



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Social Network Analysis

A.Y. 23/24

Communication Strategies

PageRank

a centrality measure based on the web

What is PageRank?

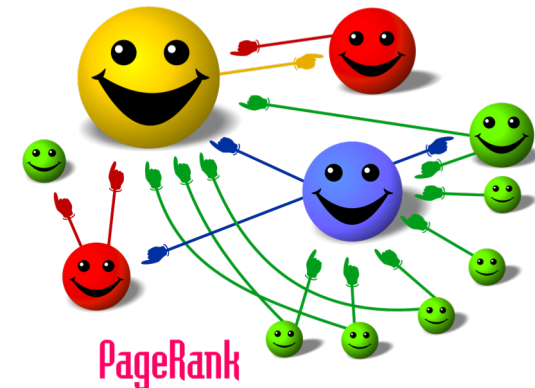
PageRank

From Wikipedia, the free encyclopedia



PageRank (PR) is an [algorithm](#) used by [Google Search](#) to rank [web pages](#) in their [search engine](#) results. PageRank was named after [Larry Page](#),^[1] one of the founders of Google. PageRank is a way of measuring the importance of website pages. According to Google:

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.^[2]



Currently, PageRank is not the only algorithm used by Google to order search results, but it is the first algorithm that was used by the company, and it is the best known.^{[3][4]} As of September 24, 2019, PageRank and all associated patents are expired.^[5]



How to organise the web?

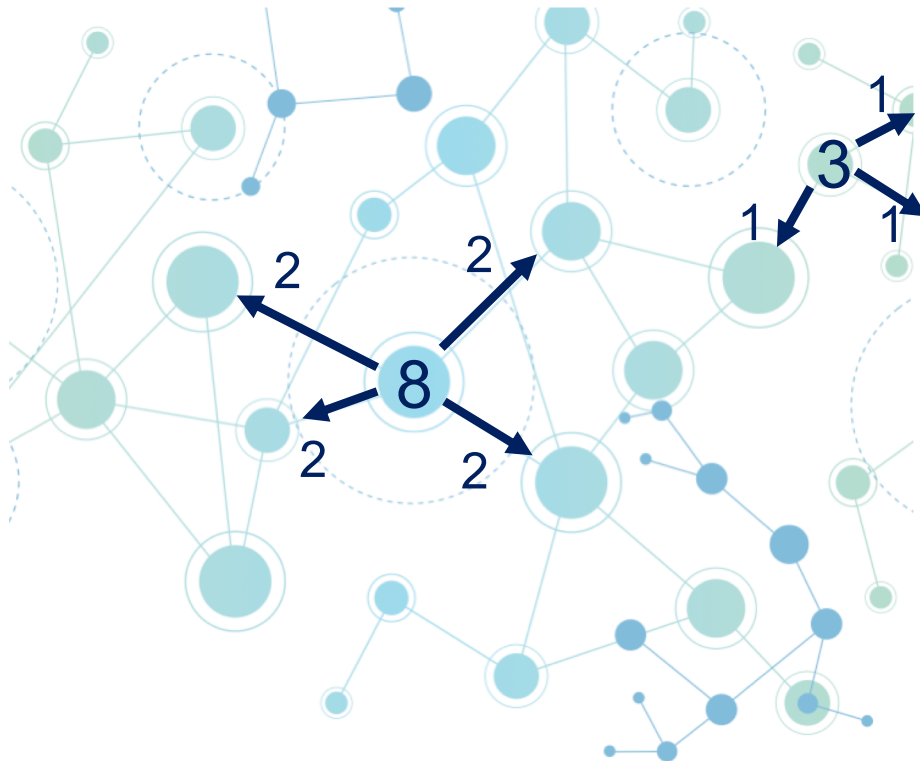
links as votes



- ❑ the higher the **number of incoming links**, the more important a node
- ❑ the more important a node, the more **valuable** the output links



Step 1: spread (evenly)
information (on centrality)
from each node

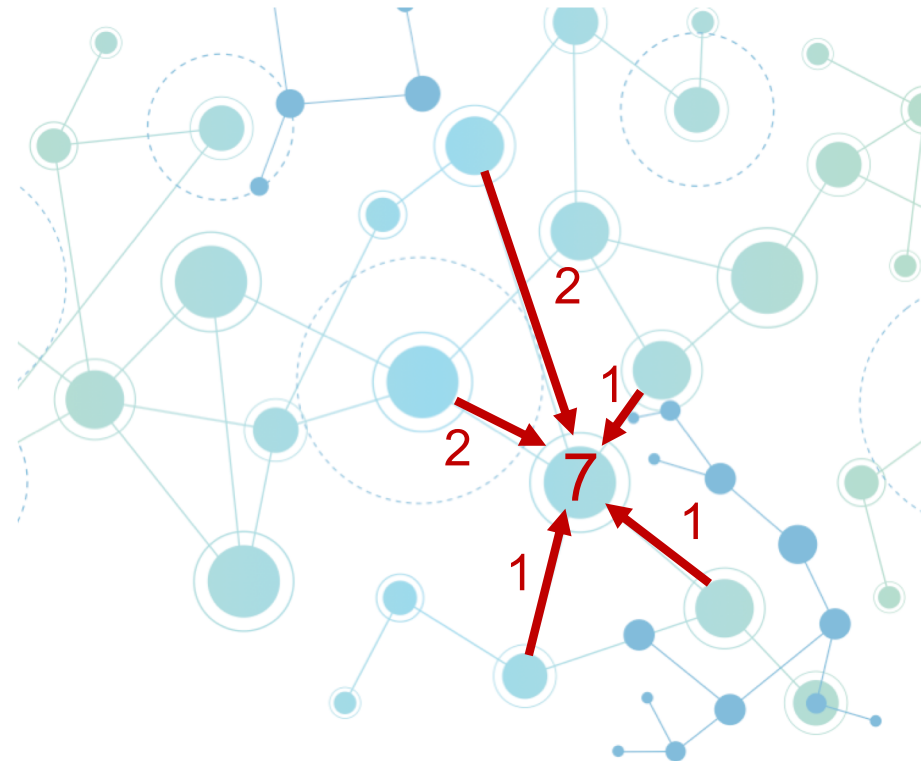


in the web this
corresponds to the idea
that starting from a web
page you choose with
equal probability one of
the sites linked by the
page



Step 2: collect spreaded information at each node
(until convergence)

in the web this roughly
corresponds to the
chance (probability) of
ending in a specific web
page



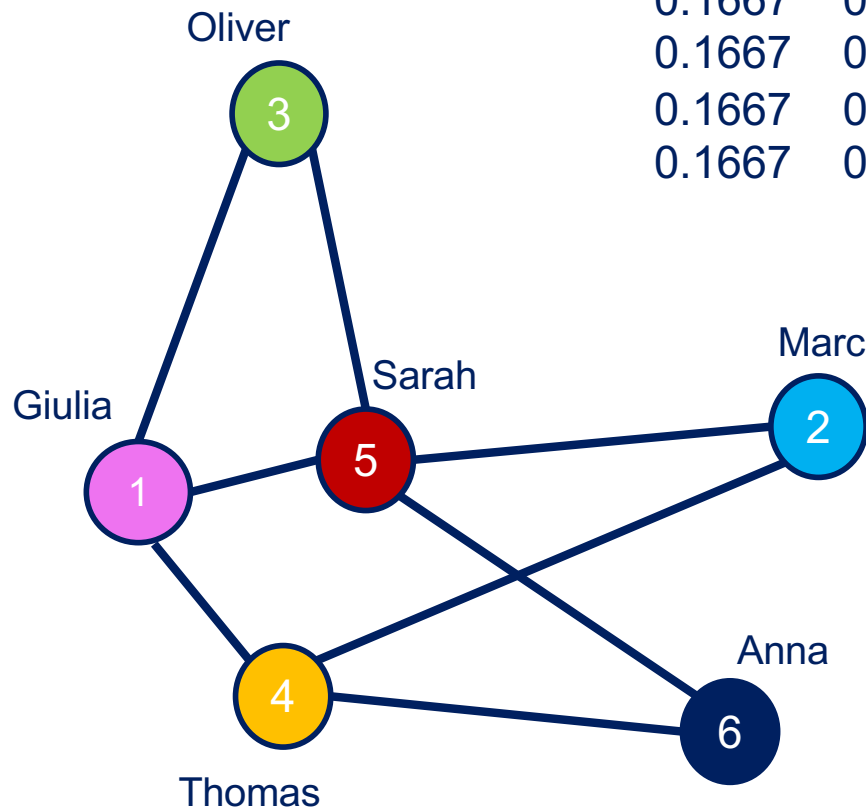


Example

random flow on a friends' network

Equally likely
assignment
to start with

| $t=1$ | 2 | 3 | 4 | 5 |
|--------|--------|--------|--------|--------|
| 0.1667 | 0.1806 | 0.1991 | 0.1723 | 0.2025 |
| 0.1667 | 0.0972 | 0.1505 | 0.1040 | 0.1436 |
| 0.1667 | 0.0972 | 0.1366 | 0.1179 | 0.1287 |
| 0.1667 | 0.2222 | 0.1574 | 0.2168 | 0.1614 |
| 0.1667 | 0.3056 | 0.2060 | 0.2851 | 0.2203 |
| 0.1667 | 0.0972 | 0.1505 | 0.1040 | 0.1436 |



Equal to
(normalized)
degree centrality
in undirected
networks !!!

| 10 | 20 | 50 | 75 | 100 | |
|--------|--------|--------|--------|--------|--------|
| 0.1783 | 0.1848 | 0.1874 | 0.1875 | 0.1875 | Giulia |
| 0.1153 | 0.1222 | 0.1249 | 0.1250 | 0.1250 | Marc |
| 0.1242 | 0.1248 | 0.1250 | 0.1250 | 0.1250 | Oliver |
| 0.2020 | 0.1917 | 0.1876 | 0.1875 | 0.1875 | Thomas |
| 0.2649 | 0.2543 | 0.2501 | 0.2500 | 0.2500 | Sarah |
| 0.1153 | 0.1222 | 0.1249 | 0.1250 | 0.1250 | Anna |

Teleportation

as a method to strengthen the result

Idea:

the surfer does not necessarily move to one of the links of the page she/he is viewing:

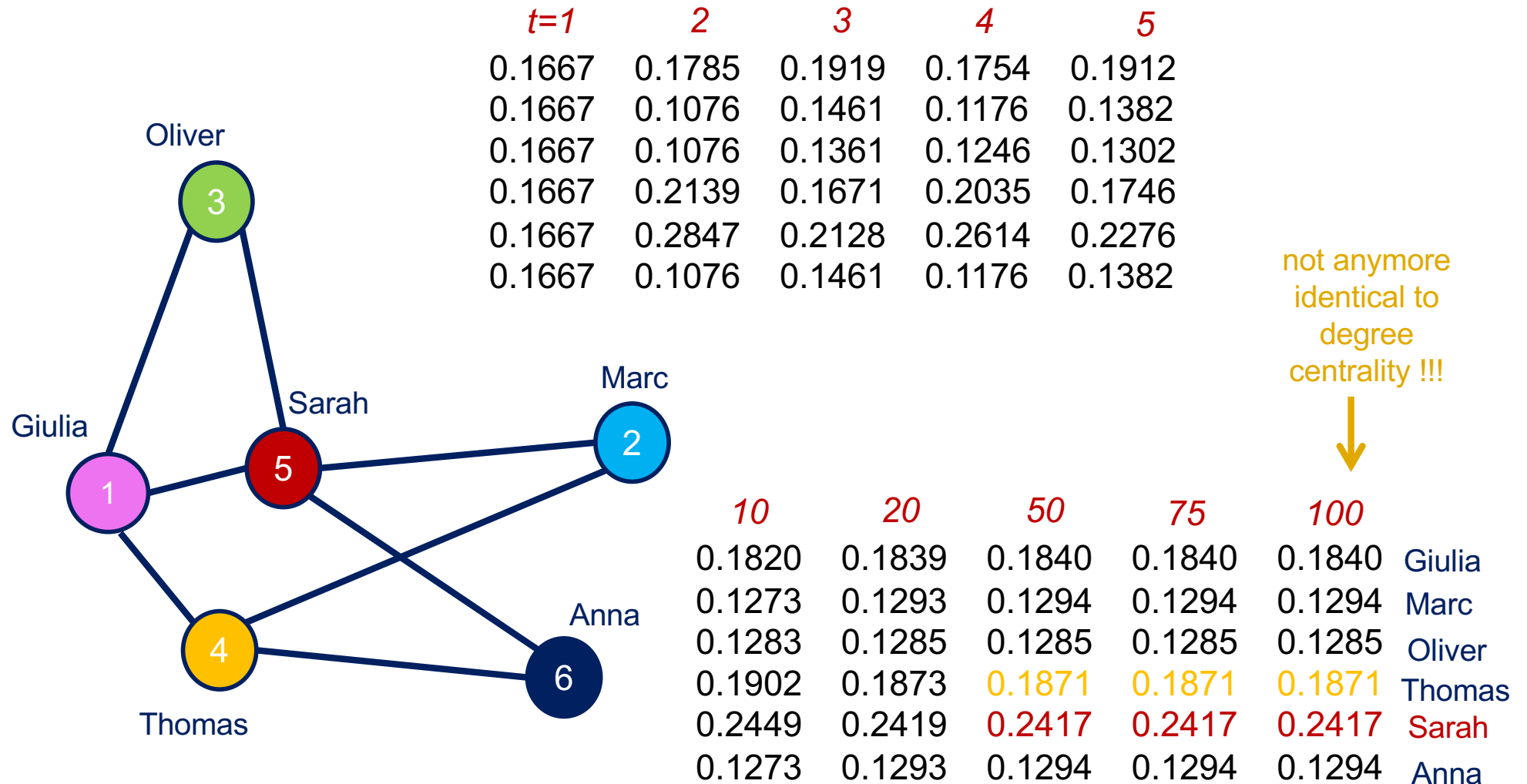
- ❑ it does with probability, say $c = 85\%$
- ❑ with probability $1 - c = 15\%$ it might jump to a **random page** (according to a predetermined **policy**)





Example

teleportation on a friends' network – random policy





- ❑ PageRank can capture the subtleties of networks
- ❑ Similar, but more reliable than degree
- ❑ Simple to implement (scalable)
- ❑ Want to see this in your projects

Visualizing PageRank

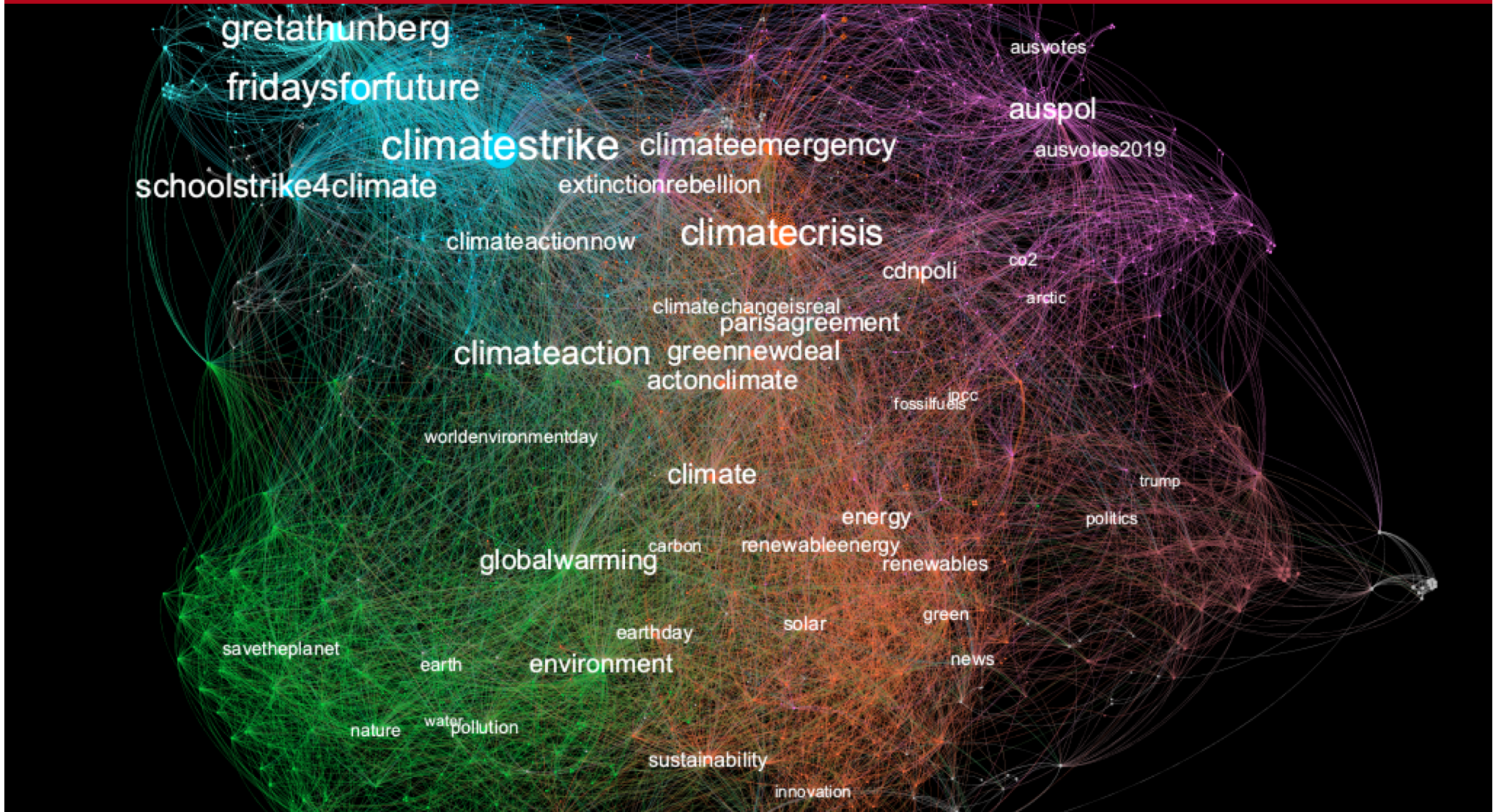
a comparison with degree centrality



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

PageRank on a semantic network

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





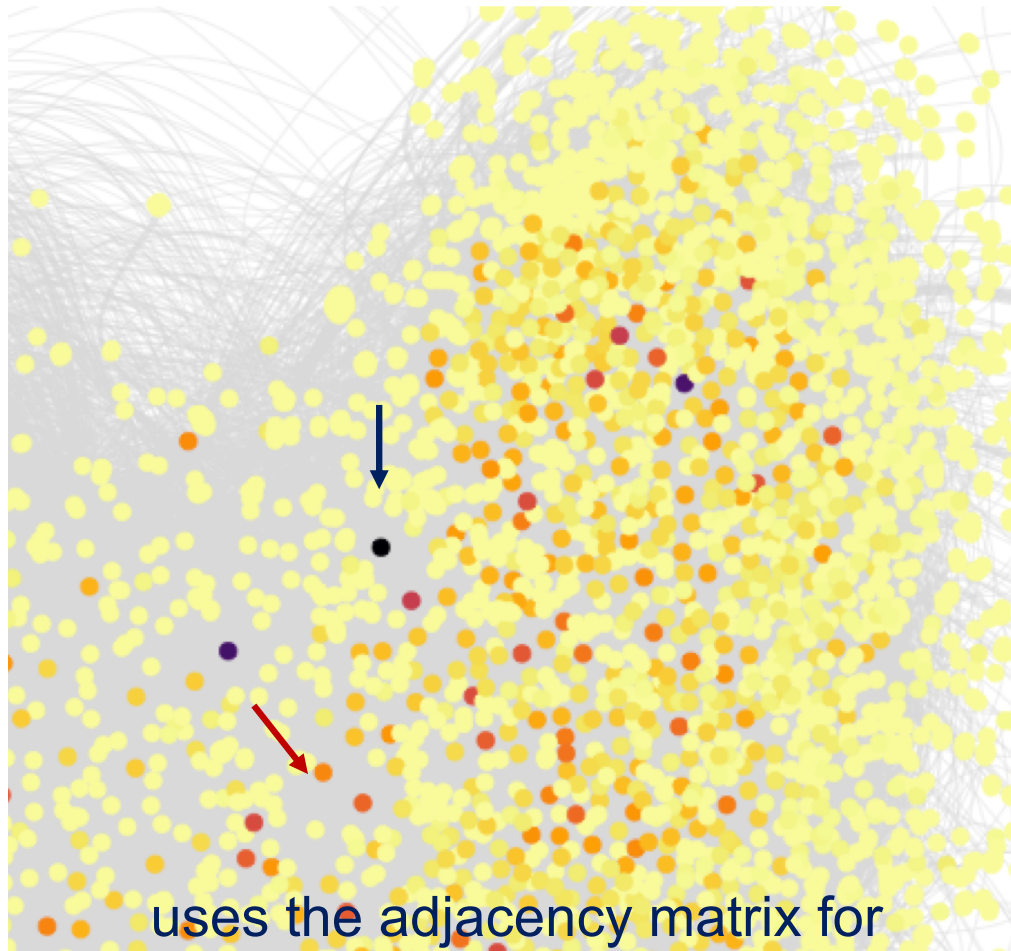
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Example of PageRank centrality

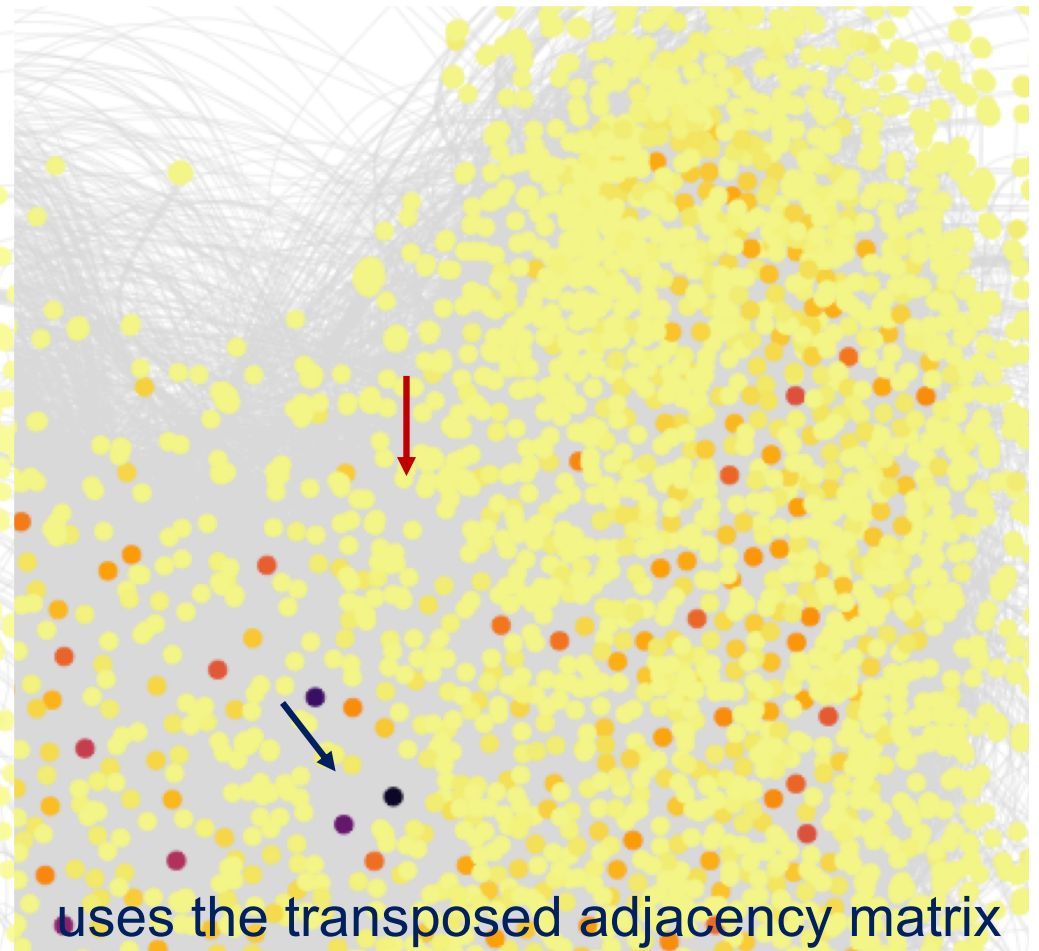
wikipedia administrator elections and vote history data

<https://snap.stanford.edu/data/wiki-Vote.html>

Authorities



Hubs



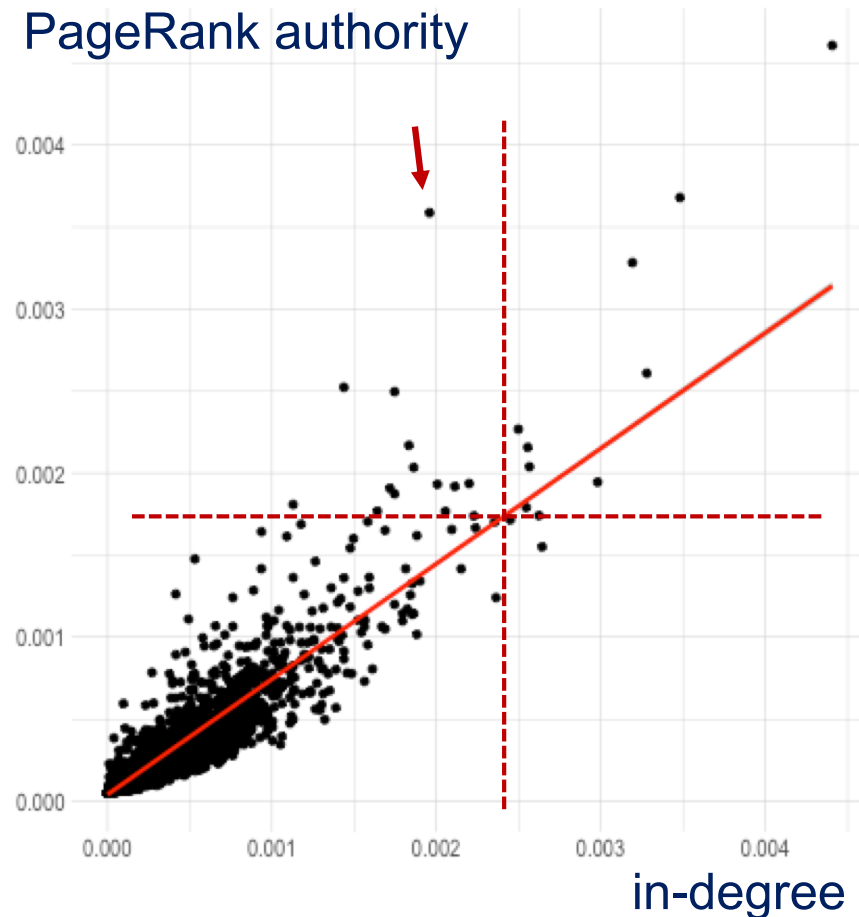


PageRank versus degree centrality

wikipedia administrator elections and vote history data

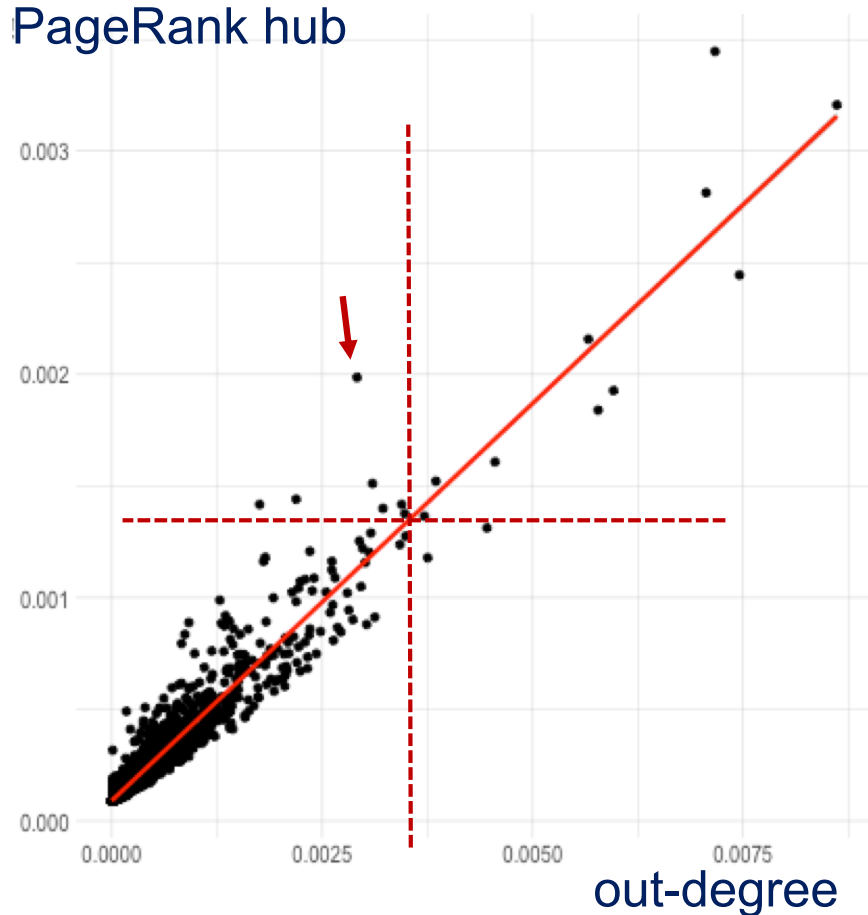
Authorities

PageRank authority



Hubs

PageRank hub



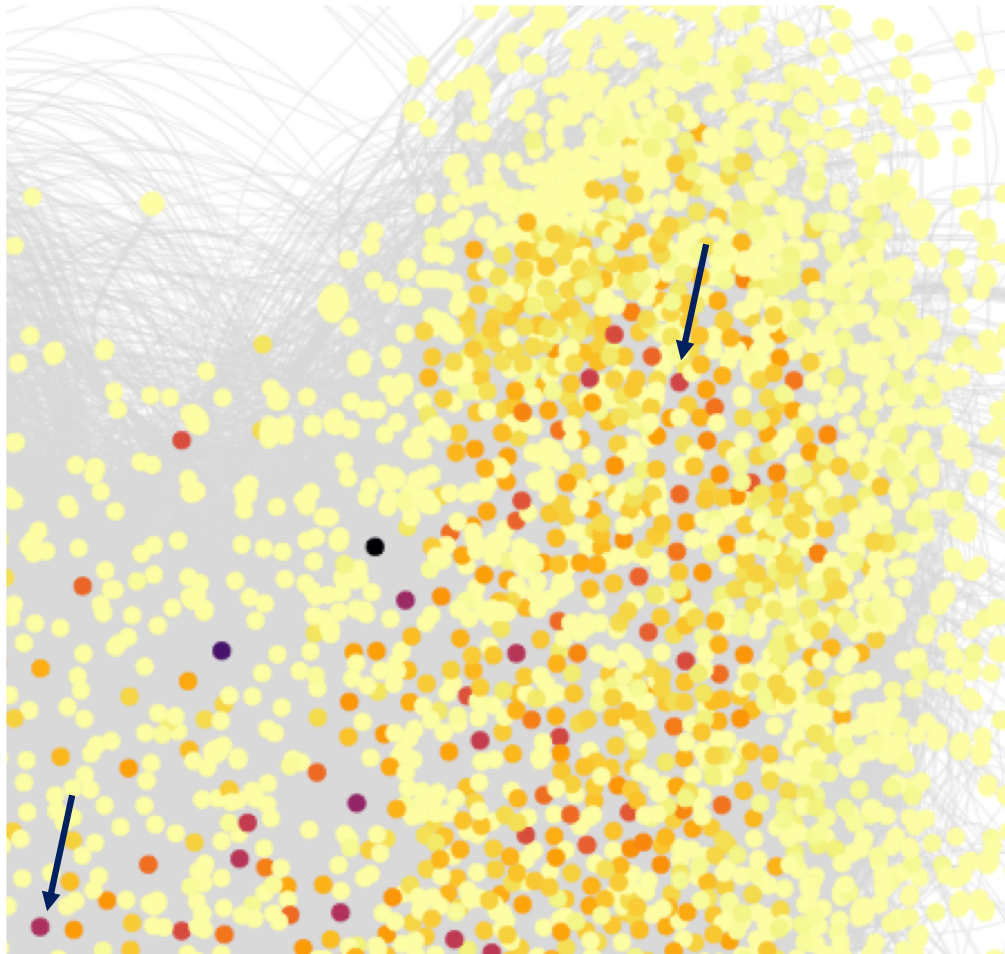


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

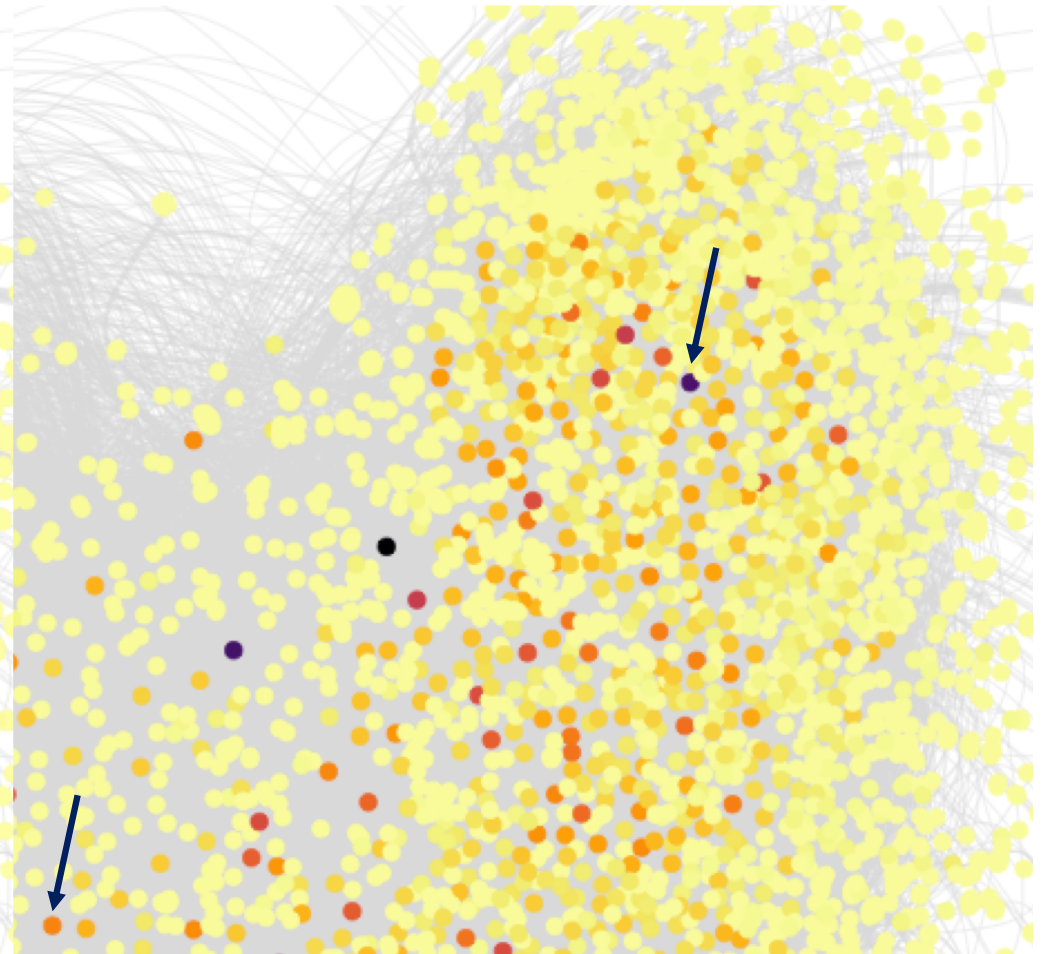
PageRank versus degree authorities

wikipedia administrator elections and vote history data

Degree



PageRank



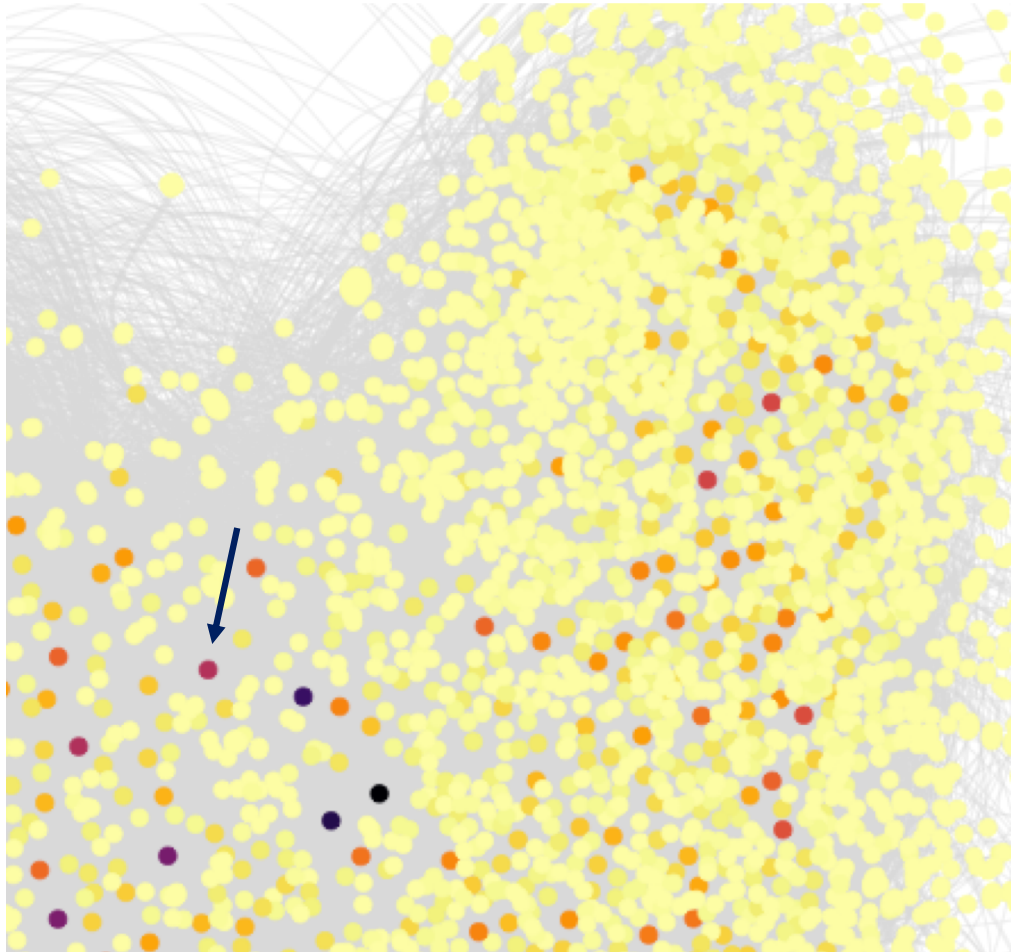


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

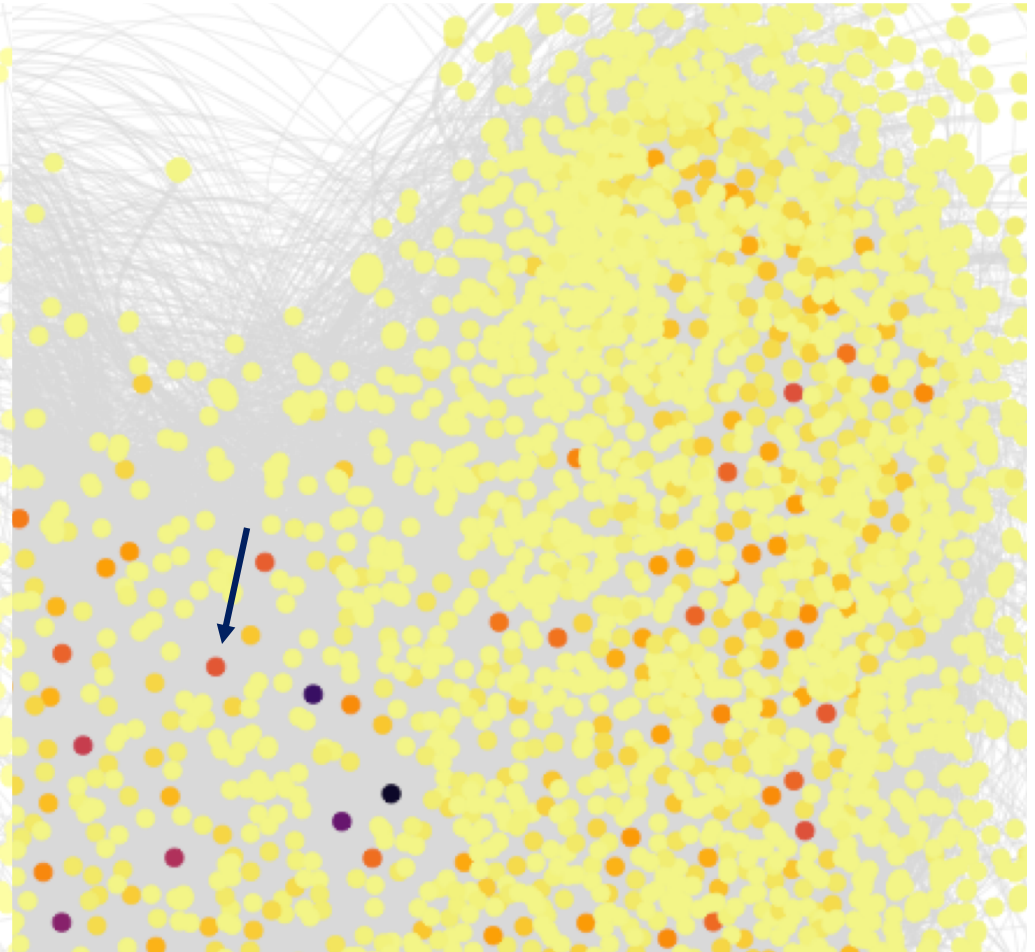
PageRank versus degree hubs

wikipedia administrator elections and vote history data

Degree



PageRank



Local PageRank

measuring closeness to a node, i.e., friendship



Measuring closeness: LocalPageRank

measure similarity to a node

Idea

- ❑ Measure **similarity** or closeness to node i by applying PageRank with teleport set **to node i only**

Result

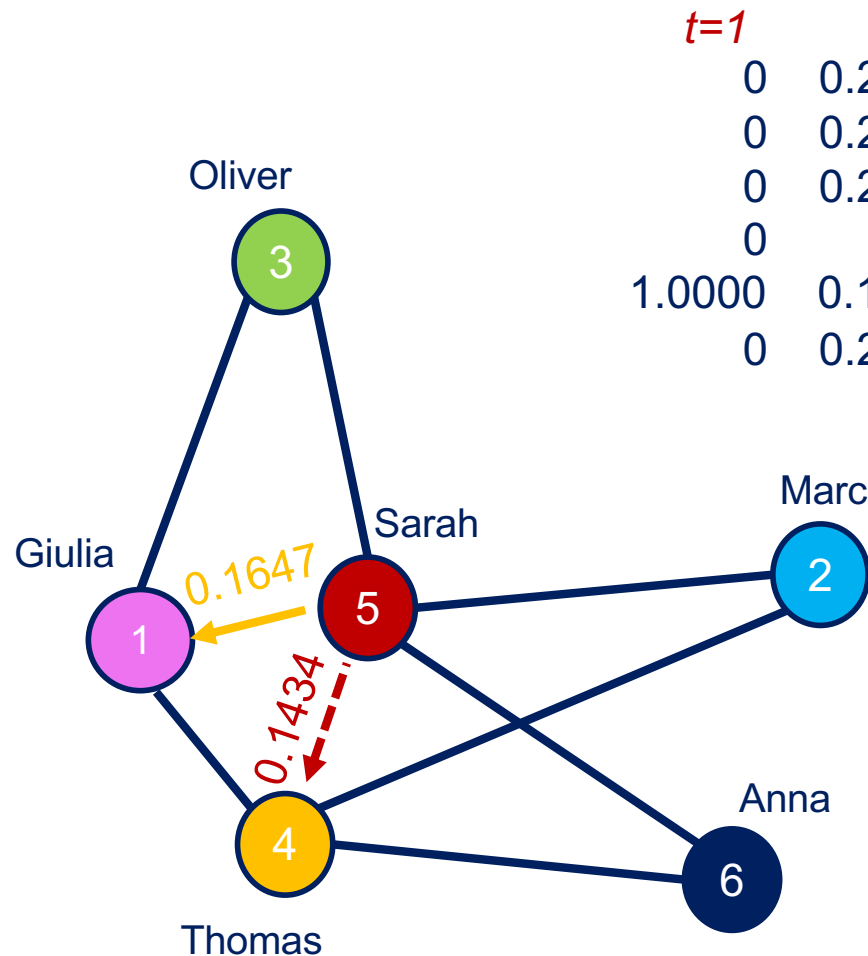
- ❑ Measures direct and indirect multiple connections, their quality, degree or weight





Example

who's Sara's best friend? Policy = jump back to Sara



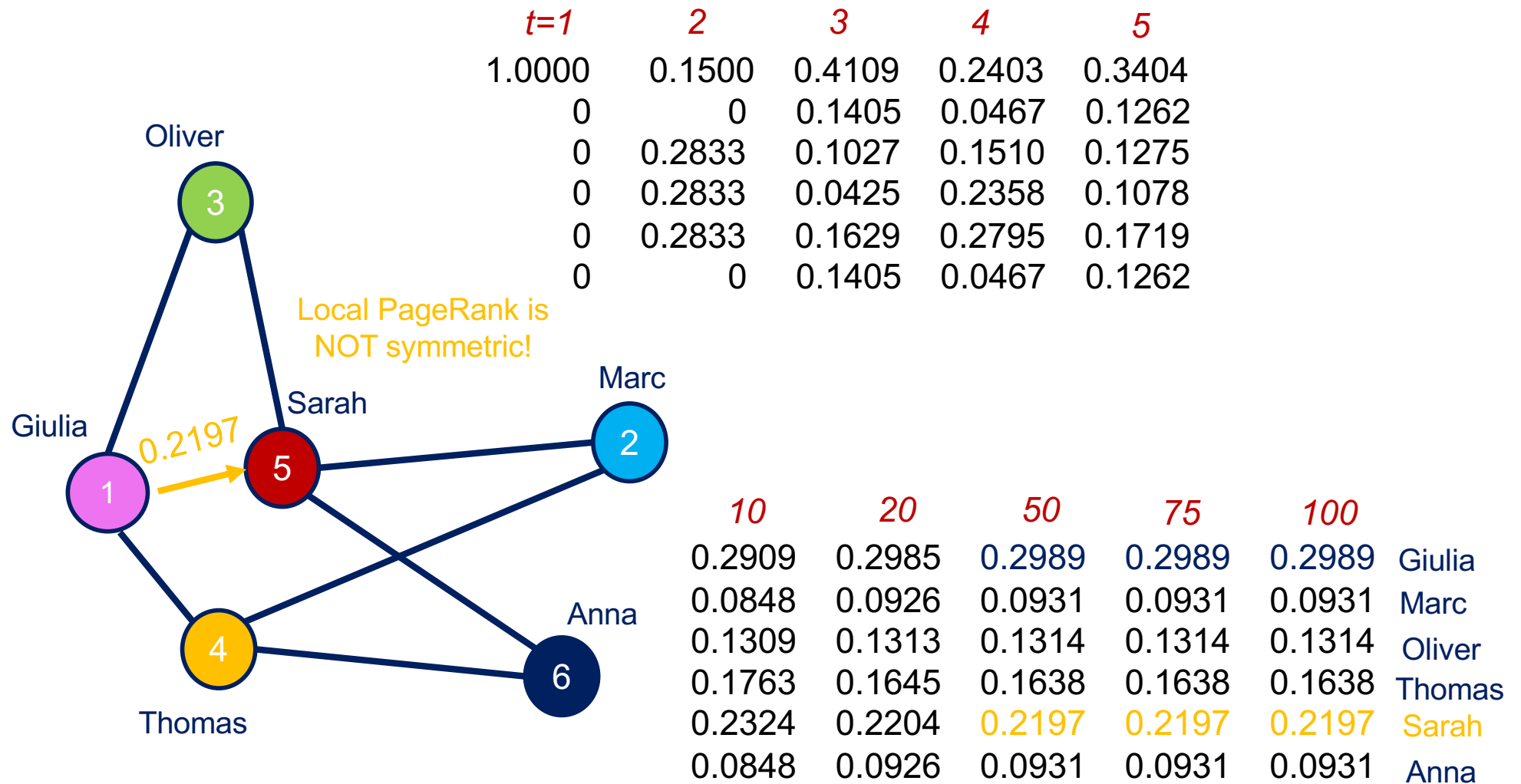
| $t=1$ | 2 | 3 | 4 | 5 |
|--------|--------|--------|--------|--------|
| 0 | 0.2125 | 0.1222 | 0.2096 | 0.1290 |
| 0 | 0.2125 | 0.0319 | 0.1705 | 0.0708 |
| 0 | 0.2125 | 0.0921 | 0.1369 | 0.1127 |
| 0 | 0 | 0.2408 | 0.0617 | 0.2043 |
| 1.0000 | 0.1500 | 0.4811 | 0.2508 | 0.4125 |
| 0 | 0.2125 | 0.0319 | 0.1705 | 0.0708 |

| 10 | 20 | 50 | 75 | 100 | |
|--------|--------|--------|--------|--------|--------|
| 0.1743 | 0.1653 | 0.1647 | 0.1647 | 0.1647 | Giulia |
| 0.1238 | 0.1144 | 0.1138 | 0.1138 | 0.1138 | Marc |
| 0.1206 | 0.1199 | 0.1199 | 0.1199 | 0.1199 | Oliver |
| 0.1285 | 0.1426 | 0.1434 | 0.1434 | 0.1434 | Thomas |
| 0.3290 | 0.3435 | 0.3444 | 0.3444 | 0.3444 | Sarah |
| 0.1238 | 0.1144 | 0.1138 | 0.1138 | 0.1138 | Anna |



Example

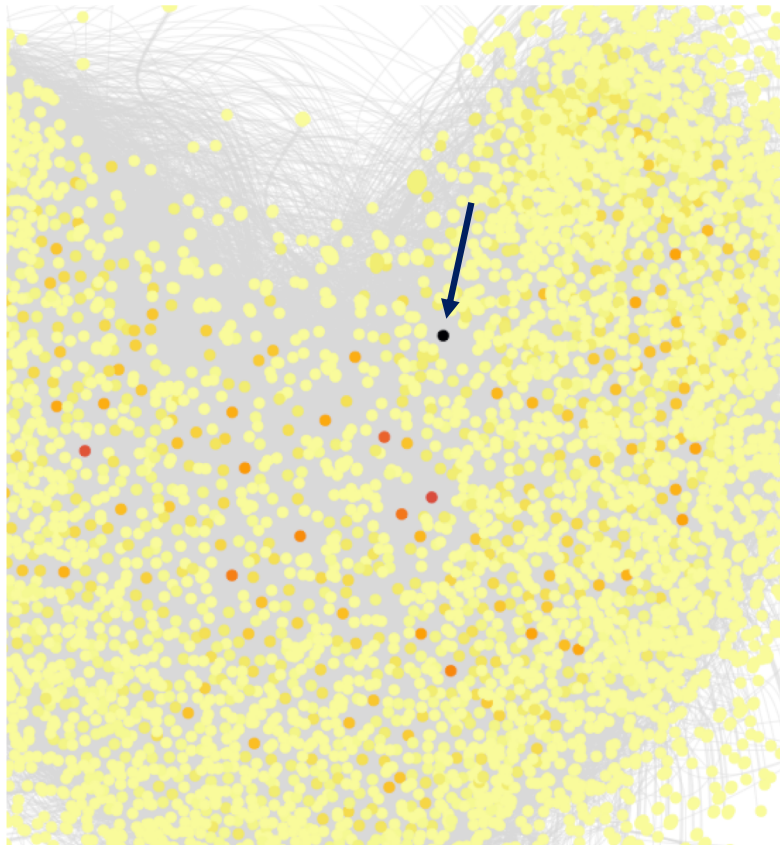
who's Giulia's best friend? Policy = jump back to Giulia





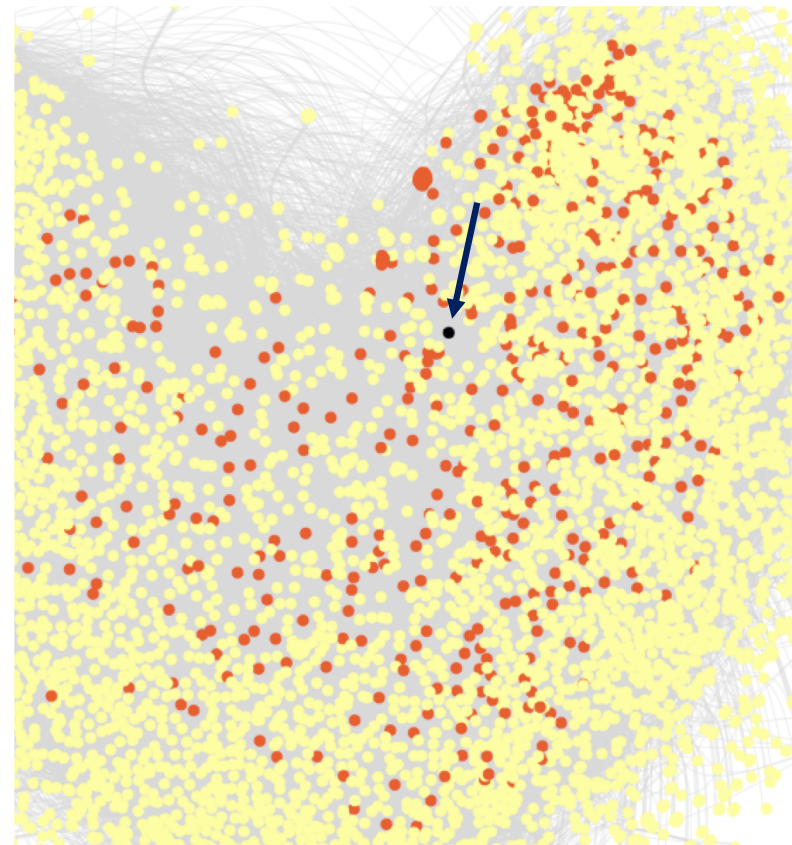
Local PageRank versus degree authorities

Local PageRank



neighbours authority score =
local node \rightarrow neighbours

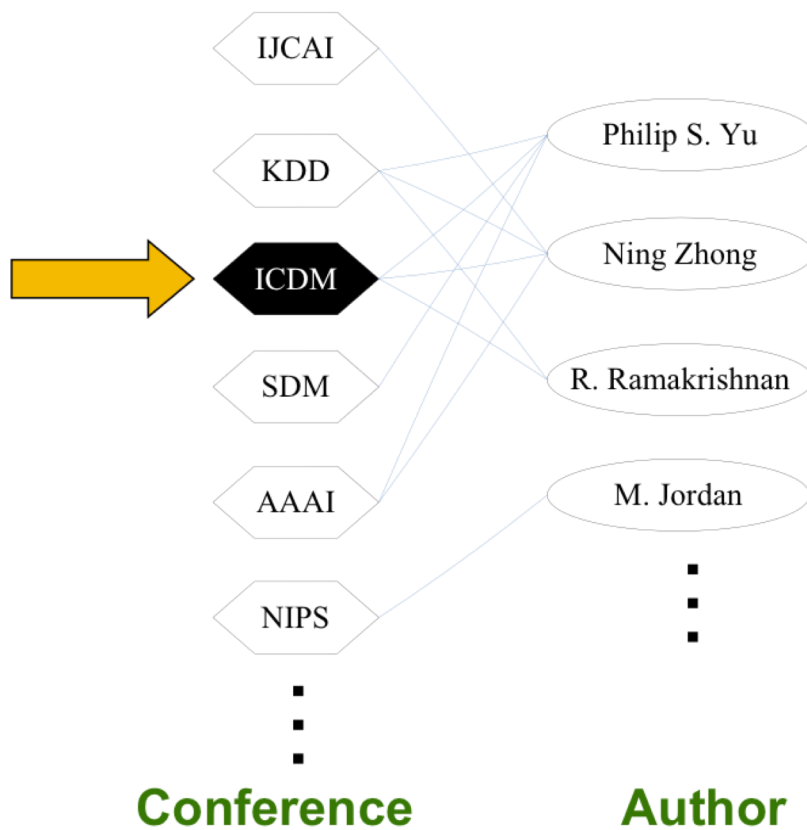
1-hop out-neighbours



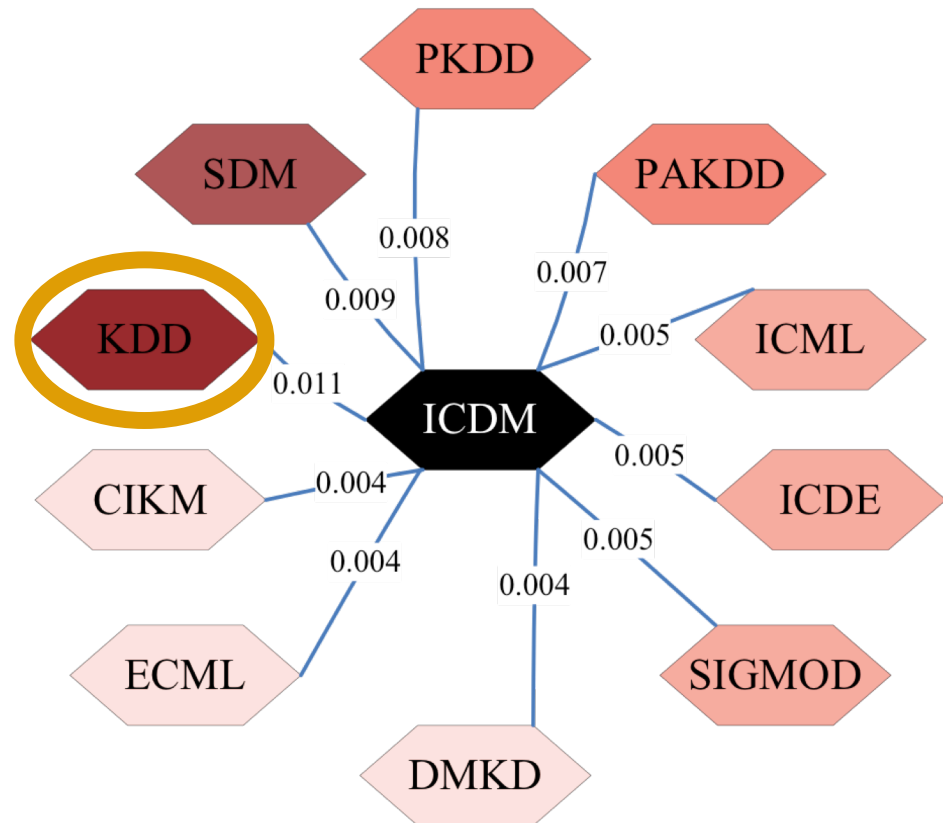


Example

what is the most related conference to ICDM?



Top 10 ranking results



ICDM = international conf. on data mining
KDD = knowledge discovery and data mining



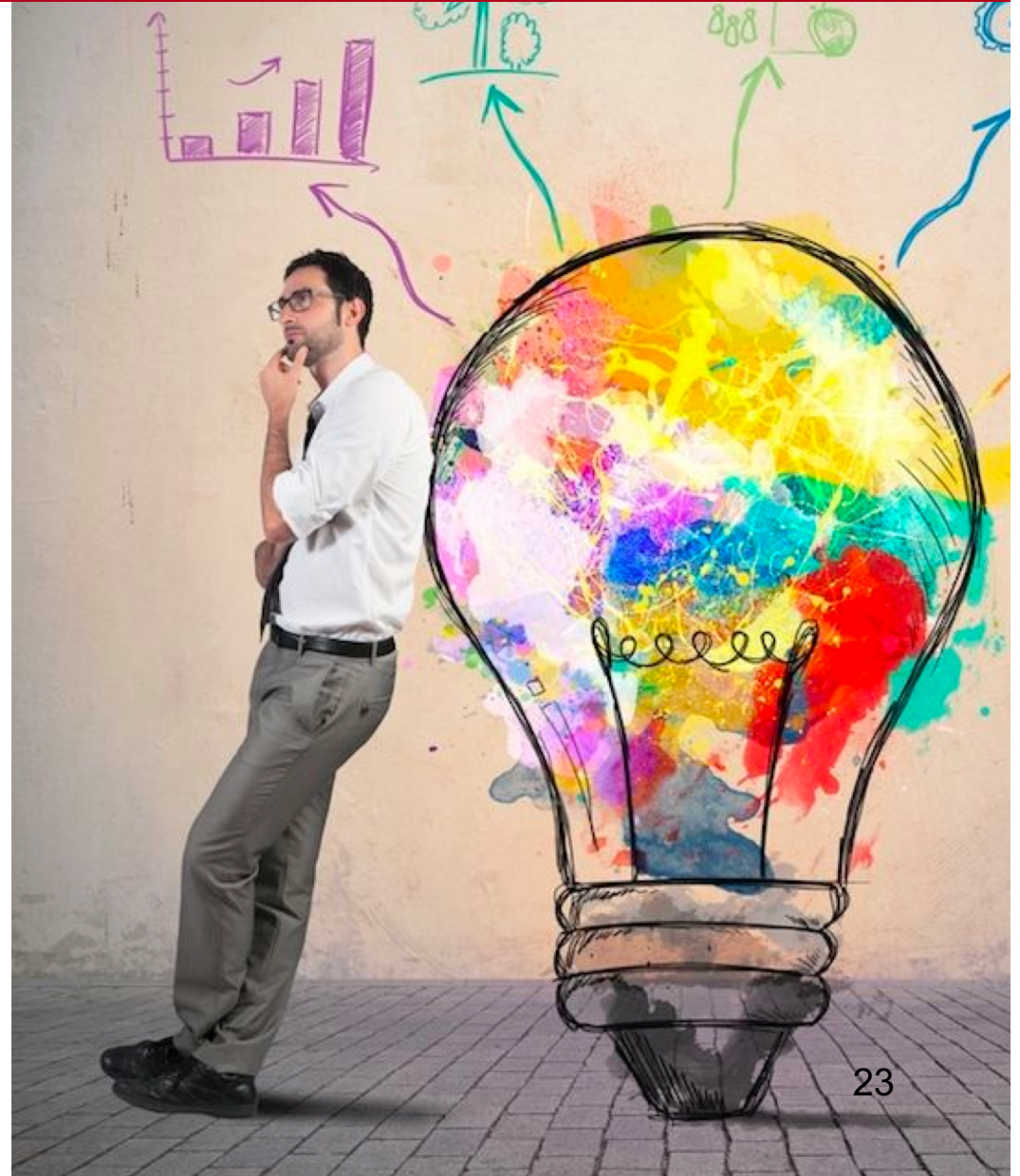
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Measuring closeness to a topic

topic specific PageRank

Want to know about a specific topic? **TopicSpecific** PageRank

PolICY = jump back, at random, to one of the nodes of the topic

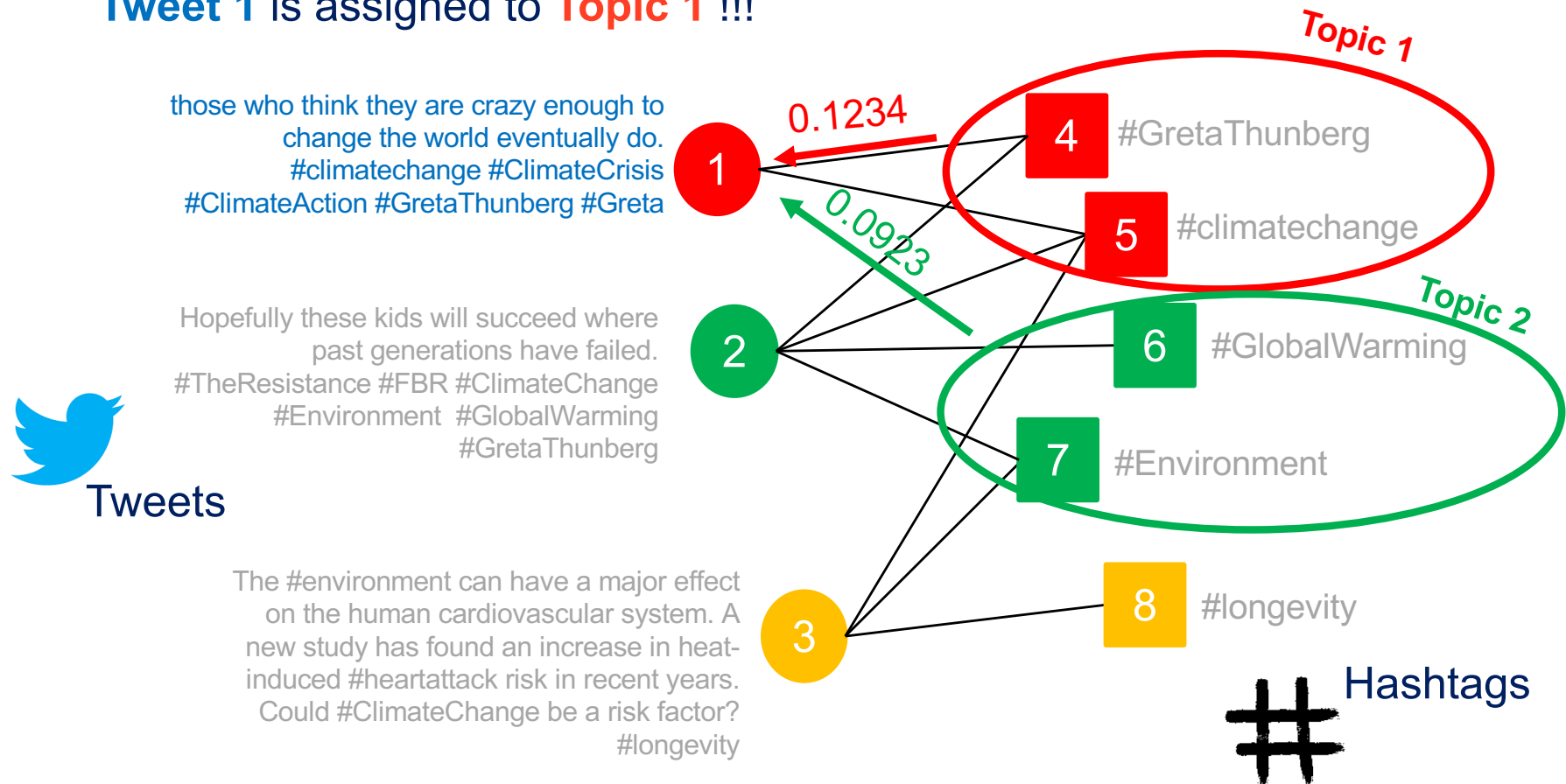




TopicSpecific PageRank example

in semantic networks

Tweet 1 is assigned to **Topic 1** !!!



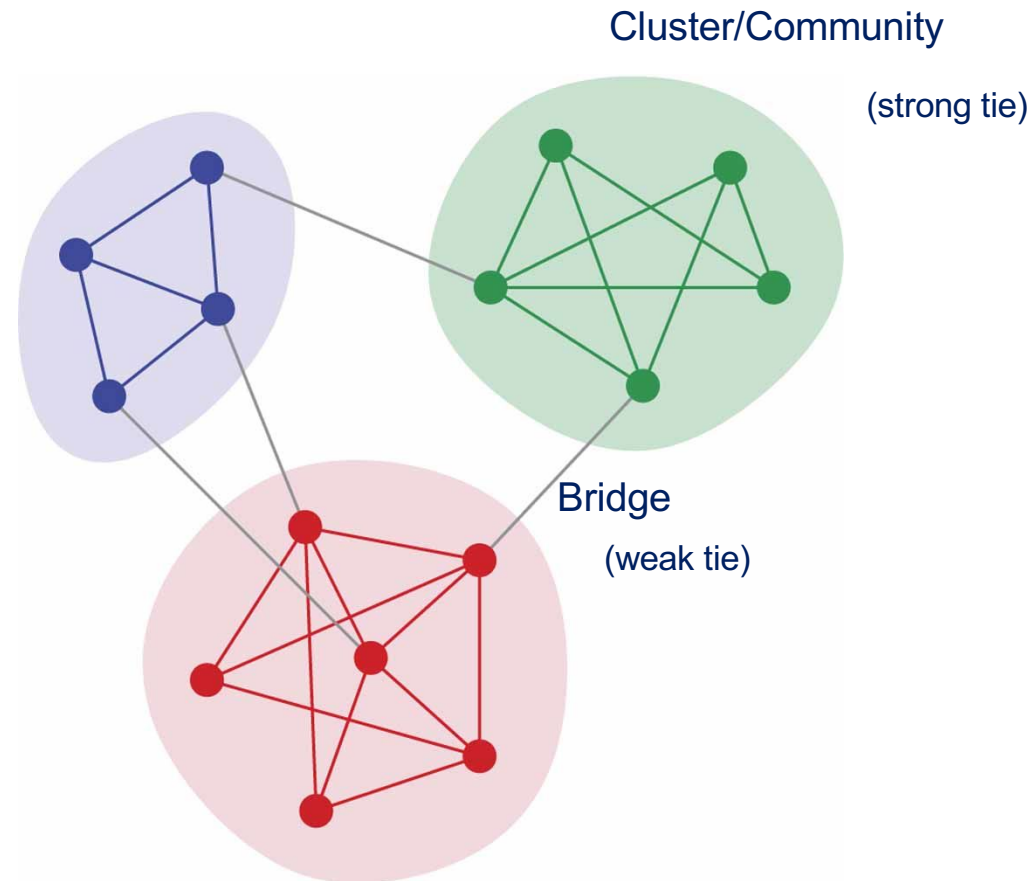
Community detection

and related concepts



Conceptual picture of a network

explaining the role of community detection



- ❑ We often think of **networks** looking like this
- ❑ But, where does this idea come from?



Q: How do people discovered their **new jobs**?

A: Through personal contacts, and mainly through **acquaintances** rather than through close friends

Remark: Good jobs are a scarce resource

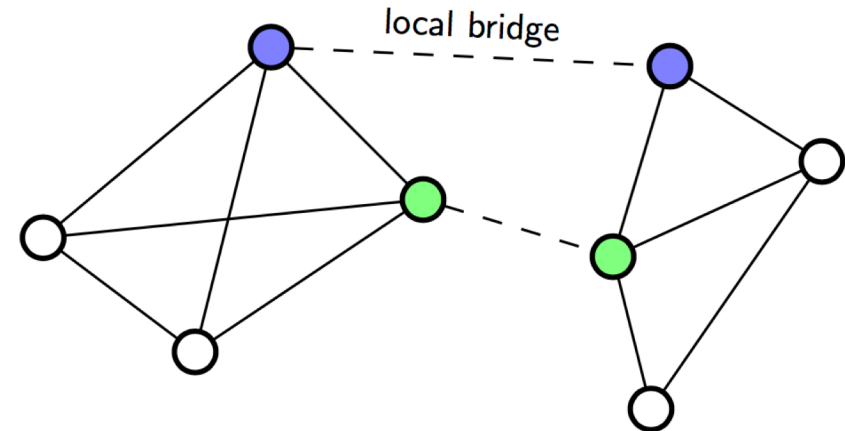
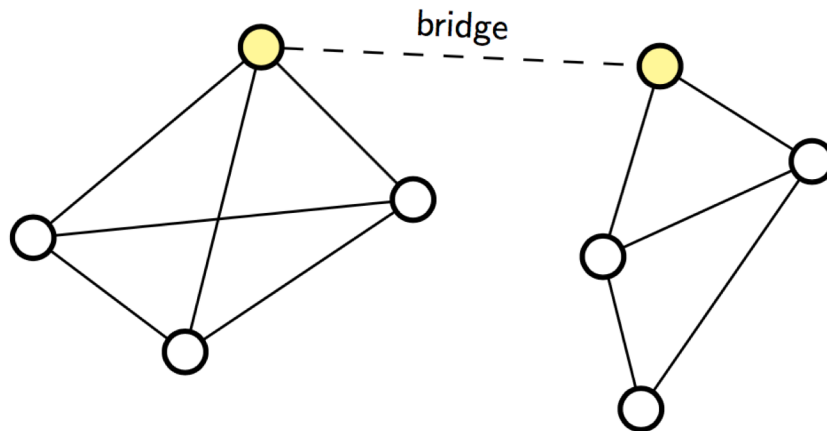
Conclusion:

- ❑ Structurally embedded edges are also socially **strong**, but are heavily redundant in terms of information access
- ❑ Long-range edges spanning different parts of the network are **socially weak**, but **allow you to gather information** from different parts of the network (and get a job)

Local cluster/community
Strong ties

Bridges
Weak ties

Local bridges



- An edge is a **bridge** if deleting it *the nodes it connects* fall into different components

this is extremely rare, e.g., because of small world properties

- An edge is a **local bridge** if, by deleting it, *the nodes it connects* have a span (distance) greater than 2, i.e., if *they do not have friends in common*

common friends imply belonging to a triadic closure



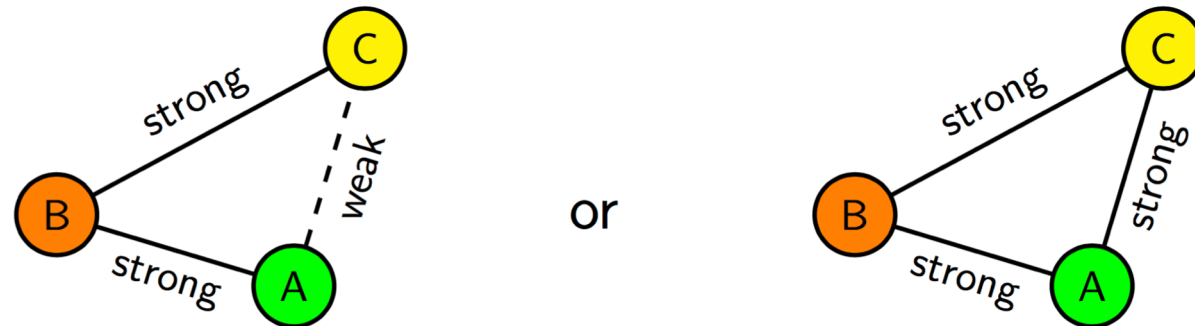
Strong triadic closure

friends/relatives and acquaintances

Assume two categories of edges:

- ❑ **Strong ties** (close friends)
- ❑ **Weak ties** (acquaintances)

Remark. If node B is strongly tied with A and C, then A and C are very likely to be connected (either weakly or strongly), that is



Strong triadic closure property – If a generic node B is strongly tied with A and C, then A and C are connected (either weakly or strongly)



Granovetter's claim

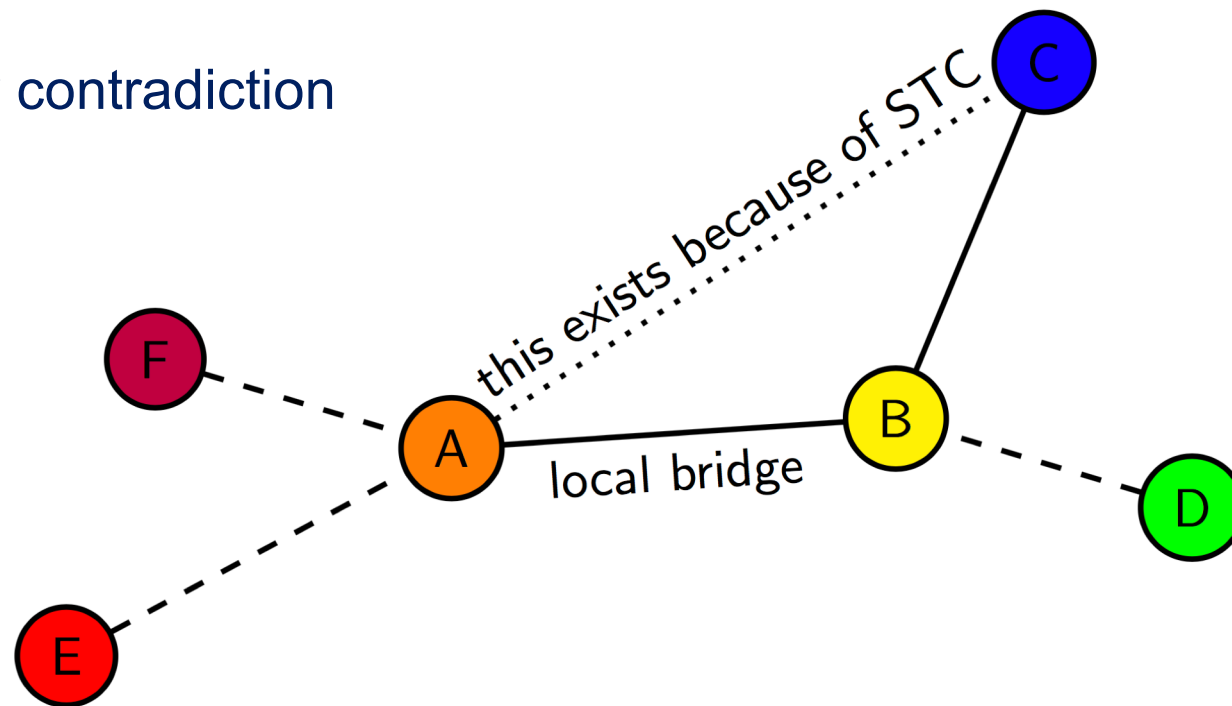
under strong triadic closure

Claim:

- Under the **strong triadic closure** property, **local bridges** are **weak ties** (if at least one of their nodes belongs to at least two strong ties)

Proof:

- By contradiction

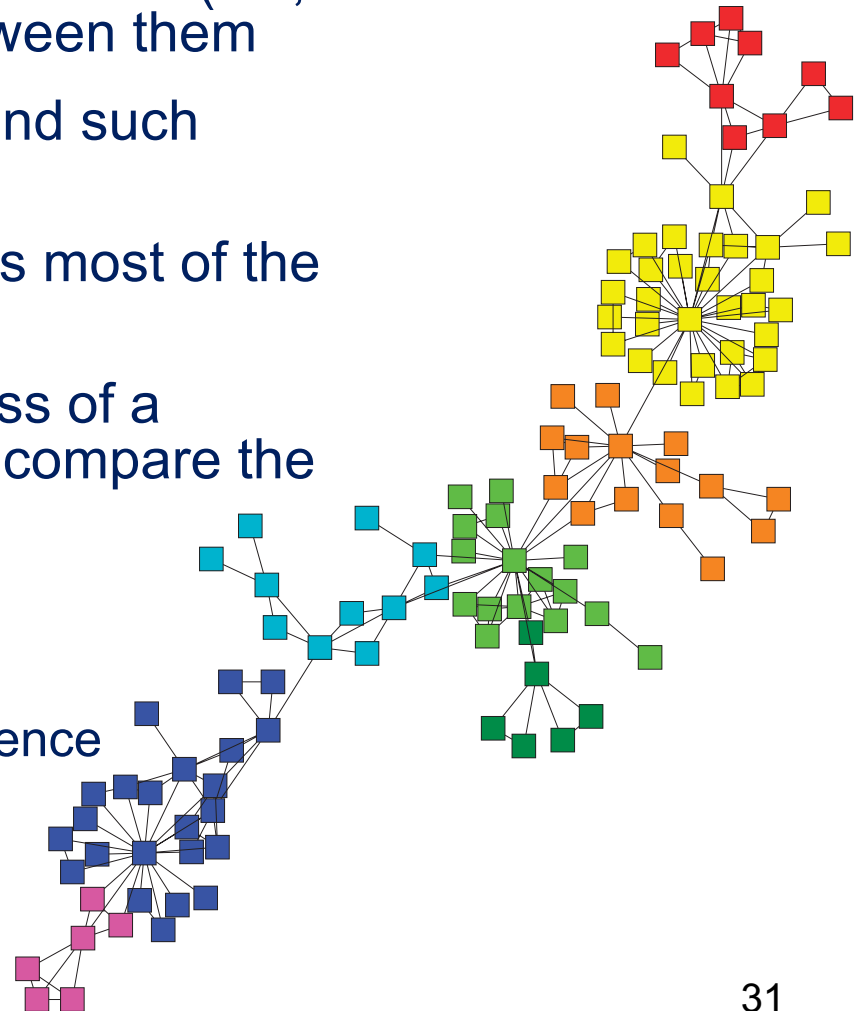




Community detection

the general approach

- ❑ Granovetter's theory suggests that networks are composed of **tightly connected sets of nodes** (i.e., communities), loosely connected between them
- ❑ We want to be able to **automatically** find such densely connected group of nodes
- ❑ We look for **unsupervised** methods, as most of the times no ground truth is available
- ❑ We look for a **measure** of the goodness of a community assignment, to be able to compare the performance of different algorithms
- ❑ Applications in:
 - social networks
 - functional brain networks in neuroscience
 - scientific interactions



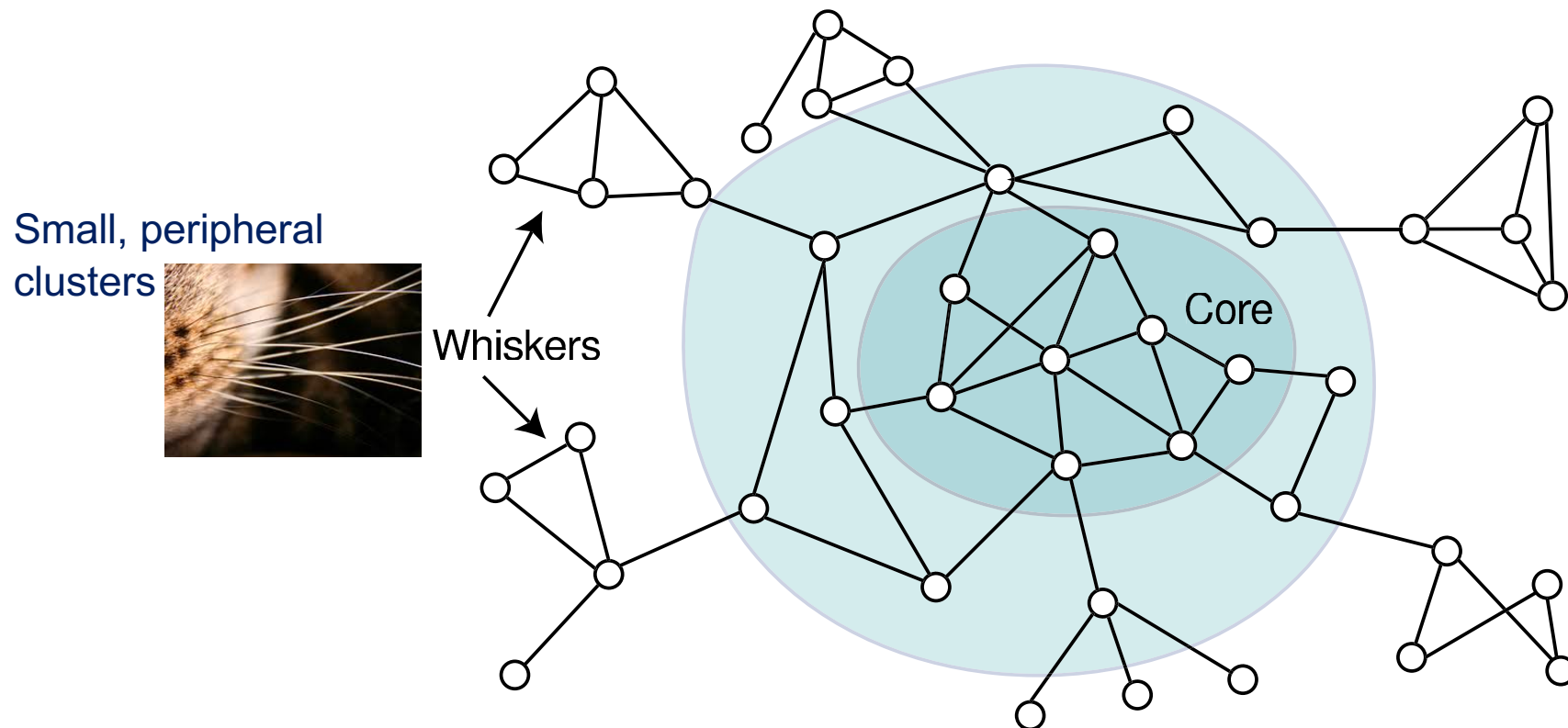


The core periphery model

Lescovec, Lang, Dasgupta, Mahoney, Community Structure in Large Networks:
Natural Cluster Sizes and the Absence of Large Well-Defined Clusters (2008)

<https://arxiv.org/abs/0810.1355>

Can we find a justification for this?

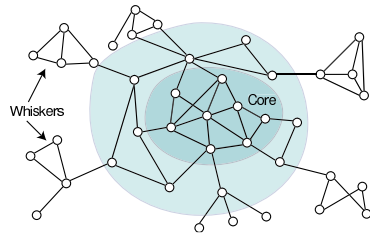


Caricature of network structure



Overlapping communities

to explain the core periphery model

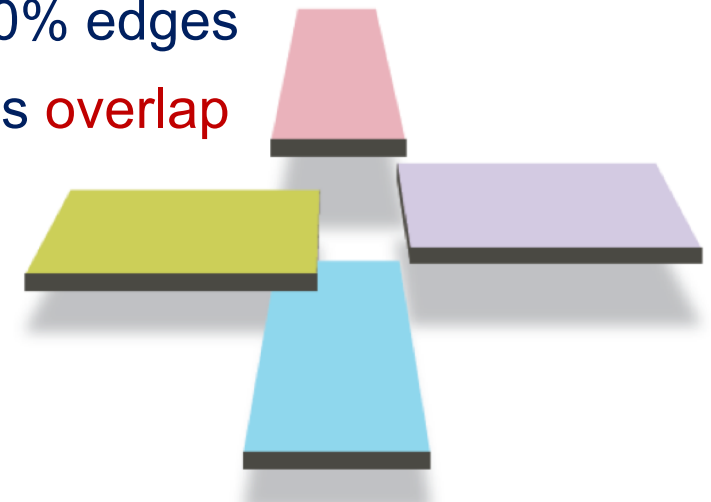
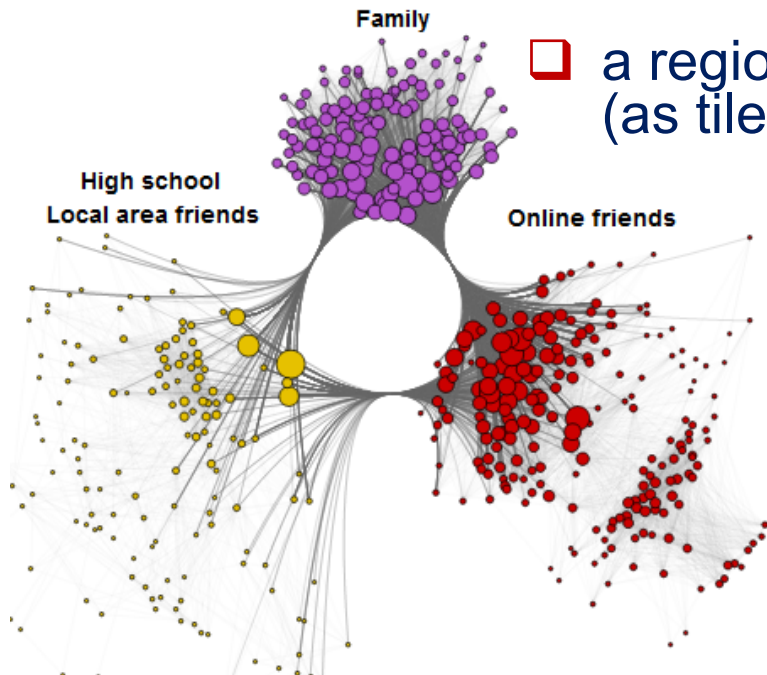


Whiskers

- are typically of size 100
- are responsible of **good** communities

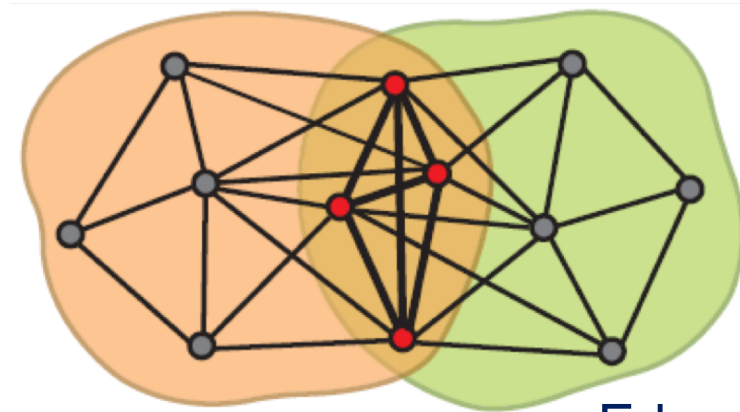
Core

- denser and denser region
- contains 60% nodes and 80% edges
- a region where communities **overlap** (as tiles)

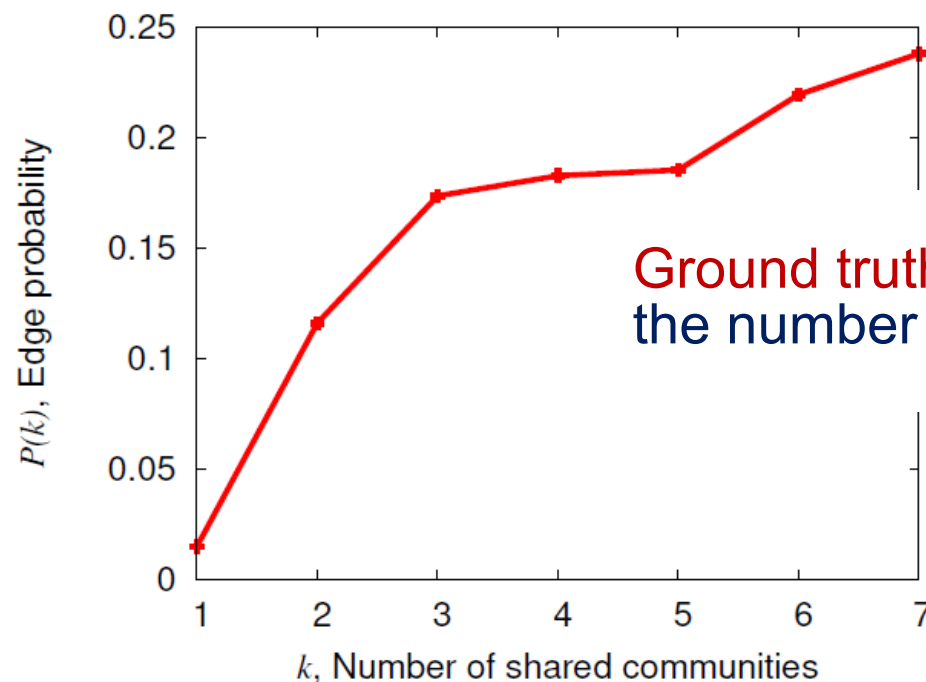


Measuring overlapping

in social networks



Edge density is
bigger in the
overlap



Ground truth - Edge probability increases with
the number of shared communities

Feld, The focused organization of social ties, [1981]
The more different communities that two individuals
share, the more likely is that they will be tied

Amazon

Algorithms

for community detection



Want to:

- ❑ measure of **how well** a network is **partitioned** into communities (i.e., sets of tightly connected nodes)

Idea:

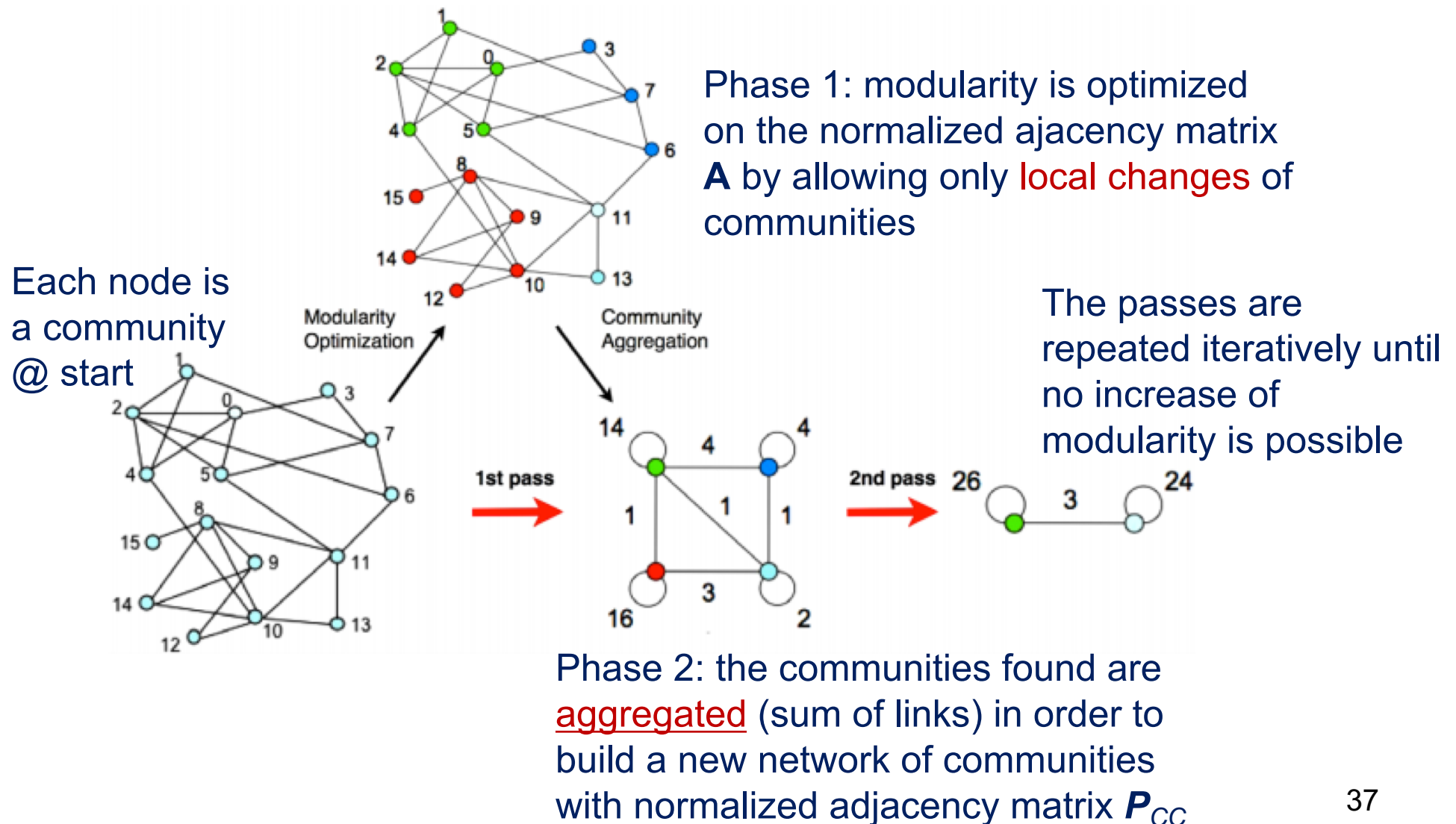
- ❑ “If the number of edges between two groups is only what one would expect on the basis of random chance, then few thoughtful observers would claim this constitutes evidence of meaningful community structure”
- ❑ **Modularity** is “the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random”
- ❑ The higher modularity, the better the community assignment



The Louvain algorithm

Blondel, Guillaume, Lambiotte, Lefebvre, Fast unfolding
of communities in large networks (2008)

<https://arxiv.org/abs/0803.0476>



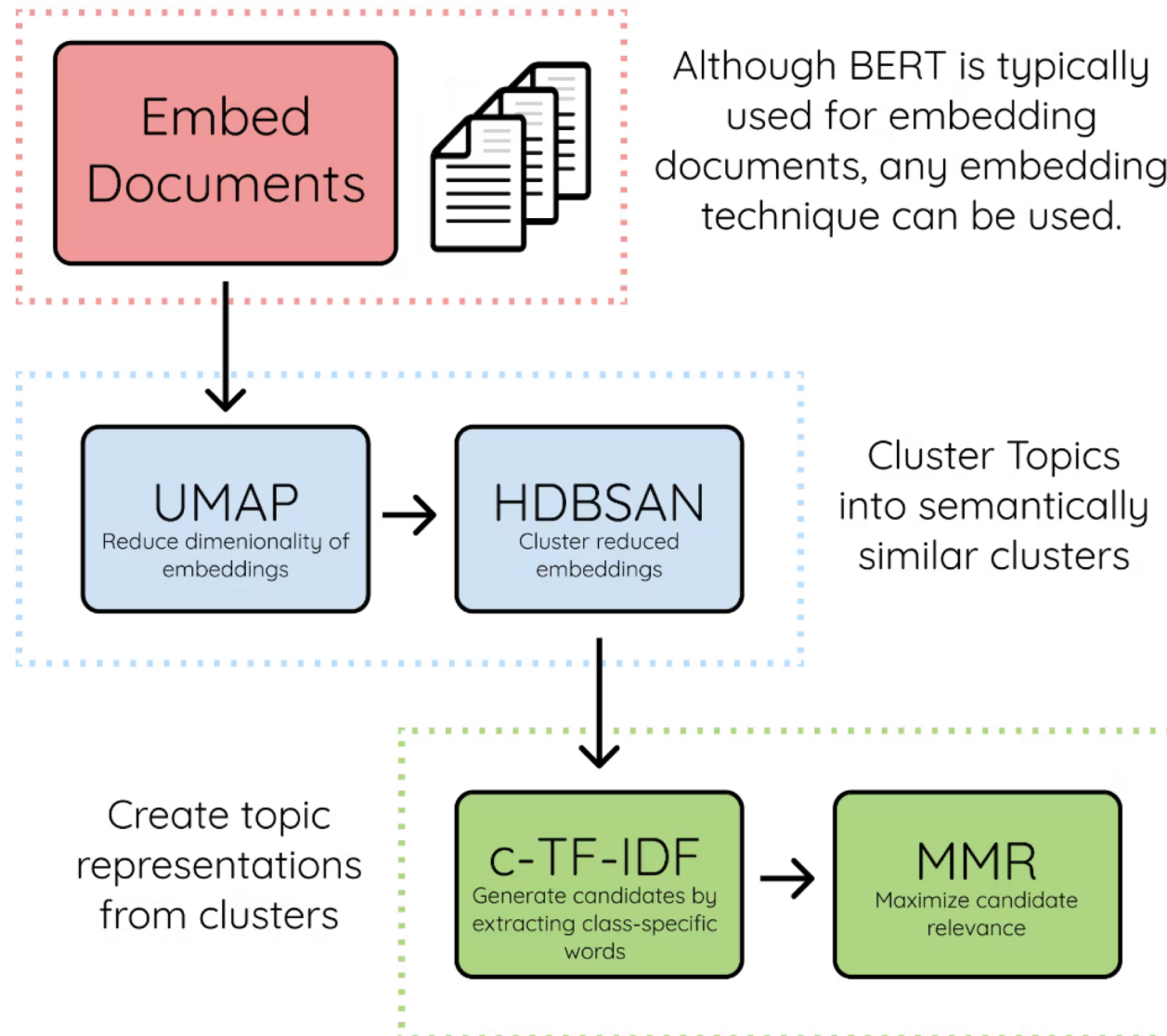


- ❑ Implements modularity optimization
- ❑ Scalable (low complexity)
- ❑ Effective
- ❑ Available as the **reference** implementation in any programming language
- ❑ A greedy technique (in the order the nodes are searched)



BERTopic

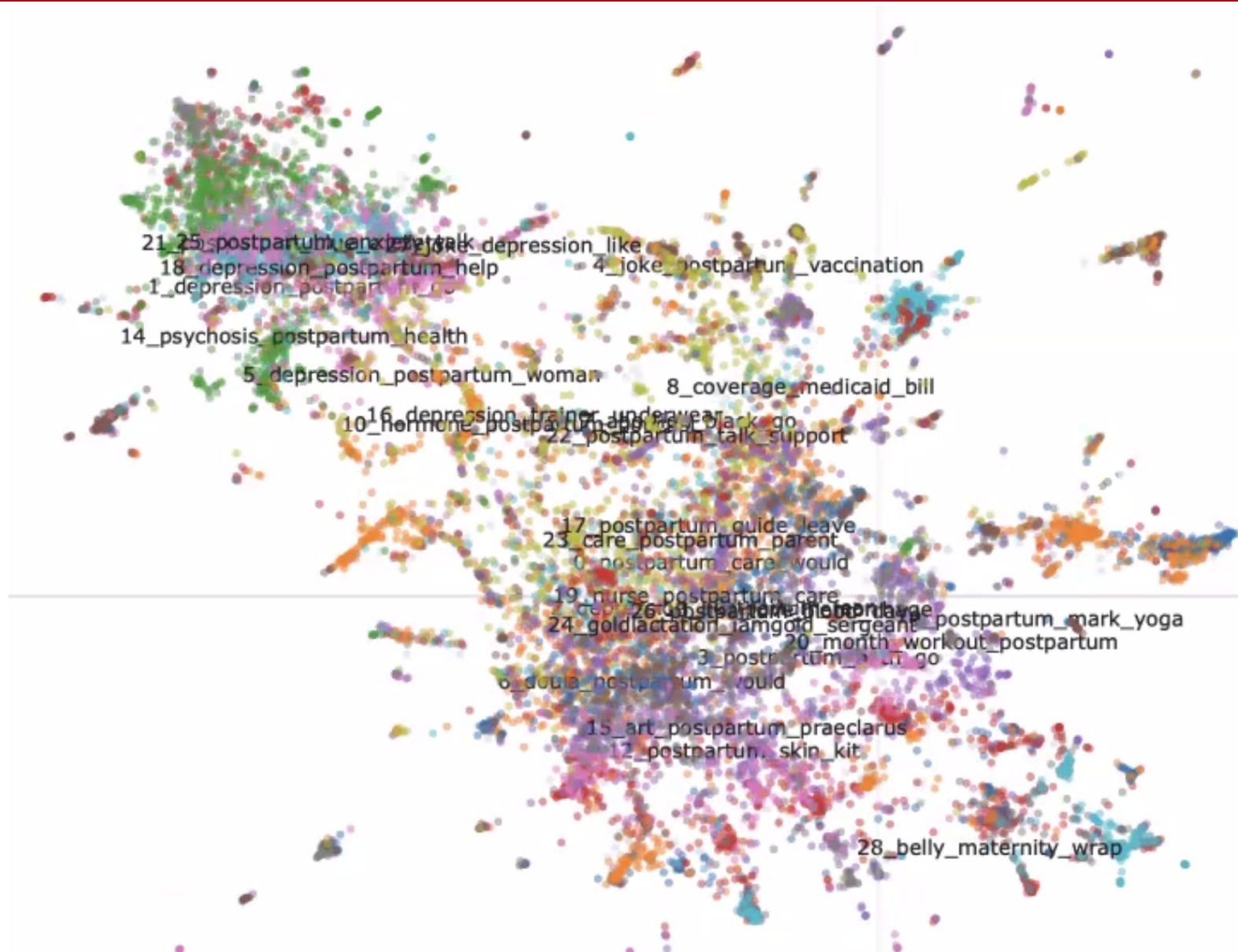
an algorithm based on ChatGPT architecture





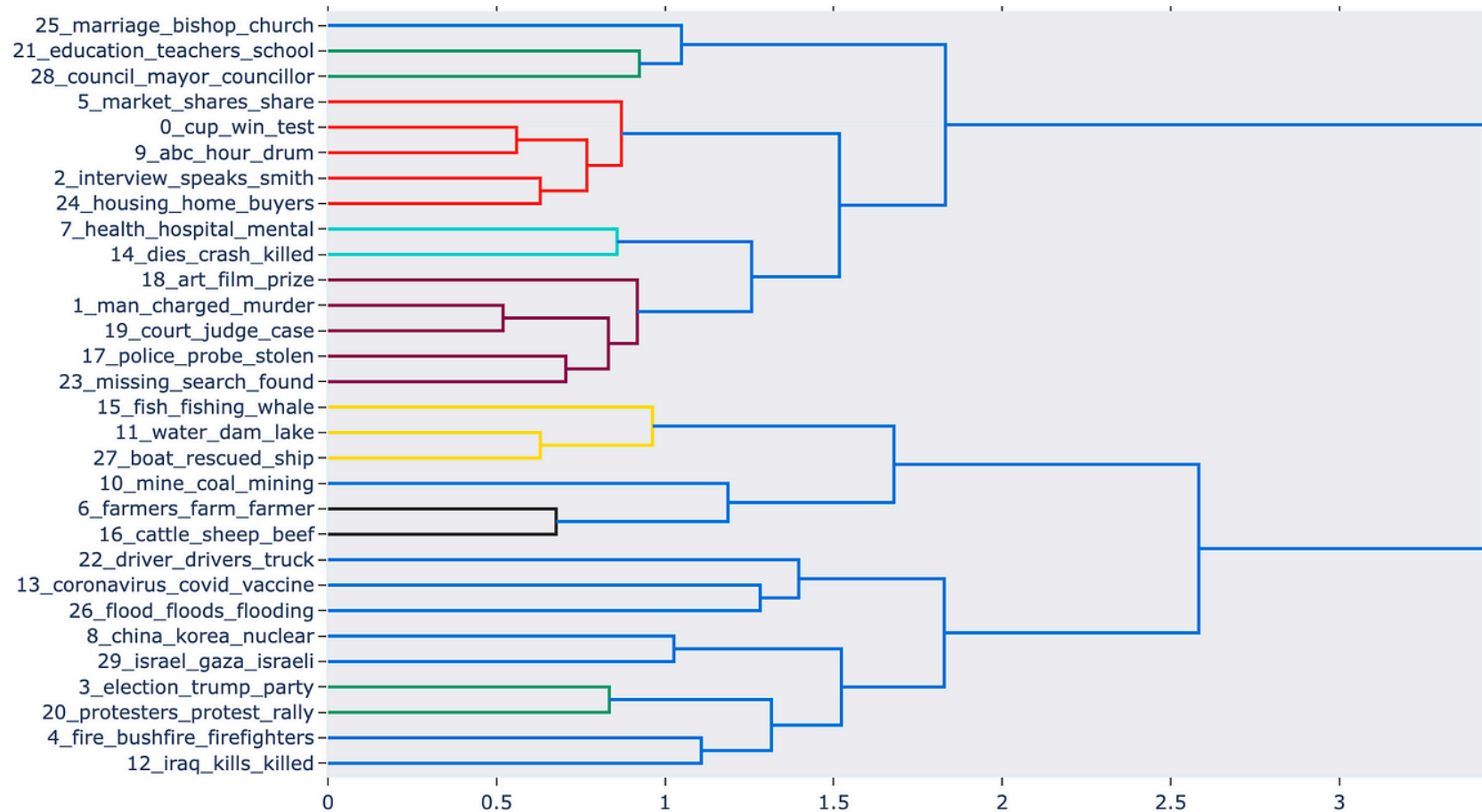
A visual example

document network using BERTopic on postpartum tweets





HDBSCAN hierarchy of topics, with those selected

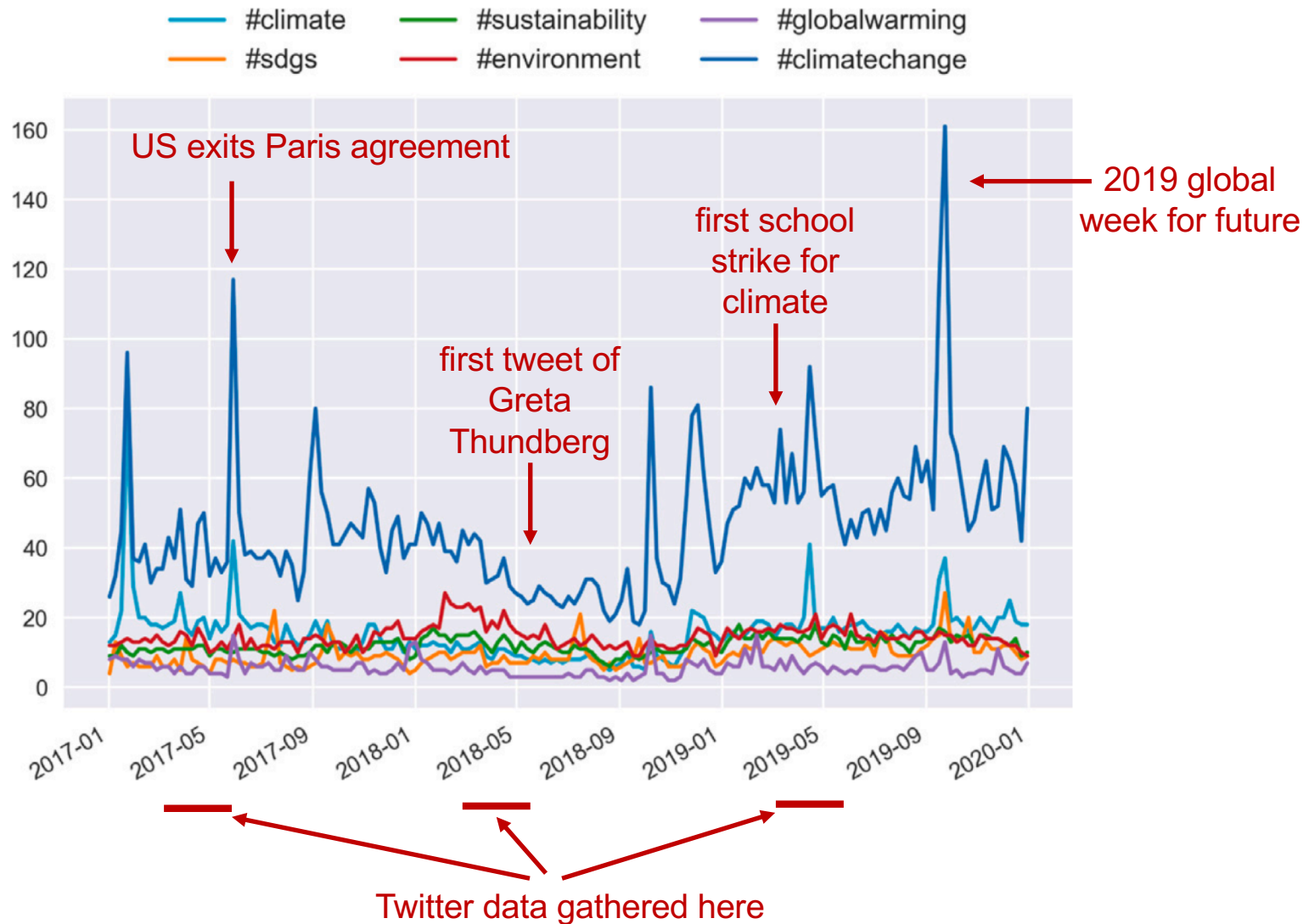


Using community detection

an overview of what it can be useful for



#climateaction on Twitter in 2017, 2018, 2019



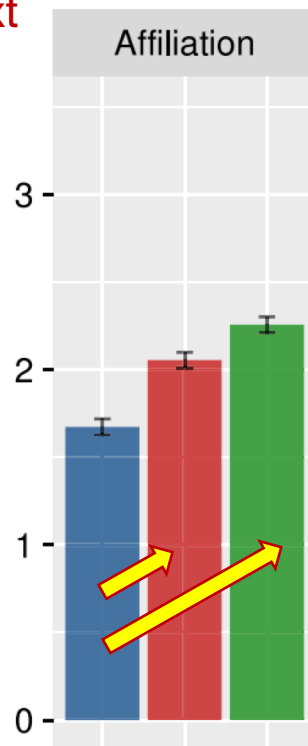


Socio-psychological linguistic markers

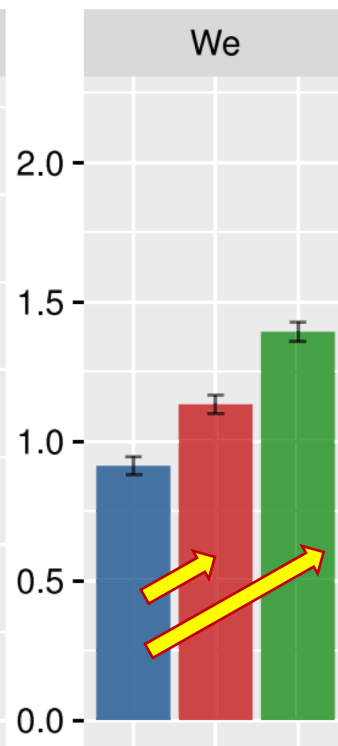
a view on the entire tweets corpus

2017 2018 2019

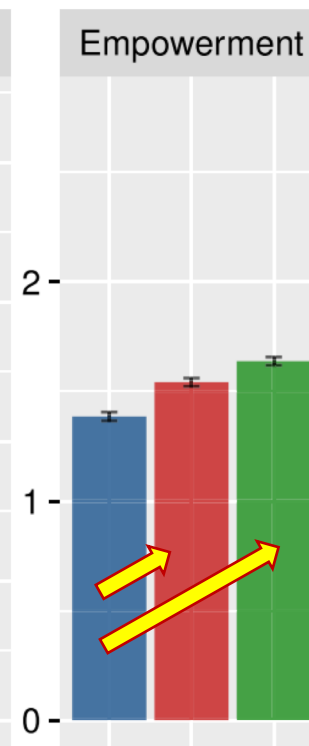
ingroup community
orientation within
the text



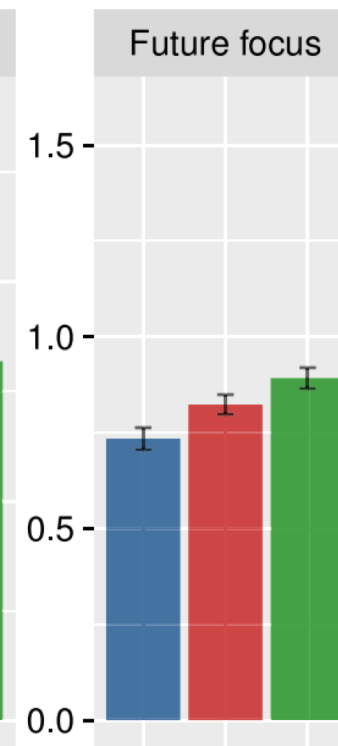
salience of
group
membership,
sense of
belonging



a person's
striving to be
independent to
assert, protect
and expand
one's self



orientation of tweets to the
past or future



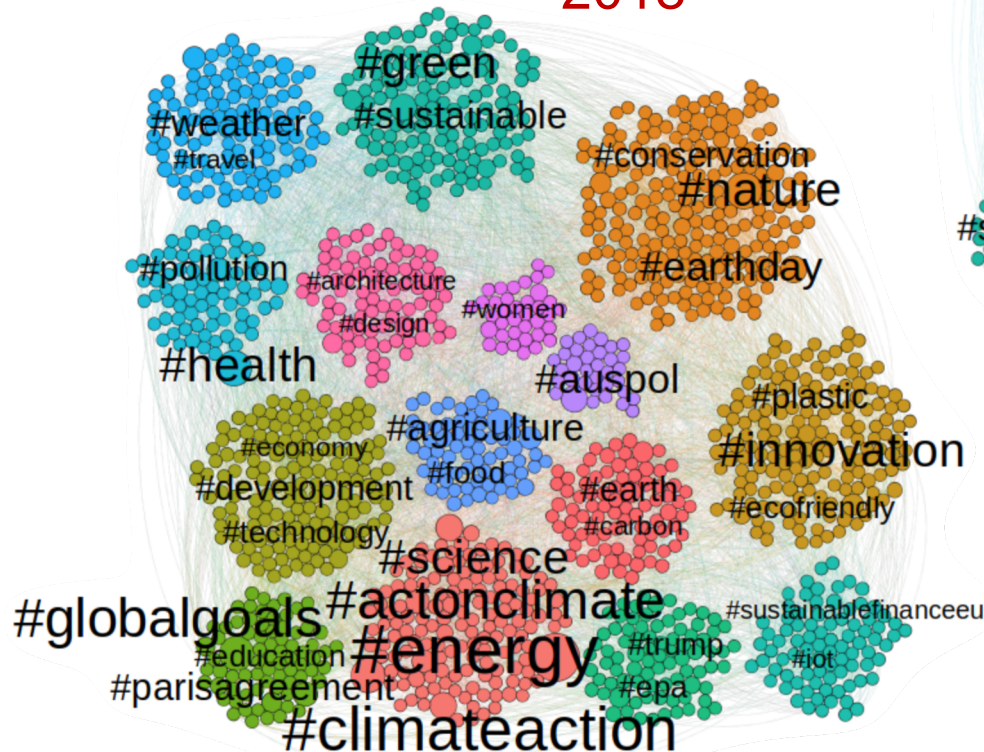
only a few
statistically
relevant
changes



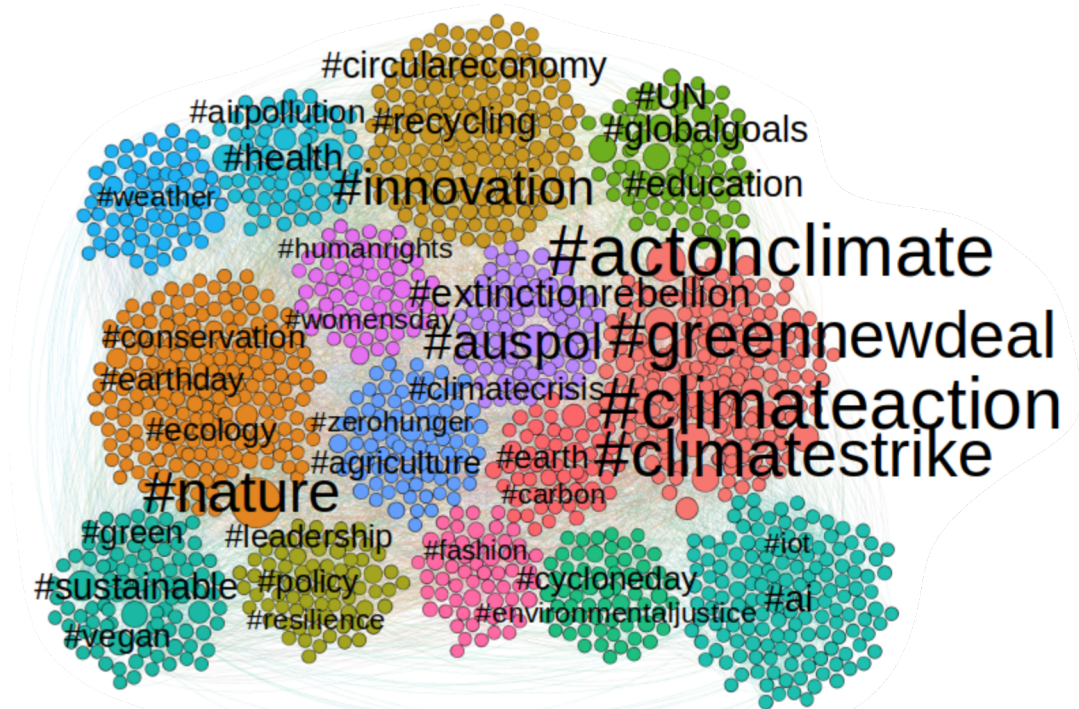
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Communities/topics in #climateaction on Twitter in 2017, 2018, 2019

2018

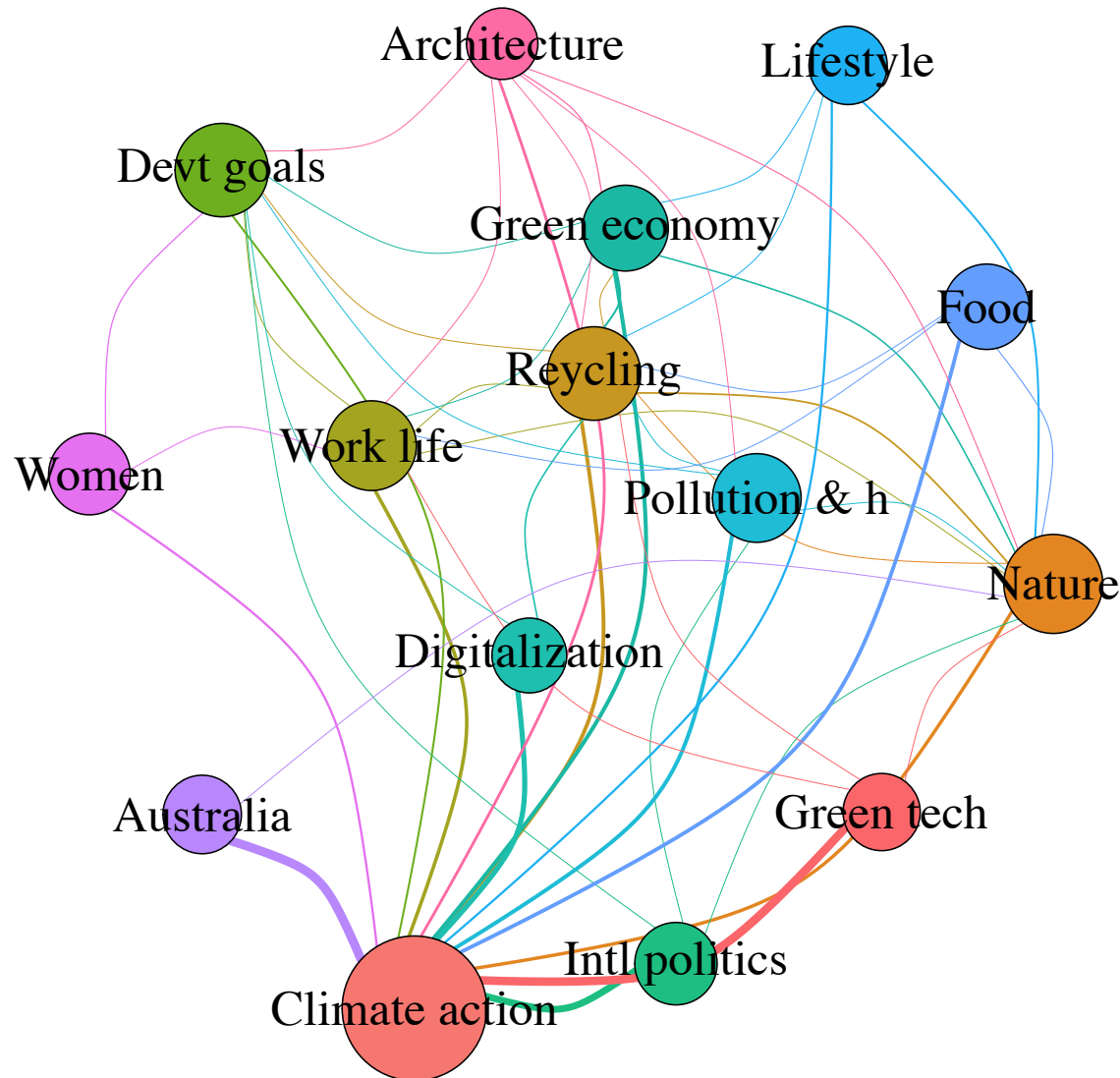


2019





Topics interdependencies



projecting the
adjacency matrix on
topics

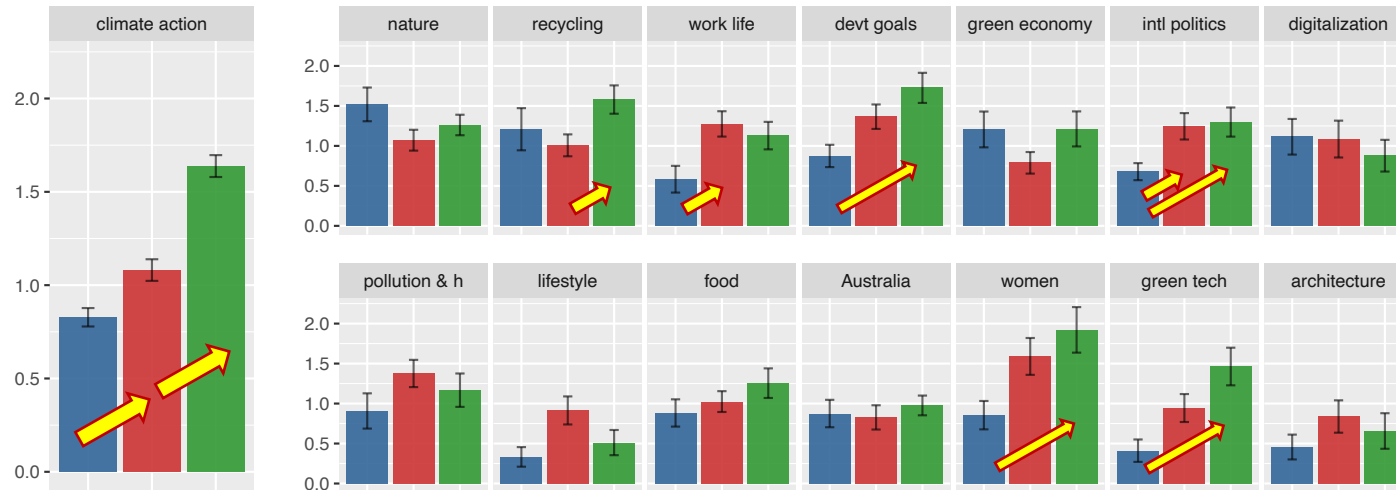
| | | |
|----------|----------|----------|
| P_{11} | P_{12} | P_{13} |
| P_{21} | P_{22} | P_{23} |
| P_{31} | P_{32} | P_{33} |



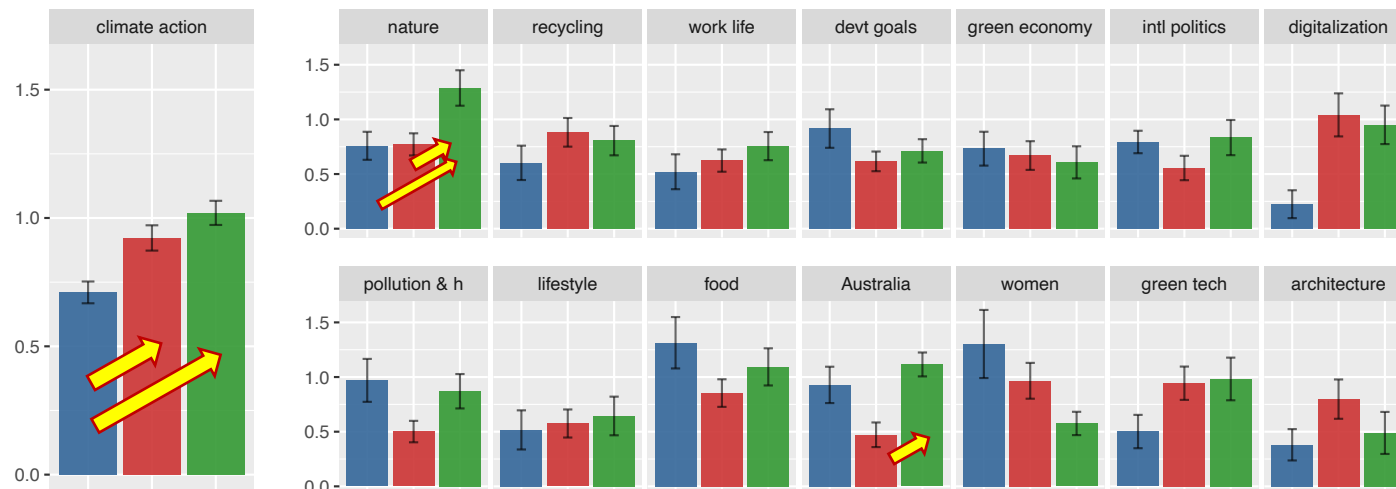
Socio-psychological linguistic markers a view inside topics

■ 2017 ■ 2018 ■ 2019

(b) We



(d) Future focus



relevant
statistically
changes of
we-future
only in the
climate
action
community

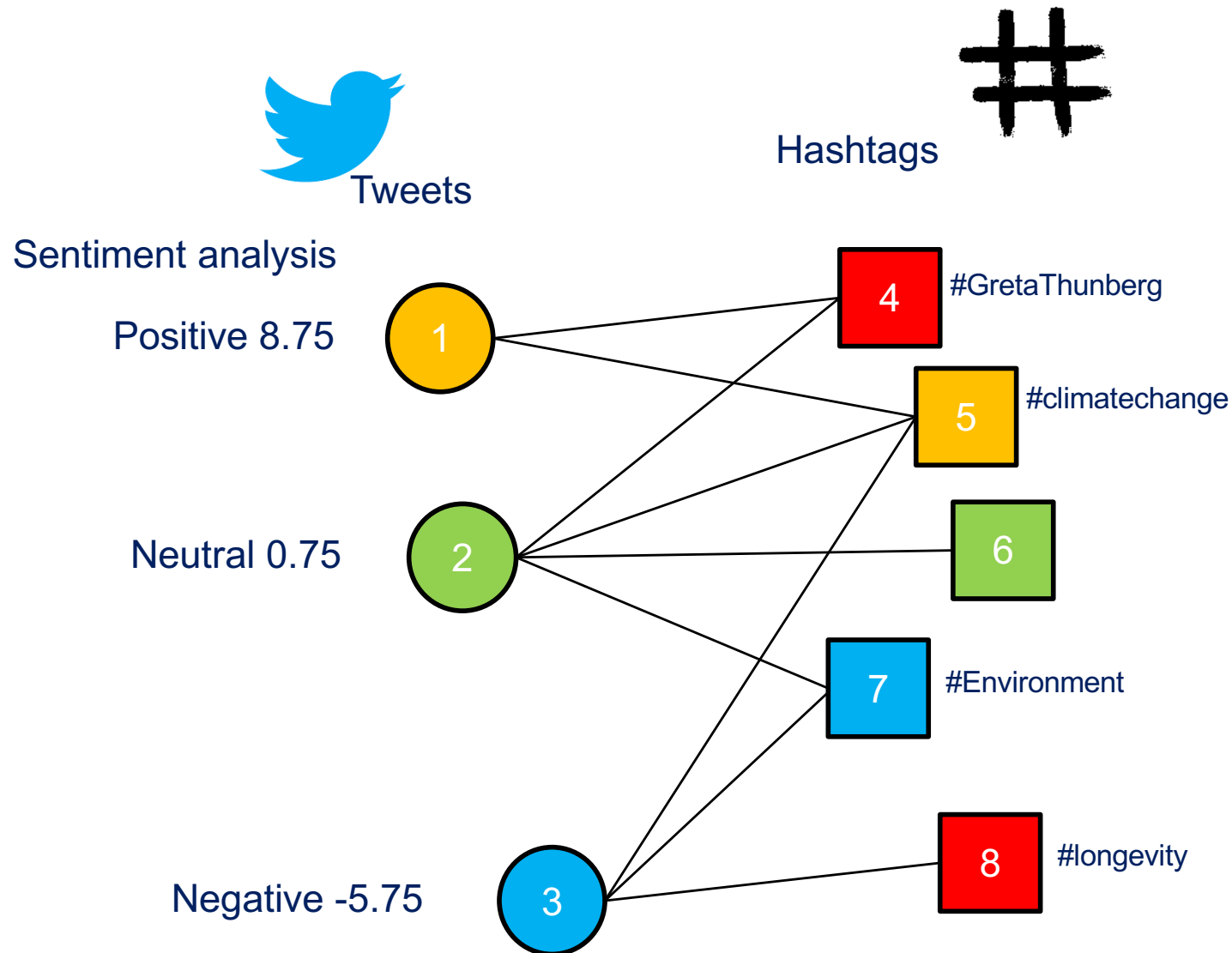
Projecting markers

on specific words, and their application



The question

how to project sentiment information from tweets to hashtags?

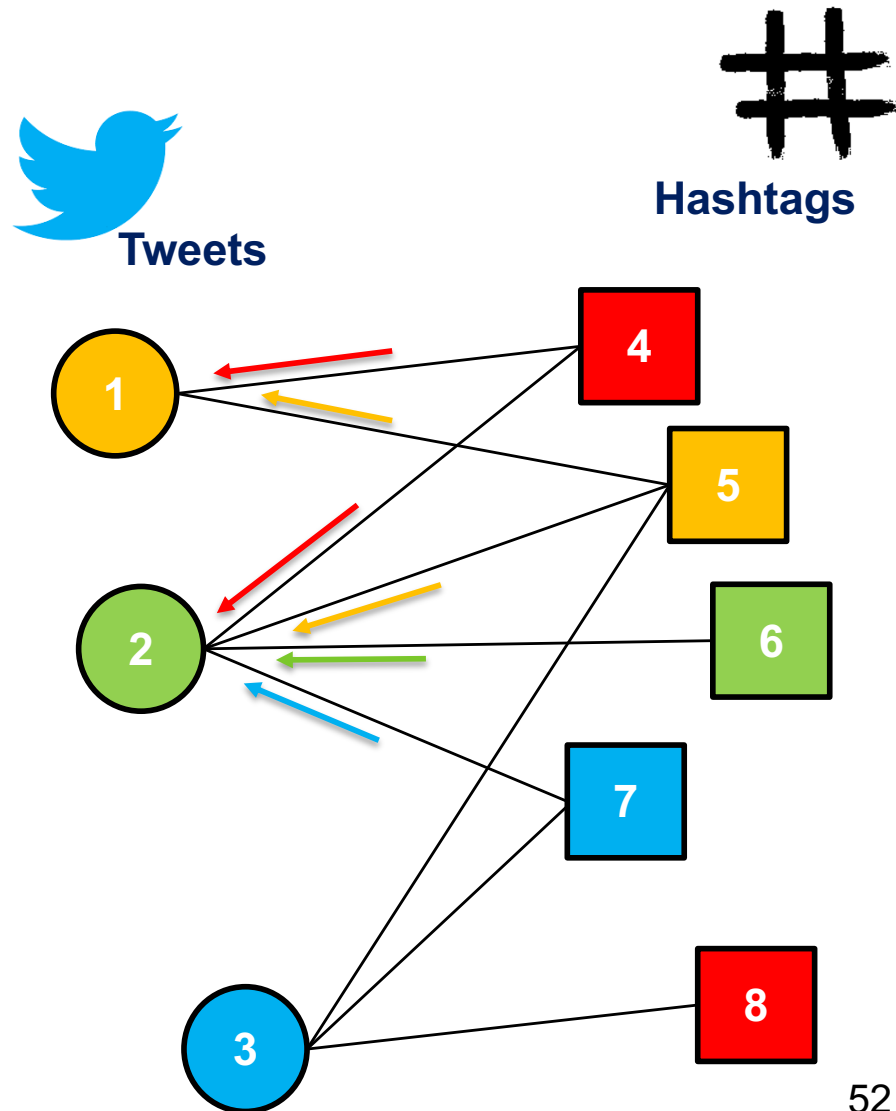


PLMP projection

a PageRank like approach

- Each hashtag captures the **average sentiment value** of the tweets it appears in
- Each tweet captures the **average sentiment** of the hashtags it contains

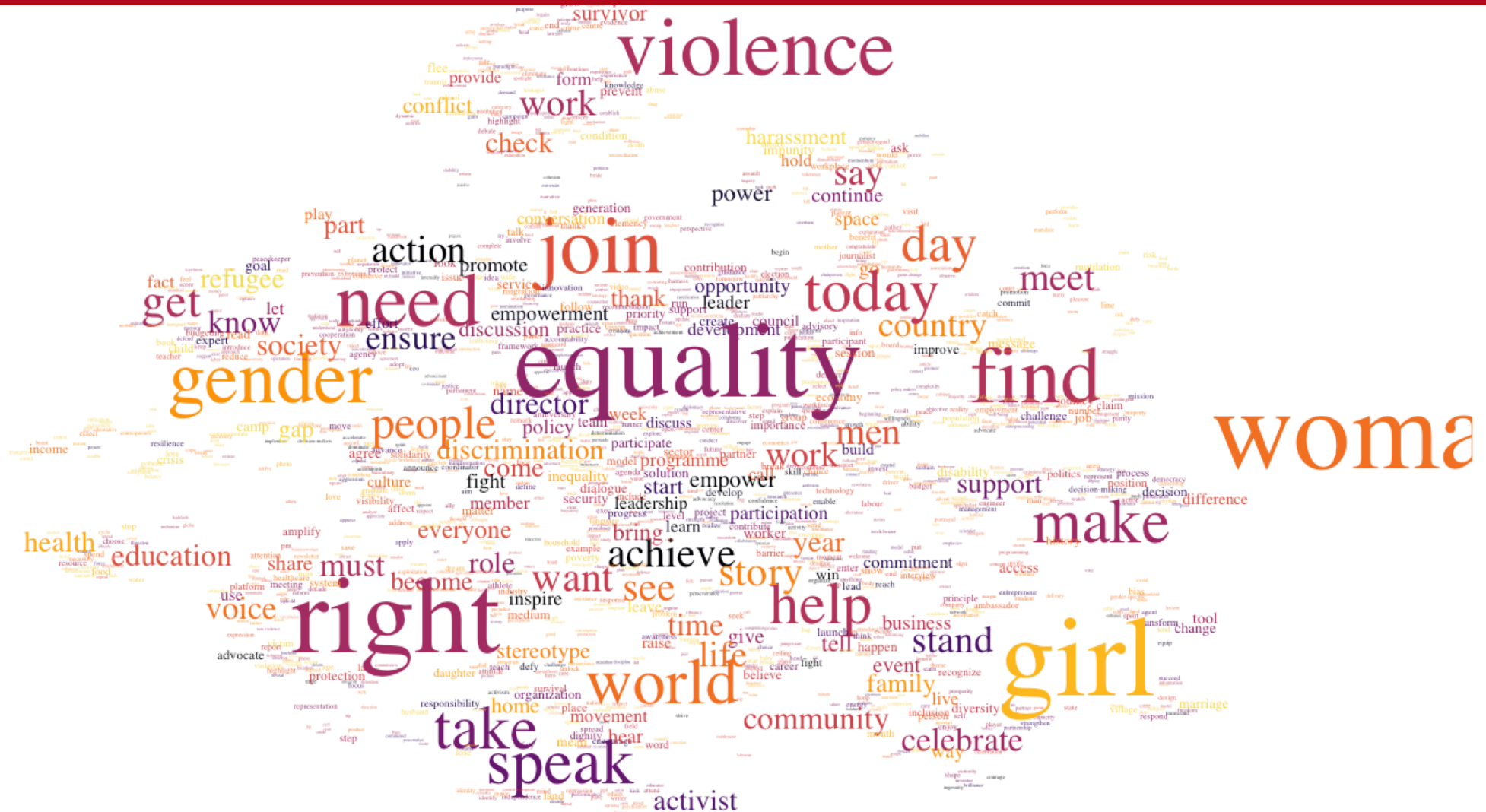
we iterate the two steps until convergence





UNIVERSITÀ
DEGLI STUDI
DI PADOVA

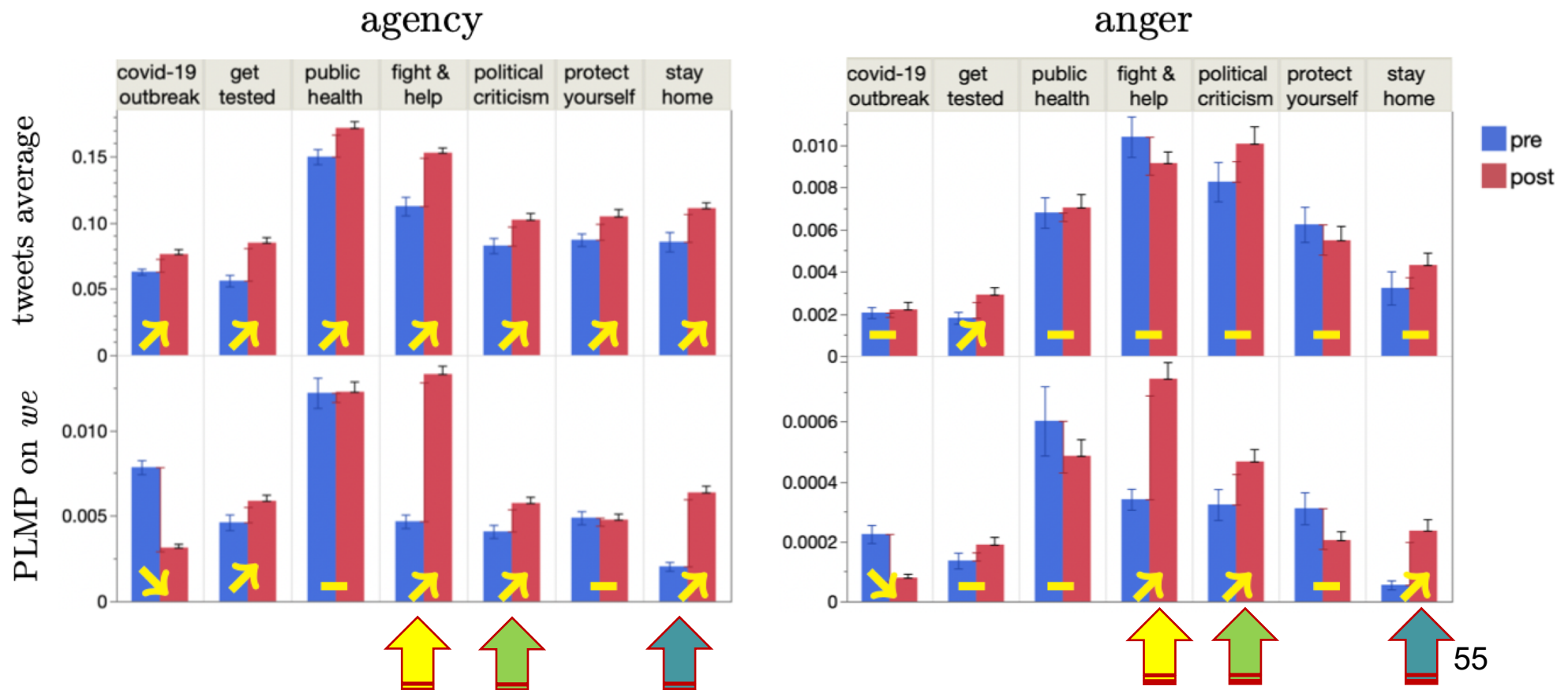
PLMP agency values





Projection of agency and anger on the target word we – #covid19

projection on the target word «we» characterises the social development of the online discourse over time and across specific topics, and capture trends that cannot be spotted without the projection





- ❑ Centrality by PageRank
- ❑ Closeness by Local PageRank
- ❑ Community detection in semantic networks is topic detection
- ❑ Usefulness of communities
- ❑ Projecting marker values for deeper insights