UNIVERSITÀ DEGLI STUDI DI PADOVA

Social Network Analysis

A.Y. 23/24

Communication Strategies

PageRank

a centrality measure based on the web

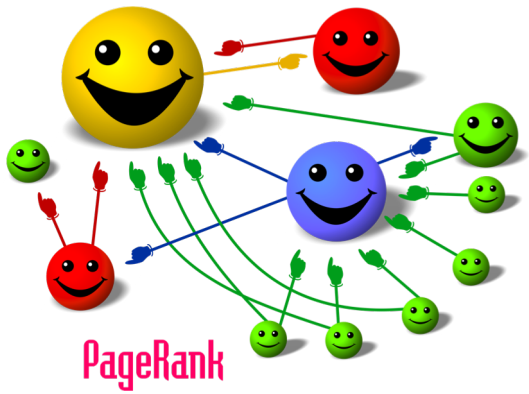
What is PageRank?

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]

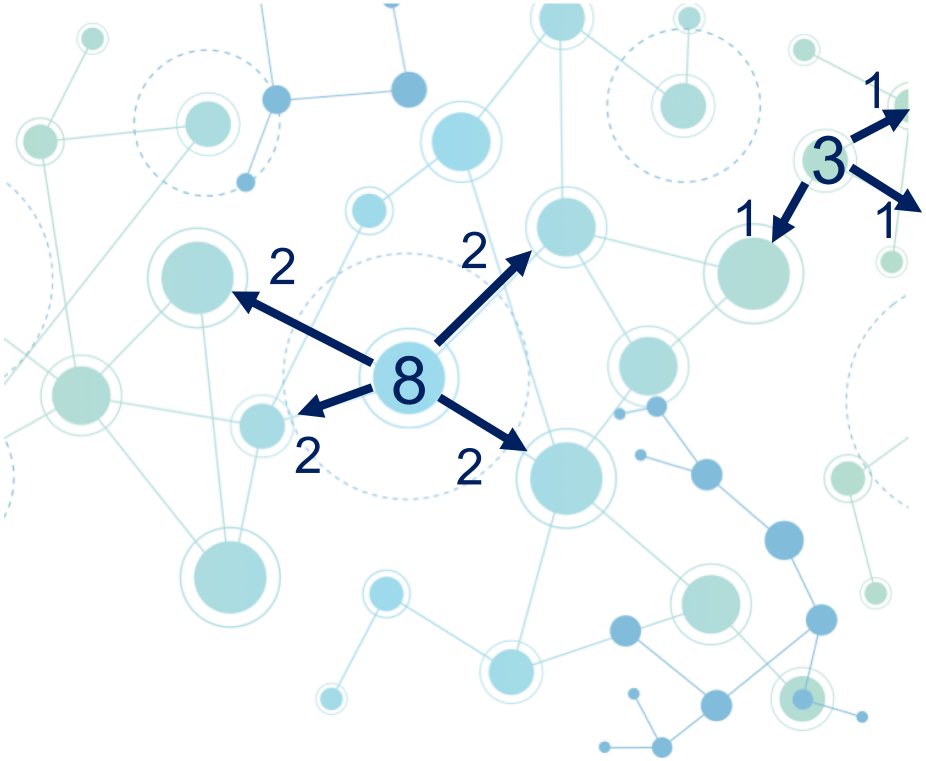
How to organise the web?

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

PageRank idea
network flow

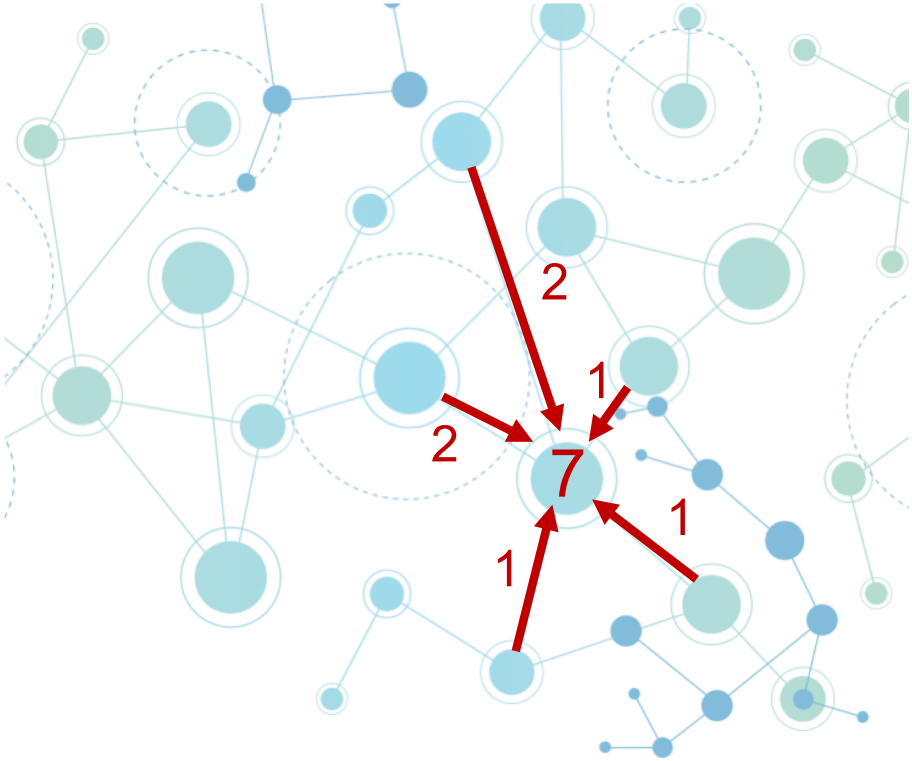
Step 1: spread (evenly) information (on centrality) from each node



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

PageRank idea
network flow

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



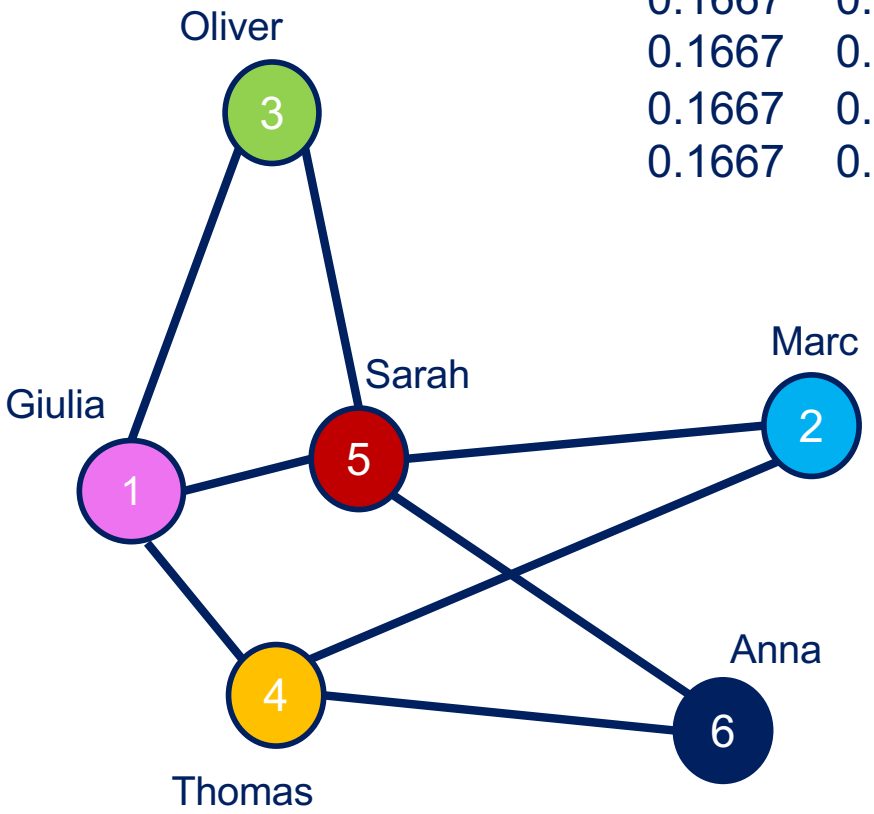
in the web this roughly corresponds to the chance (probability) of ending in a specific web page

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 strenghten the result

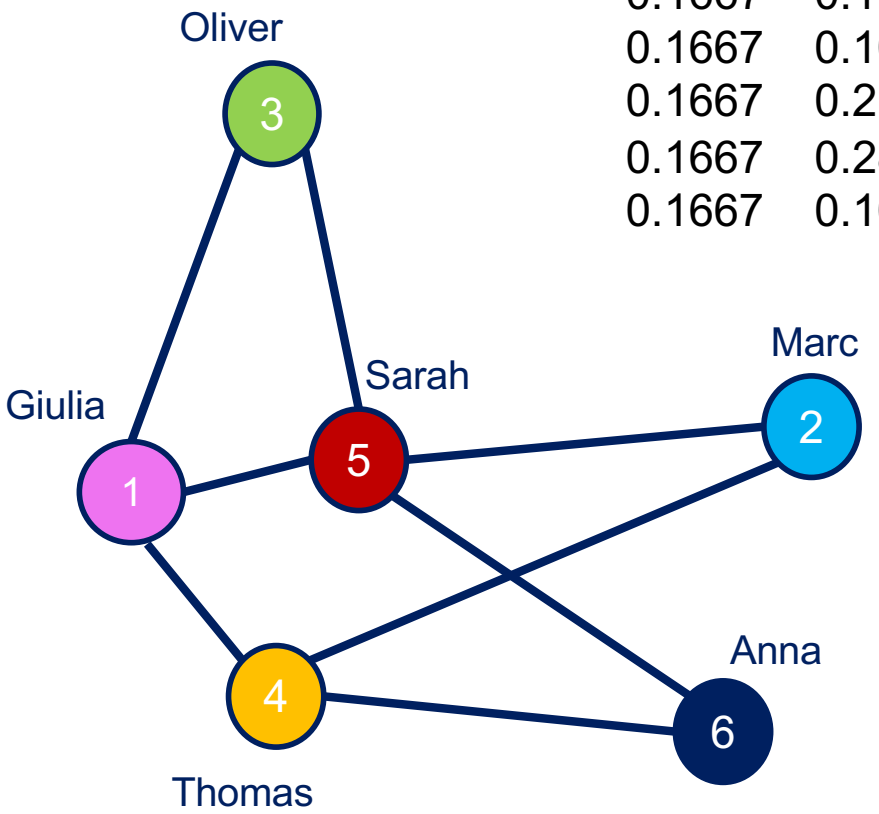
Idea:

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



t=1	*2*	*3*	*4*	*5*
0.1667	0.1785	0.1919	0.1754	0.1912
0.1667	0.1076	0.1461	0.1176	0.1382
0.1667	0.1076	0.1361	0.1246	0.1302
0.1667	0.2139	0.1671	0.2035	0.1746
0.1667	0.2847	0.2128	0.2614	0.2276
0.1667	0.1076	0.1461	0.1176	0.1382

not anymore identical to degree centrality !!!

10	*20*	*50*	*75*	*100*	
0.1820	0.1839	0.1840	0.1840	0.1840	Giulia
0.1273	0.1293	0.1294	0.1294	0.1294	Marc
0.1283	0.1285	0.1285	0.1285	0.1285	Oliver
0.1902	0.1873	0.1871	0.1871	0.1871	Thomas
0.2449	0.2419	0.2417	0.2417	0.2417	Sarah
0.1273	0.1293	0.1294	0.1294	0.1294	Anna

Discussion

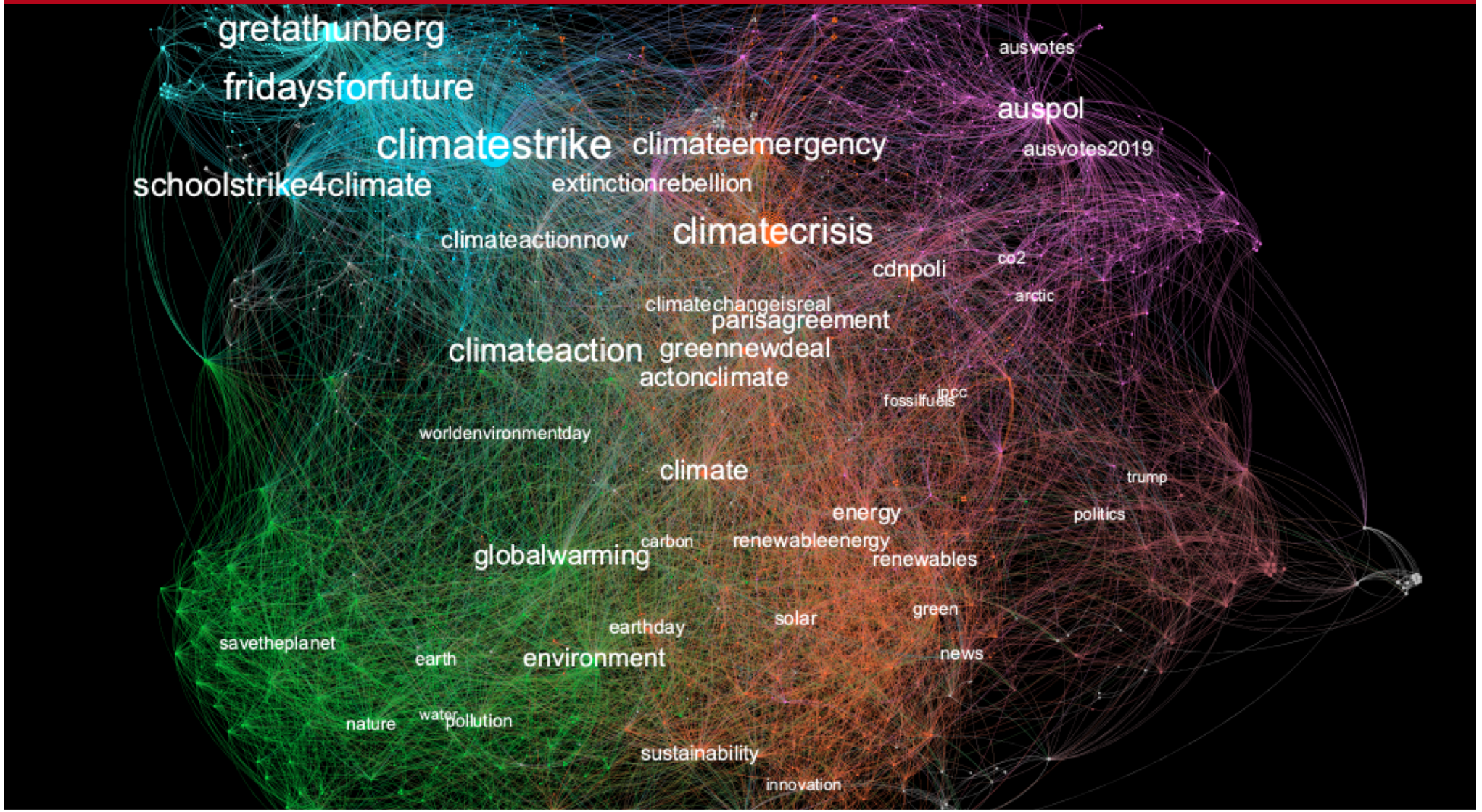
- PageRank can capture the subtleties of networks
- Similar, but more reliable than degree
- Simple to implement (scalable)
- Want to see this in your projects

Visualizing PageRank

a comparison with degree centrality

PageRank on a semantic network

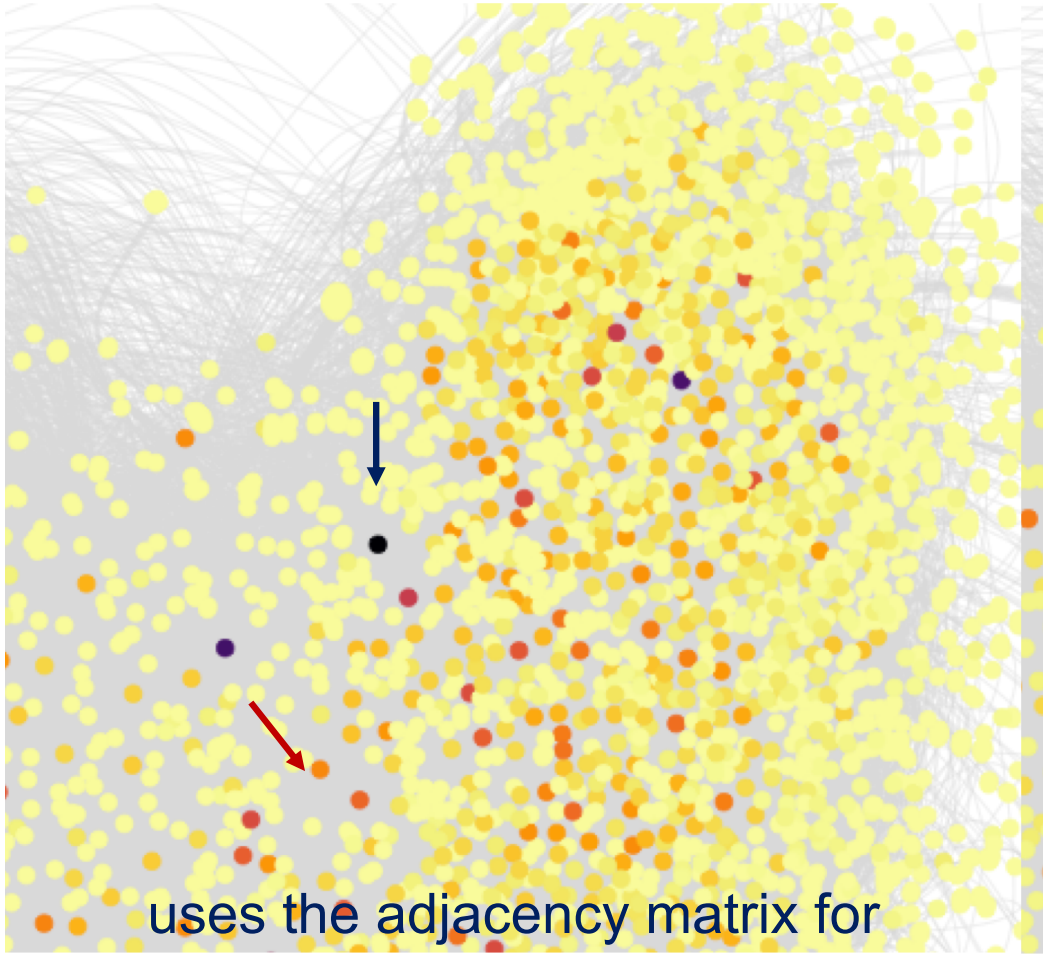
2019 hashtag network related to #climatechange
(from Twitter, after #gretathunberg)



Example of PageRank centrality

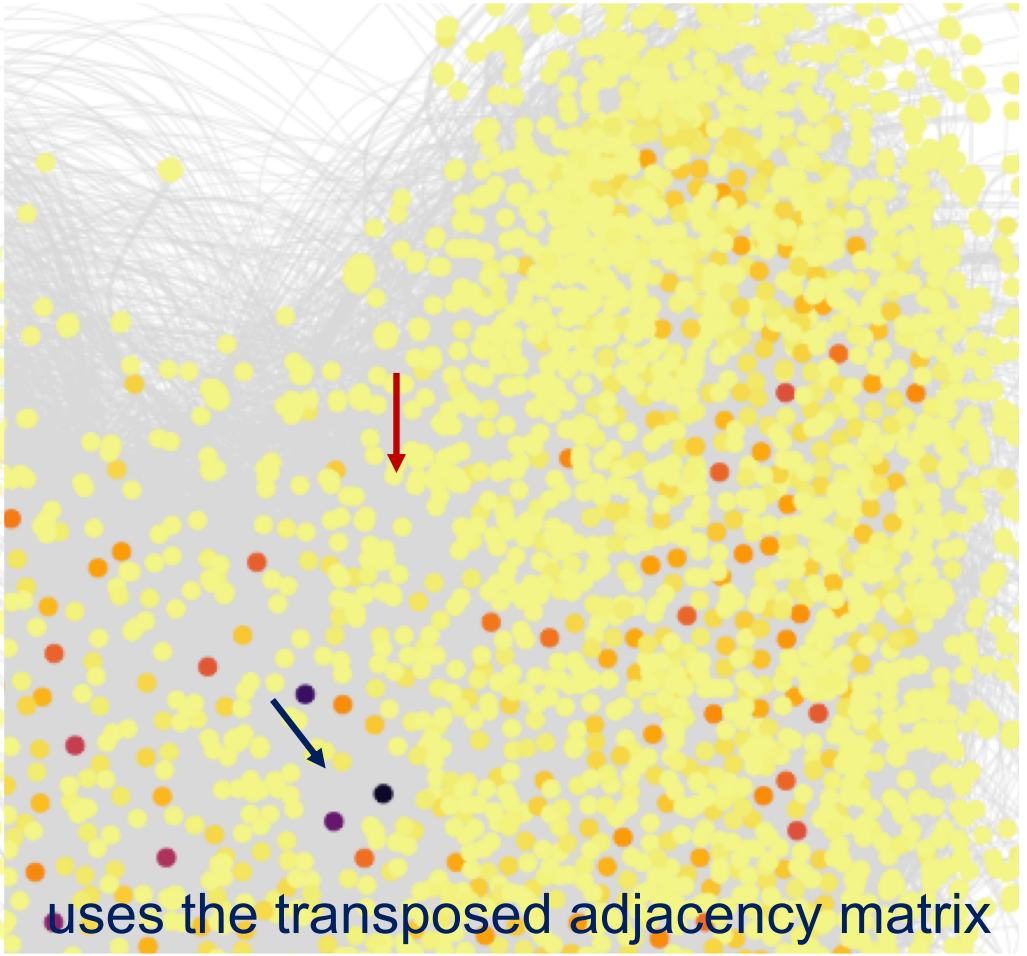
wikipedia administrator elections and vote history data
https://snap.stanford.edu/data/wiki-Vote.html

Authorities



uses the adjacency matrix for spreading

Hubs

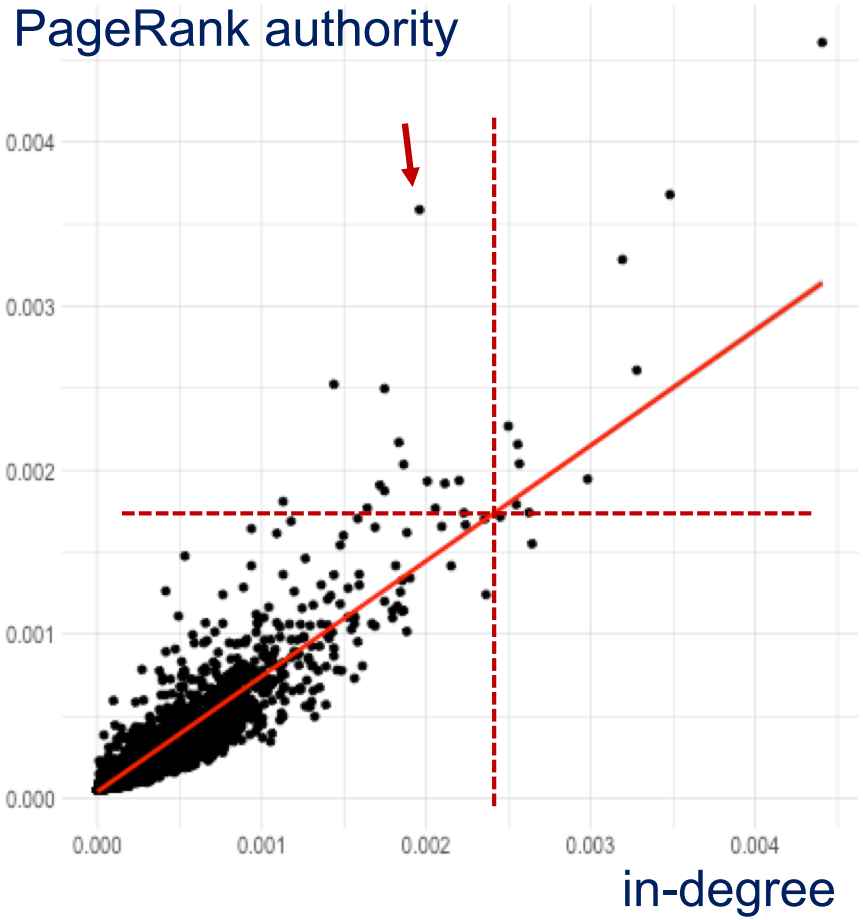


uses the transposed adjacency matrix for spreading (spreading backwards)

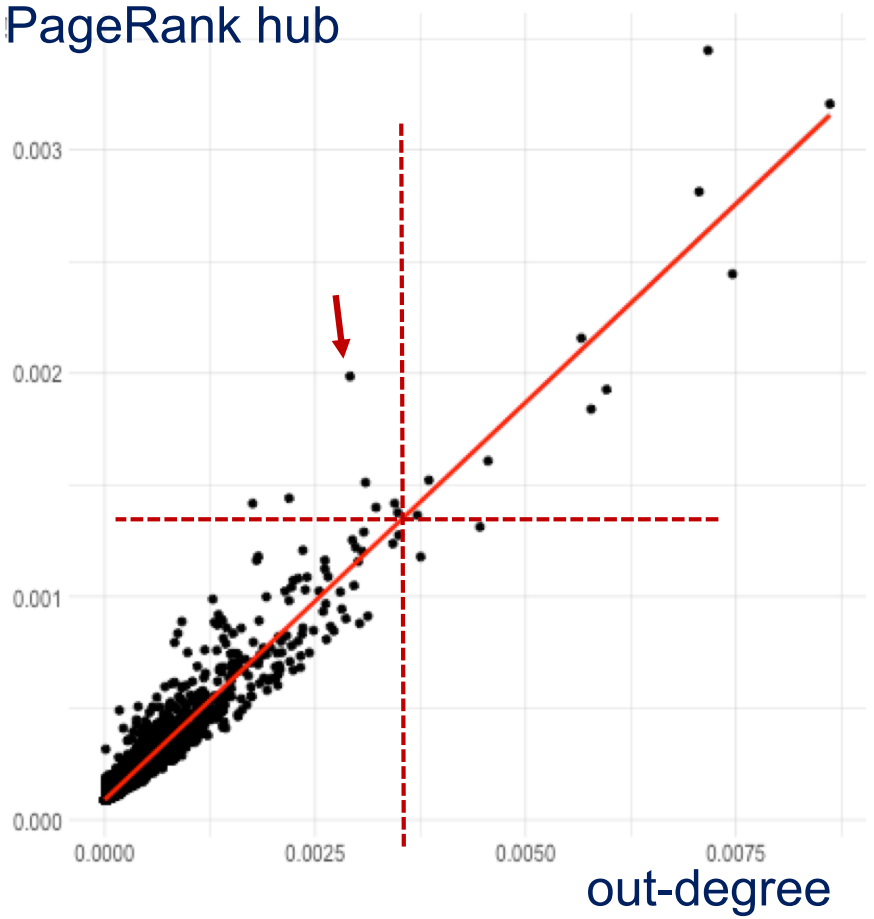
PageRank versus degree centrality

wikipedia administrator elections and vote history data

Authorities



Hubs

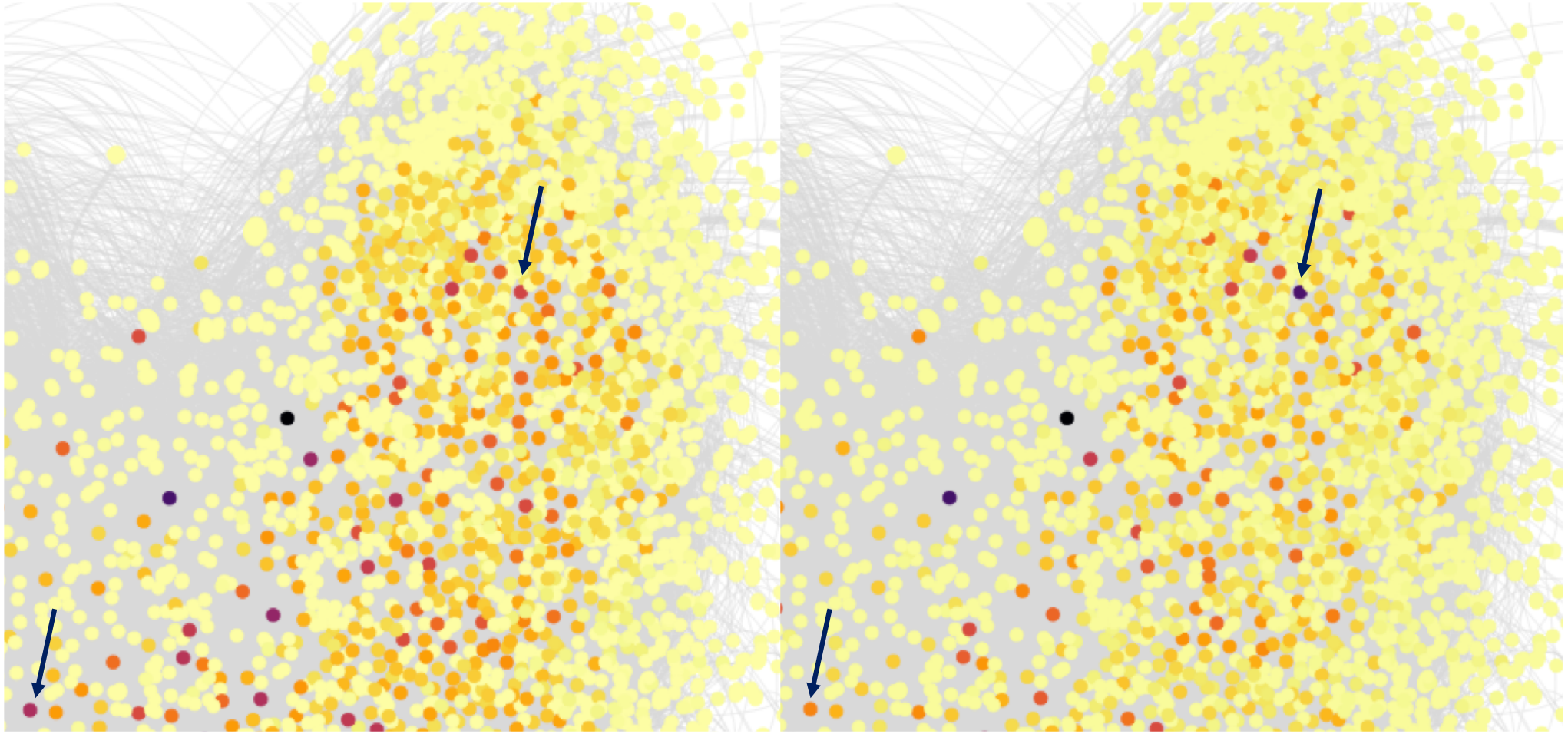


PageRank versus degree authorities

wikipedia administrator elections and vote history data

Degree

PageRank

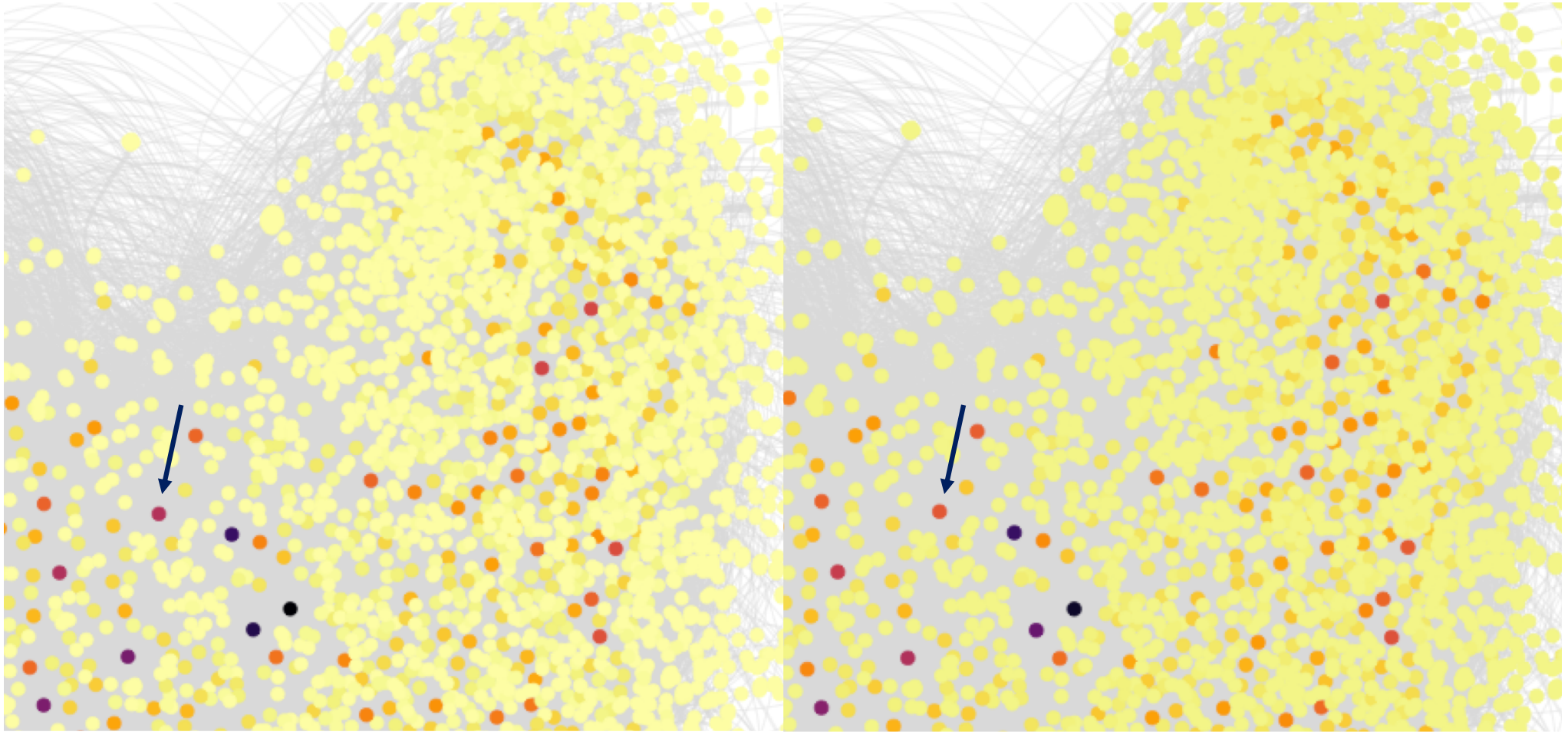


PageRank versus degree hubs

wikipedia administrator elections and vote history data

Degree

PageRank



Local PageRank

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

UNIVERSITÀ DEGLI STUDI DI PADOVA

Measuring closeness: LocalPageRank

measure similarity to a node

Idea

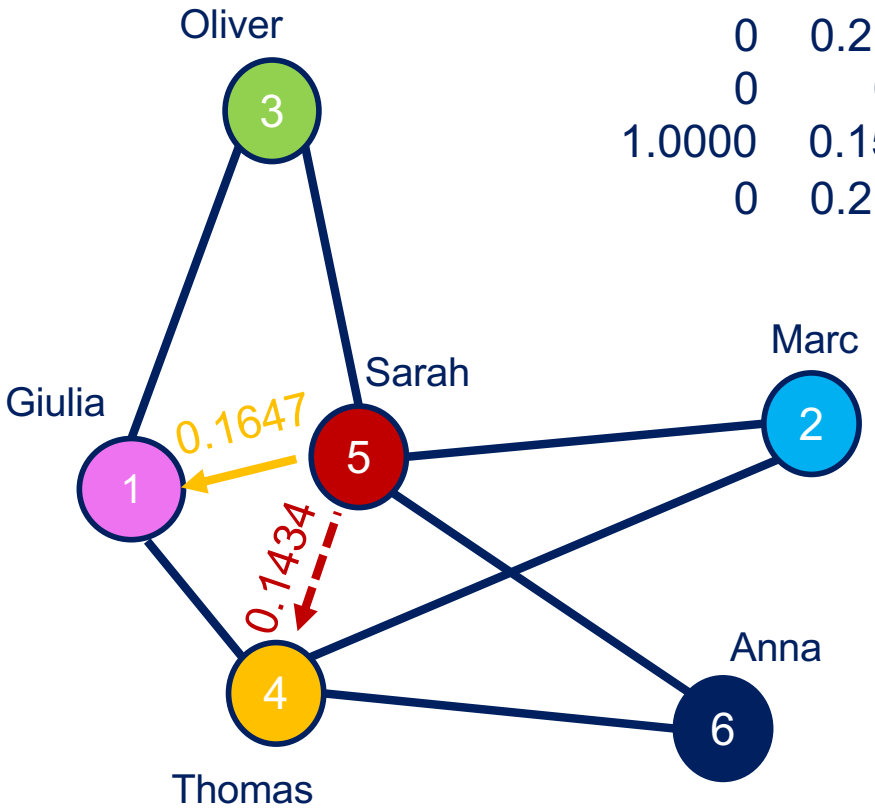
- Measure similarity or closeness to node *i* by applying PageRank with teleport set *to node i only*

Result

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

Example

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

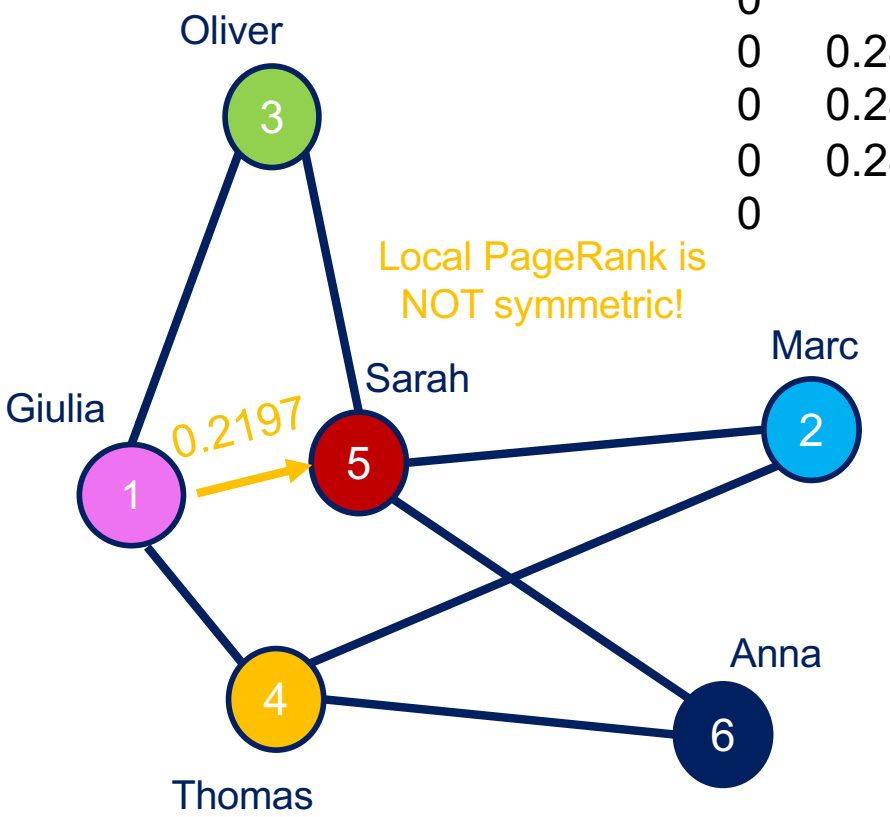


t=1	*2*	*3*	*4*	*5*
0	0.2125	0.1222	0.2096	0.1290
0	0.2125	0.0319	0.1705	0.0708
0	0.2125	0.0921	0.1369	0.1127
0	0	0.2408	0.0617	0.2043
1.0000	0.1500	0.4811	0.2508	0.4125
0	0.2125	0.0319	0.1705	0.0708

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



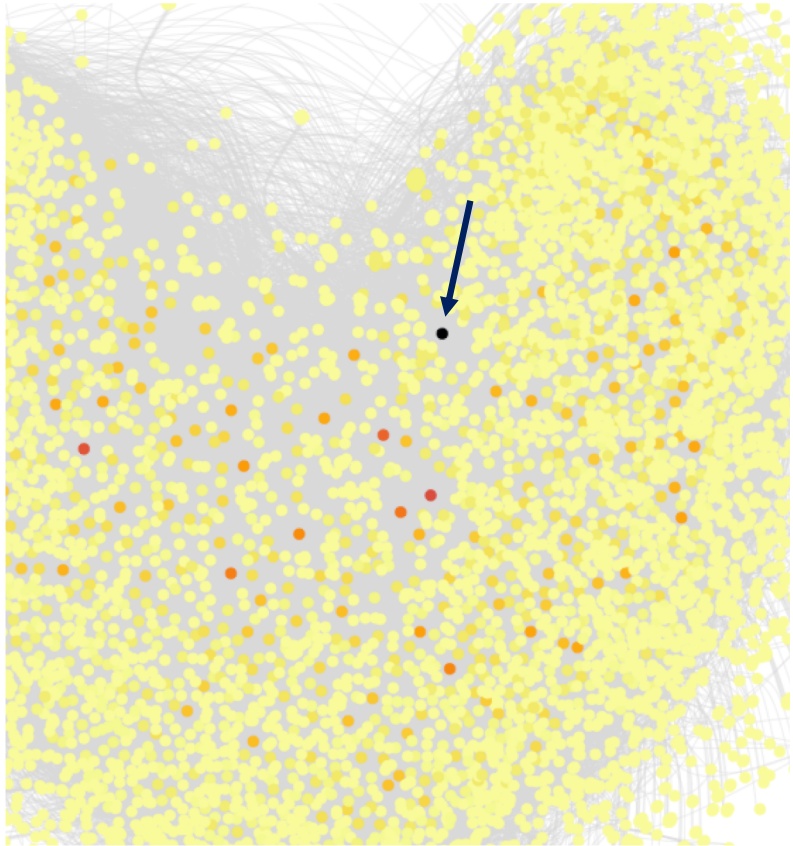
t=1	*2*	*3*	*4*	*5*
1.0000	0.1500	0.4109	0.2403	0.3404
0	0	0.1405	0.0467	0.1262
0	0.2833	0.1027	0.1510	0.1275
0	0.2833	0.0425	0.2358	0.1078
0	0.2833	0.1629	0.2795	0.1719
0	0	0.1405	0.0467	0.1262

10	*20*	*50*	*75*	*100*	
0.2909	0.2985	0.2989	0.2989	0.2989	Giulia
0.0848	0.0926	0.0931	0.0931	0.0931	Marc
0.1309	0.1313	0.1314	0.1314	0.1314	Oliver
0.1763	0.1645	0.1638	0.1638	0.1638	Thomas
0.2324	0.2204	0.2197	0.2197	0.2197	Sarah
0.0848	0.0926	0.0931	0.0931	0.0931	Anna

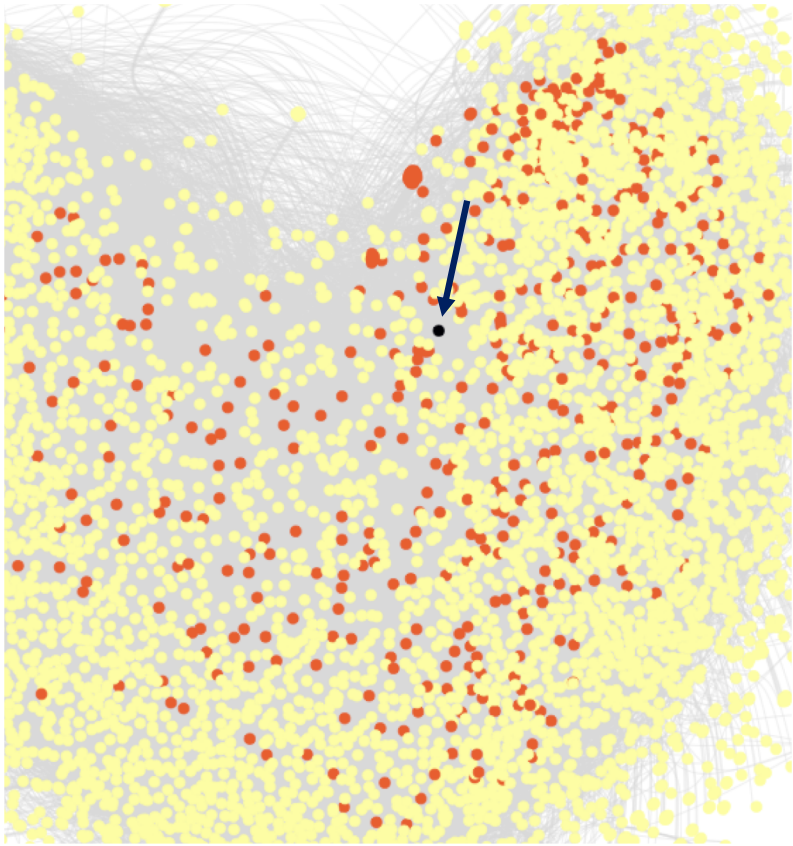
Local PageRank versus degree

authorities

Local PageRank



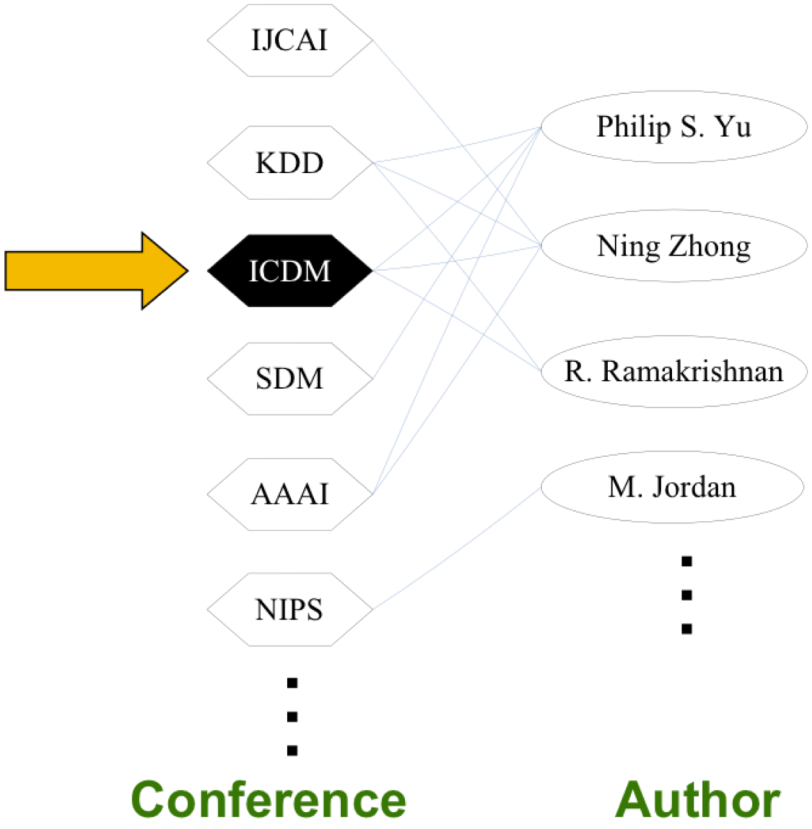
1-hop out-neighbours



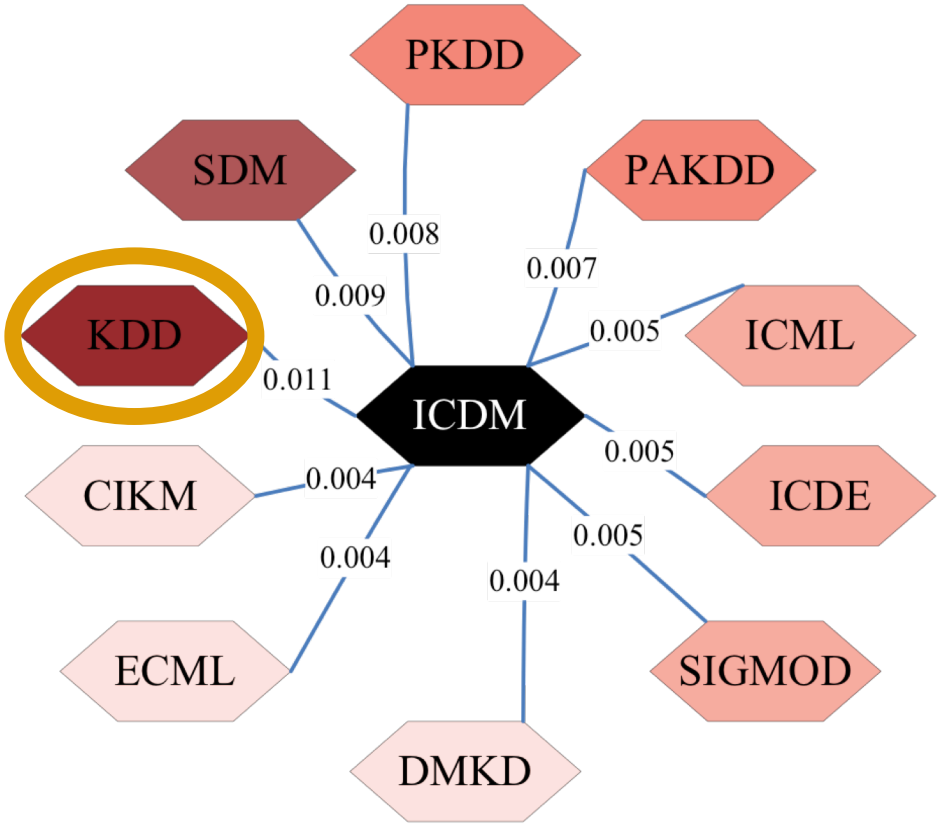
neighbours authority score =
local node → neighbours

Example

what is the most related copnference to ICDM?



Top 10 ranking results



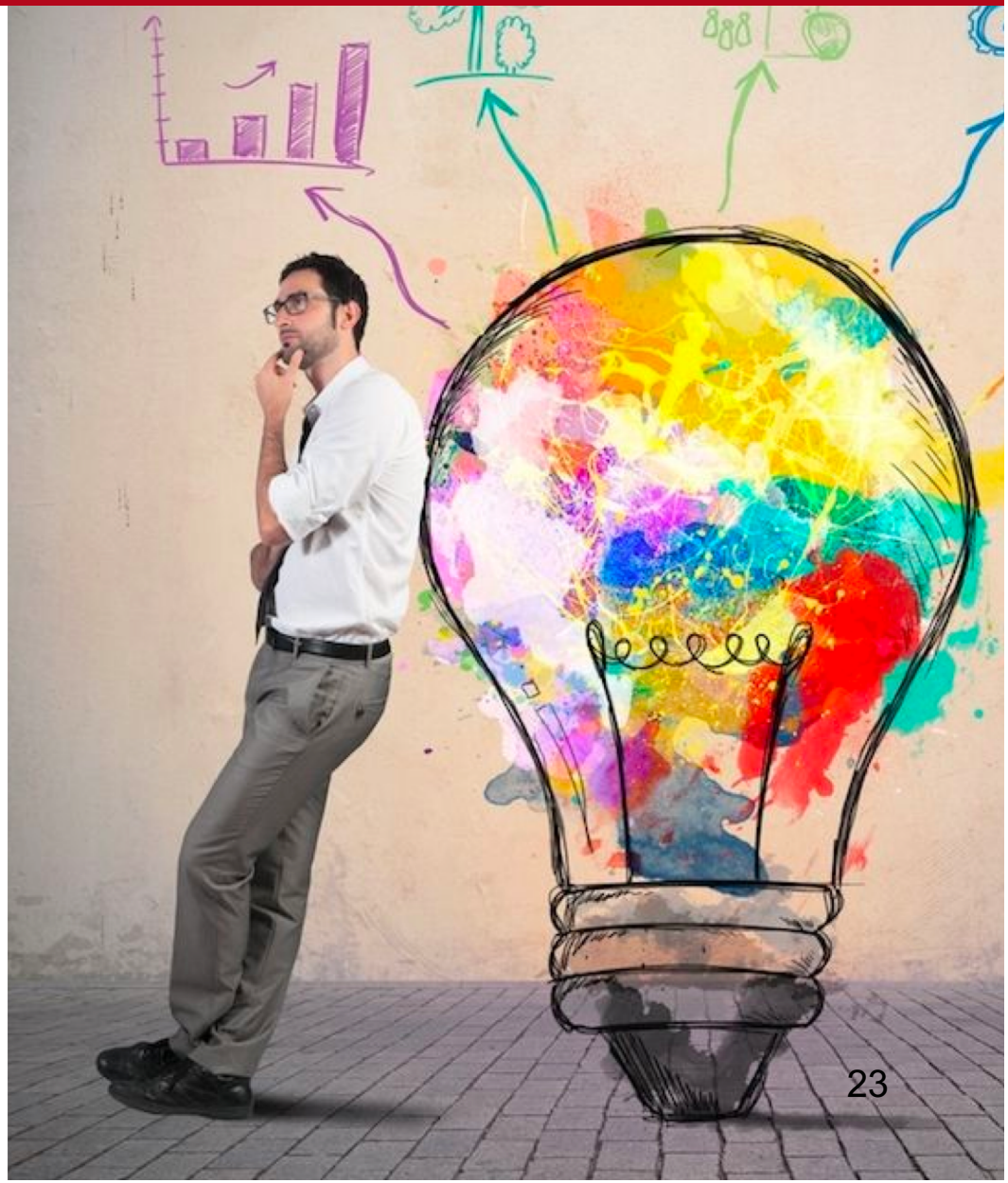
ICDM = international conf. on data mining
KDD = knowledge discovery and data mining

Measuring closeness to a topic

topic specific PageRank

Want to know about a specific topic? TopicSpecific PageRank

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



TopicSpecific PageRank example

in semantic networks

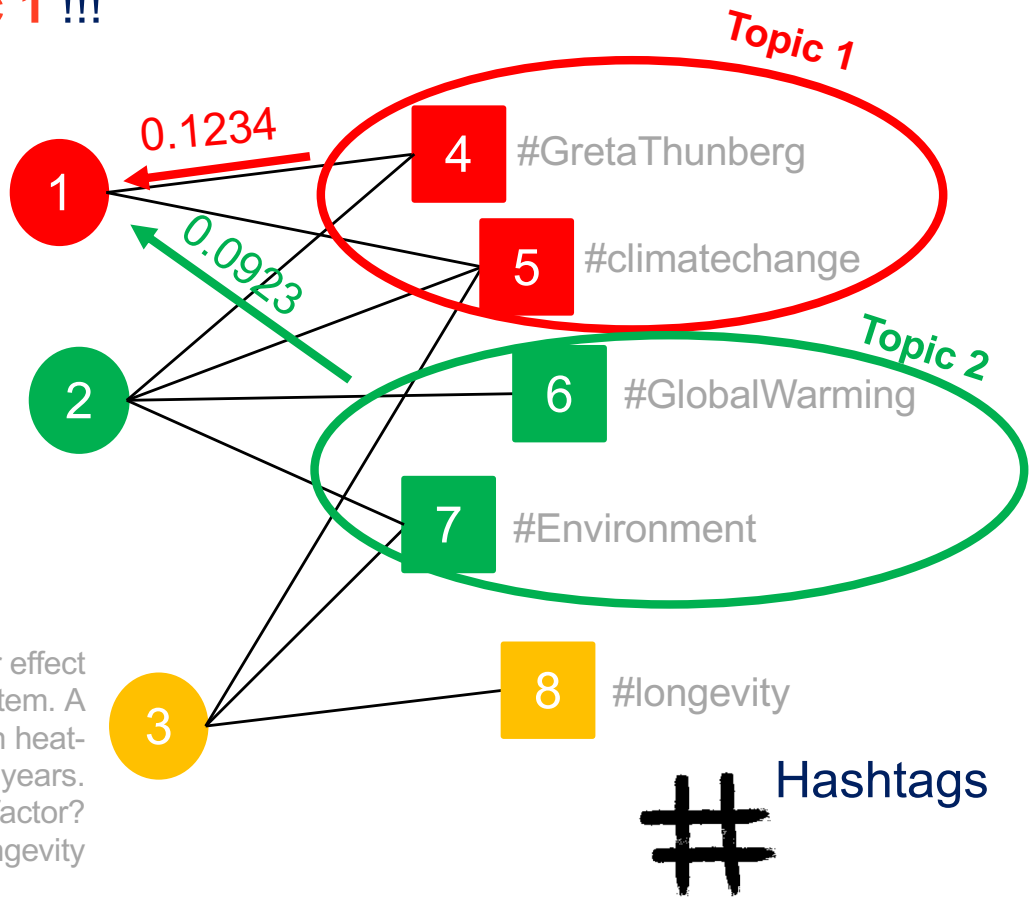
Tweet 1 is assigned to **Topic 1** !!!

those who think they are crazy enough to change the world eventually do. #climatechange #ClimateCrisis #ClimateAction #GretaThunberg #Greta

Hopefully these kids will succeed where past generations have failed. #TheResistance #FBR #ClimateChange #Environment #GlobalWarming #GretaThunberg

The #environment can have a major effect on the human cardiovascular system. A new study has found an increase in heat-induced #heartattack risk in recent years. Could #ClimateChange be a risk factor? #longevity

Tweets

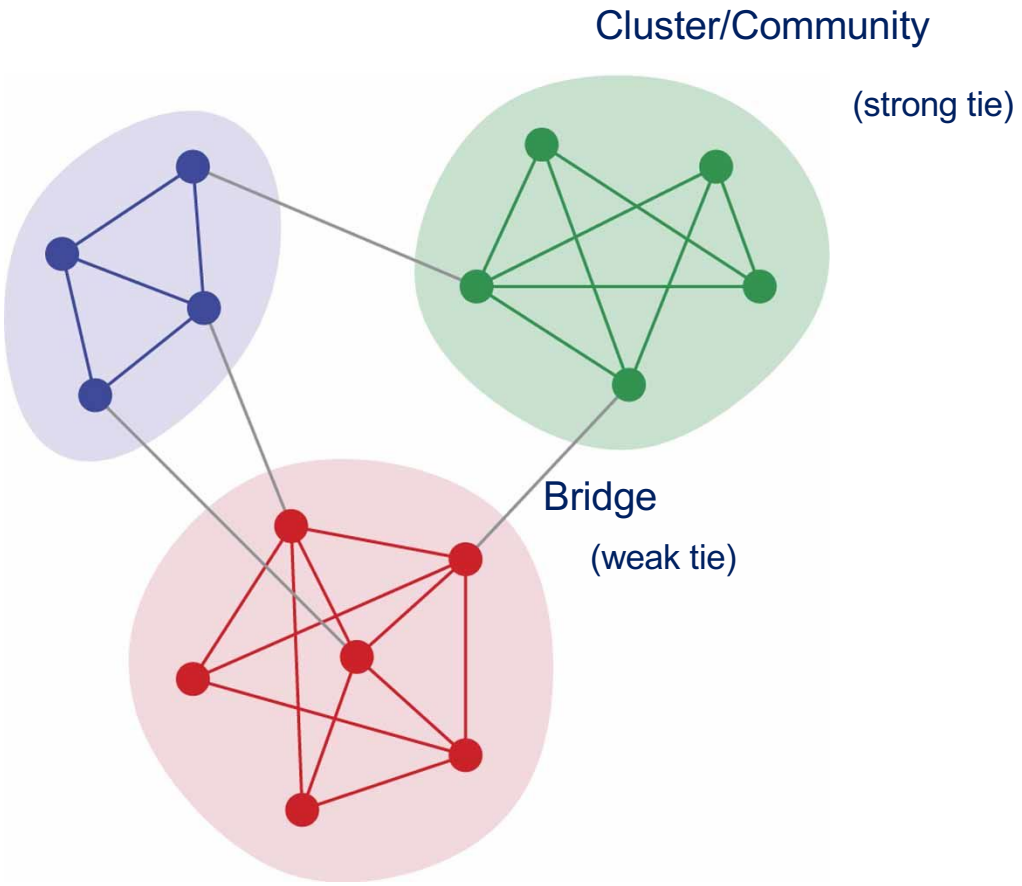


Community detection

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?

Granovetter's explanation

Granovetter, The strength of weak ties [1973]
https://www.jstor.org/stable/pdf/2776392.pdf

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

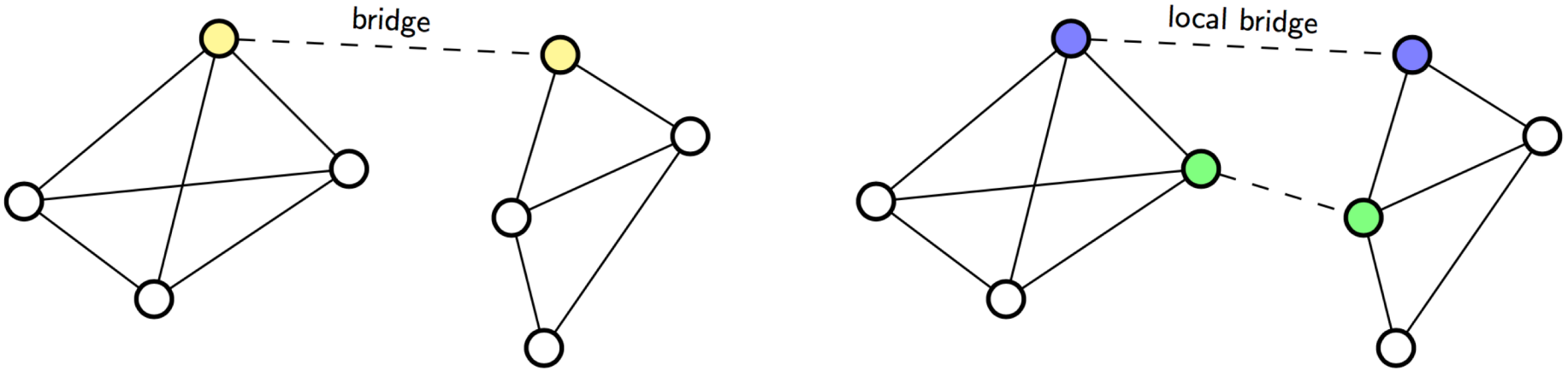
Conclusion:

- 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

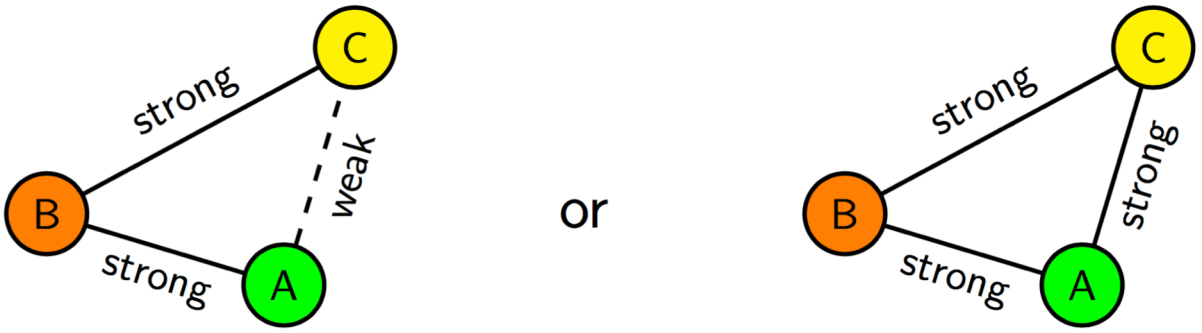
Strong triadic closure

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

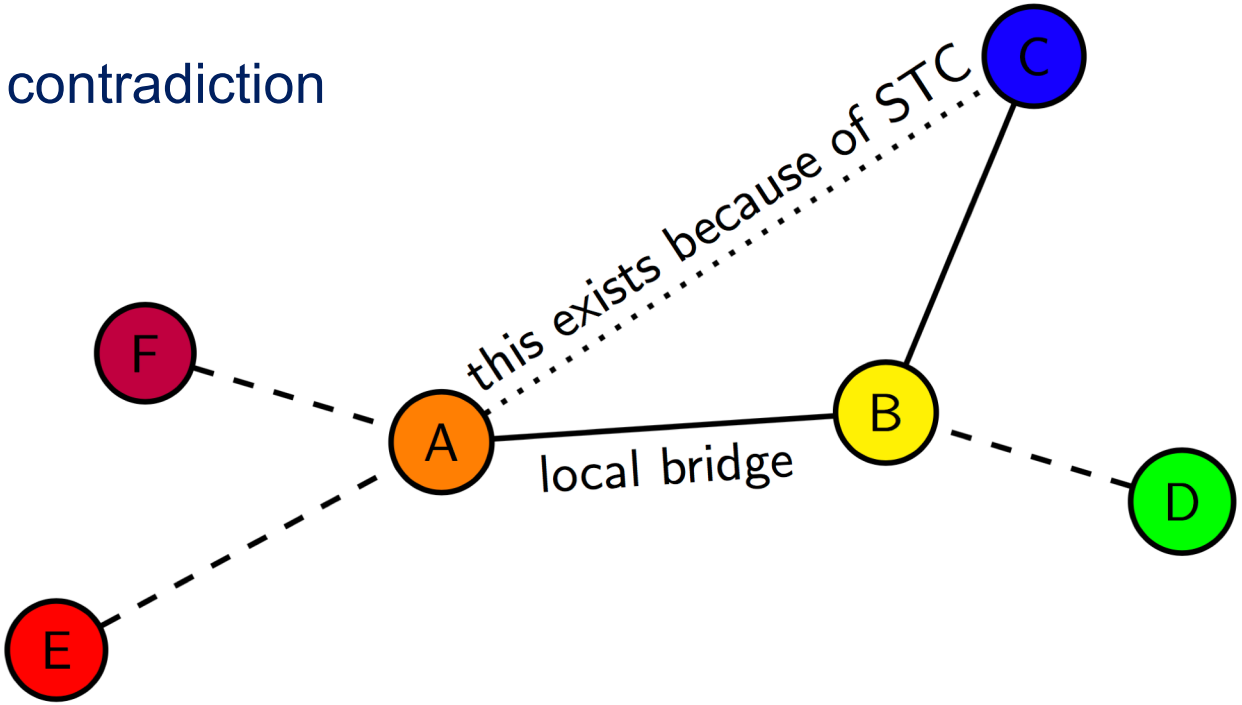
under strong triadic closure

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

Figure 2 | A network of collaborations among scientists at a research

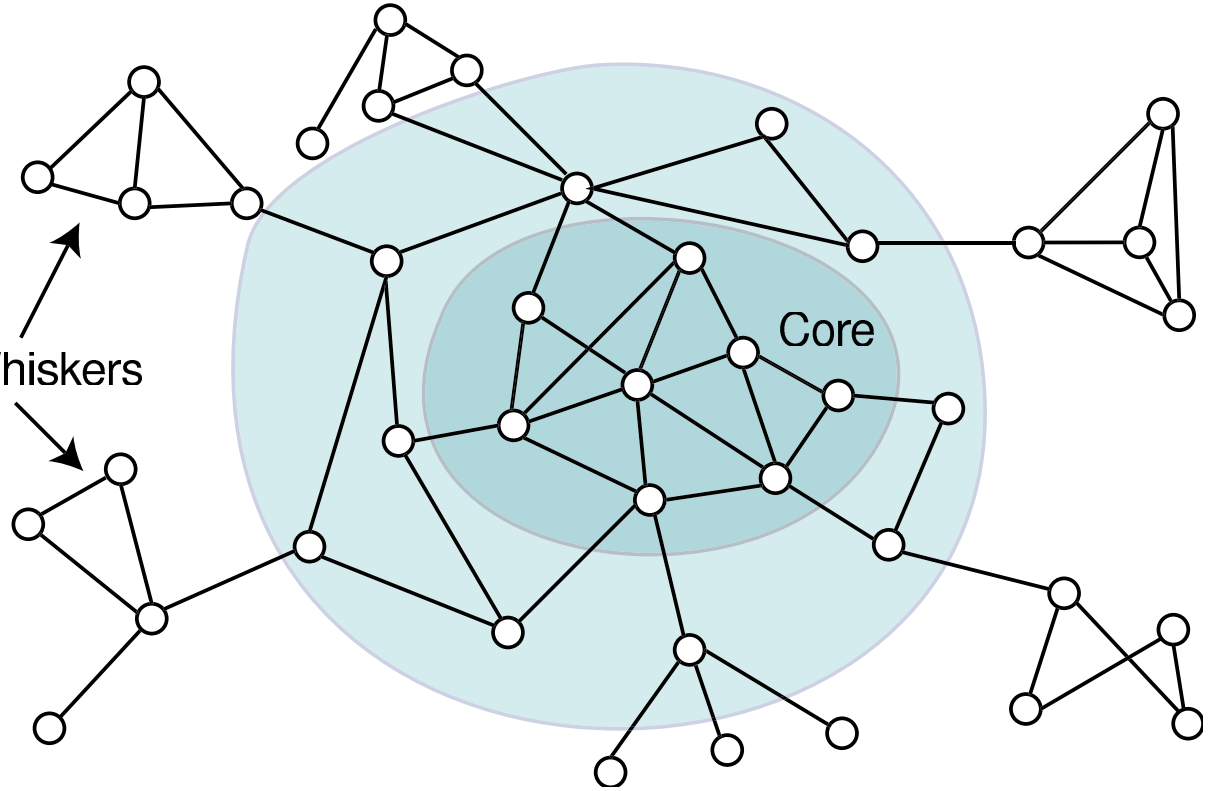
The core periphery model

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?

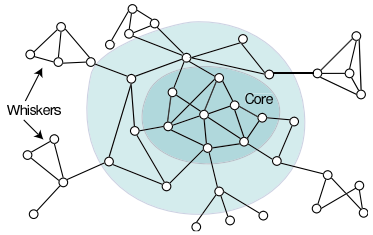
Small, peripheral clusters



Caricature of network structure

Overlapping communities

to explain the core periphery model

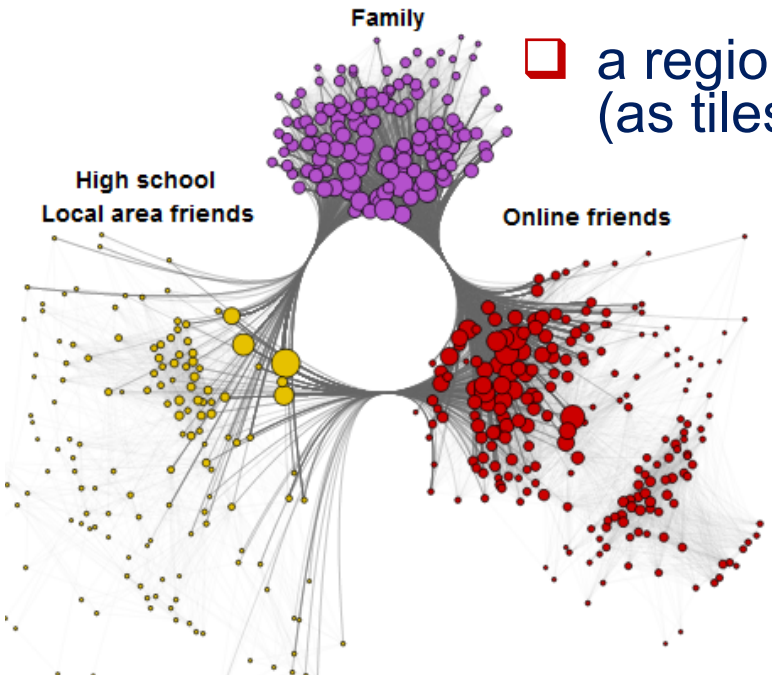


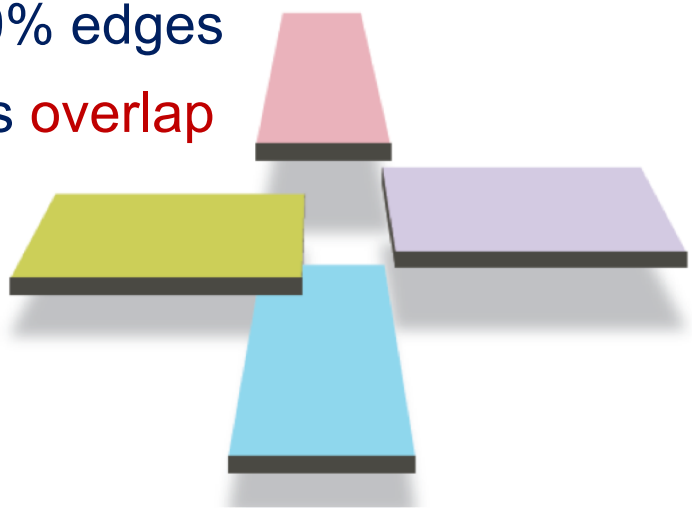
Wiskers

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

Core

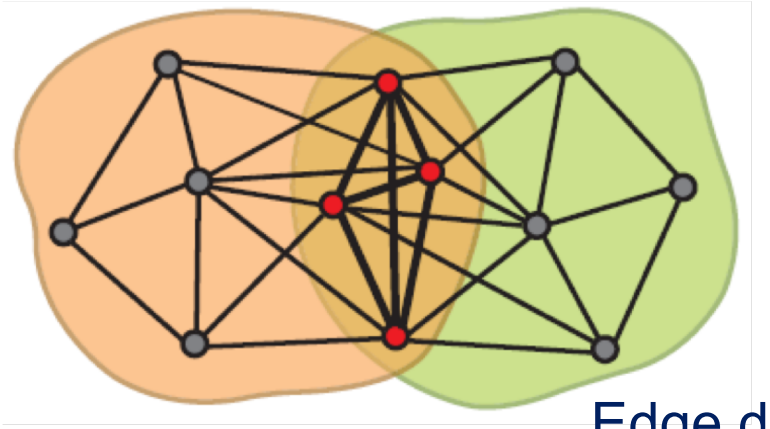
- denser and denser region
- contains 60% nodes and 80% edges
- a region where communities overlap (as tiles)



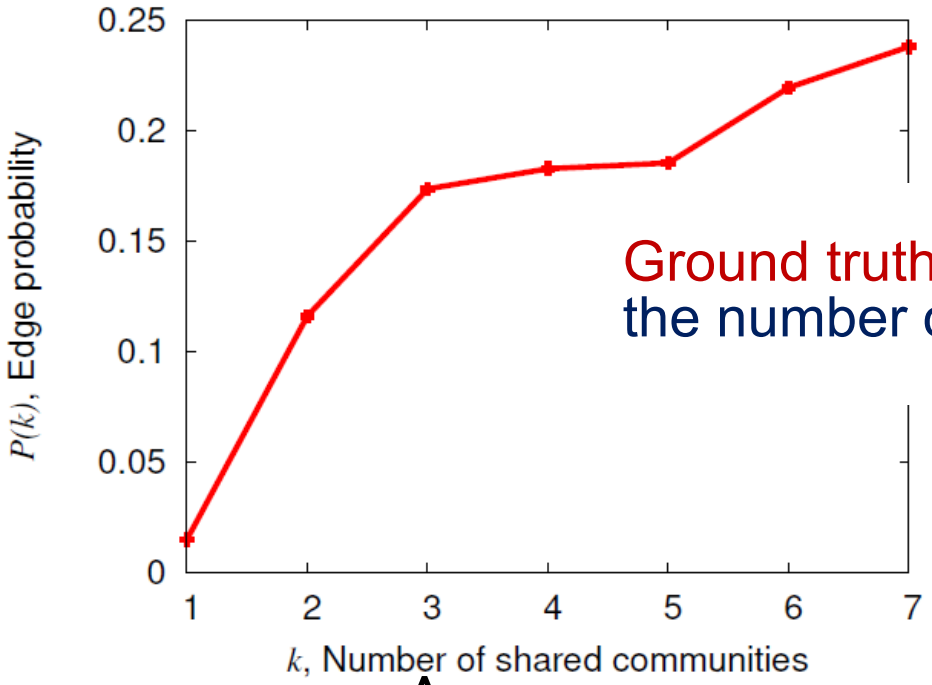


Measuring overlapping

in social networks



Edge density is bigger in the overlap



Amazon

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

Algorithms

for community detection

Modularity

Newman, Modularity and community structure in networks (2006)
https://www.pnas.org/content/pnas/103/23/8577.full.pdf

Want to:

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

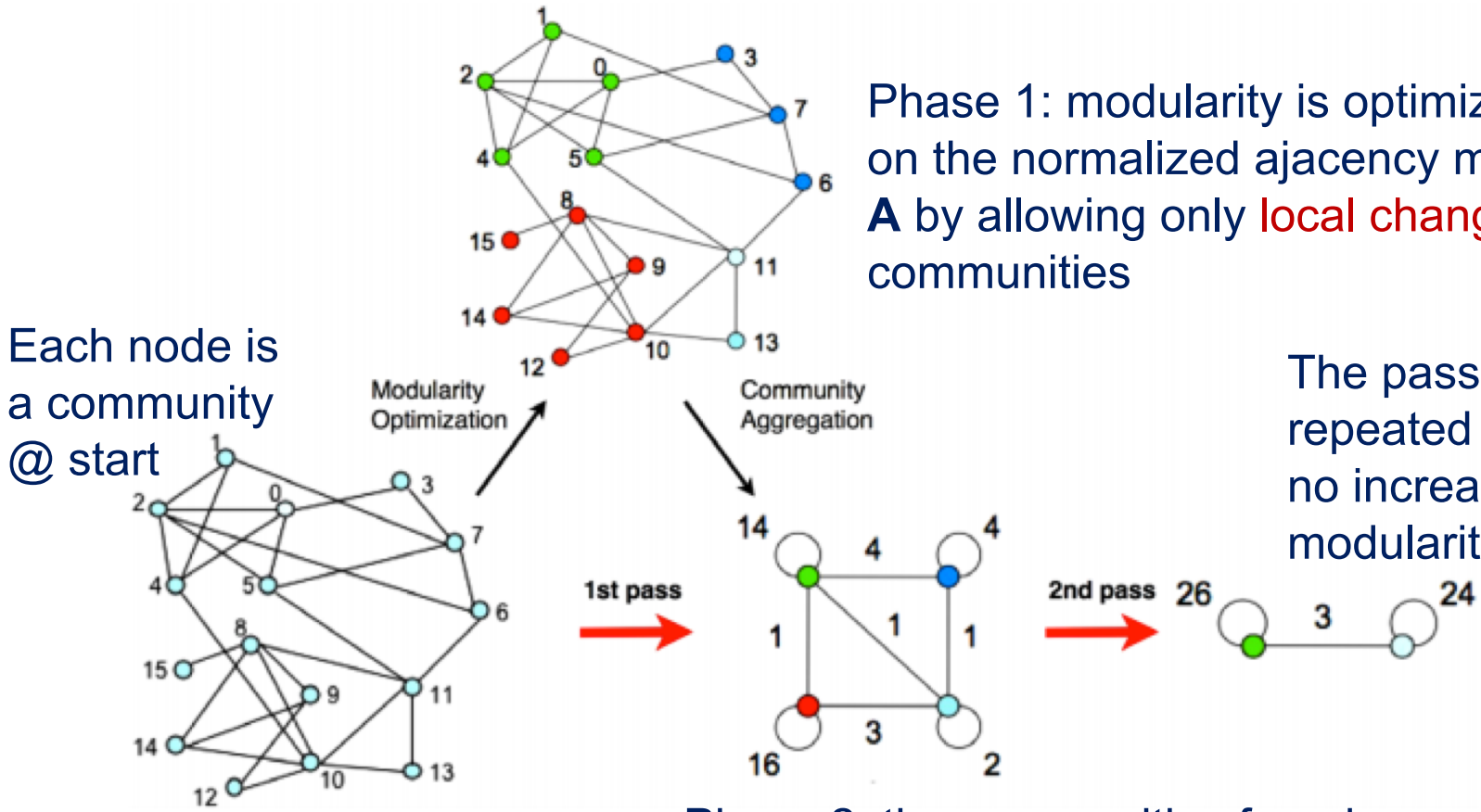
Idea:

- "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



Each node is a community @ start

Phase 1: modularity is optimized on the normalized ajacency matrix **A** by allowing only local changes of communities

The passes are repeated iteratively until no increase of modularity is possible

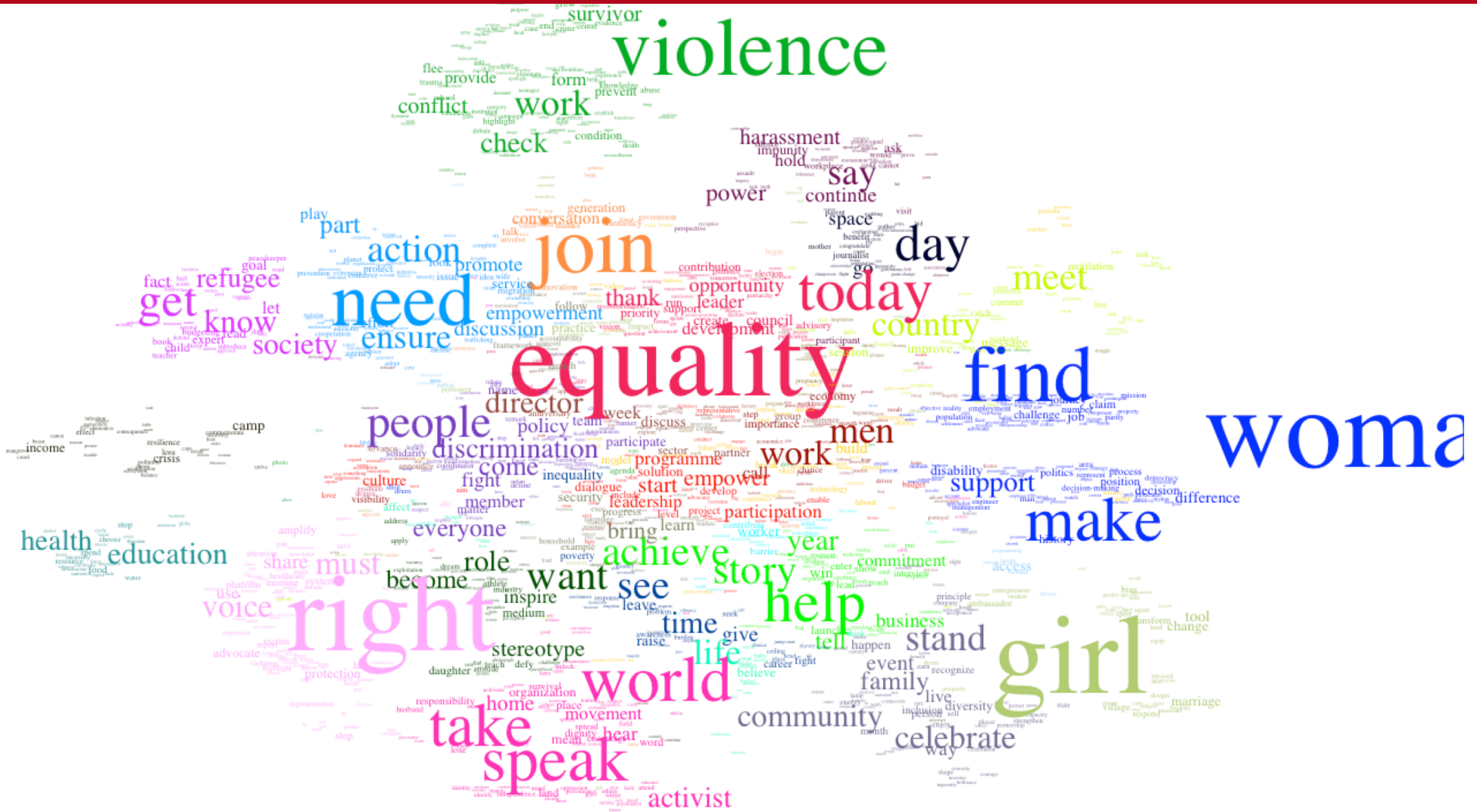
Phase 2: the communities found are aggregated (sum of links) in order to build a new network of communities with normalized adjacency matrix $\boldsymbol{P}_{CC}$

Characteristics of Louvain

- 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)

An example

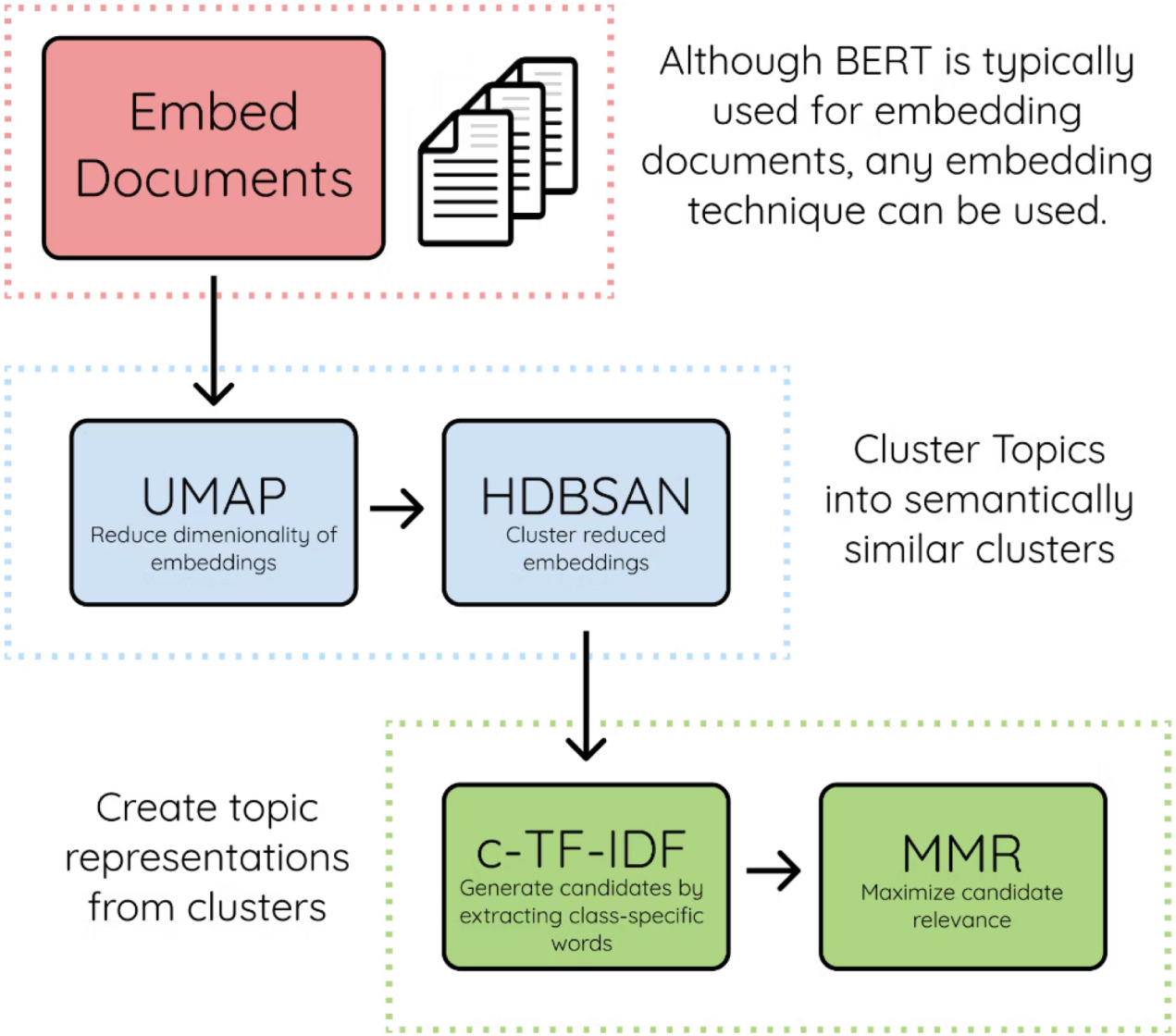
community detection applied to semantic networks = topic detection



#metoo tweets

BERTopic

an algorithm based on ChatGPT architecture



HDBSCAN in BERTopic

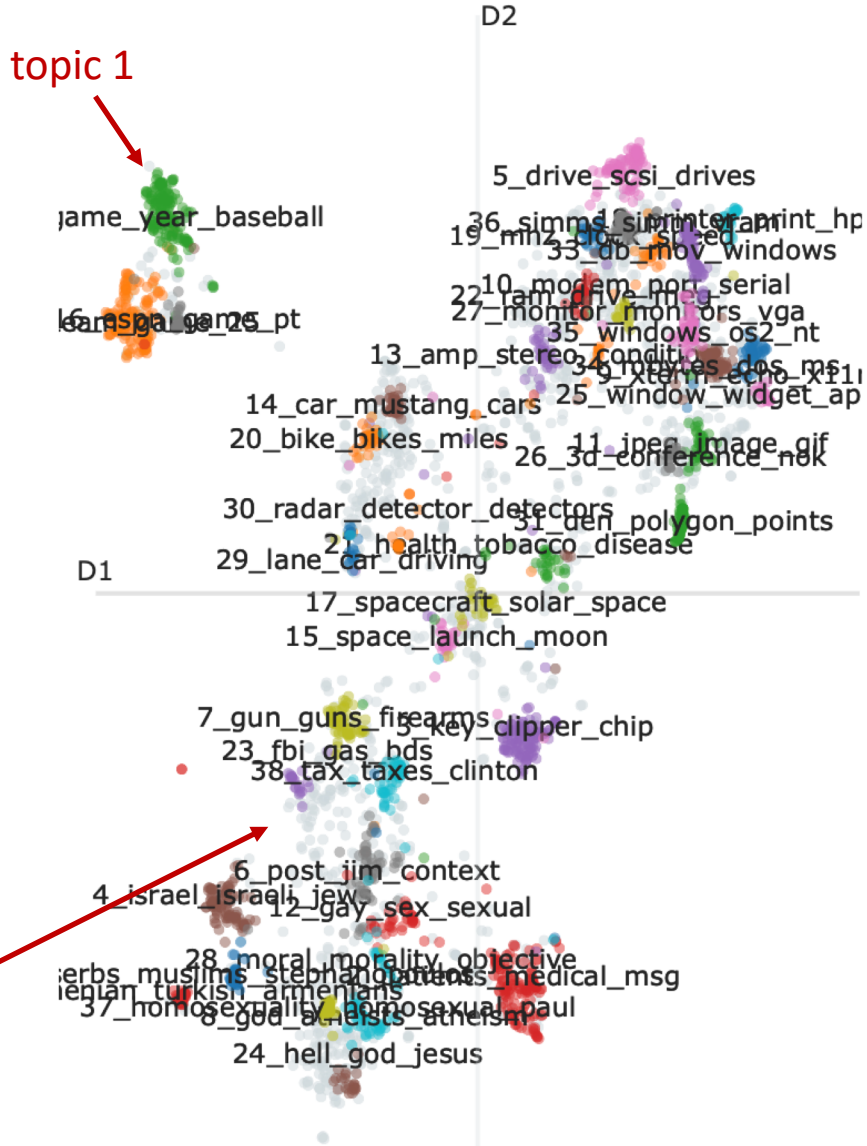
clustering documents into different topics

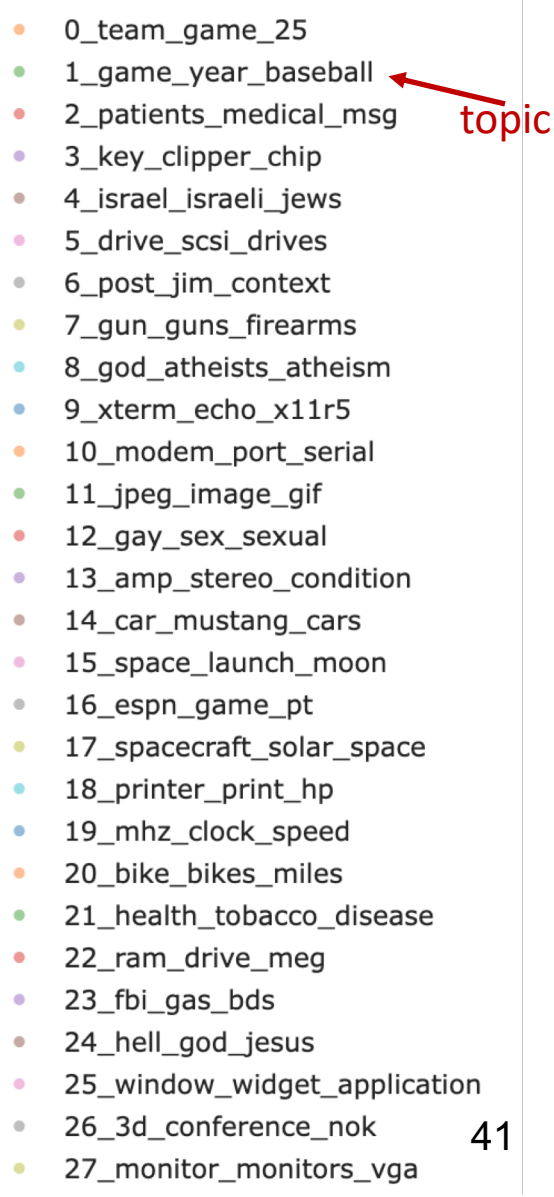
1. each document is mapped into an embedding (vector) by BERT

2. cosine metric is used to identify distances among documents

3. HDBSCAN is run to identify topics

outliers in gray

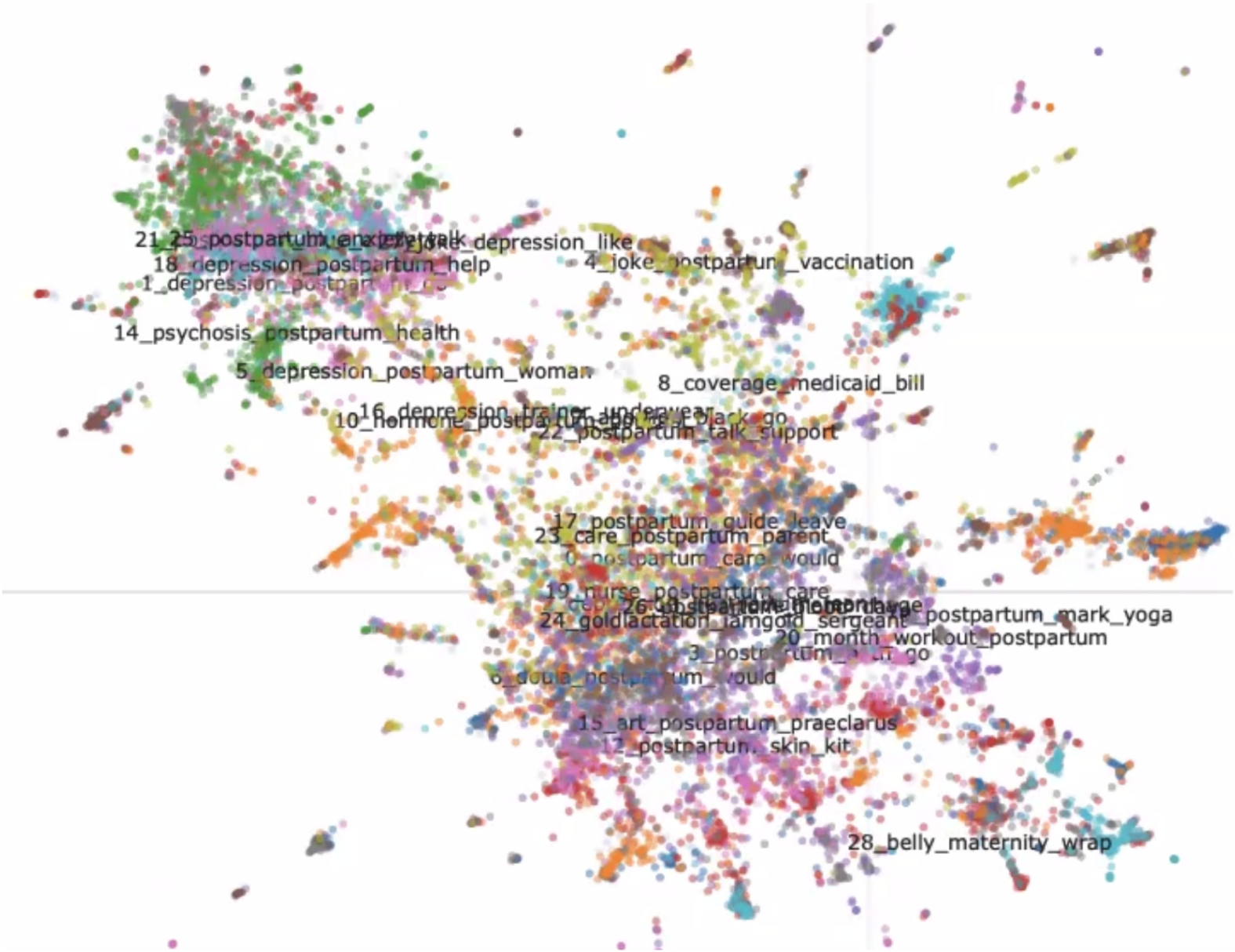




topic 1

A visual example

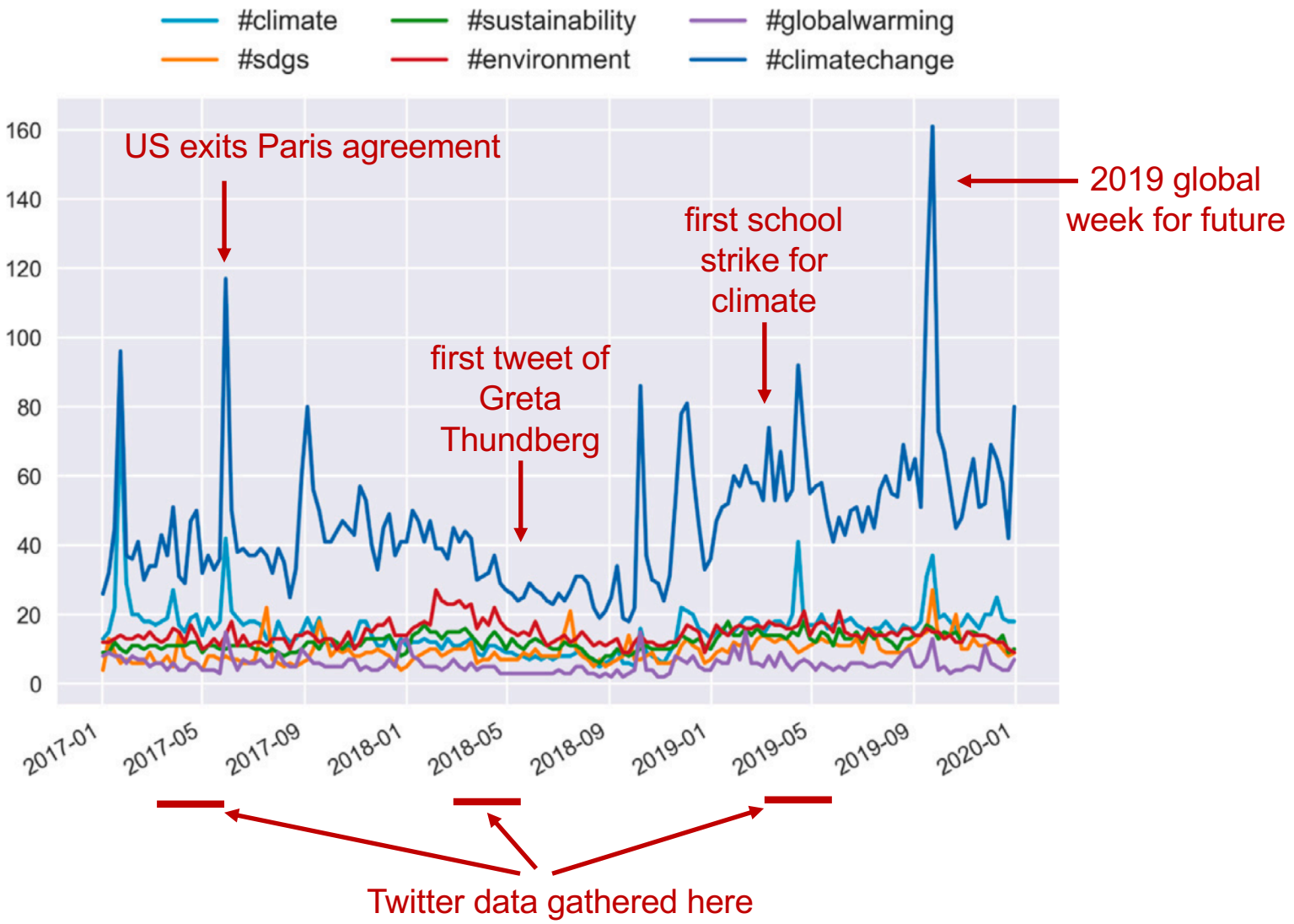
document network using BERTopic on postpartum tweets



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

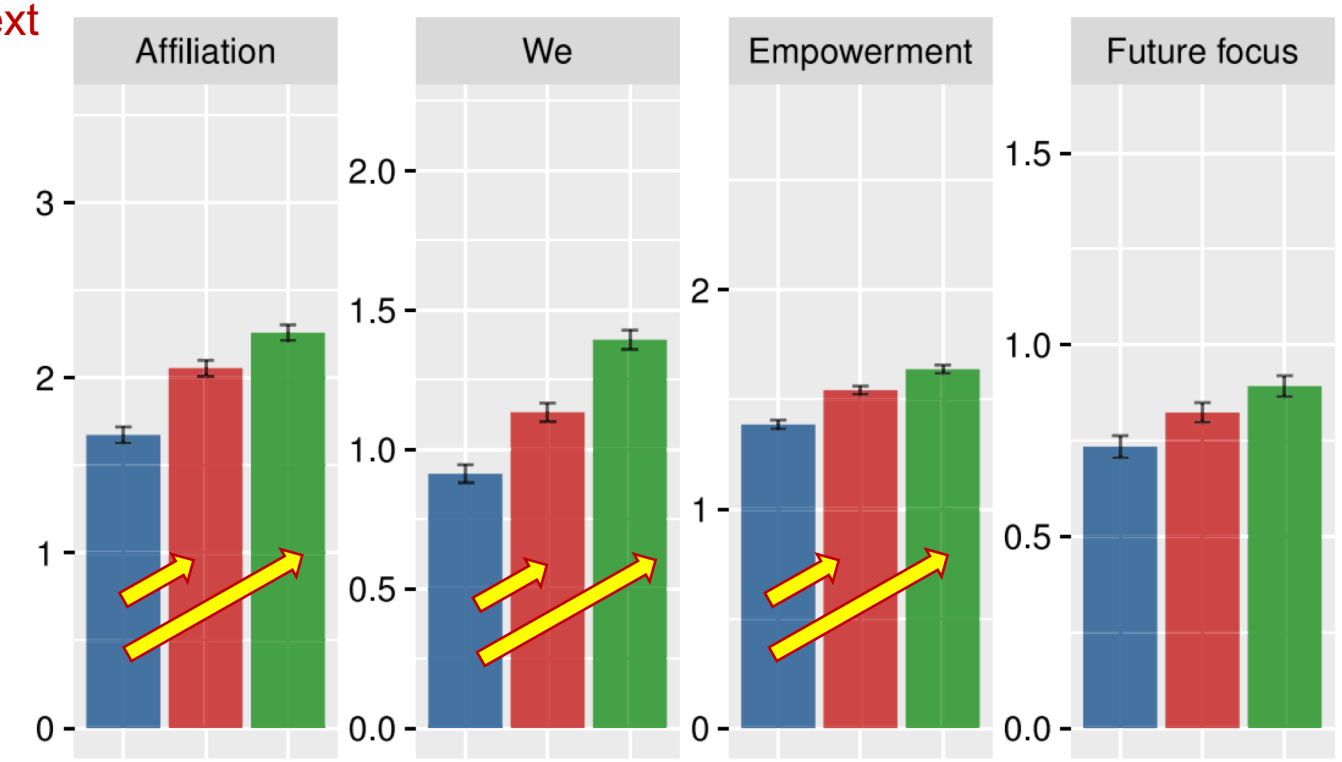
2017 2018 2019

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

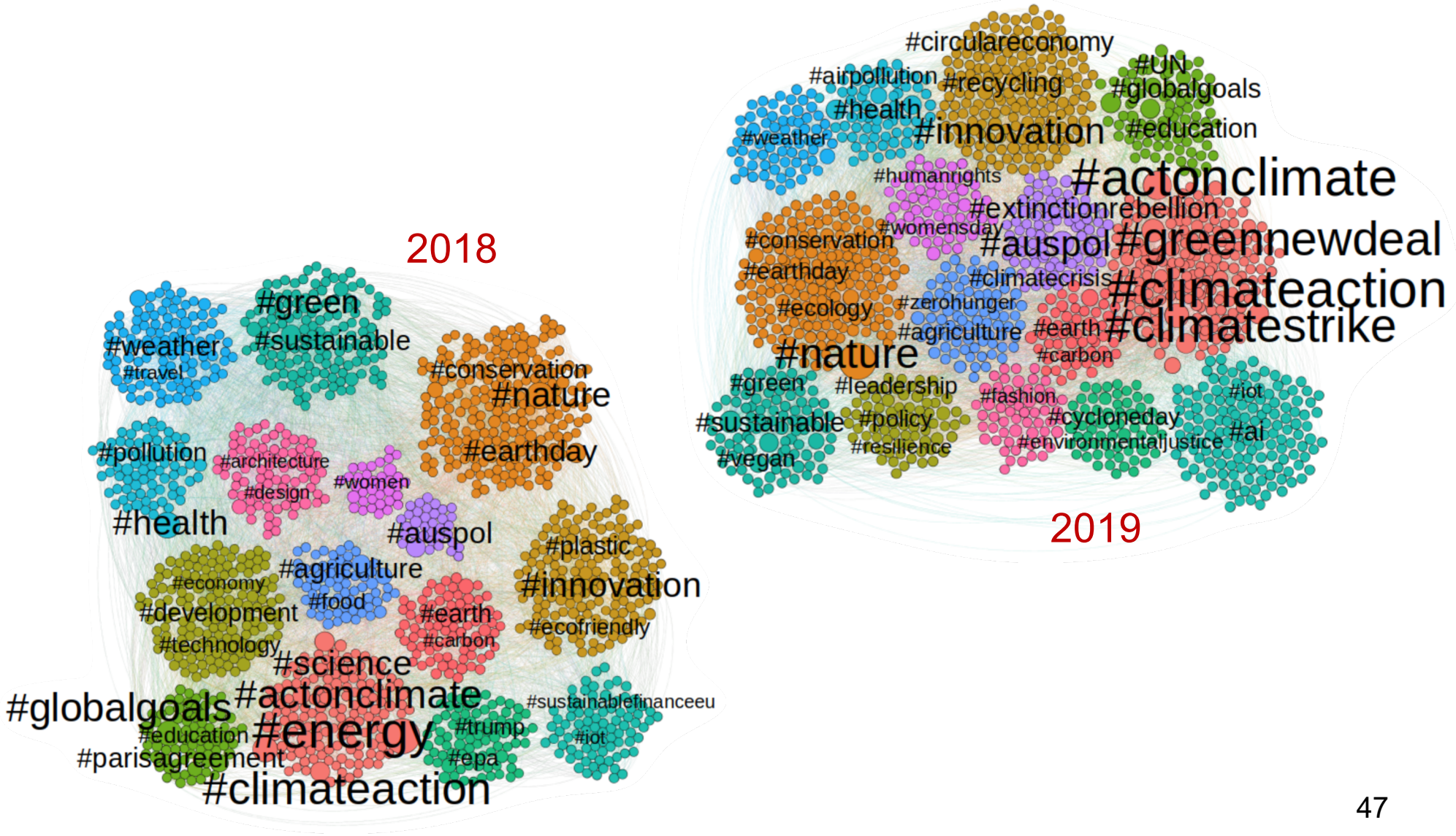
orientation of tweets to the past or future



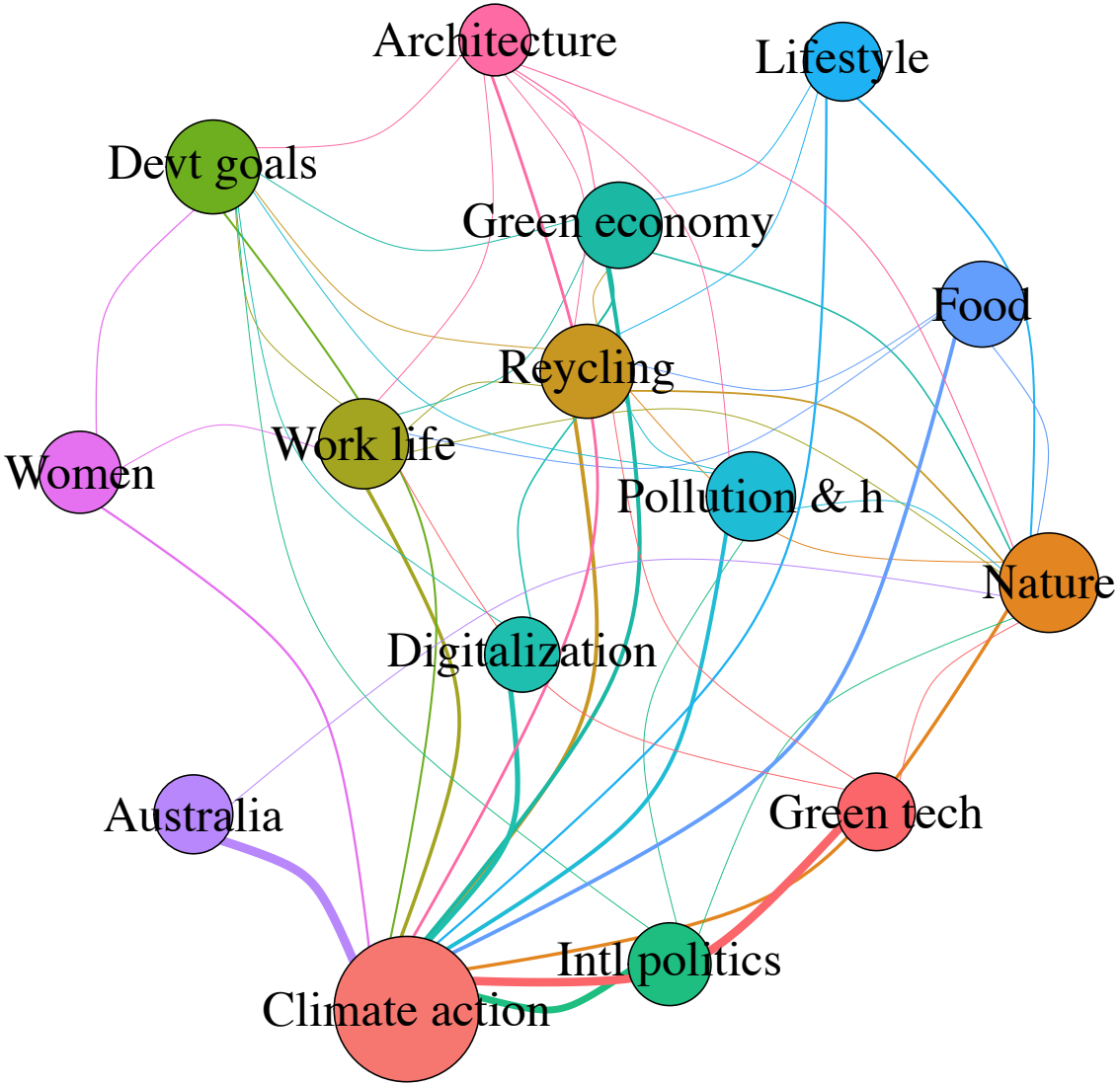
only a few statistically relevant changes

Communities/topics in #climateaction

on Twitter in 2017, 2018, 2019



Topics interdependencies

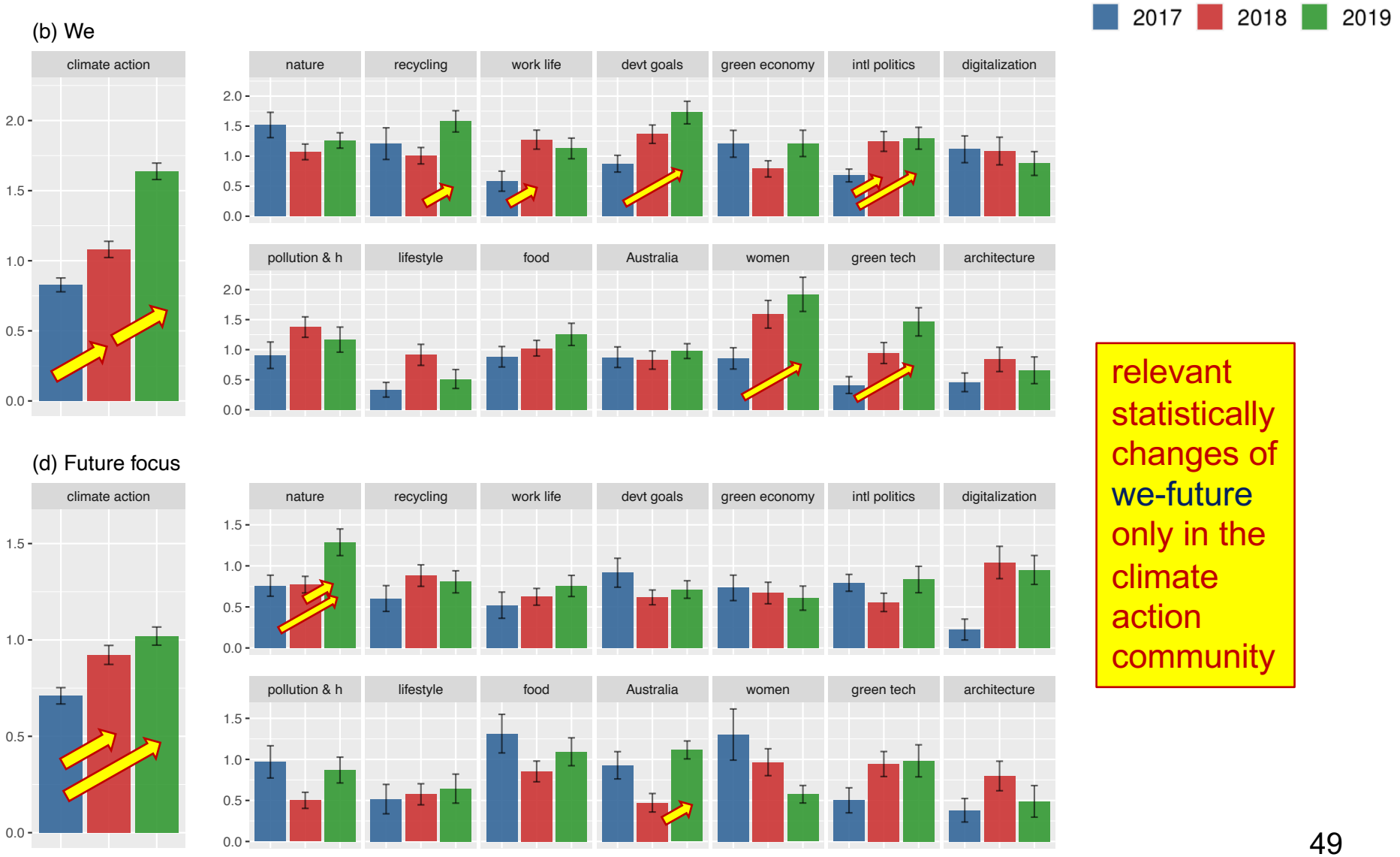


projecting the adjacency matrix on topics

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



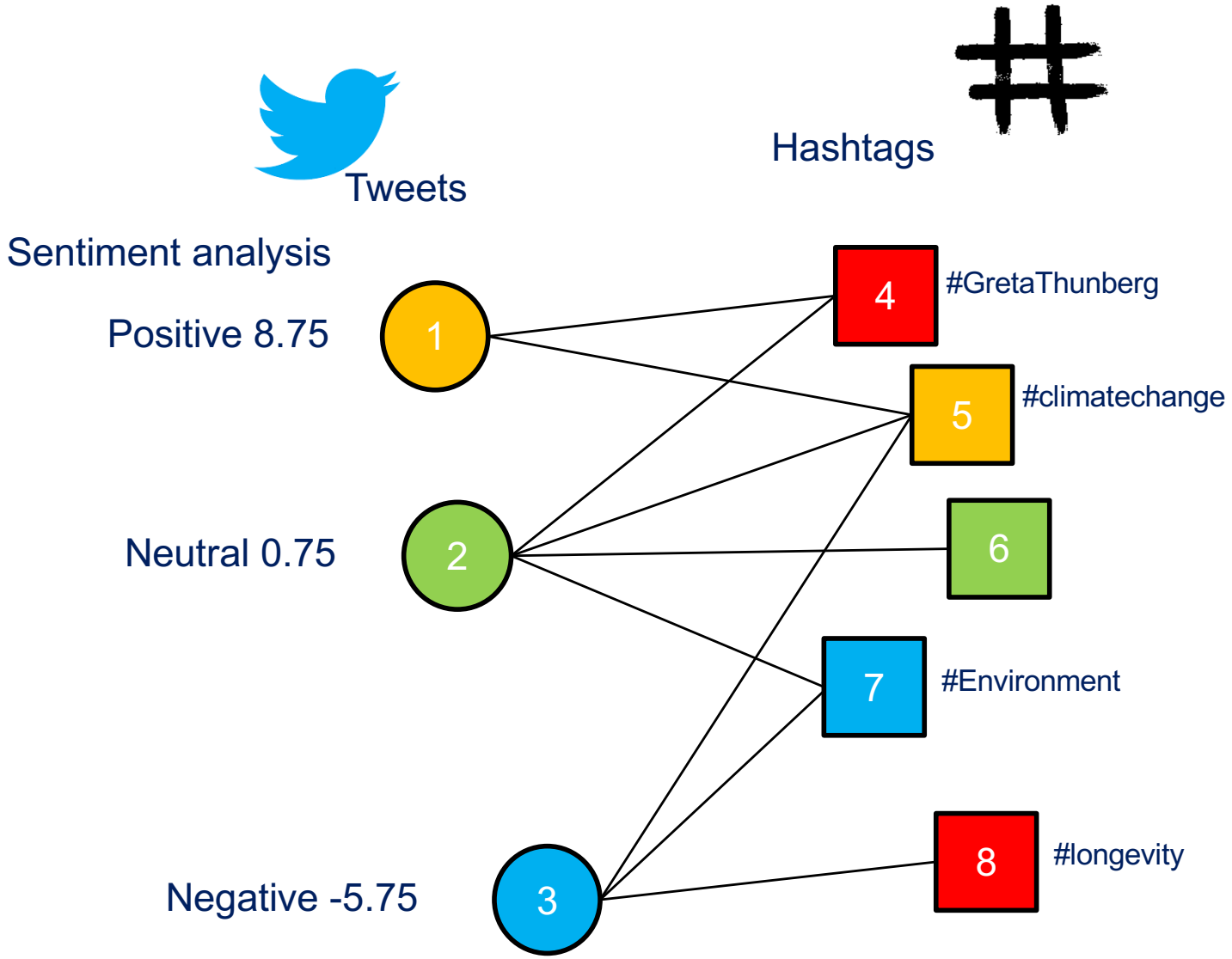
relevant statistically changes of we-future only in the climate action community

Projecting markers

on specific words, and their application

The question

how to project sentiment information from tweets to hashtags?

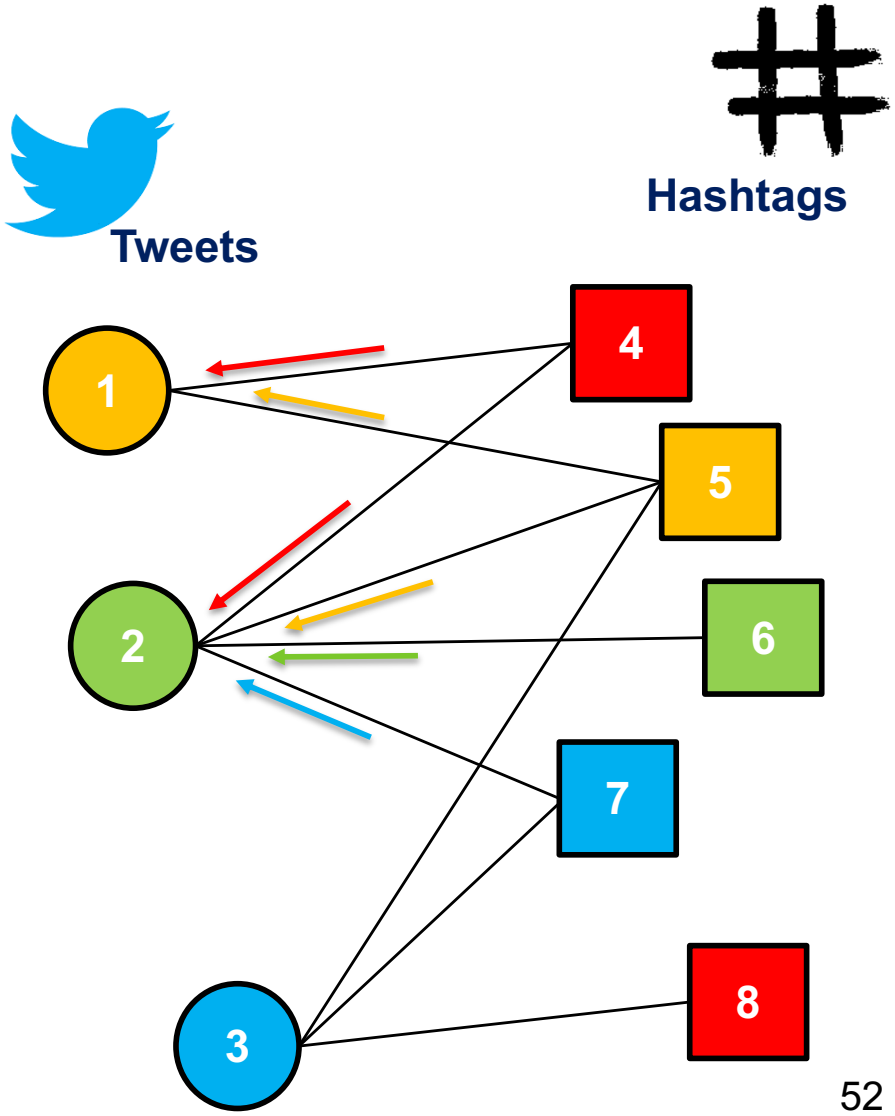


PLMP projection

a PageRank like approach

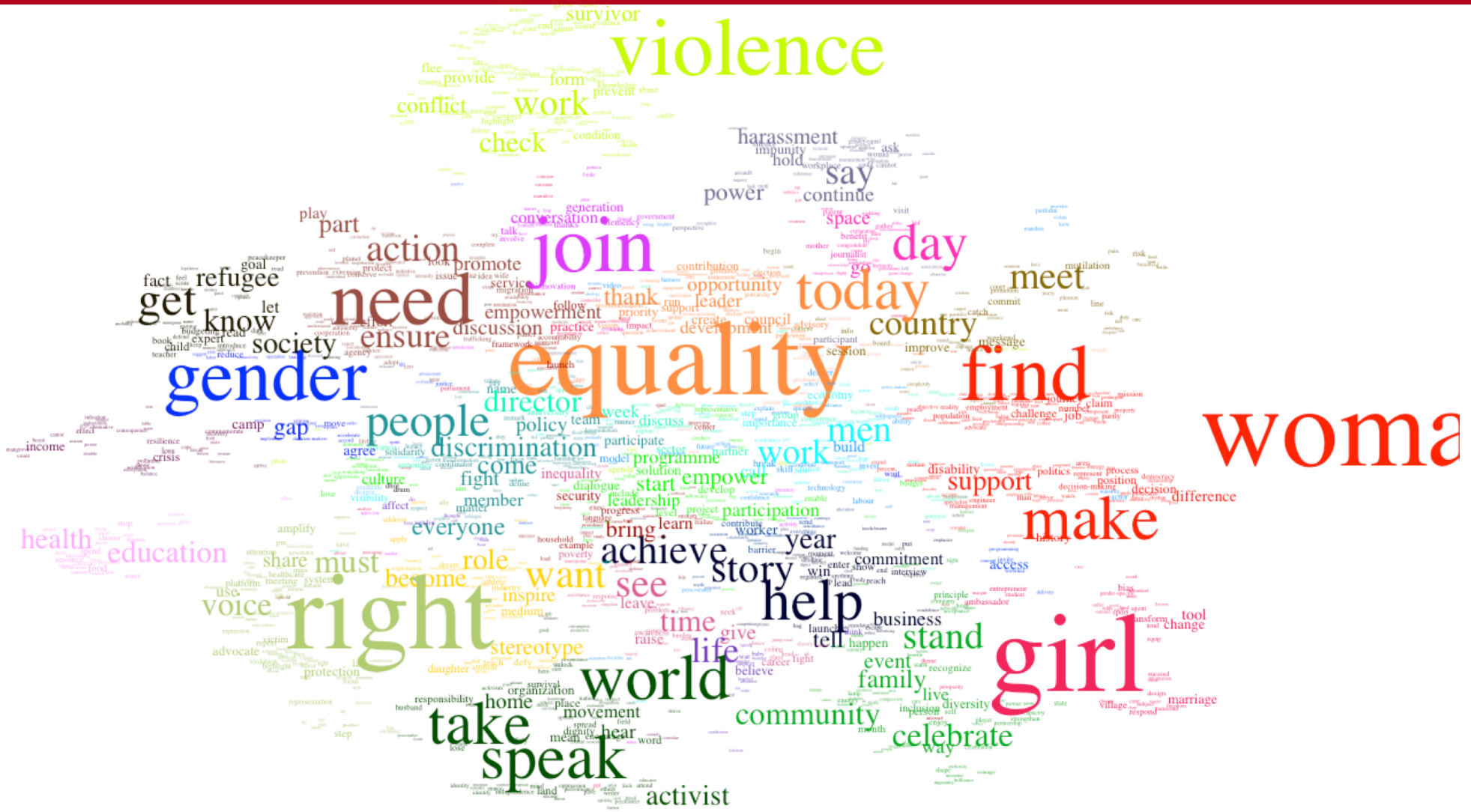
- Each hashtag captures the average sentiment value of the tweets it appears in
- Each tweet captures the average sentiment of the hashtags it contains

we iterate the two steps until convergence



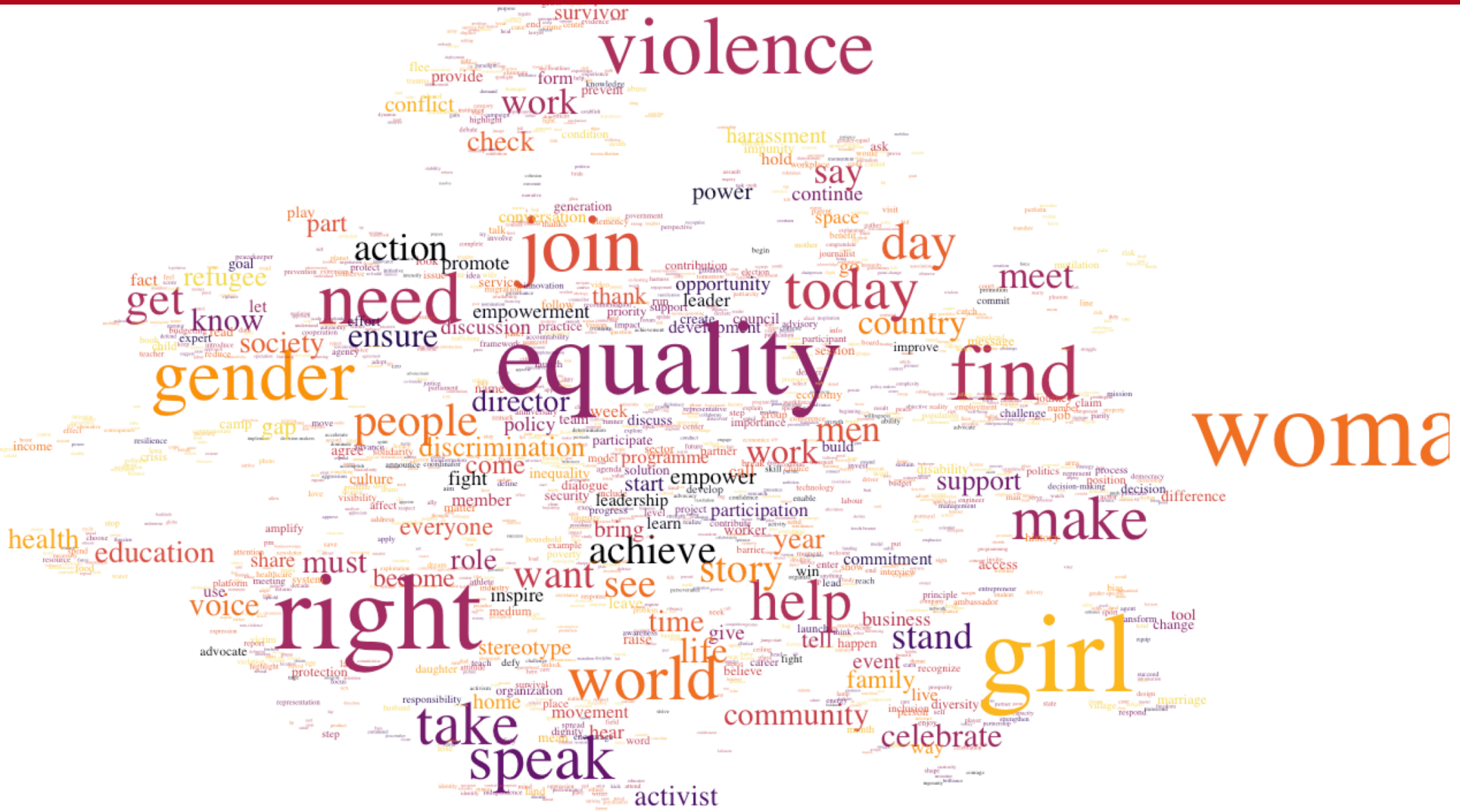
An example

community detection applied to semantic networks = topic detection



#metoo tweets

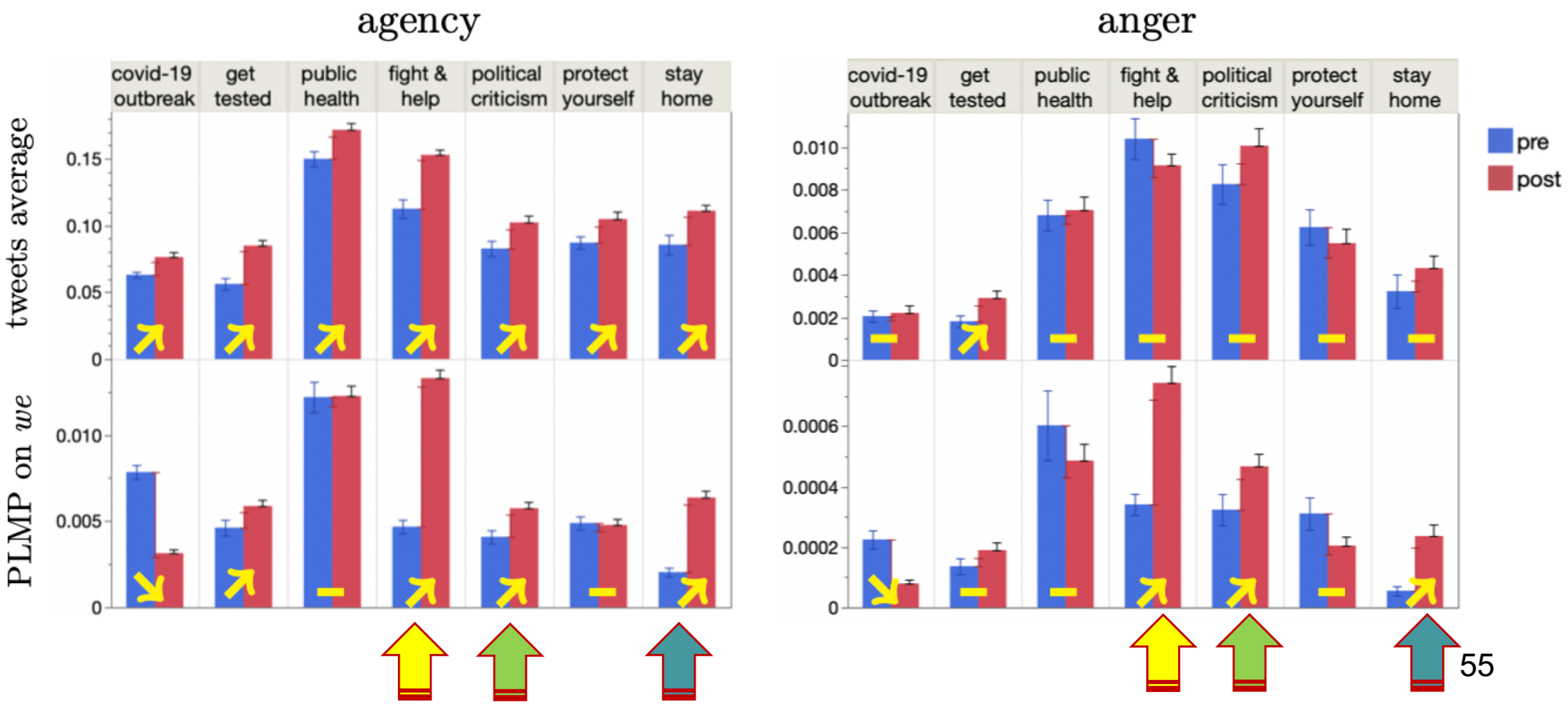
PLMP agency values



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



Takeaways

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