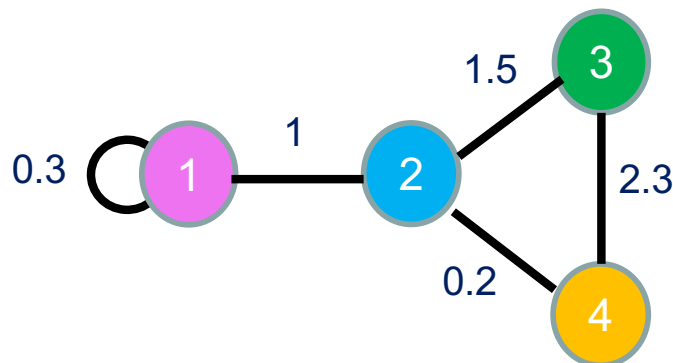


□ An adjacency matrix $A = [a_{ij}]$ associated to graph $\mathcal{G} (\mathcal{V}, \mathcal{E})$ has

entries $a_{ij} = 0$ for $(i,j) \notin \mathcal{E}$ (not a connection)

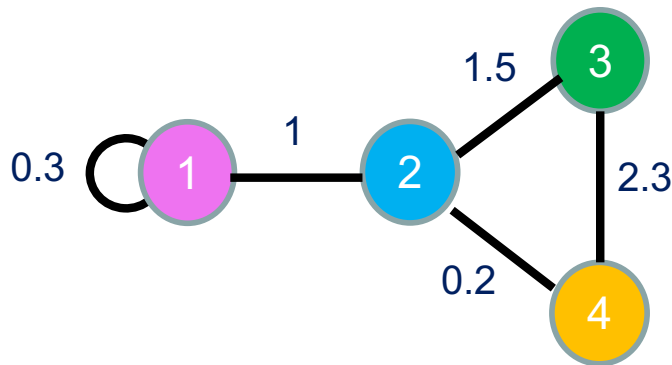
if nodes i and j are **connected** then $a_{ij} \neq 0$

in **plain** graphs $a_{ij} = 1$ for $(i,j) \in \mathcal{E}$



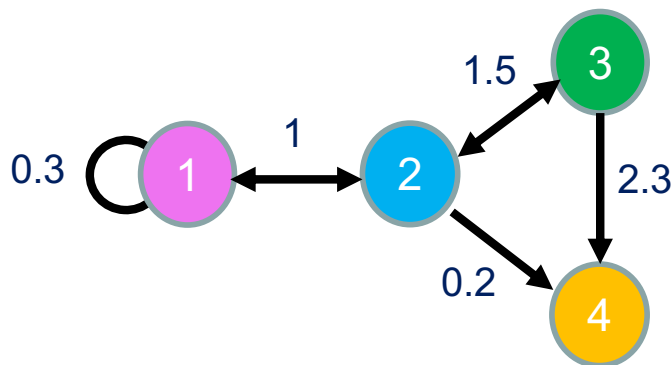
$$A = \begin{bmatrix} 0.3 & 1 & 0 & 0 \\ 1 & 0 & 1.5 & 0.2 \\ 0 & 1.5 & 0 & 2.3 \\ 0 & 0.2 & 2.3 & 0 \end{bmatrix}$$

□ Undirected graph = **symmetric** matrix



$$A = \begin{bmatrix} 0.3 & 1 & 0 & 0 \\ 1 & 0 & 1.5 & 0.2 \\ 0 & 1.5 & 0 & 2.3 \\ 0 & 0.2 & 2.3 & 0 \end{bmatrix}$$

□ Directed graph = **asymmetric** matrix

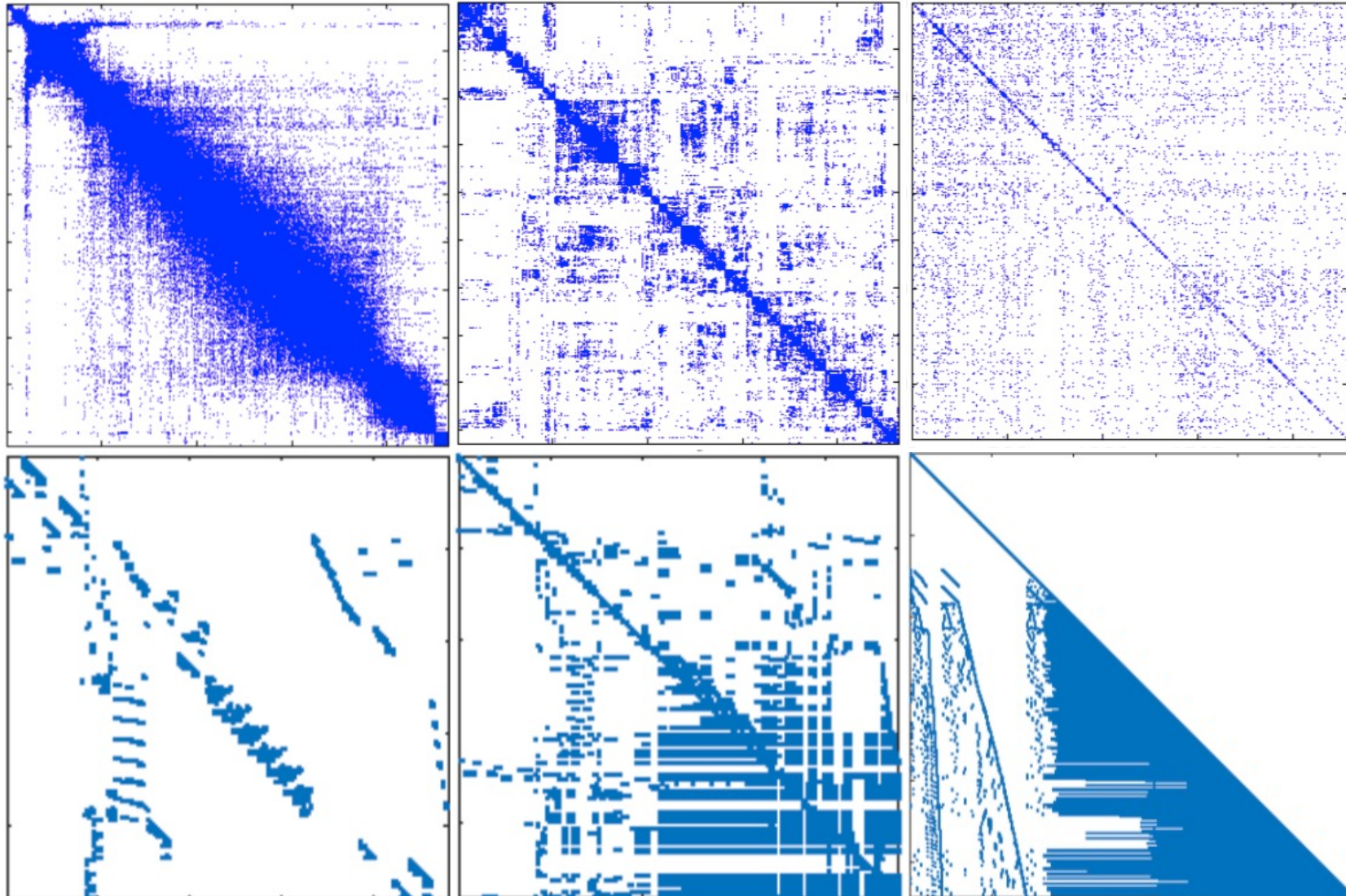


$$A = \begin{bmatrix} 0.3 & 1 & 0 & 0 \\ 1 & 0 & 1.5 & 0 \\ 0 & 1.5 & 0 & 0 \\ 0 & 0.2 & 2.3 & 0 \end{bmatrix}$$



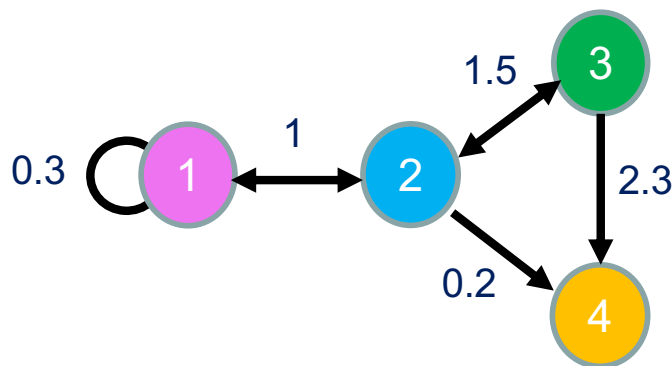
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Symmetries





- The weight a_{ij} is associated to
- i th row
 - j th column
 - directed edge $j \rightarrow i$ starting from node j and leading to node i



$$A = \begin{bmatrix} 0.3 & 1 & 0 & 0 \\ 1 & 0 & 1.5 & 0 \\ 0 & 1.5 & 0 & 0 \\ 0 & 0.2 & 2.3 & 0 \end{bmatrix}$$

Labels for the matrix elements:

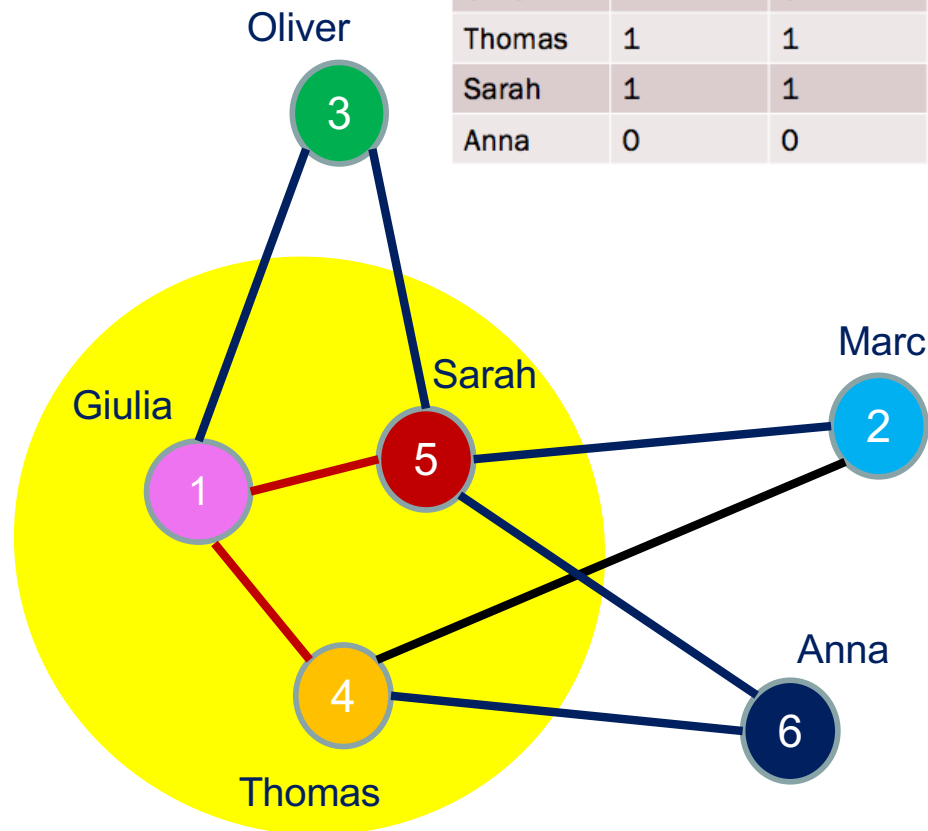
- a_{24} (top right, red)
- a_{34} (middle right, red)
- a_{42} (bottom left, red)
- a_{43} (bottom middle, red)



Question

which of these representations do you like best?

	Giulia	Marc	Oliver	Thomas	Sarah	Anna
Giulia	X					
Marc	0	X				
Oliver	1	0	X			
Thomas	1	1	0	X		
Sarah	1	1	1	0	X	
Anna	0	0	0	1	1	x



edge list

1 → 3
1 → 4
1 → 5
2 → 4
2 → 5
3 → 5
4 → 6
5 → 6

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

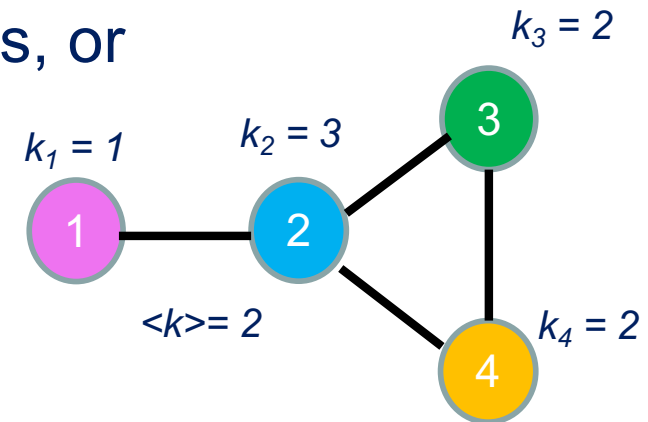
Node degree

in directed and undirected networks



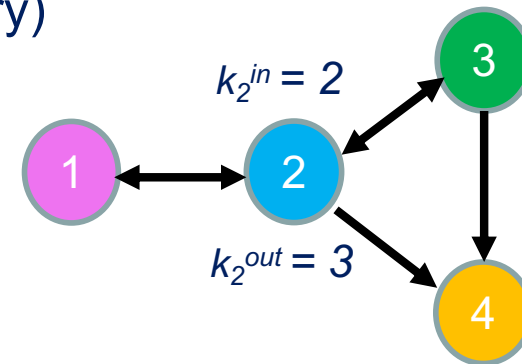
- ❑ The **degree** k_i of node i in an **undirected** networks is

the # of links i has to other nodes, or
the # of nodes i is linked to



- ❑ The # of nodes is $N = |\mathcal{V}|$
- ❑ The # of edges is $L = |\mathcal{E}| = \frac{1}{2} \sum_i k_i$
- ❑ The **average** degree is $\langle k \rangle = \sum_i k_i / N = 2L / N$

- For **directed** networks we distinguish between
 in-degree k_i^{in} = # of entering links
 out-degree k_i^{out} = # of exiting links
 (undirected: $k_i^{in} = k_i^{out}$ due to the symmetry)



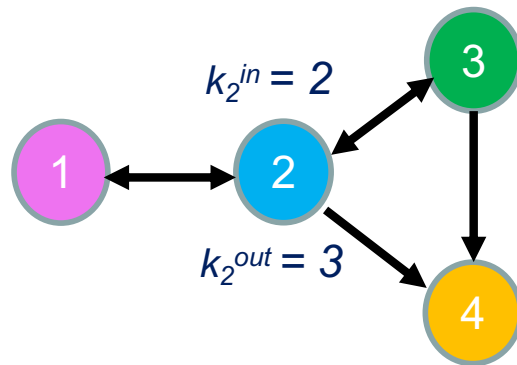
- The # of links is $L = \sum_i k_i^{in} = \sum_i k_i^{out}$
 (no need for factor $\frac{1}{2}$)

The average # of links is $\langle k \rangle = L / N$



Adjacency matrix and degree

- The in (out) degree can be obtained by **summing** the adjacency matrix over rows (columns)



no self-loops in this case!!!

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

$k_2^{in} = 2$
 $k_2^{out} = 3$

- A few useful **linear algebra** expressions

$$\mathbf{k}^{in} = \mathbf{A} \cdot \mathbf{1}$$

$$\mathbf{k}^{out} = \mathbf{A}^T \cdot \mathbf{1} = (\mathbf{1}^T \cdot \mathbf{A})^T$$

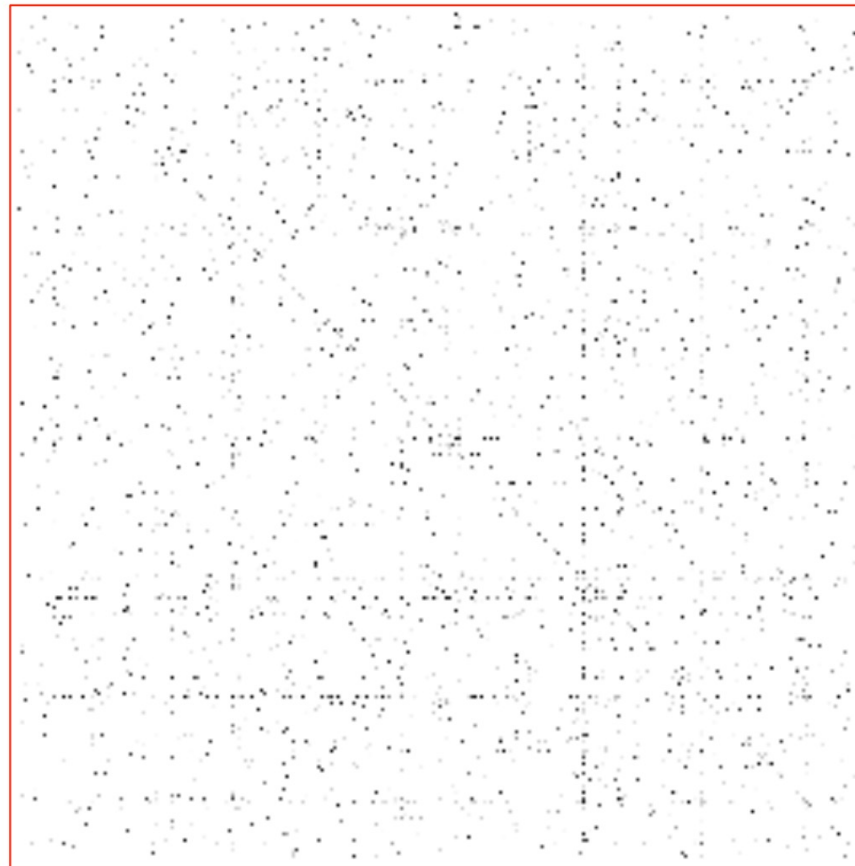
valid also for weighted graphs



□ The adjacency matrix is typically sparse

good for tractability !

$A =$



protein
interaction
network



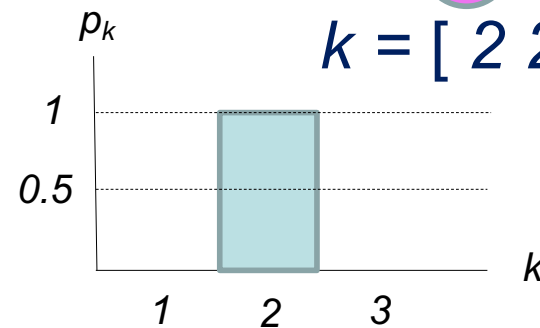
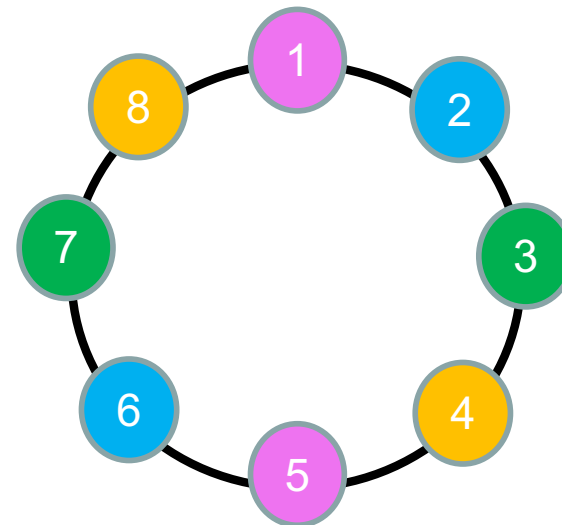
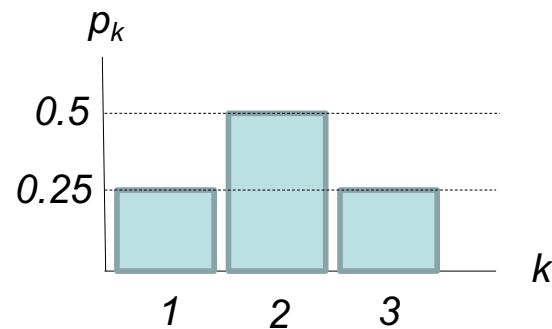
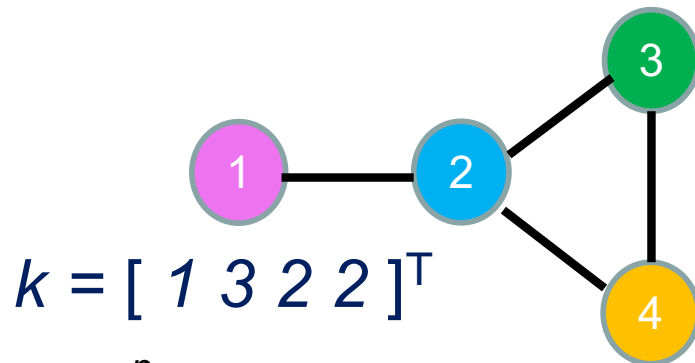
Real networks are sparse

- ❑ The maximum degree is $\langle k \rangle_{\max} = N - 1$
- ❑ In real networks $\langle k \rangle \ll N - 1$

network	type	N	L	$\langle k \rangle$
www	directed	3.2×10^5	1.5×10^6	4.60
Protein	directed	1870	4470	2.39
Co-authorships	undirected	23133	93439	8.08
Movie actors	undirected	7×10^5	29×10^6	83.7

□ Degree distribution p_k , a probability distr.

p_k is the fraction of nodes that have degree exactly equal to k (i.e., # of nodes with that degree / N)



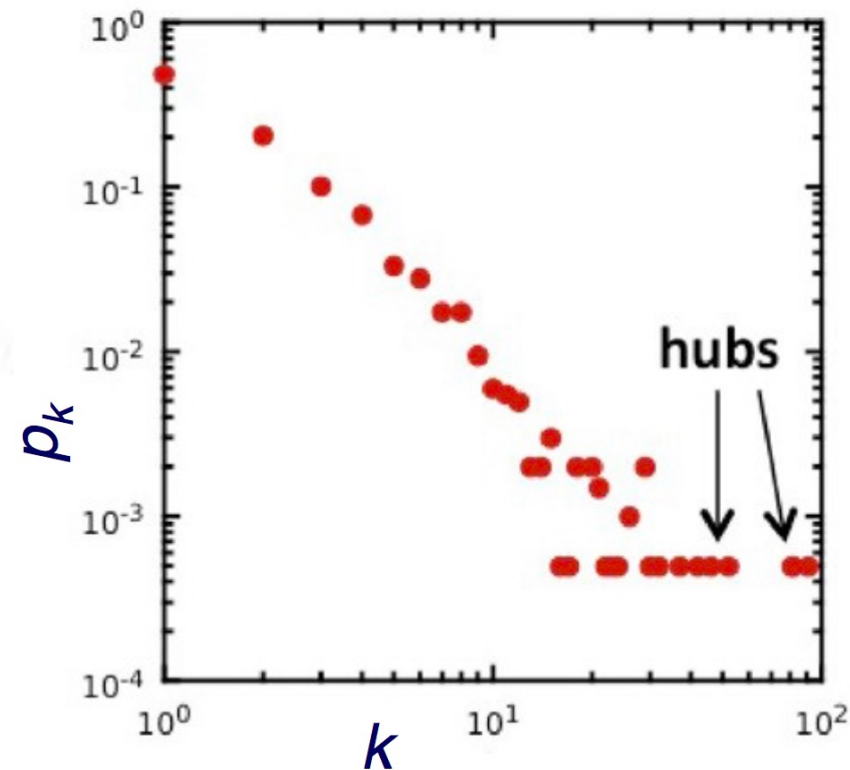
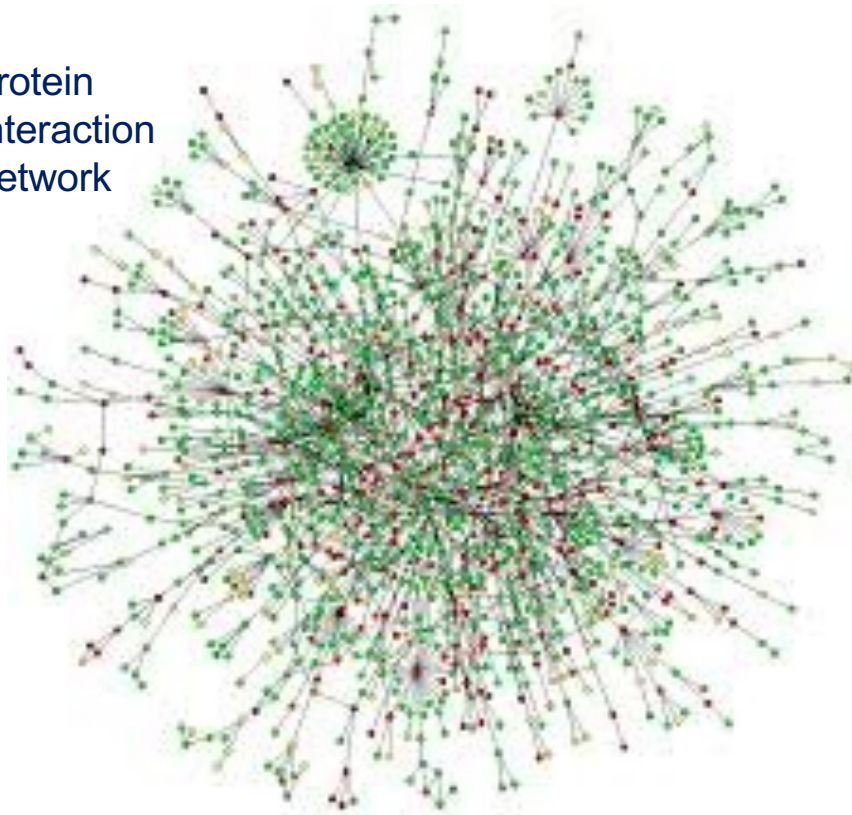


Degree distribution

- In real world (large) networks, degree distribution is typically **heavy-tailed**

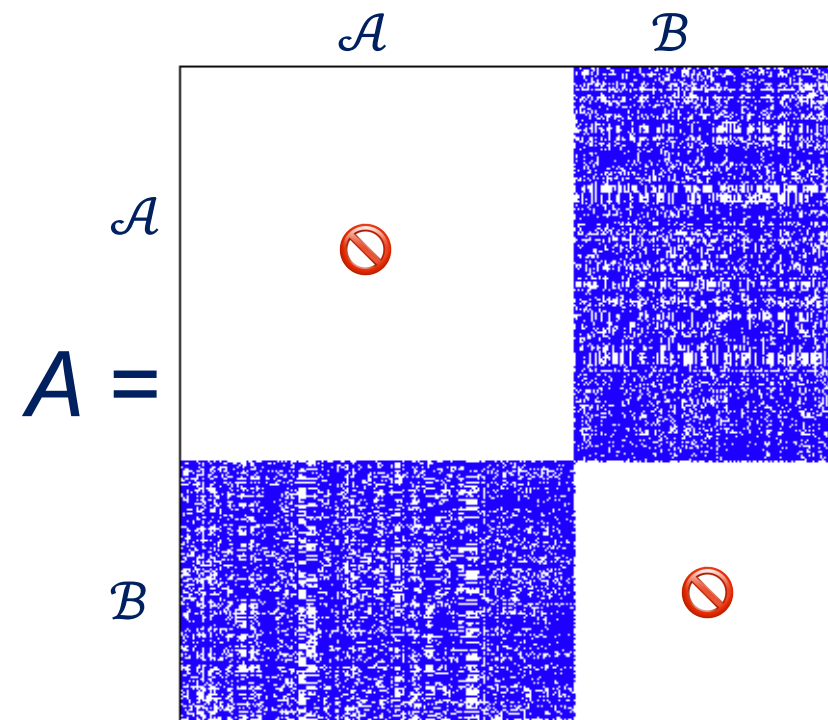
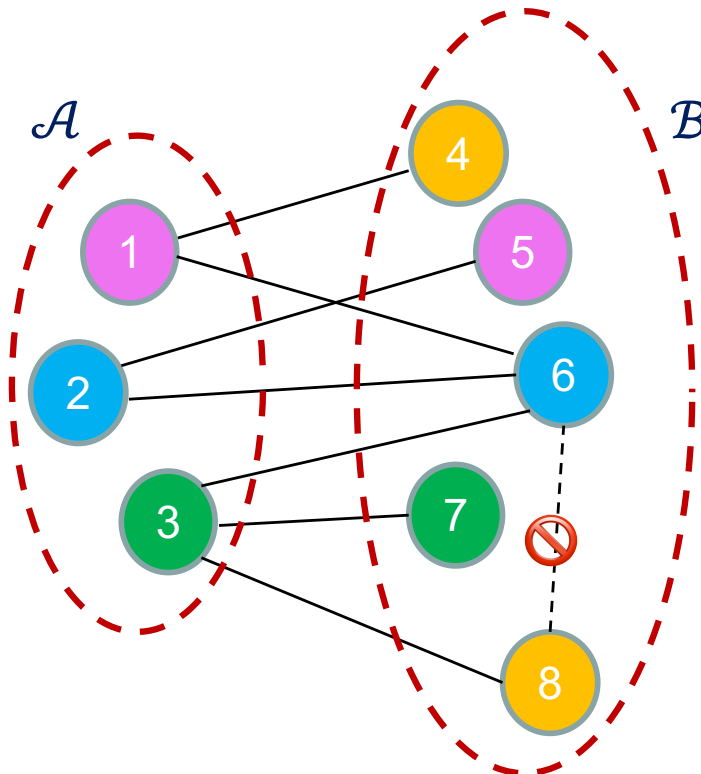
nodes with high degree = **hubs**

protein
interaction
network

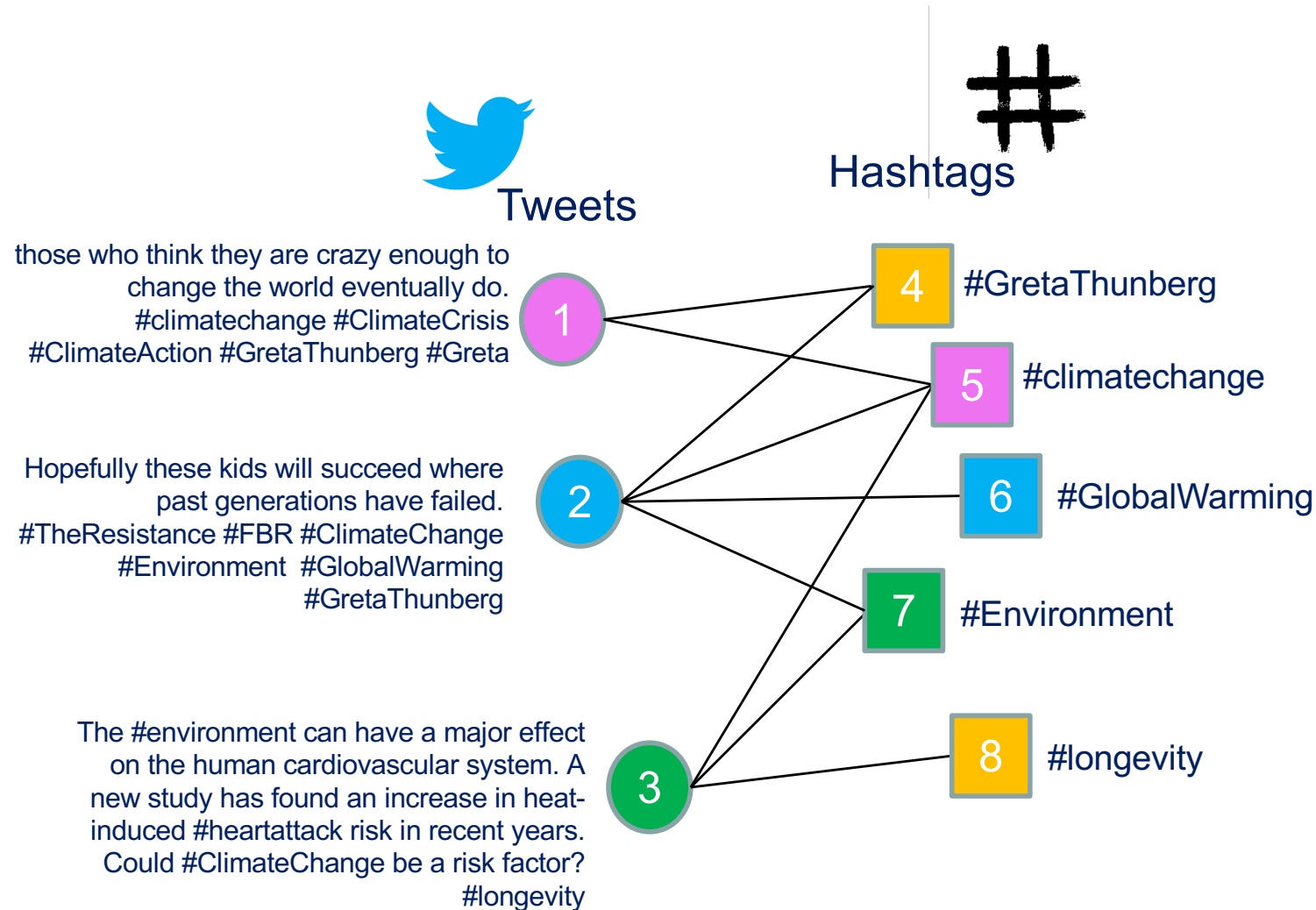


Bipartite graphs

- Connections are available only between the groups $\mathcal{A}$ and $\mathcal{B}$



Bipartite graph example





- ❑ Bipartite graphs represent **memberships**/relationships, e.g., groups ($\mathcal{A}$) to which people ($\mathcal{B}$) belong

examples: movies/actors, classes/students, conferences/authors

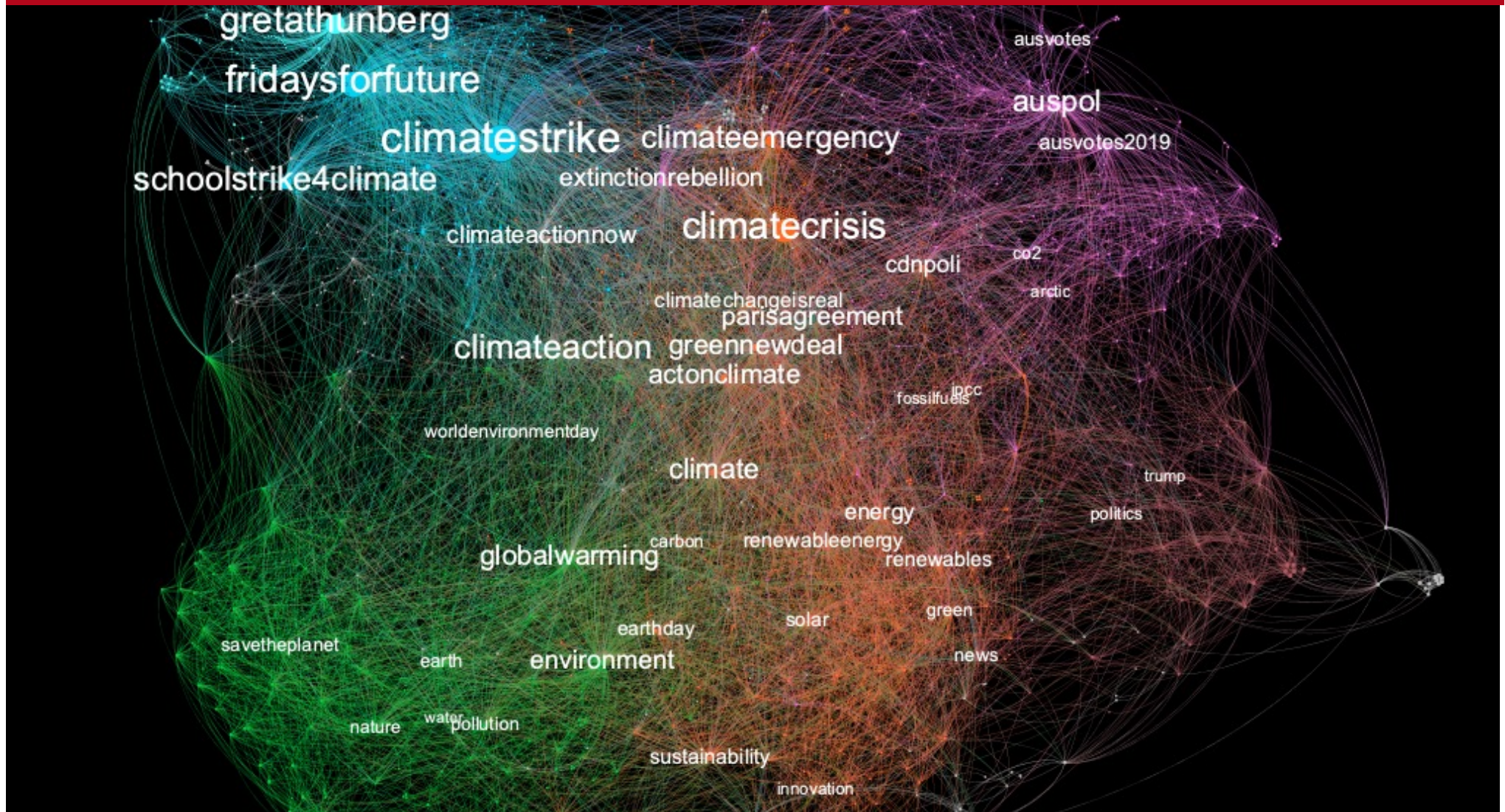
- ❑ We can build separate networks (**projections**) for $\mathcal{A}$ and $\mathcal{B}$ (sometimes this is useful)

in the **movies/actors** example being linked can be interpreted in two ways: “**actors in the same movie**” (projection on $\mathcal{B}$), or “**movies sharing the same actor**” (projection on $\mathcal{A}$)



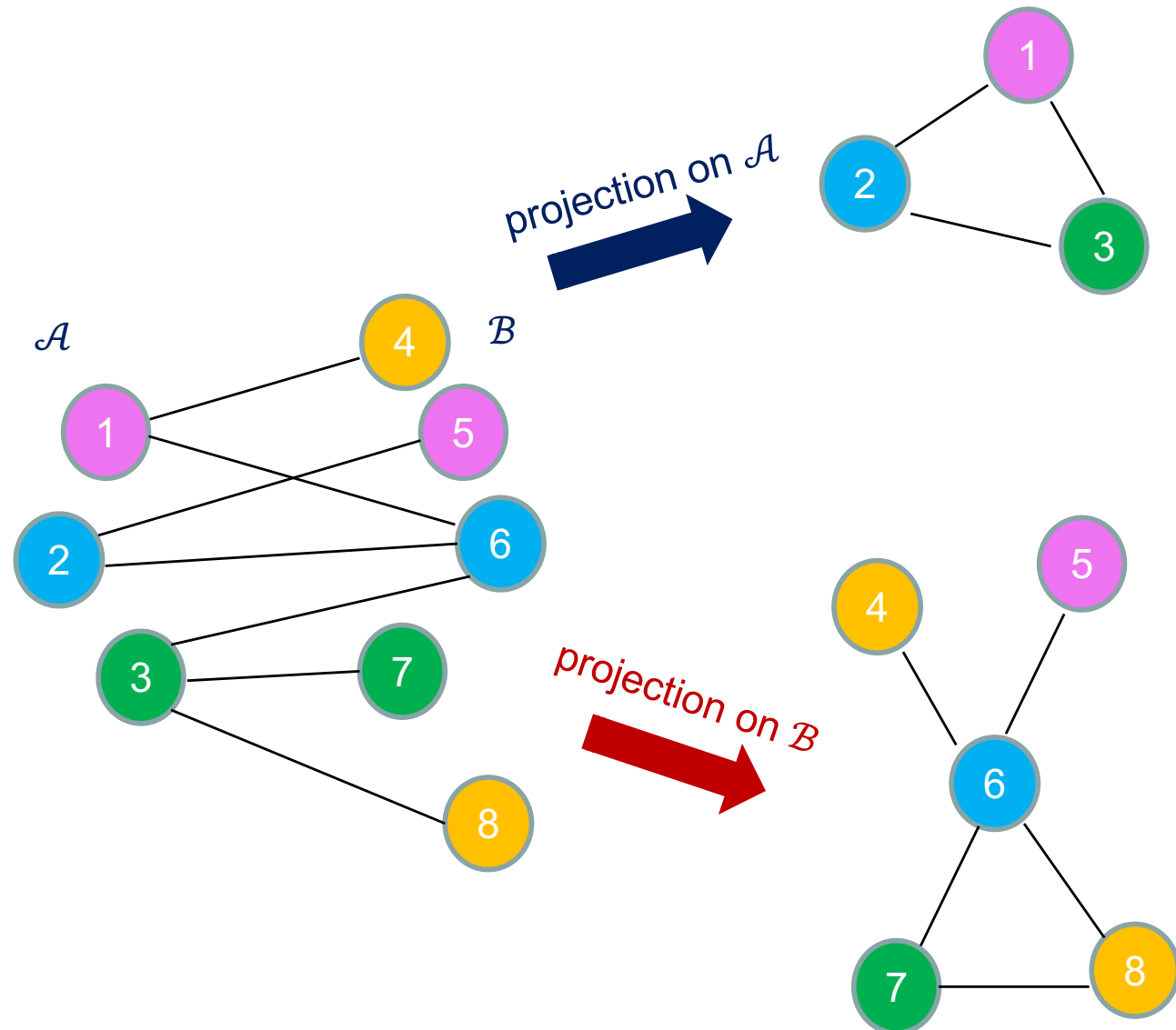
Projection on a semantic network

#hashtags that appear in the same tweet are linked





Abstract example



Nodes are linked
if they have a
common
neighbour in $\mathcal{B}$

PS: we say that
nodes i and j have a
common neighbour k
if both i and j are
connected to k

Nodes are linked
if they have a
common
neighbour in $\mathcal{A}$

- The two **projections on $\mathcal{A}$ and $\mathcal{B}$** can be obtained by inspecting the squared adjacency matrix A^2

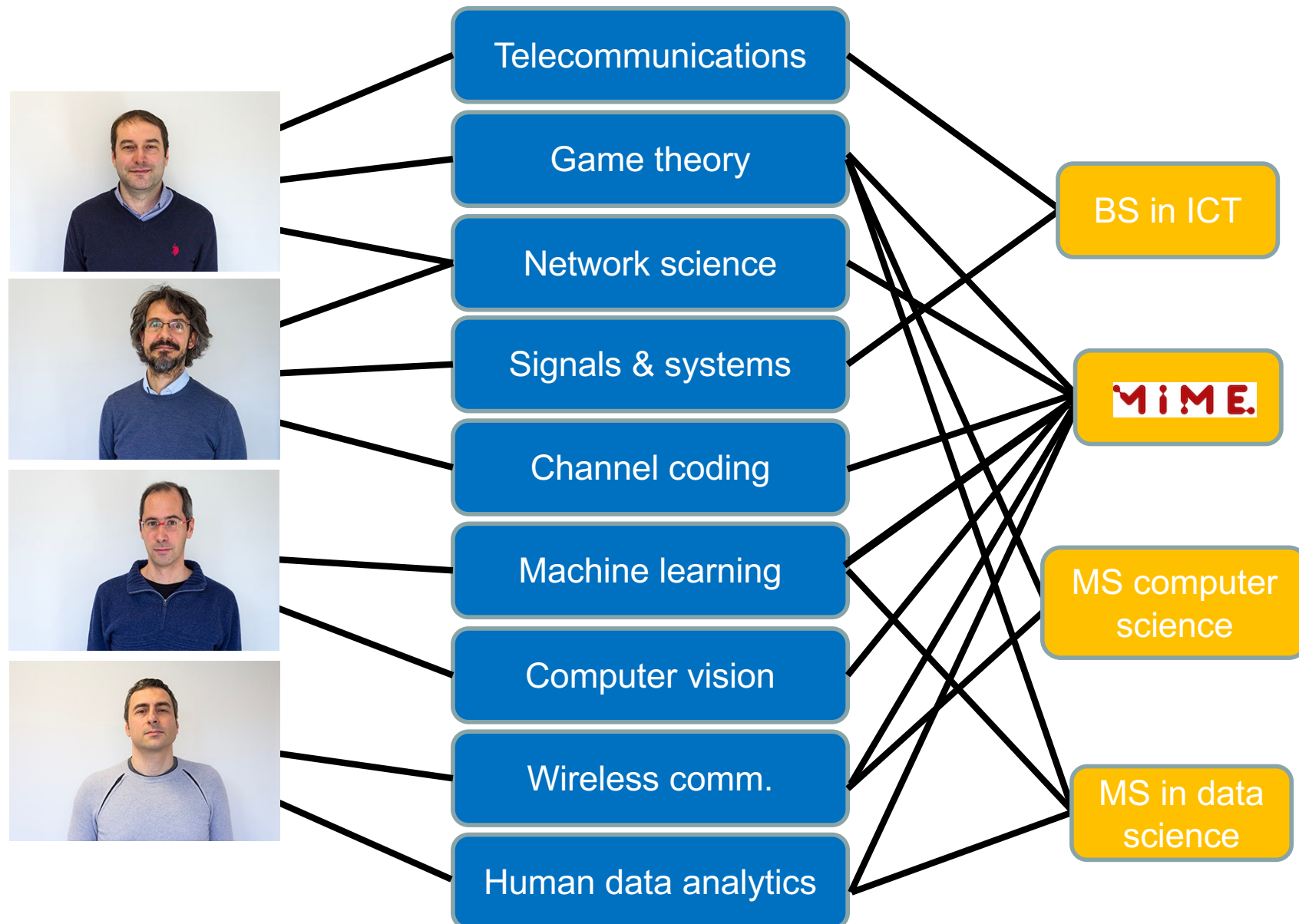
$$A^2 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 3 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} A_1 \\ A_2 \end{matrix}$$

of common neighbors of $i=6$ and $j=5$

of neighbors of $i=6$



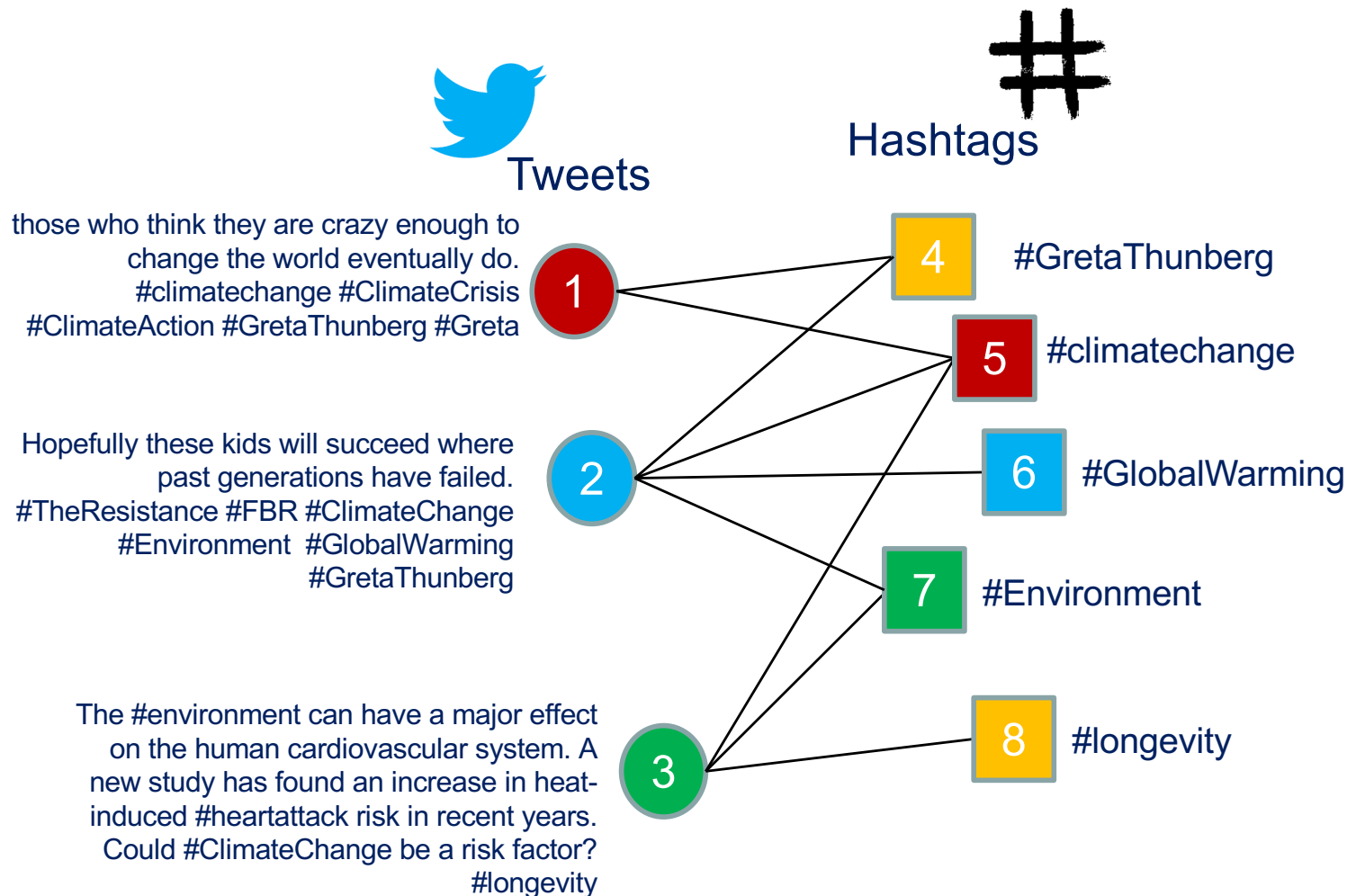
Tri-partite graphs



Semantic networks



Conceptual picture of a semantic network on Twitter





Probability matrices linking words to documents

number of occurrences
of words in documents

$N_{wd} =$

0	1	1	1	#globalwarming
1	1	1	1	#climatechange
1	0	1	0	#climateaction
0	0	1	1	#gretathunberg
1	0	1	1	#environment

probability of words
given a documents

$P_{w|d} =$

0	$\frac{1}{2}$	$\frac{1}{5}$	$\frac{1}{4}$
$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{5}$	$\frac{1}{4}$
$\frac{1}{3}$	0	$\frac{1}{5}$	0
0	0	$\frac{1}{5}$	$\frac{1}{4}$
$\frac{1}{3}$	0	$\frac{1}{5}$	$\frac{1}{4}$

we identify a
document
probability

$p_d = \begin{cases} \frac{1}{D} & \text{equally likely} \\ \frac{n_d}{\sum_d n_d} & \text{custom} \end{cases}$

we capture the **statistical**
properties by
normalizing by columns



Probability matrices projecting to words or documents

bipartite
network

joint probability of words
and documents

$$\mathbf{P}_{wd} = \mathbf{P}_{w|d} \text{diag}(\mathbf{p}_d)$$

0	1/8	1/20	1/16
1/12	1/8	1/20	1/16
1/12	0	1/20	0
0	0	1/20	1/16
1/12	0	1/20	1/16



marginal probabilities

$$\mathbf{p}_w = \mathbf{P}_{wd} \mathbf{1} \quad \mathbf{p}_d = \mathbf{P}_{wd}^T \mathbf{1}$$

$$p_{w_1, w_2} = \sum_d p_{w_1|d, \cancel{w_2}} p_{w_2, d}$$

$$\mathbf{P}_{ww} = \mathbf{P}_{wd} \text{diag}(\mathbf{p}_d)^{-1} \mathbf{P}_{wd}^T$$

$$\mathbf{p}_w = \mathbf{P}_{ww} \mathbf{1}$$

projection on words

projection on documents

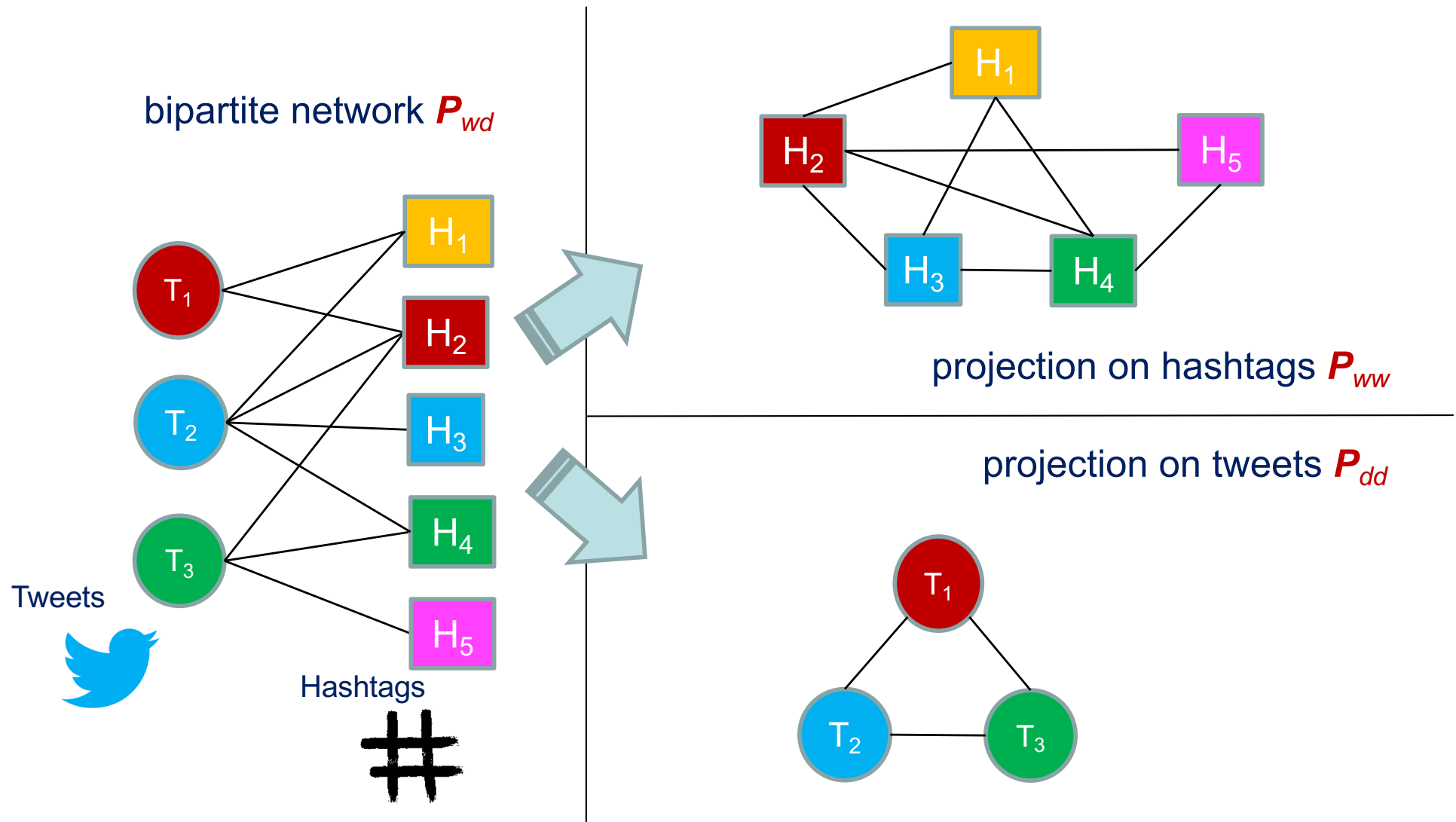
$$p_{d_1, d_2} = \sum_w p_{d_1|w, \cancel{d_2}} p_{d_2, w}$$

$$\mathbf{P}_{dd} = \mathbf{P}_{wd}^T \text{diag}(\mathbf{p}_w)^{-1} \mathbf{P}_{wd}$$

$$\mathbf{p}_d = \mathbf{P}_{dd} \mathbf{1}$$

Bipartite and projected networks

a comparison





The role of TF-IDF

term frequency – inverse document frequency

term frequency =
frequency (probability) of
the word in the document

inverse document frequency =
(log) fraction of documents
that contain the word

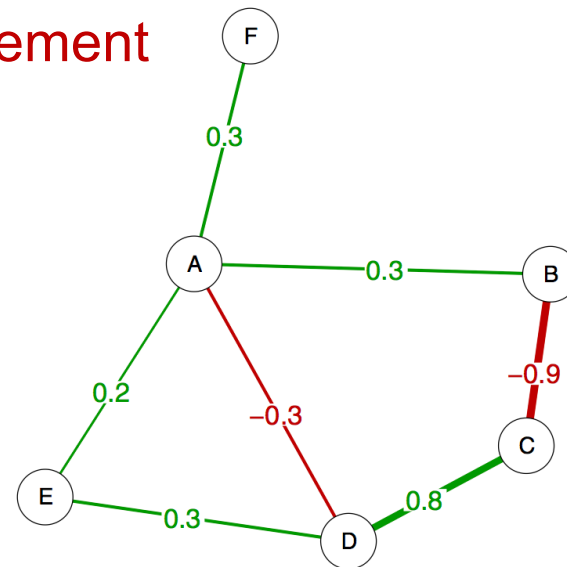
$$\text{TF-IDF}_{w|d} = p_{w|d} \cdot -\log \left(\frac{\sum_d (n_{wd} > 0)}{D} \right)$$

- ❑ An heuristic
- ❑ **Punishes** words that appear in many documents
- ❑ **Enhances** words that are document specific

Signed graphs

□ Edges can have signed values

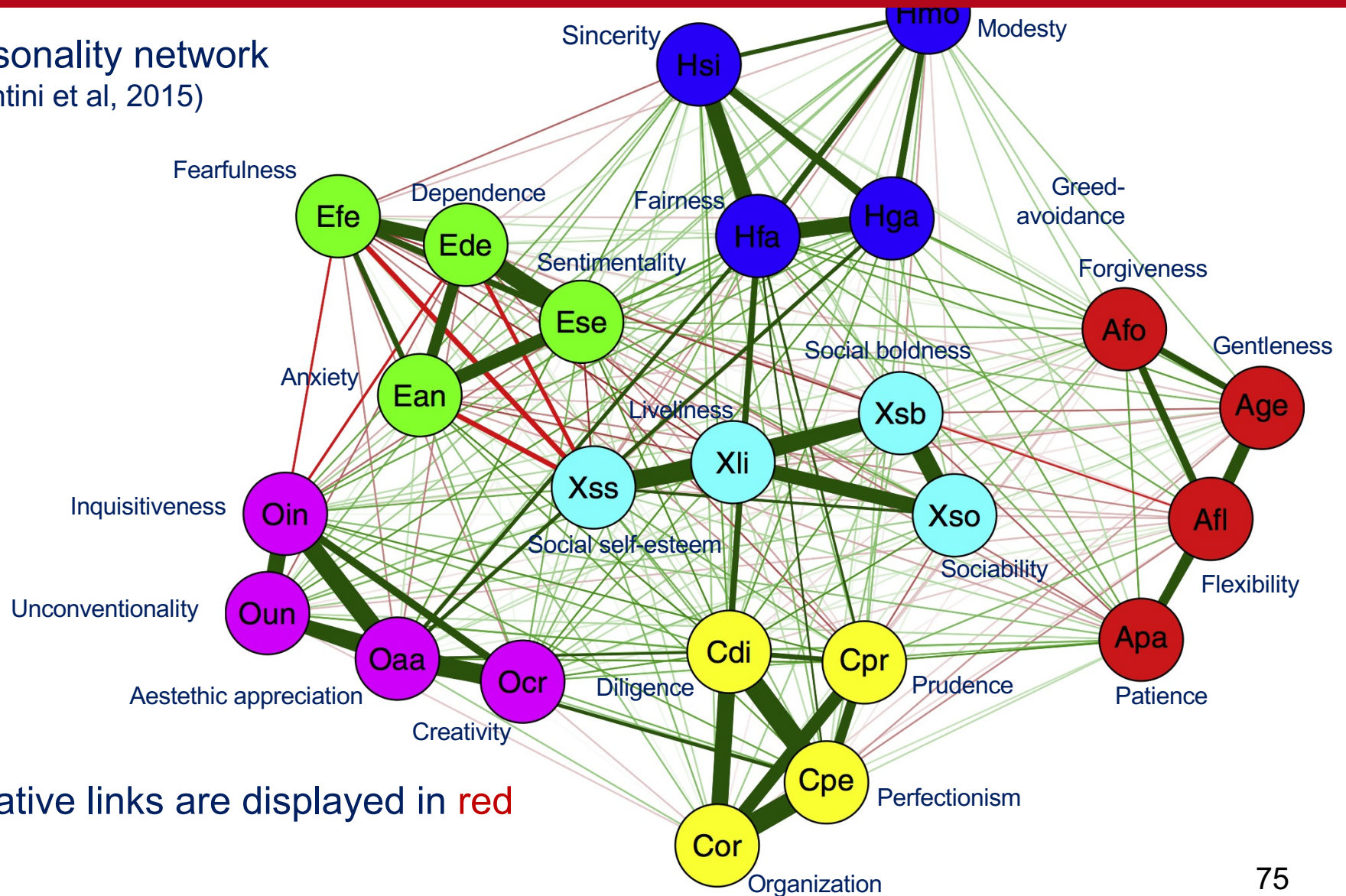
positive if there is an agreement between nodes
negative if there's a disagreement



□ This is typical of correlation networks

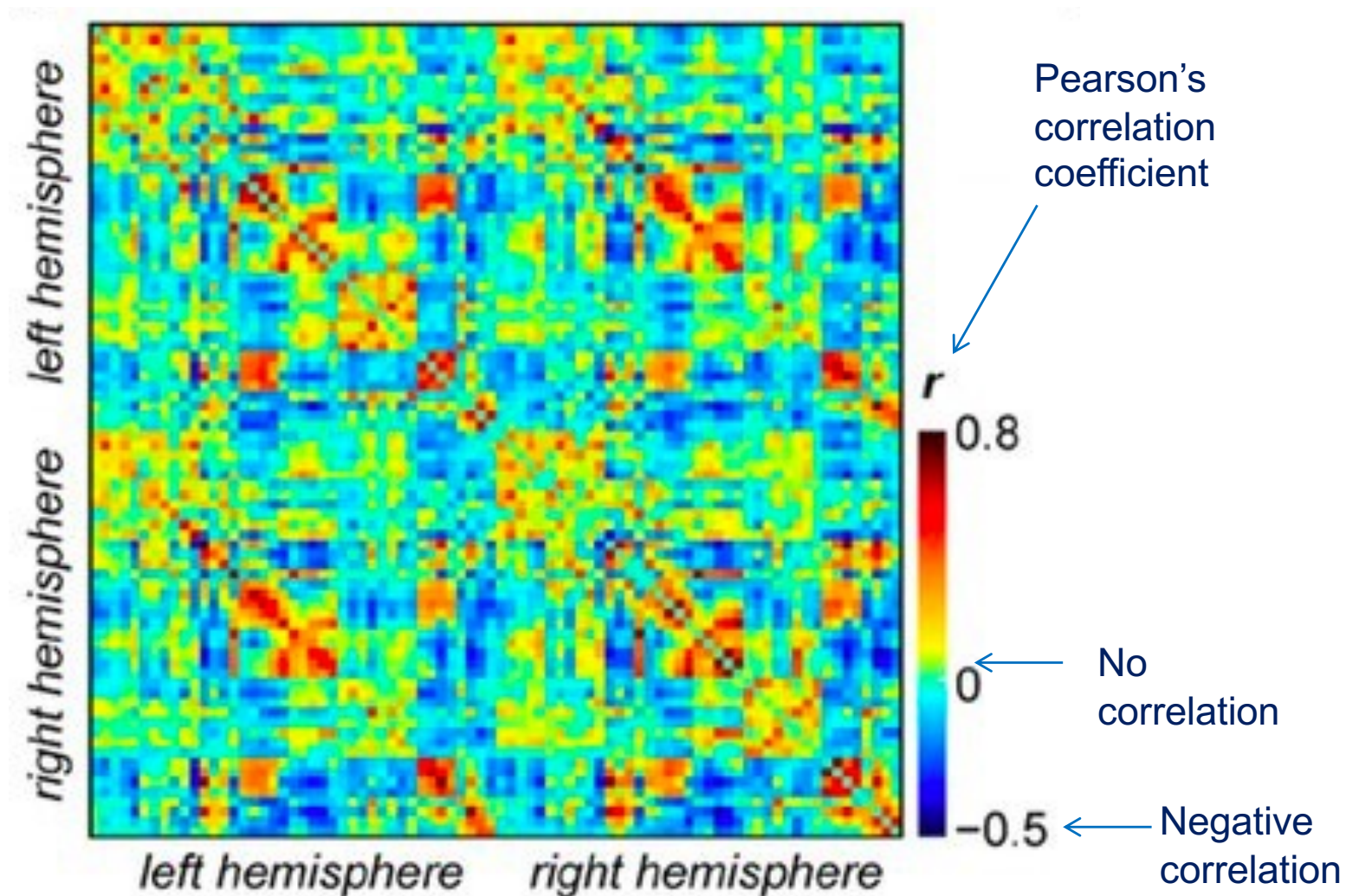
Signed graph example

A personality network
(Costantini et al, 2015)



Signed graph example

An fMRI adjacency matrix (fMRI = functional magnetic resonance imaging)





How can you correlate data? an overview

Cosine similarity

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{|\mathbf{x}|_2 |\mathbf{y}|_2}$$

positive for positive
valued data $\mathbf{x}$ and $\mathbf{y}$

norm 2

Pearson correlation coefficient

$$r(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} - m\mathbf{x})^T (\mathbf{y} - m\mathbf{y})}{\sigma_x \sigma_y}$$

always with a sign

Sørensen-Dice coefficient

$$\text{dice}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\frac{1}{2}|\mathbf{x}|_1 + \frac{1}{2}|\mathbf{y}|_1}$$

for binary data
(it is an F1 score)
always positive



Partial correlation

to remove confounding contributions

partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed

when determining the numerical relationship between two variables of interest, using their correlation coefficient will give misleading results if there is another confounding variable that is numerically related to both variables of interest

$$\text{partial}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{e}_x^T \mathbf{e}_y}{|\mathbf{e}_x|_2 |\mathbf{e}_y|_2}$$

collection of data
vectors, other than $\mathbf{x}$
and $\mathbf{y}$, plus the
constant vector $\mathbf{1}$

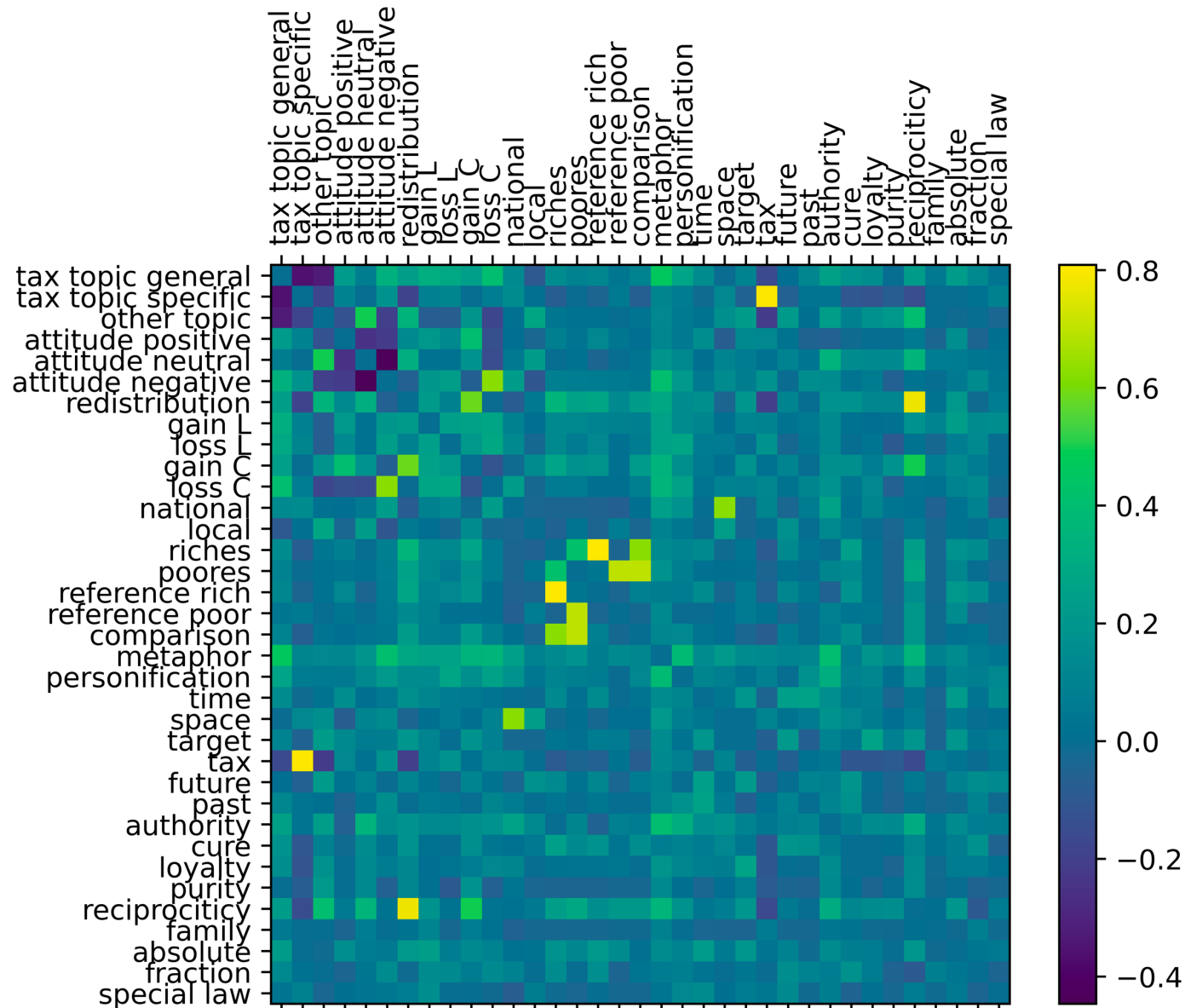
$$\mathbf{e}_x = (\mathbf{I} - \underbrace{\mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T}) \mathbf{x}$$

projection on the space
orthogonal to $\text{span}(\mathbf{Z})$



Tax questionnaire example

Pearson correlation used



x and **y** data are values (binary or not) for **different categories** among a wide set of **people** interviewed

(can do the same with posts and **sentiment** analysis over the **posts**)



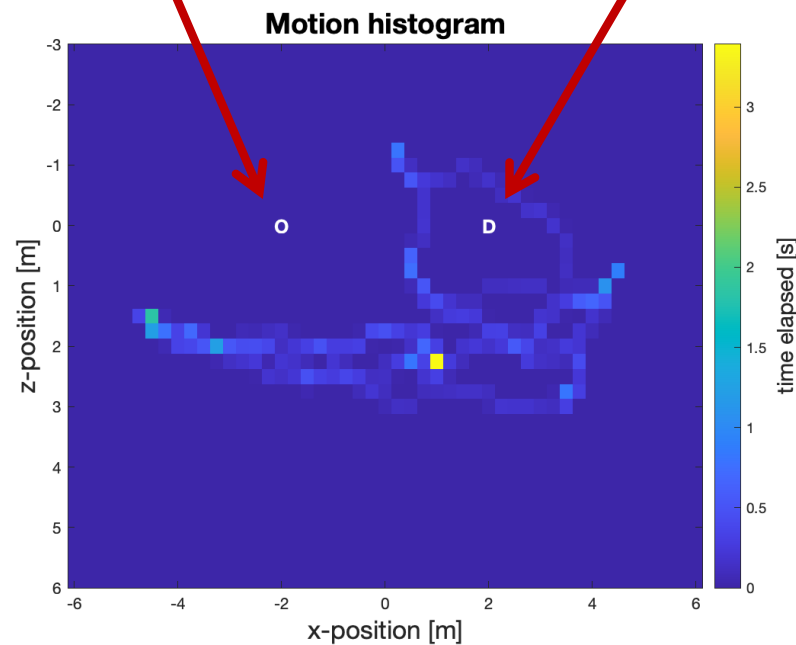
Motion patterns in VR example

studying immersive environments

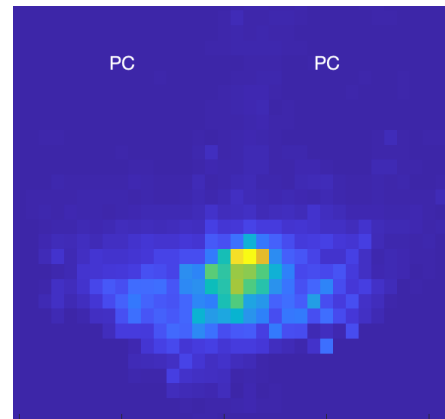


original
pointcloud

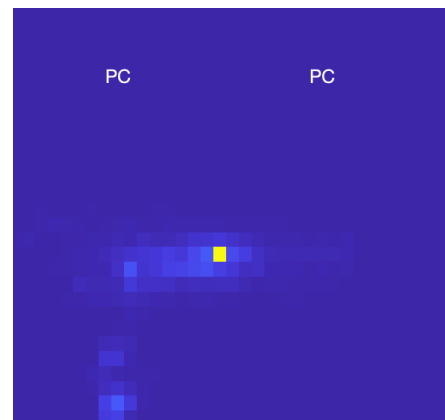
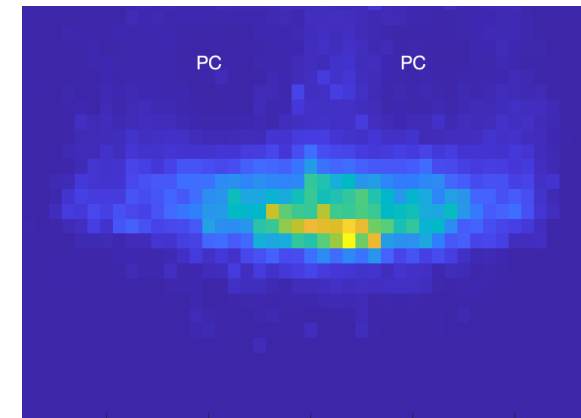
distorted
pointcloud



Cluster 1: walking
from a distance



Cluster 2: walking
closely

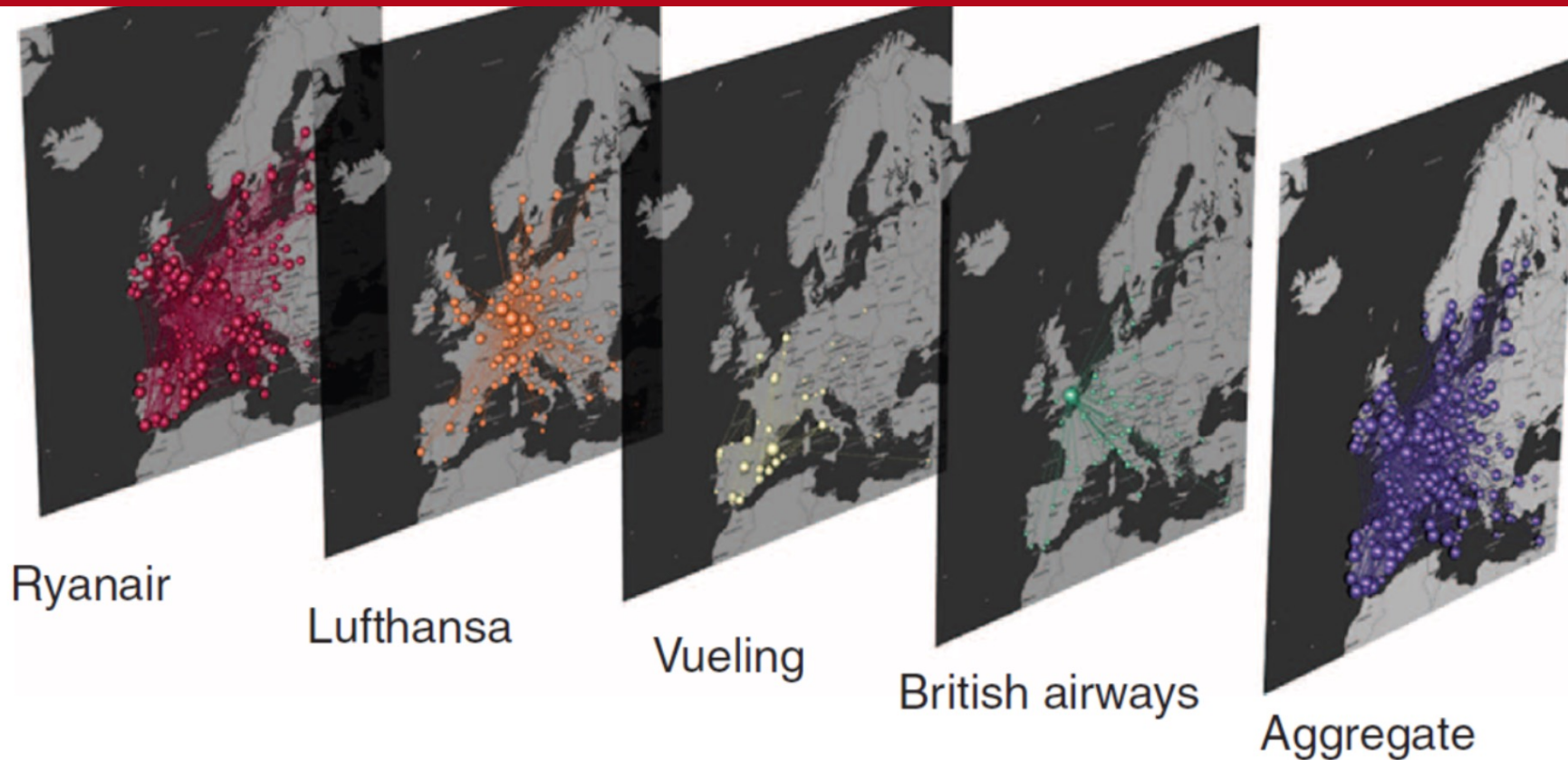


Cluster 3:
standing still

motion behaviours
detected by Louvain
on **Pearson**
correlations over
(filtered) motion
patterns

Multi-layer networks

Multi-layer networks



described by a **set** of adjacency matrices $\mathbf{A}_\ell$

average connection $\mathbf{A} = \sum_\ell \mathbf{A}_\ell$



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

US republicans and democrats interactions on Twitter (2020)

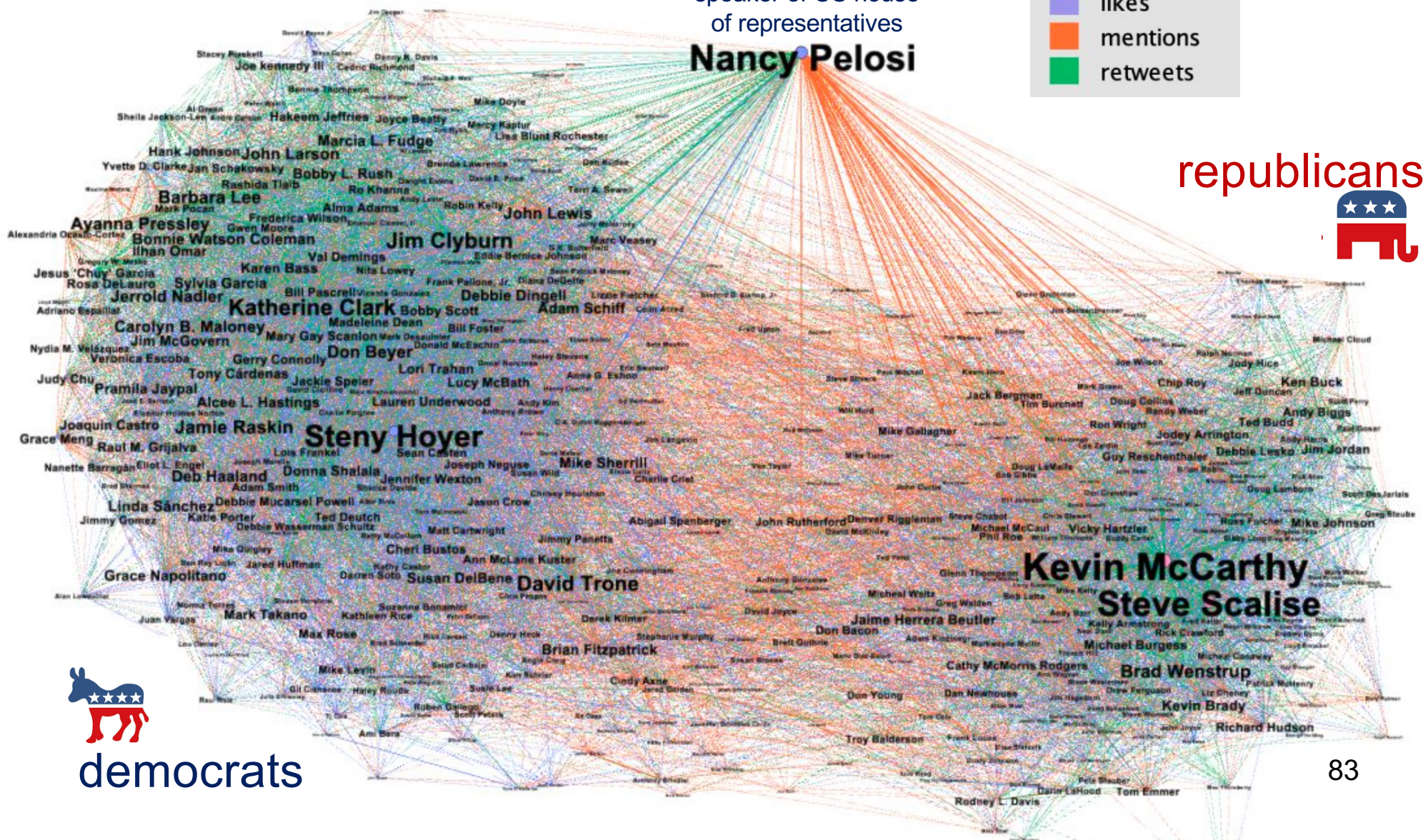
An example

speaker of US house
of representatives

Nancy Pelosi

- likes
- mentions
- retweets

republicans



democrats

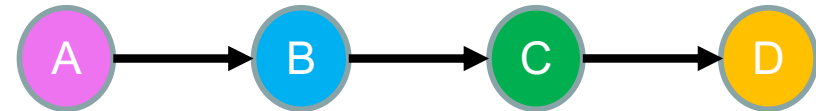
Paths and connectivity

in graphs



□ Path

a sequence of interconnected nodes (meaning each pair of nodes adjacent in the sequence are connected by a link)

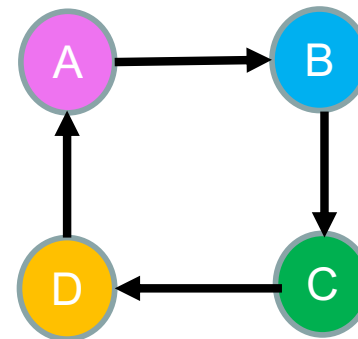


□ Path length

of links involved in the path (if the path involves n nodes then the path link is $n-1$)

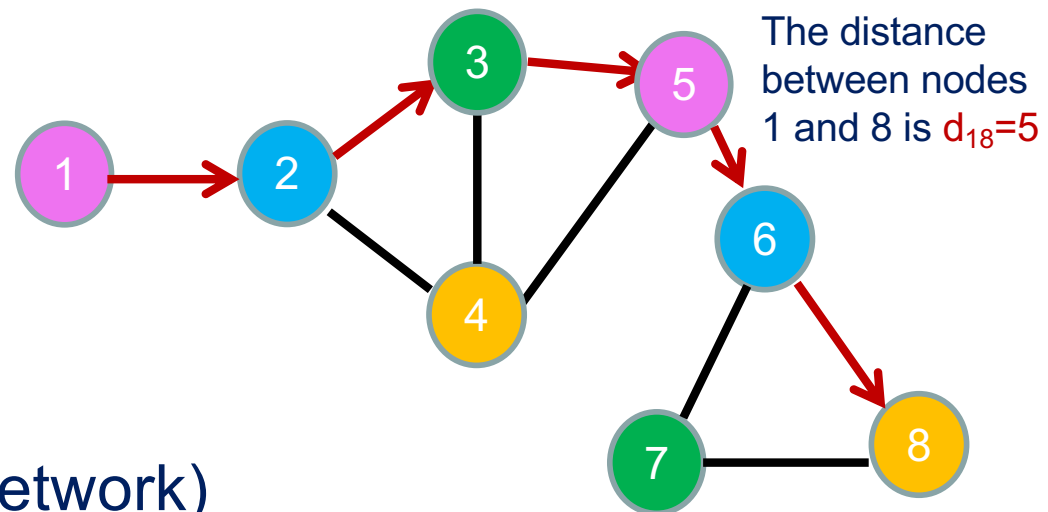
□ Cycle

path where starting and ending nodes coincide



□ Shortest path (between any two nodes)

the path with the minimum length, which is called the **distance**



it is **not** unique!

□ Diameter (of the network)

the highest distance in the network

The diameter is $d=5$

□ Algorithms

available to compute distances: **Dijkstra**,
Bellman-Ford, **BFS**

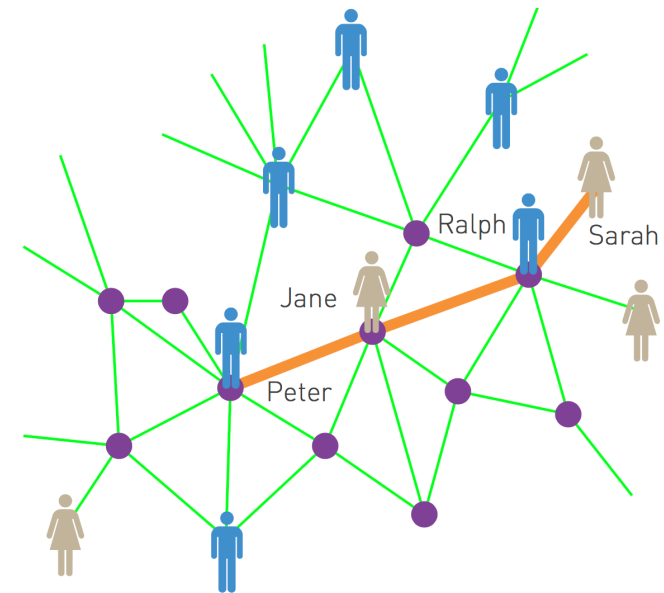
□ Average path length

average distance between all nodes pairs (apply an algorithm to all node couples, and take the average)

□ In real networks distance between two randomly chosen nodes is generally short

□ Milgram [1967]: *6 degrees of separation*

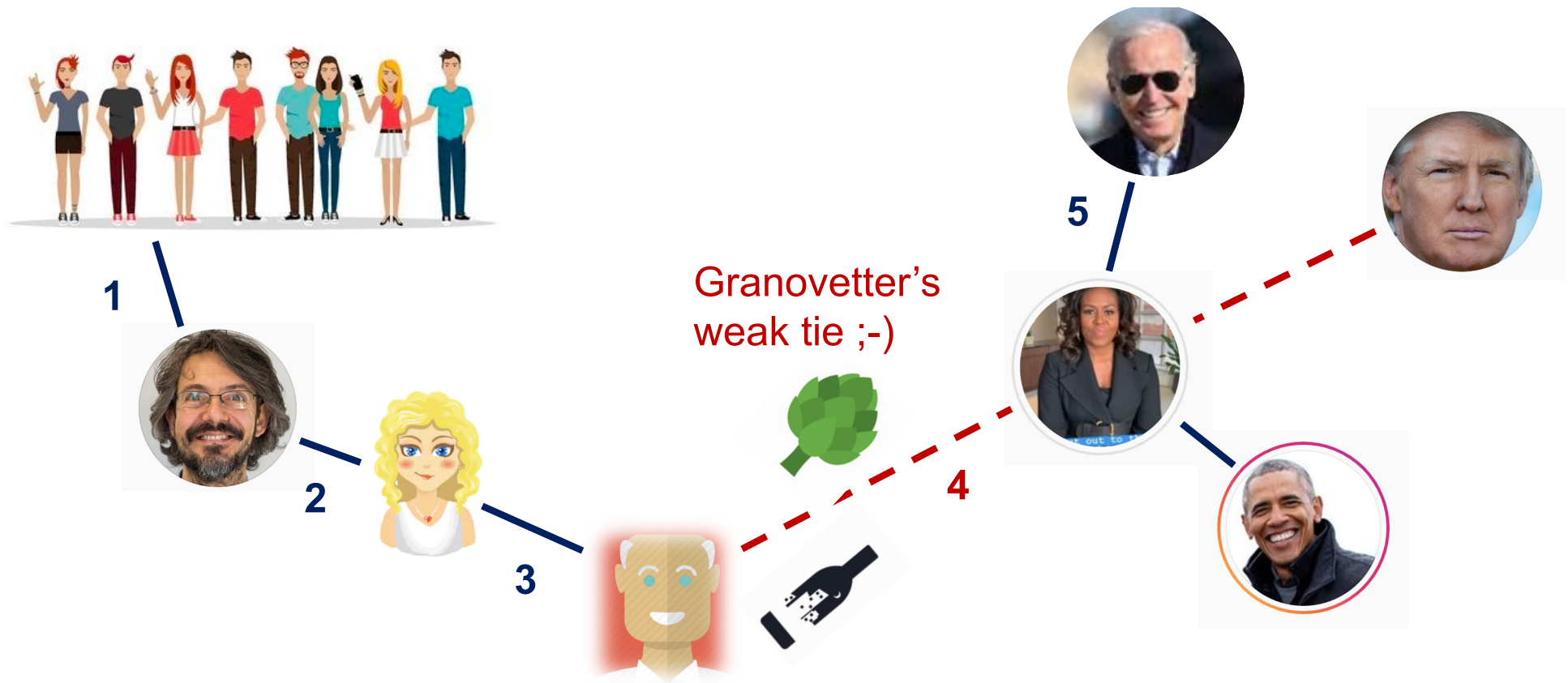
□ What does this mean?
We are more connected than we think





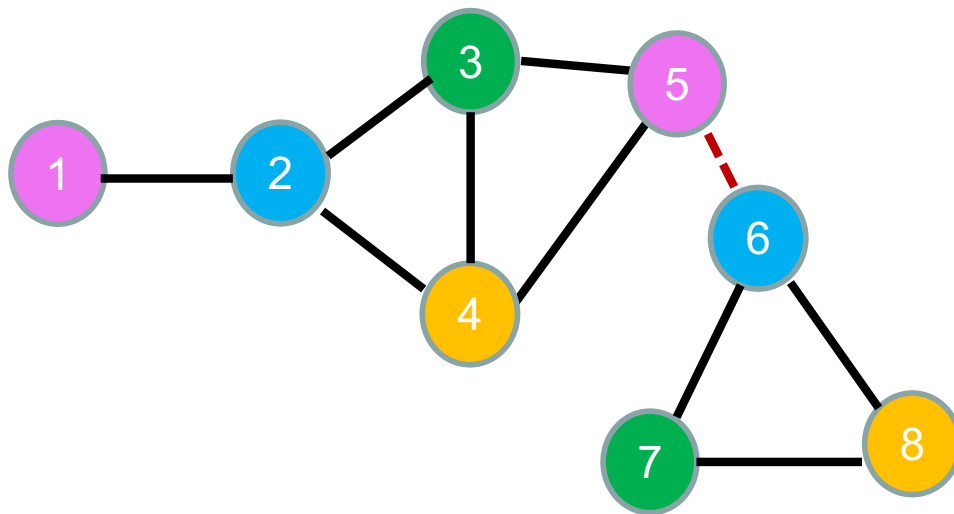
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Small world we and the US presidents



□ A **bridge** is a link between two connected components

its removal would make the network disconnected



$A =$

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



For directed networks we distinguish between

- ❑ **Strongly** connected components

where $i \rightarrow j$ and $j \rightarrow i$ for all choices of (i, j) in the component

- ❑ **Weakly** connected components

connected in the undirected sense (i.e., disregard link directions)

Condensation graph

- Strong connectivity induces a **partition** in disjoint **strongly connected** sets $\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_k$
- By reinterpreting the sets as nodes we obtain a **condensation graph** $\mathcal{G}^*$ where $i \rightarrow j$ is an edge if a connection exists between sets $\mathcal{V}_i \rightarrow \mathcal{V}_j$

