UNIVERSITÀ DEGLI STUDI DI PADOVA

Social Network Analysis

A.Y. 23/24

Communication Strategies

PageRank

a centrality measure based on the web

UNIVERSITÀ DEGLI STUDI DI PADOVA

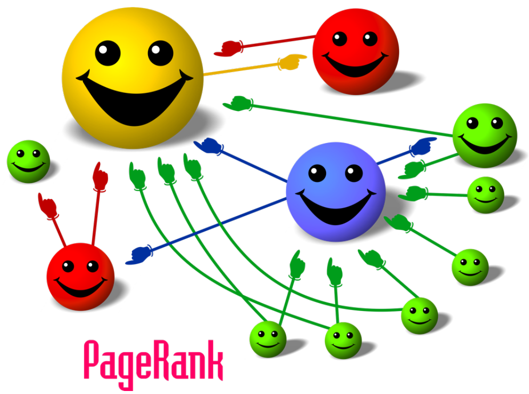
What is PageRank?

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]

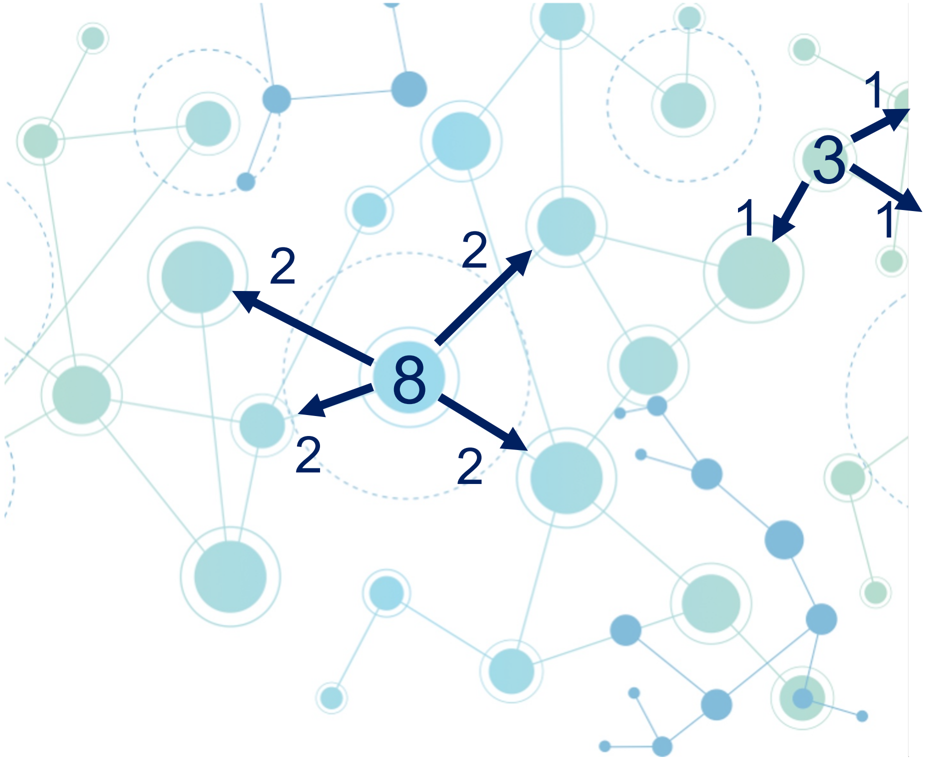
How to organise the web?

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



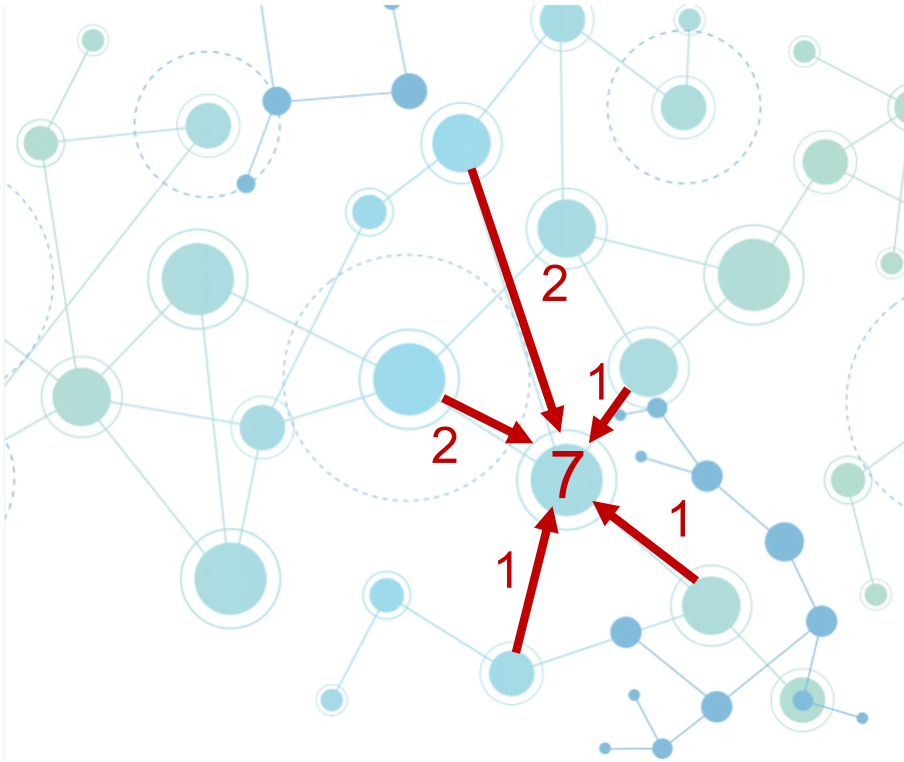
in the web this corresponds to the idea that starting from a web page you choose with equal probability one of the sites linked by the page

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

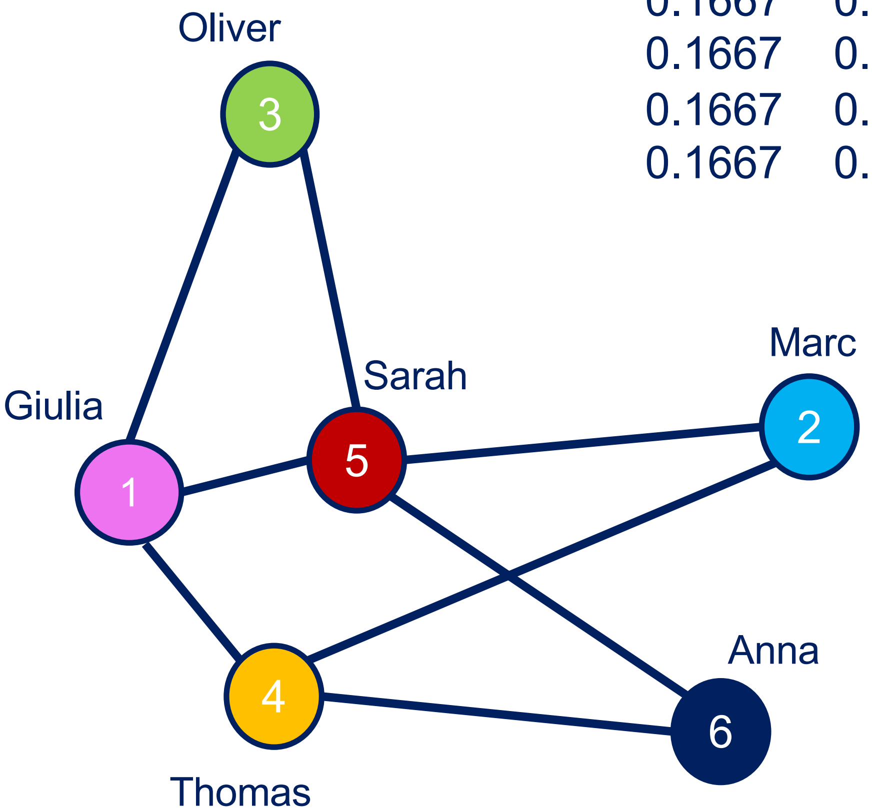


Example

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 !!!

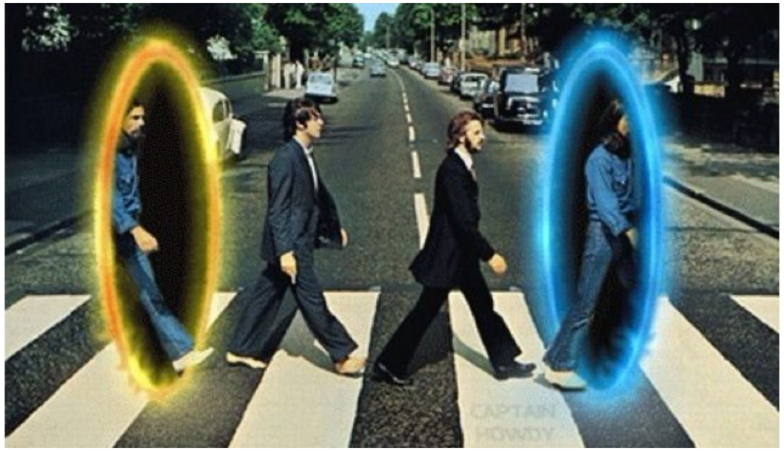
10	20	50	75	100	
0.1783	0.1848	0.1874	0.1875	0.1875	Giulia
0.1153	0.1222	0.1249	0.1250	0.1250	Marc
0.1242	0.1248	0.1250	0.1250	0.1250	Oliver
0.2020	0.1917	0.1876	0.1875	0.1875	Thomas
0.2649	0.2543	0.2501	0.2500	0.2500	Sarah
0.1153	0.1222	0.1249	0.1250	0.1250	Anna

Teleportation

as a method to strenghten the result

Idea:

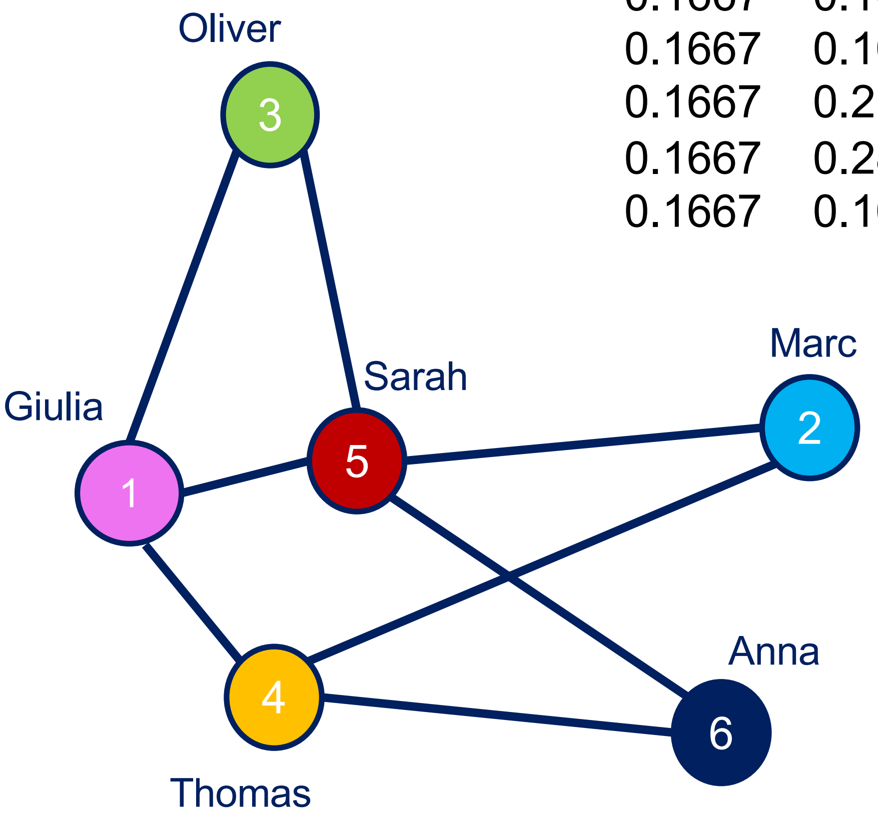
the surfer does not necessarily move to one of the links of the page she/he is viewing:



- 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

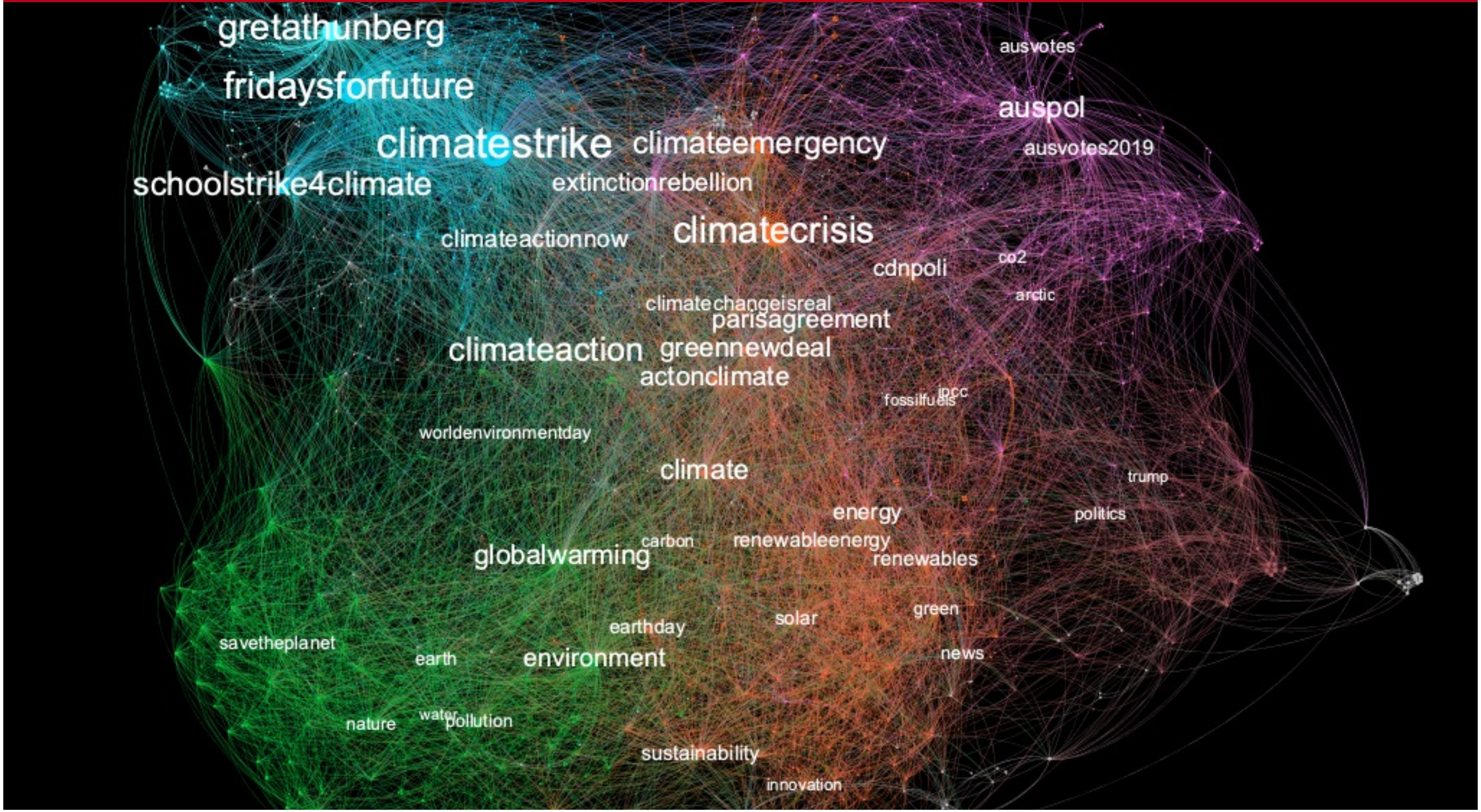
Visualizing PageRank

a comparison with degree centrality

UNIVERSITÀ DEGLI STUDI DI PADOVA

PageRank on a semantic network

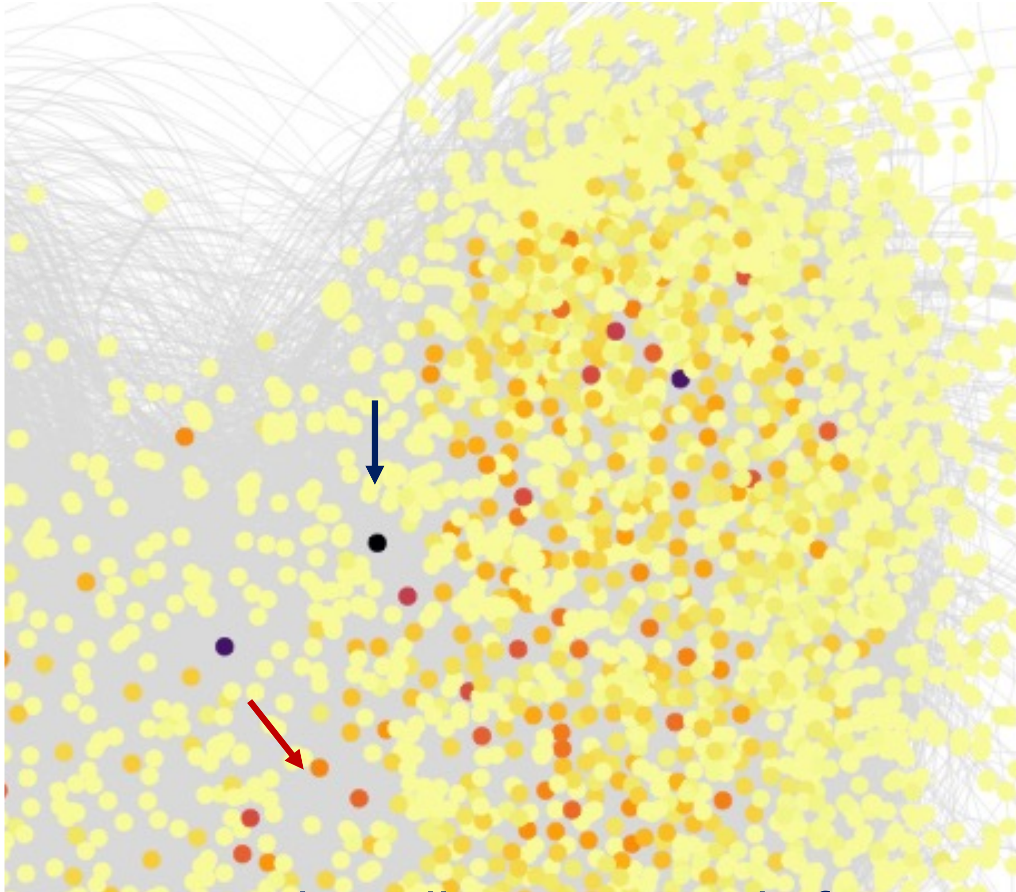
2019 hashtag network related to #climatechange
(from Twitter, after #gretathunberg)



Example of PageRank centrality

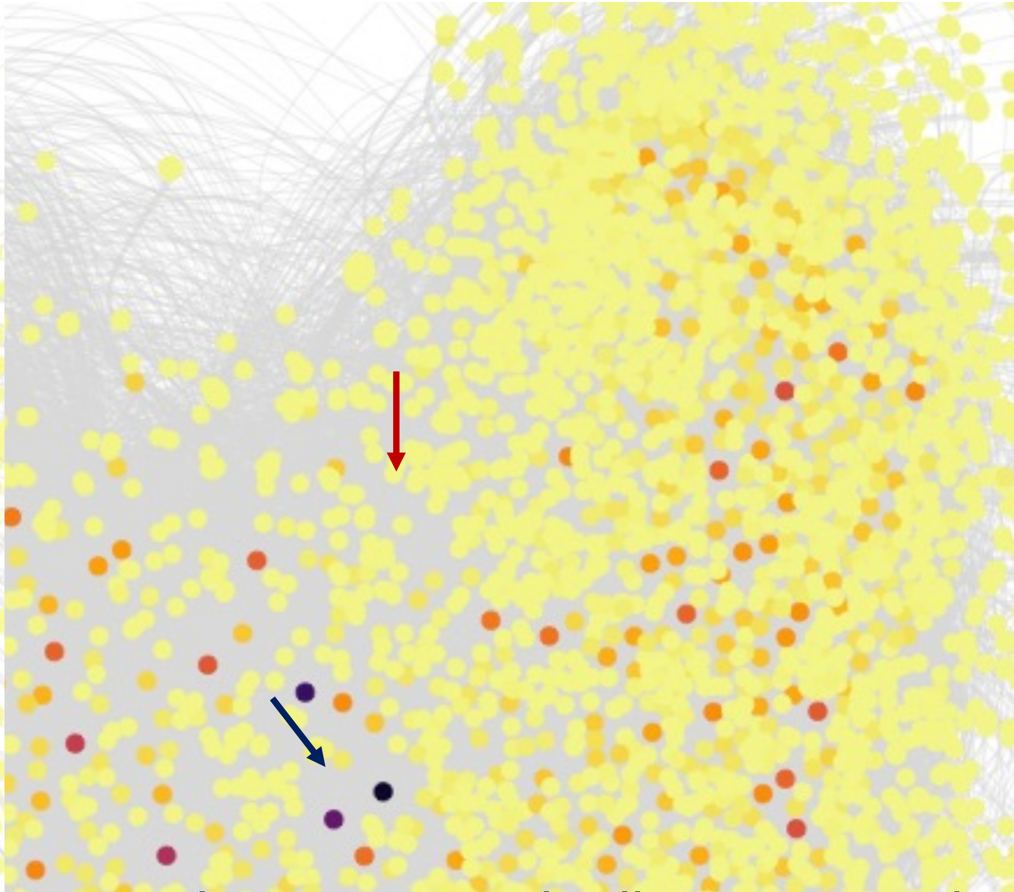
wikipedia administrator elections and vote history data
https://snap.stanford.edu/data/wiki-Vote.html

Authorities



uses the adjacency matrix for spreading

Hubs

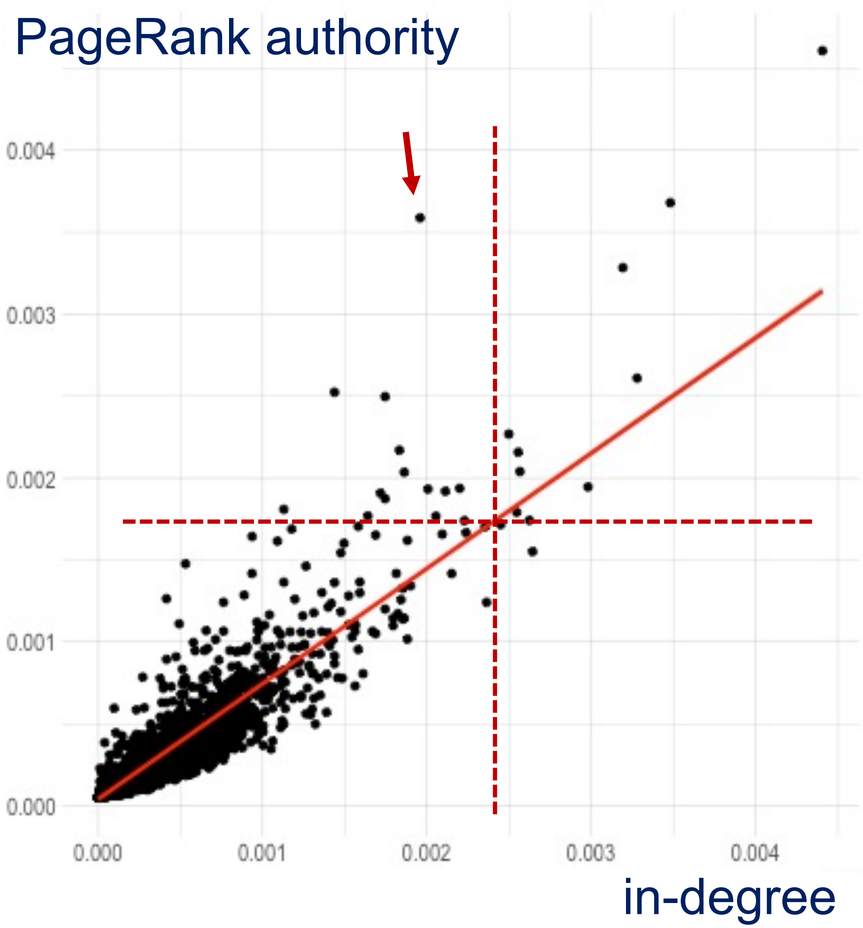


uses the transposed adjacency matrix for spreading (spreading backwards)

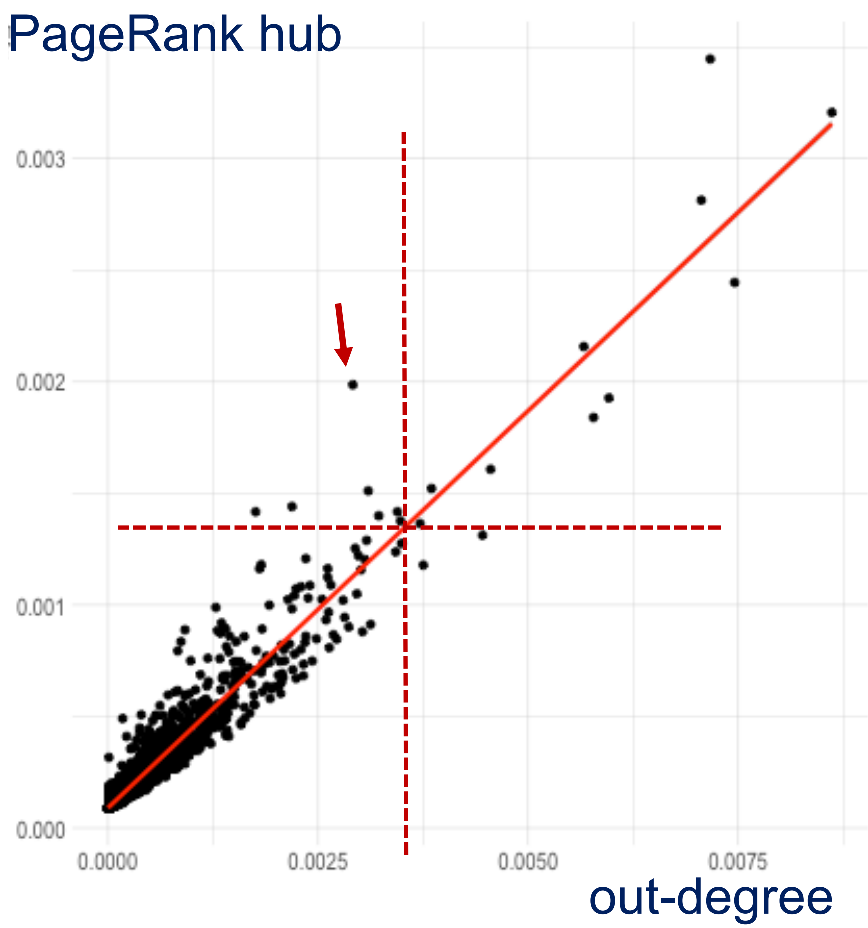
PageRank versus degree centrality

wikipedia administrator elections and vote history data

Authorities



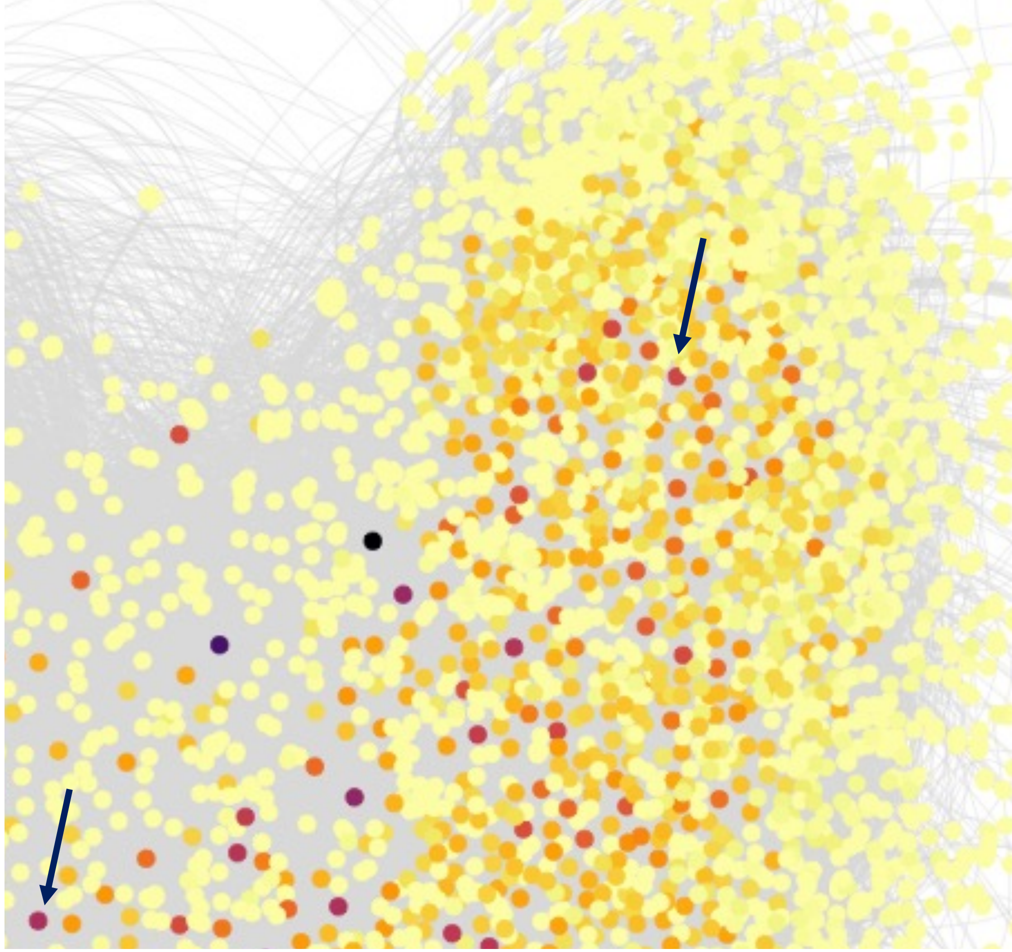
Hubs



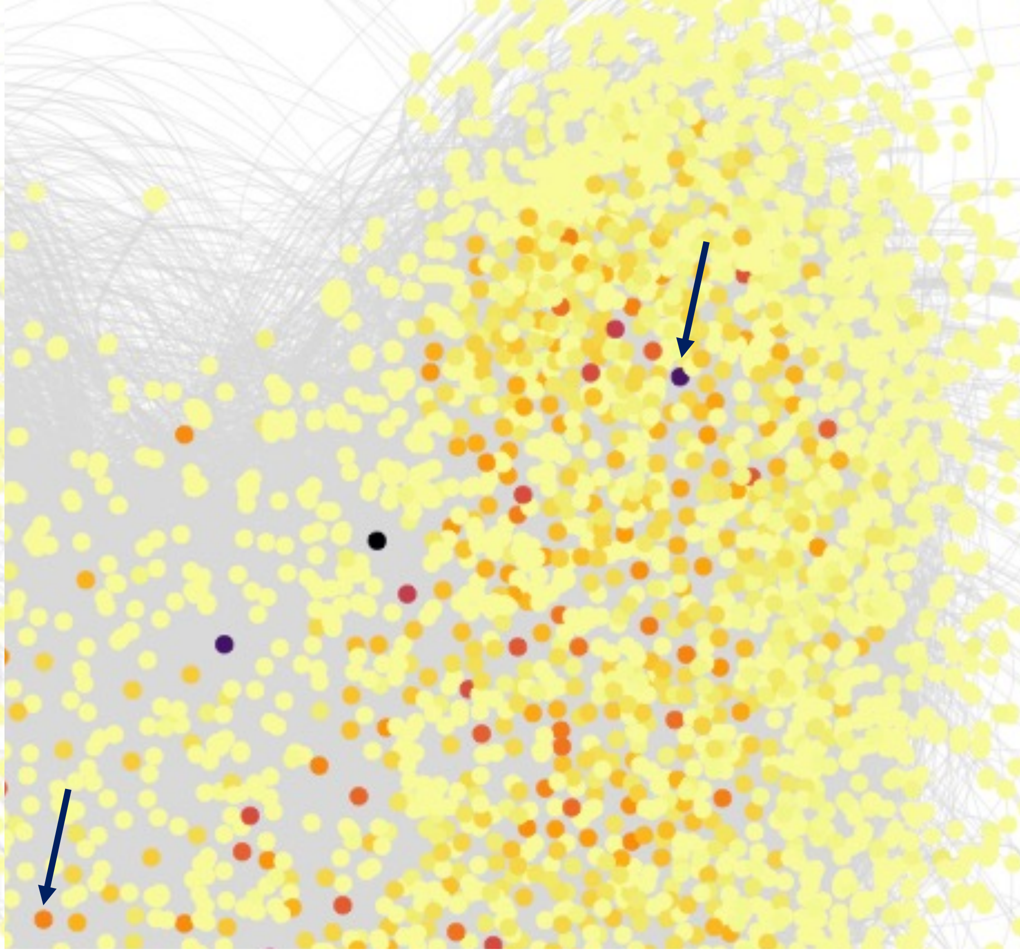
PageRank versus degree authorities

wikipedia administrator elections and vote history data

Degree



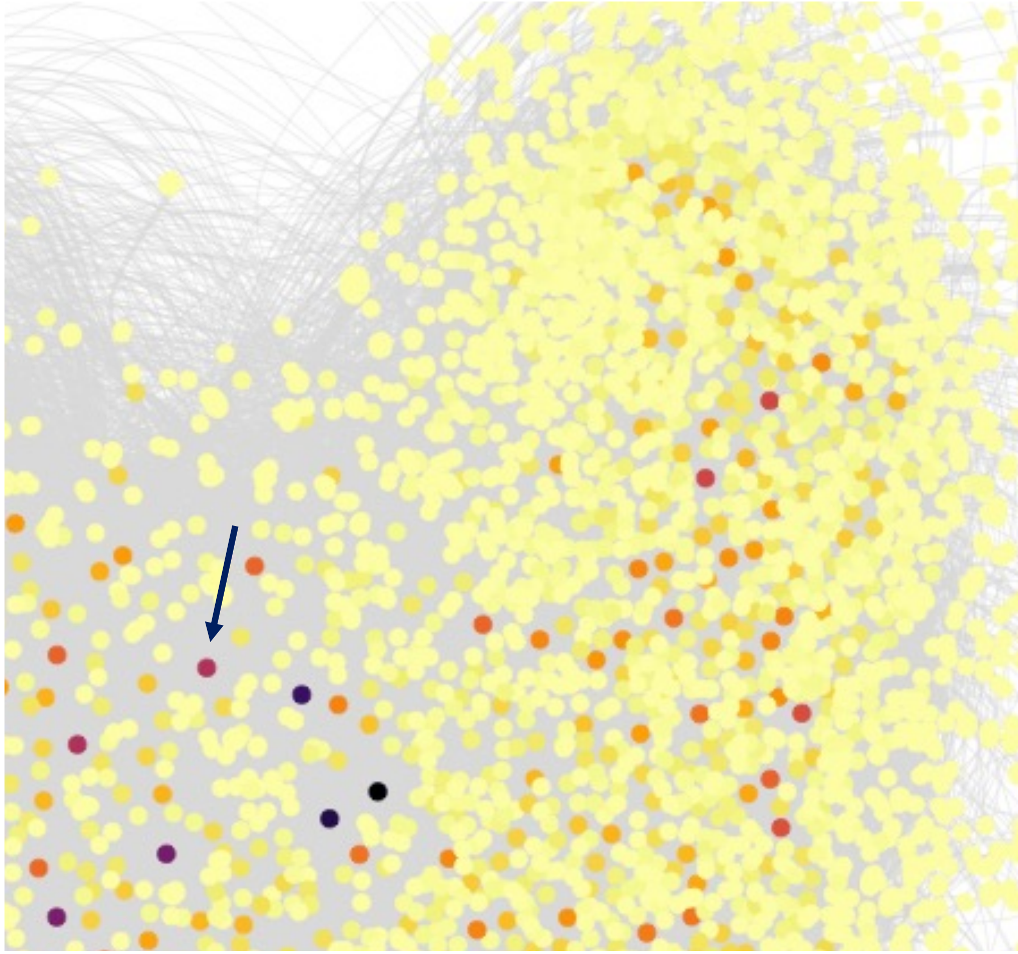
PageRank



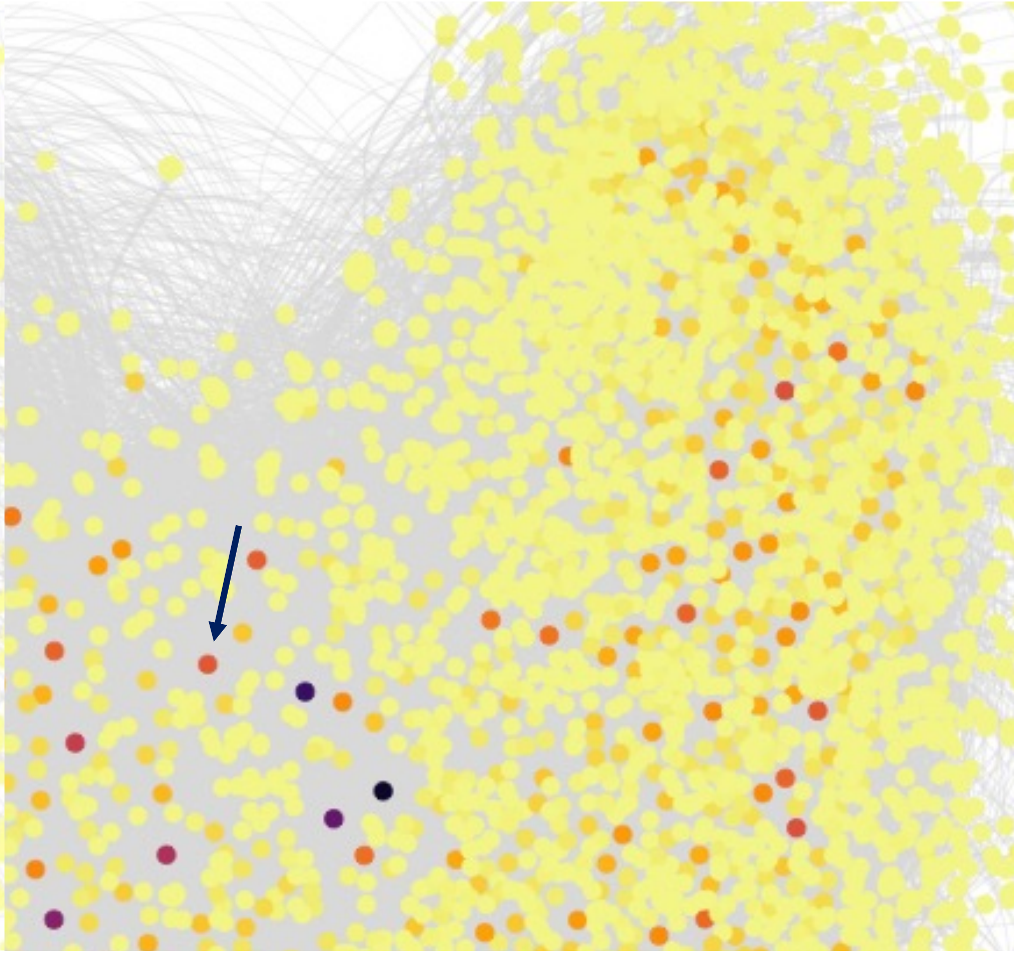
PageRank versus degree hubs

wikipedia administrator elections and vote history data

Degree



PageRank



Local PageRank

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

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*

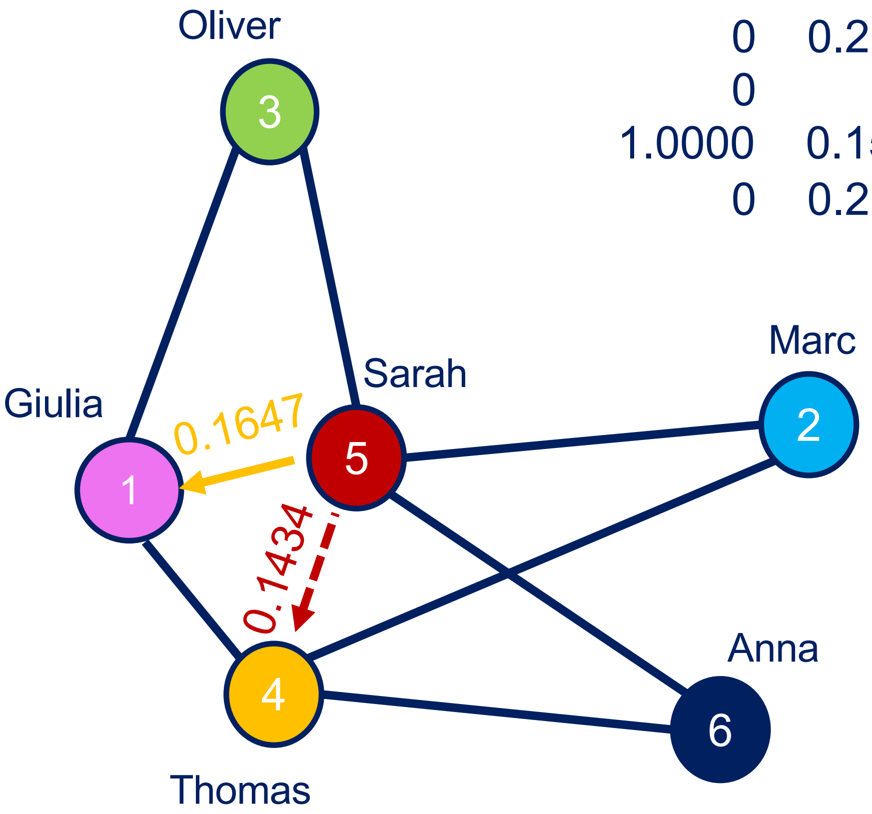
Result

- 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

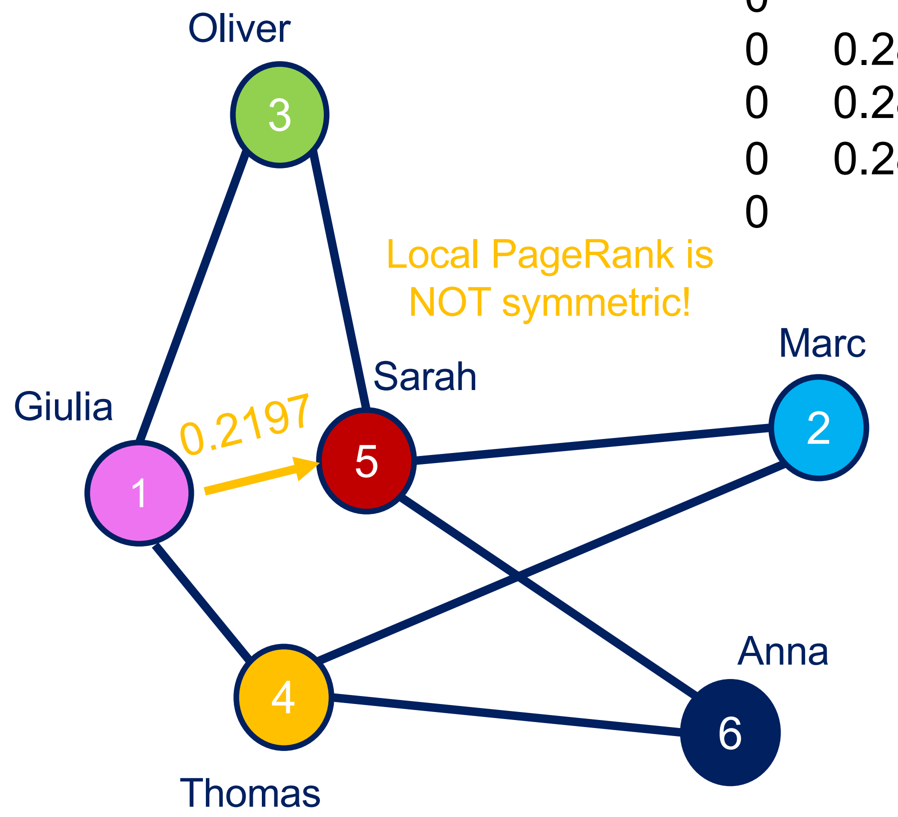


10	*20*	*50*	*75*	*100*	
0.1743	0.1653	0.1647	0.1647	0.1647	Giulia
0.1238	0.1144	0.1138	0.1138	0.1138	Marc
0.1206	0.1199	0.1199	0.1199	0.1199	Oliver
0.1285	0.1426	0.1434	0.1434	0.1434	Thomas
0.3290	0.3435	0.3444	0.3444	0.3444	Sarah
0.1238	0.1144	0.1138	0.1138	0.1138	Anna

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



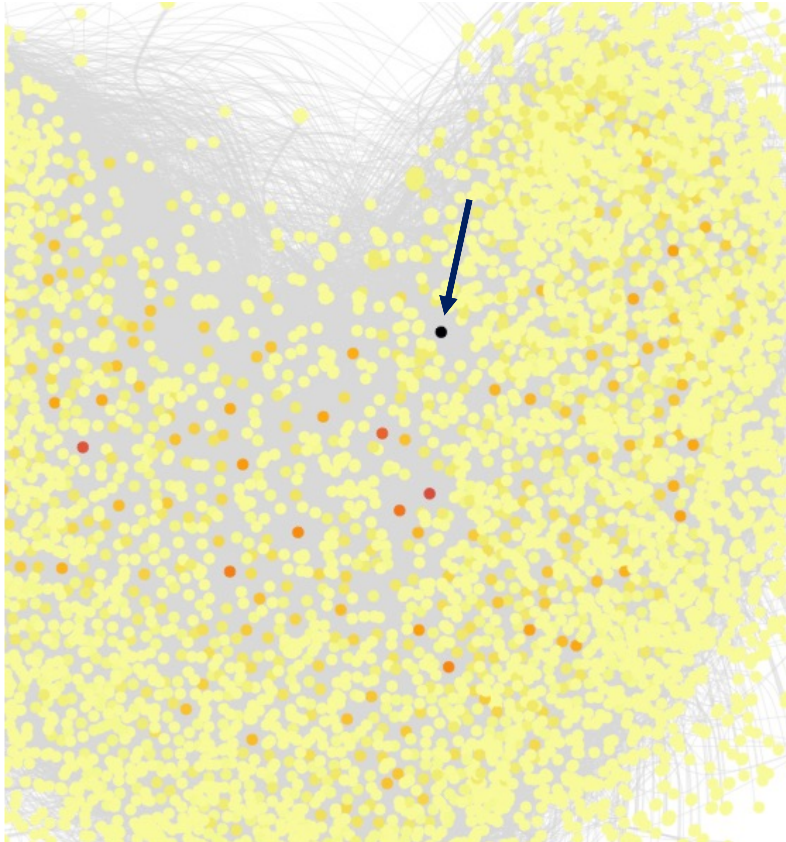
10	20	50	75	100	
0.2909	0.2985	0.2989	0.2989	0.2989	Giulia
0.0848	0.0926	0.0931	0.0931	0.0931	Marc
0.1309	0.1313	0.1314	0.1314	0.1314	Oliver
0.1763	0.1645	0.1638	0.1638	0.1638	Thomas
0.2324	0.2204	0.2197	0.2197	0.2197	Sarah
0.0848	0.0926	0.0931	0.0931	0.0931	Anna

UNIVERSITÀ DEGLI STUDI DI PADOVA

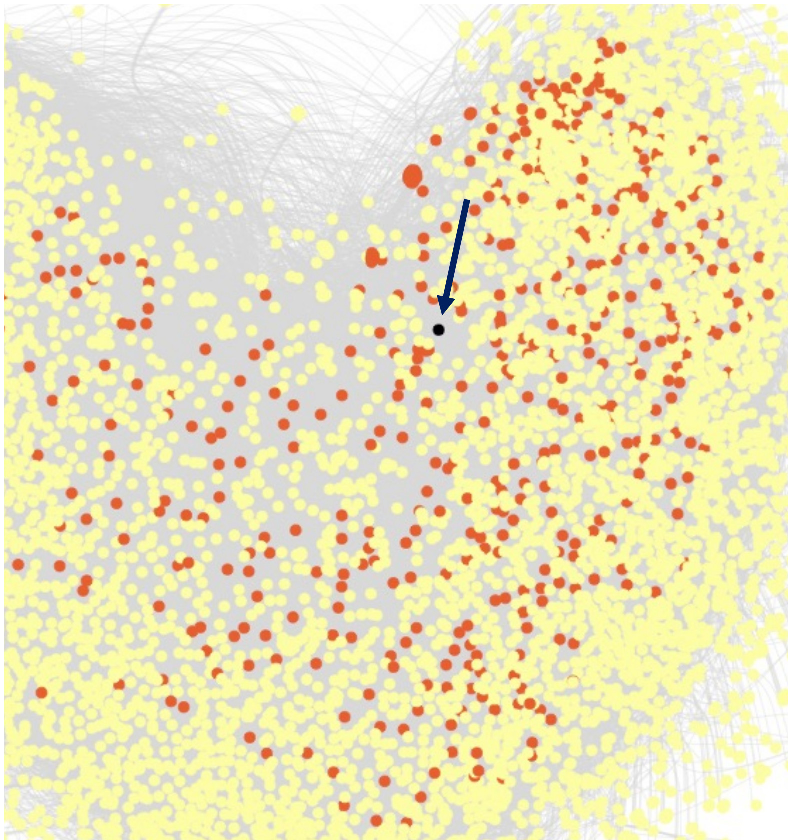
Local PageRank versus degree

authorities

Local PageRank



1-hop out-neighbours

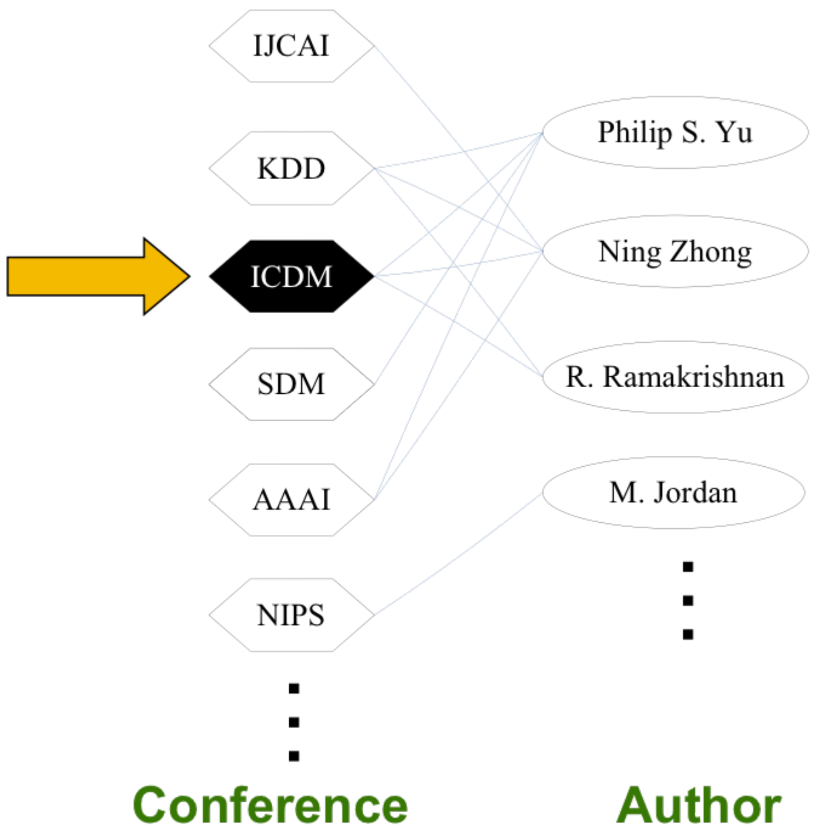


neighbours authority score =
local node → neighbours

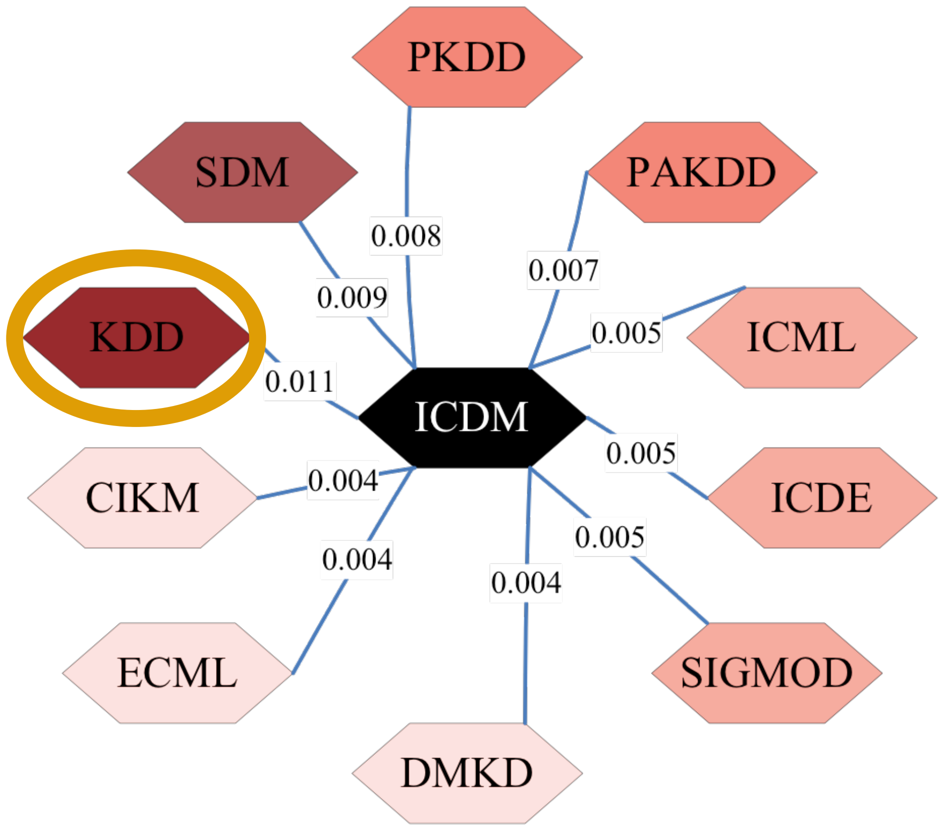
UNIVERSITÀ DEGLI STUDI DI PADOVA

Example

what is the most related copnference to ICDM?



Top 10 ranking results



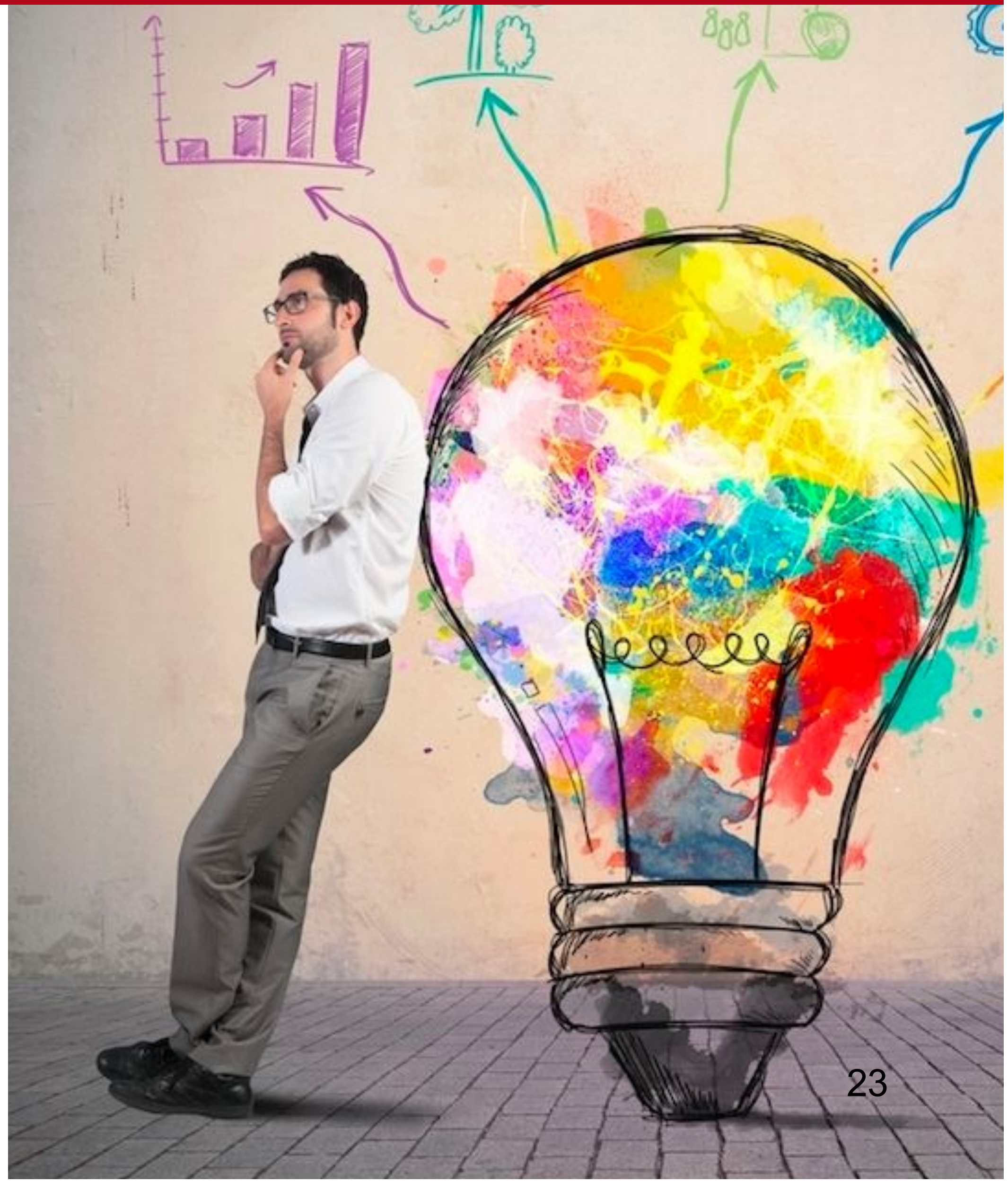
ICDM = international conf. on data mining
KDD = knowledge discovery and data mining

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



UNIVERSITÀ DEGLI STUDI DI PADOVA

TopicSpecific PageRank example

in semantic networks

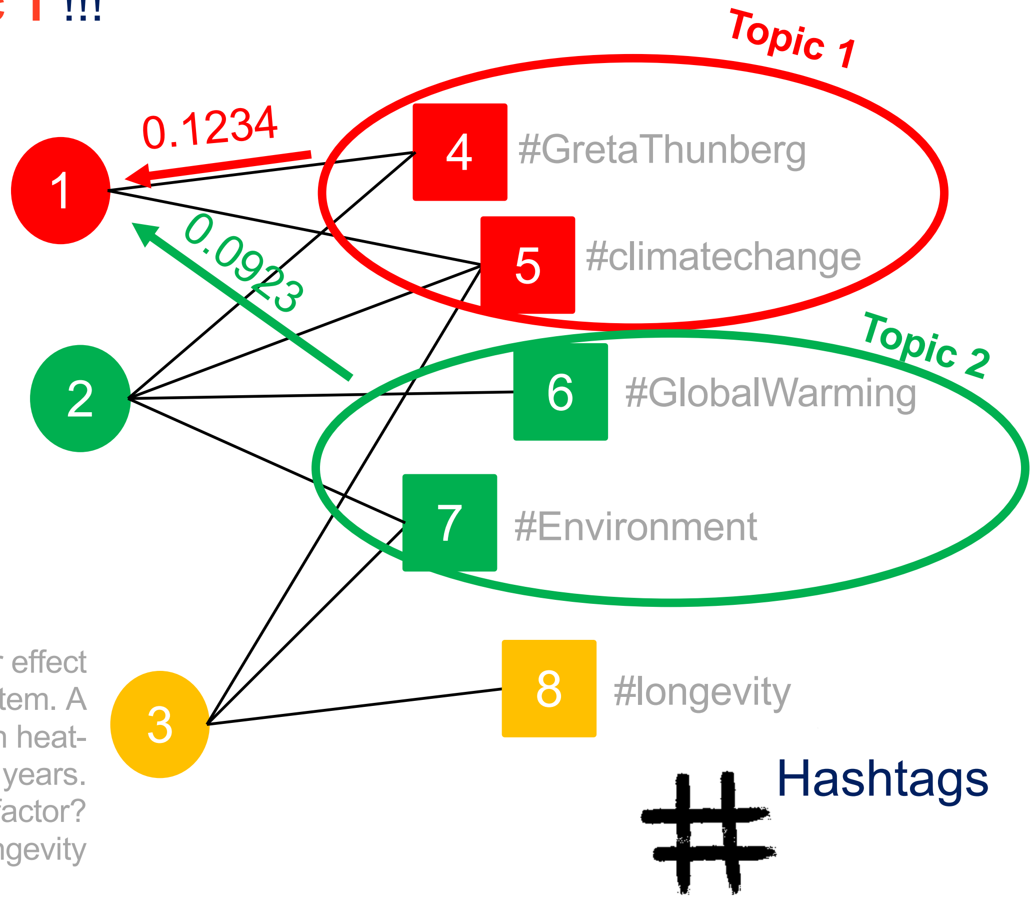
Tweet 1 is assigned to Topic 1 !!!

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

Tweets

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



Closeness and Harmonic centralities

importance of nodes as spreaders of information

Closeness centrality

a definition

Closeness centrality

From Wikipedia, the free encyclopedia

In a connected graph, **closeness centrality** (or **closeness**) of a node is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the *closer* it is to all other nodes.

Closeness was defined by Bavelas (1950) as the reciprocal of the **farness**,[1][2] that is:

$$C(x) = \frac{1}{\sum_y d(y,x)}.$$

where $d(y,x)$ is the distance between vertices x and y. When

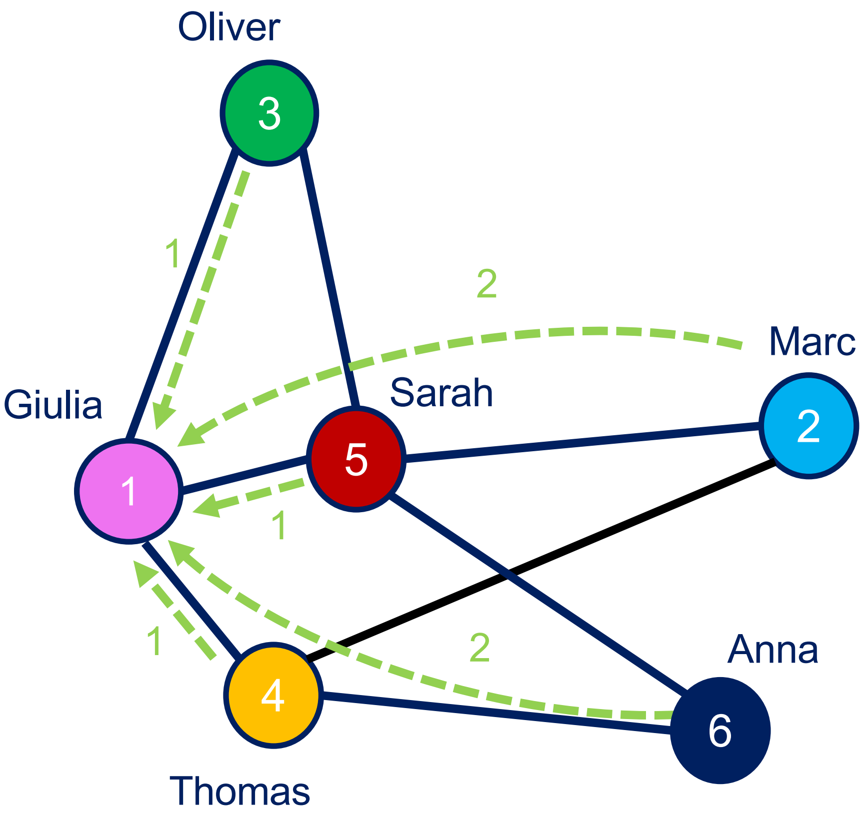
Rationale: the node which is the easiest to reach, the one which is the best for spreading information

An example

on how to calculate closeness centrality

count the lengths of the shortest paths leading to Giulia

1 + 2 + 1 + 2 + 1 = 7



Closeness

0.1429	Giulia
0.1250	Marc
0.1250	Oliver
0.1429	Thomas
0.1667	Sarah
0.1250	Anna

Sarah is the preferred node for spreading information

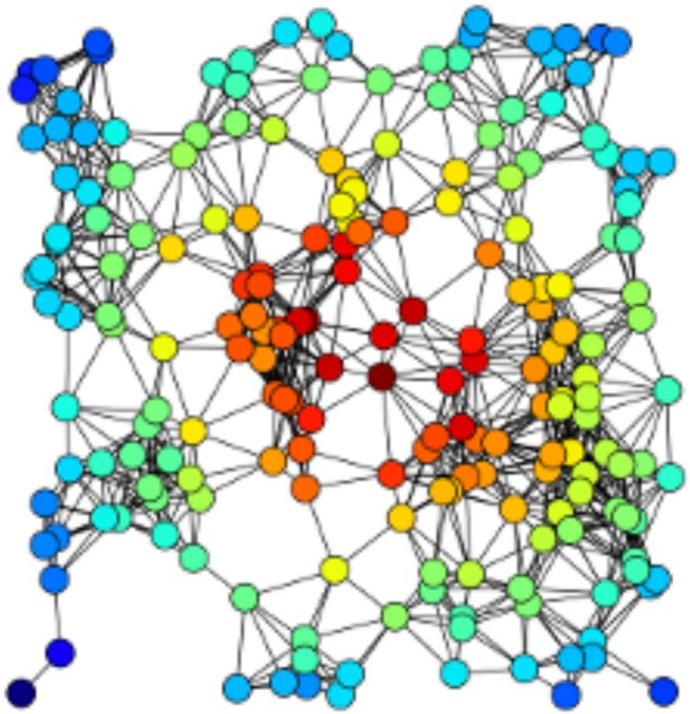
C(Giulia) = 1/7
= 0.1429

Università degli Studi di Padova

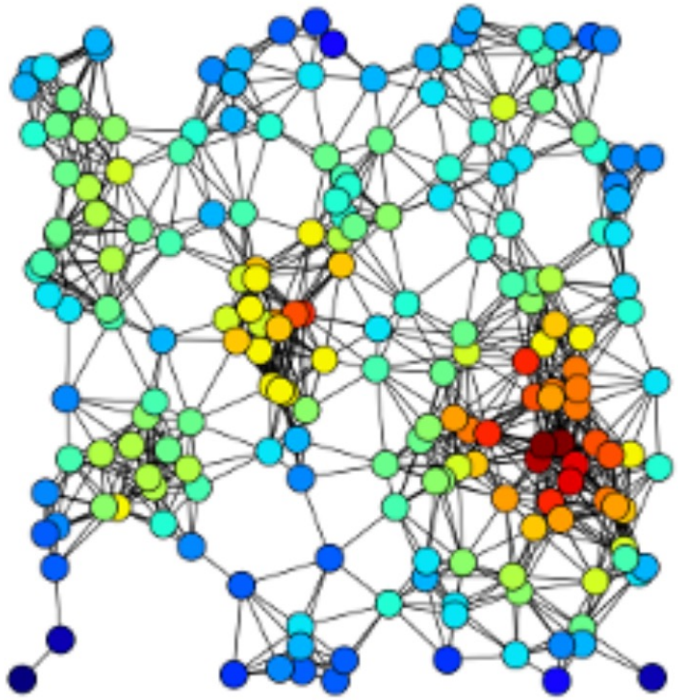
Closeness versus degree centrality

a graphical interpretation

Closeness



Degree



Harmonic centrality

a definition

In disconnected graphs [edit]

When a graph is not strongly connected, a widespread idea is that of using the sum of reciprocal of distances, instead of the reciprocal of the sum of distances, with the convention $1/\infty = 0$:

$$H(x) = \sum_{y \neq x} \frac{1}{d(y, x)}.$$

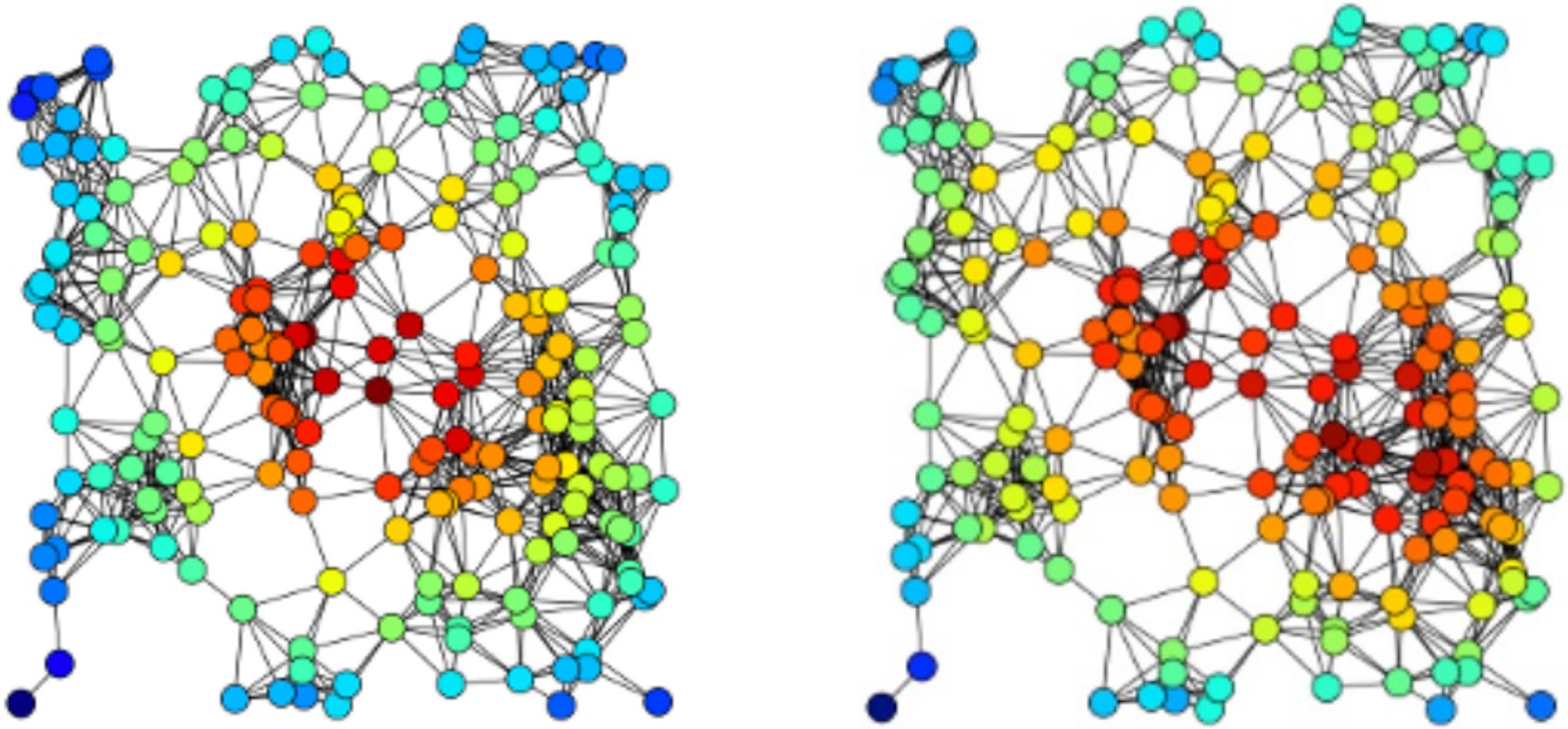
The most natural modification of Bavelas's definition of closeness is following the general principle proposed by Marchiori and Latora (2000)[3] that in graphs with infinite distances the harmonic mean behaves better than the arithmetic mean. Indeed, Bavelas's closeness can be described as the denormalized reciprocal of the arithmetic mean of distances, whereas harmonic centrality is the denormalized reciprocal of the harmonic mean of distances.

Closeness versus harmonic centrality

a graphical interpretation

Closeness

Harmonic



Betweenness centrality

importance of nodes as bridges or brokers

Betweenness centrality

a definition

Betweenness centrality

From Wikipedia, the free encyclopedia

In graph theory, **betweenness centrality** is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.

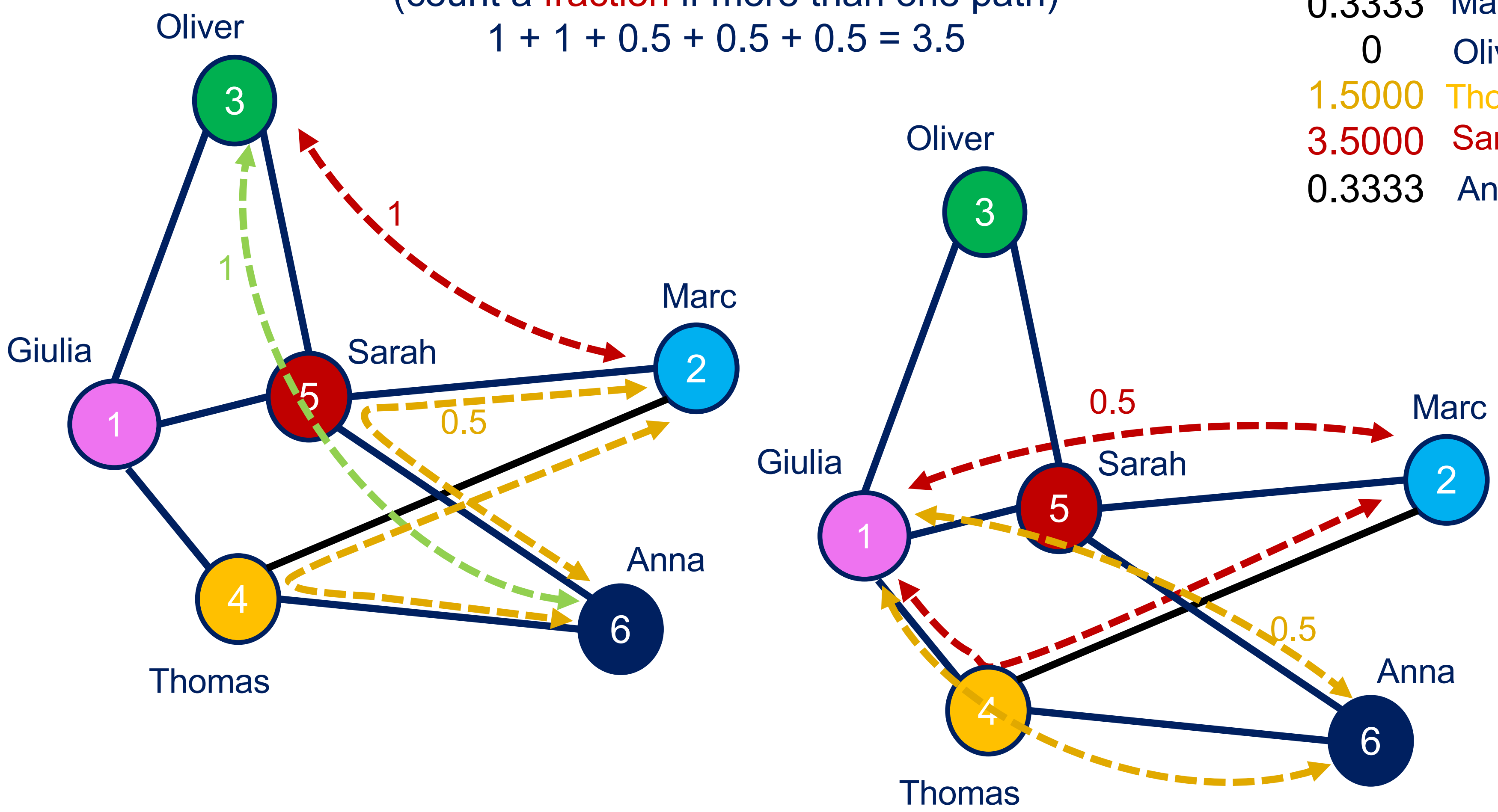
Betweenness centrality was devised as a general measure of centrality:[1] it applies to a wide range of problems in network theory, including problems related to social networks, biology, transport and scientific cooperation. Although earlier authors have intuitively described centrality as based on betweenness, Freeman (1977) gave the first formal definition of betweenness centrality.

Rationale: the node which takes you elsewhere (bridge, broker)

An example

on how to calculate betweenness centrality

count the # of shortest paths
passing through Sarah
(count a fraction if more than one path)
1 + 1 + 0.5 + 0.5 + 0.5 = 3.5



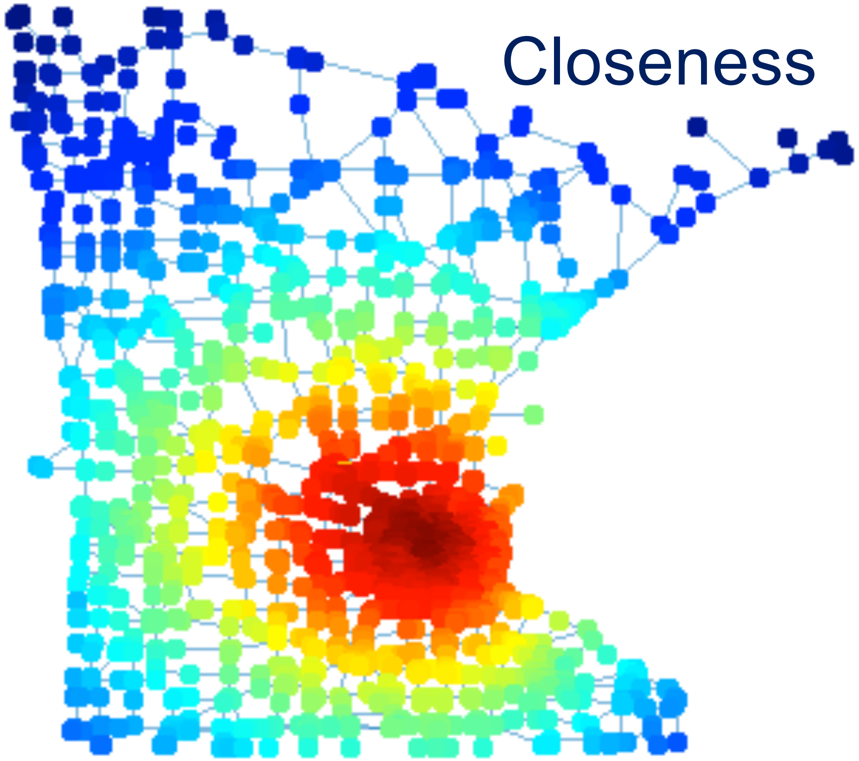
Betweenness

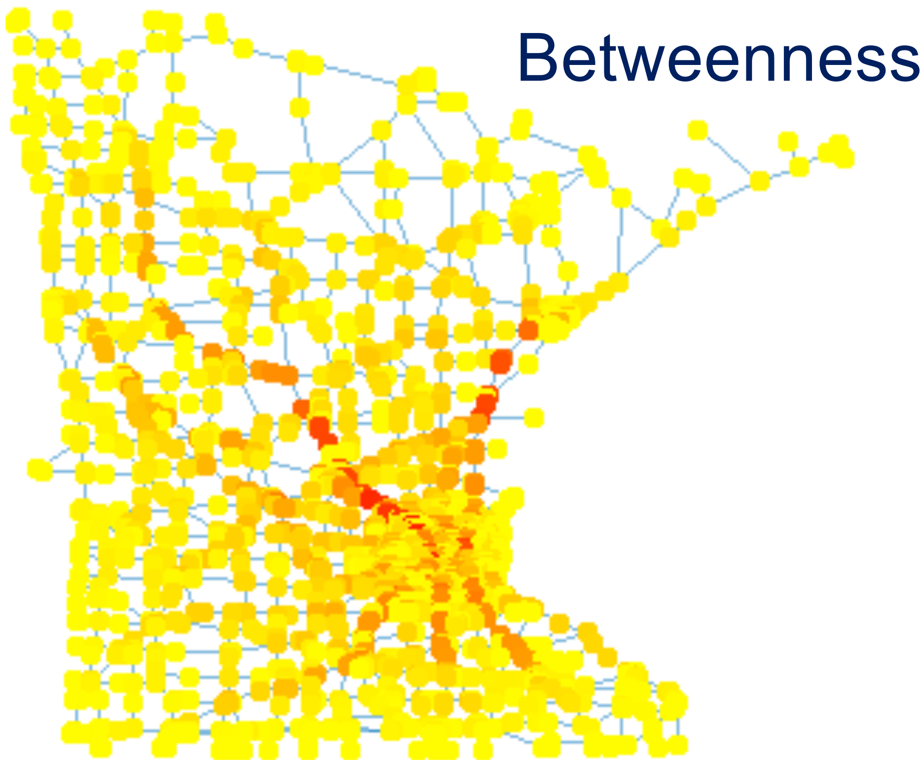
1.3333	Giulia
0.3333	Marc
0	Oliver
1.5000	Thomas
3.5000	Sarah
0.3333	Anna

Closeness vs betweenness centrality

a graphical interpretation

Minnesota road network



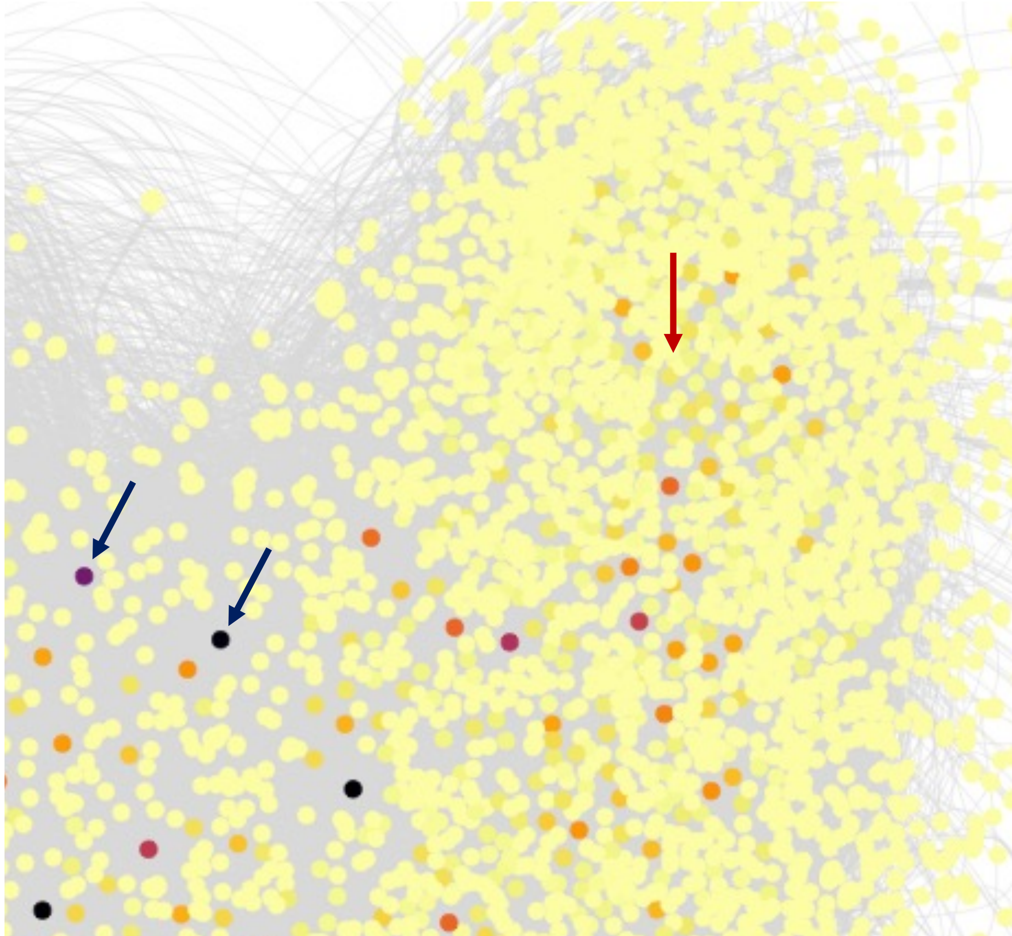


Closeness is a measure of center of gravity (best node to spread info)

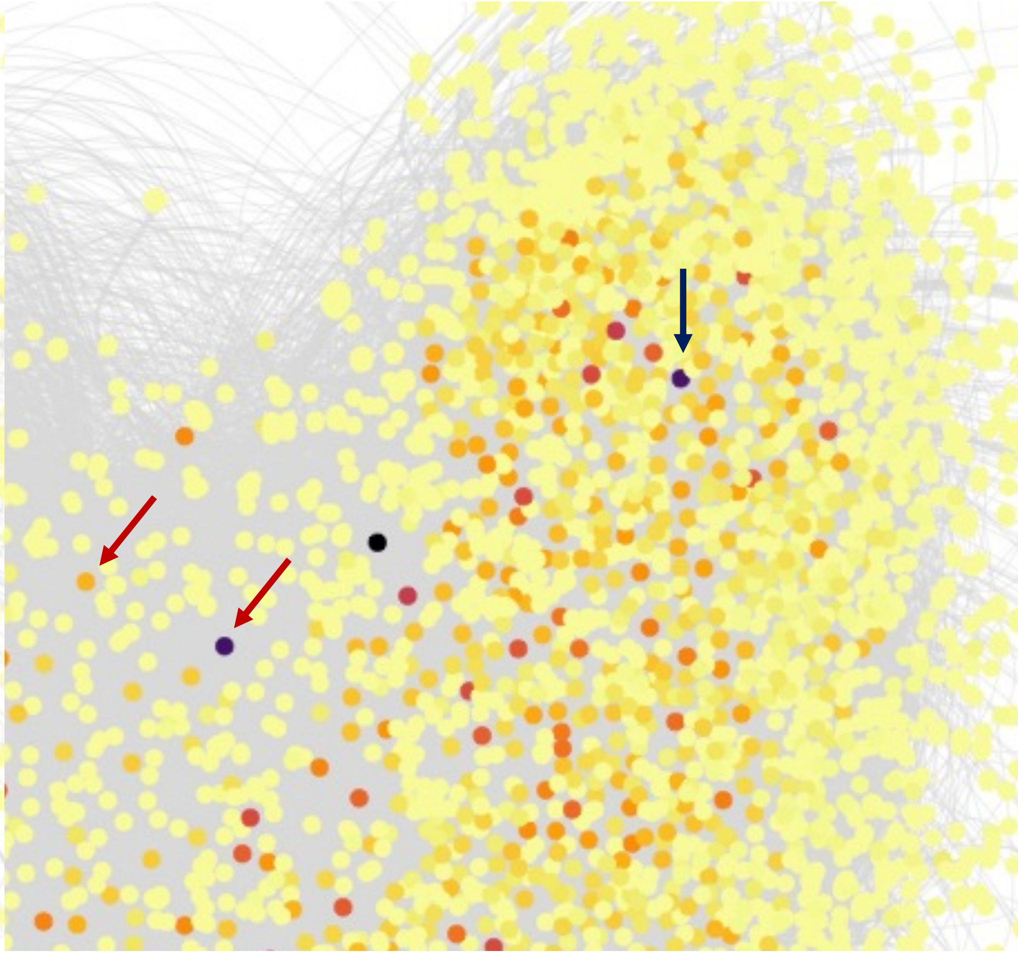
Betweenness is a measure of brokerage (i.e., being a bridge)

Betweenness vs PageRank centrality
wiki vote network

Betweenness



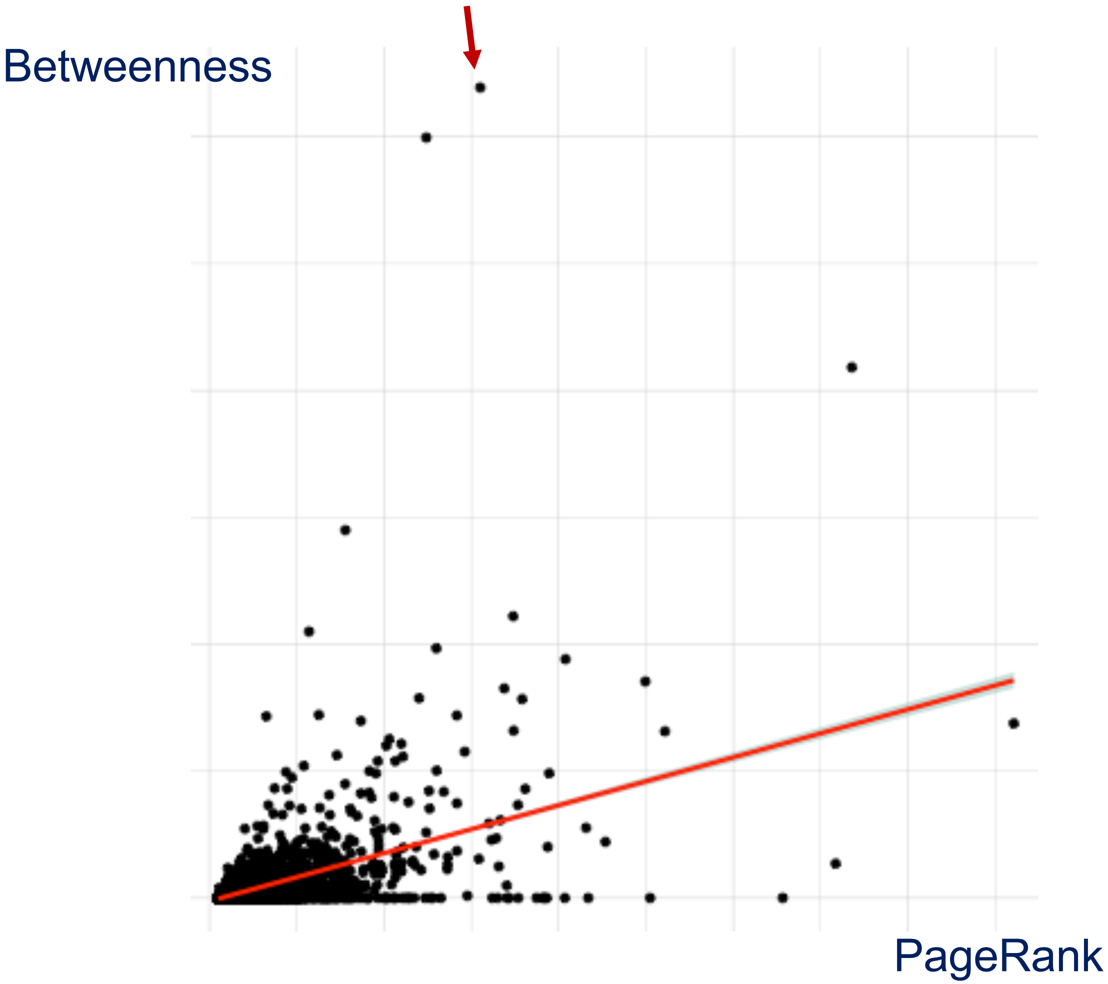
PageRank



UNIVERSITÀ DEGLI STUDI DI PADOVA

Betweenness vs PageRank centrality

a correlation view



Clustering coefficient

how tightly linked is the network locally

UNIVERSITÀ DEGLI STUDI DI PADOVA

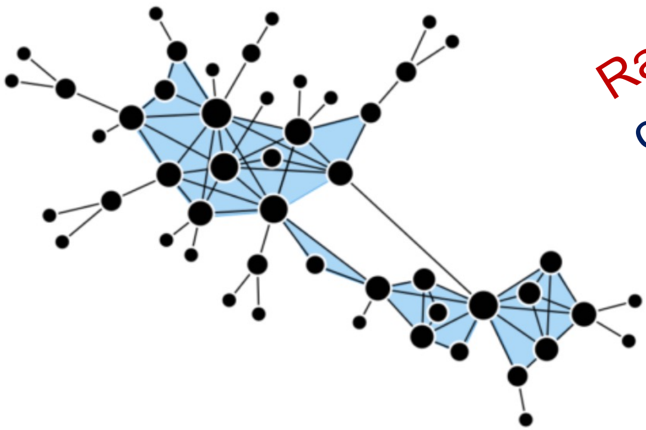
Clustering coefficient

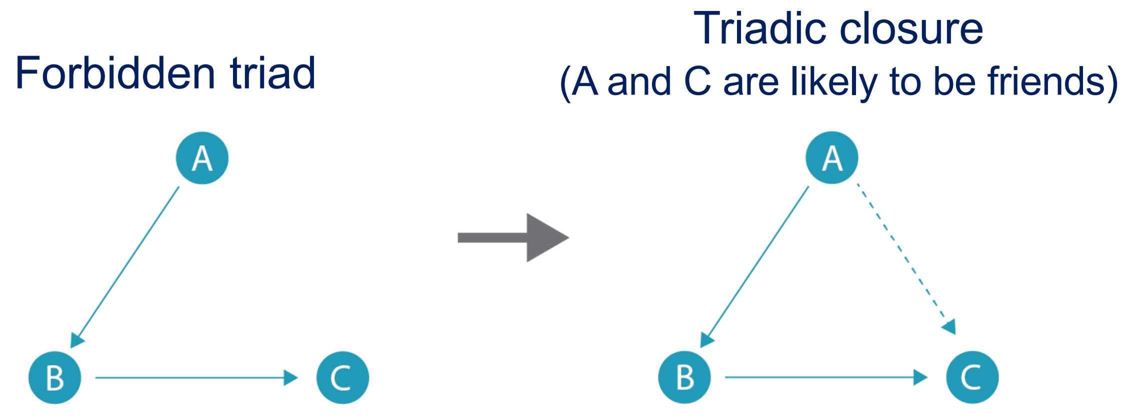
a definition

Local clustering coefficient [edit]

The **local clustering coefficient** of a vertex (node) in a graph quantifies how close its neighbours are to being a clique (complete graph). Duncan J. Watts and Steven Strogatz introduced the measure in 1998 to determine whether a graph is a small-world network.

Rationale: how strongly connected is the network locally / general indication of the graph's tendency to be organized into clusters





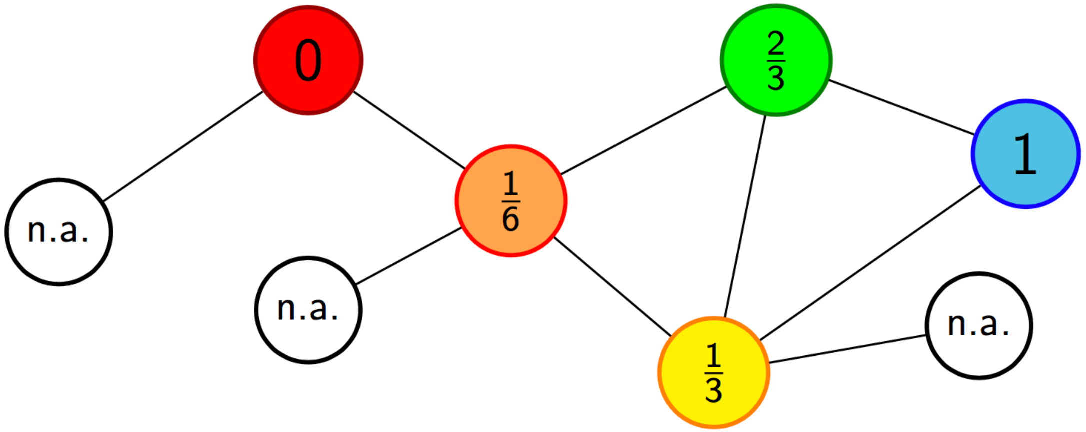
Triadic closure

- A and C are likely to have the opportunity to meet because they have a common friend B
- The fact that A and C is friends with B gives them the basis of trusting each other
- B may have the incentive to bring A and C together, as it may be hard for B to maintain disjoint relationships

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

Local clustering coefficient

a measure of triadic closures



Local Clustering coefficient C_i counts the fraction of pairs of neighbours N_i which form a triadic closure with node i

$$C_i = \frac{1}{|\mathcal{N}_i|(|\mathcal{N}_i|-1)} \sum_{\substack{(j,k)\in\mathcal{N}_i^2 \\ j\neq k}} \mathrm{tc}_{i,j,k}$$

equal to diag($\mathbf{A}^3$)

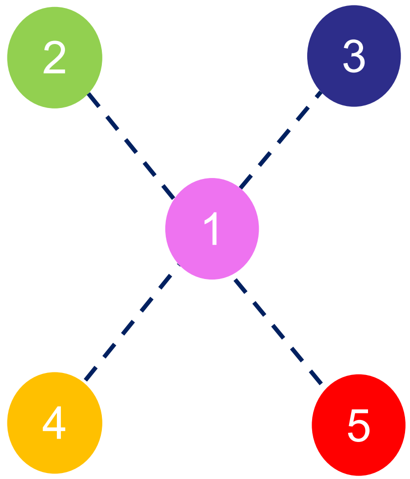
where $tc_{ijk} = 1$ if the triplet (i,j,k) forms a triadic closure, and zero otherwise

Local clustering coefficient

examples

not connected neighbourhood

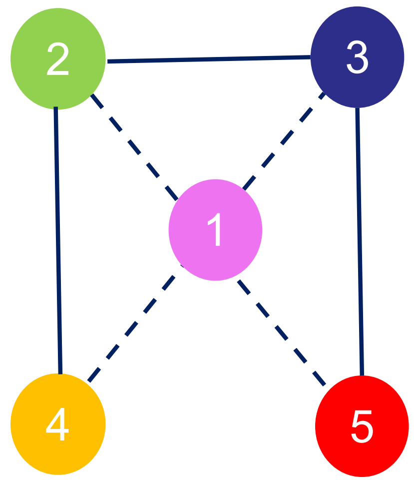
$<C> = 0$



$C_1 = 0$

weakly connected neighbourhood

$<C> = 0.766$



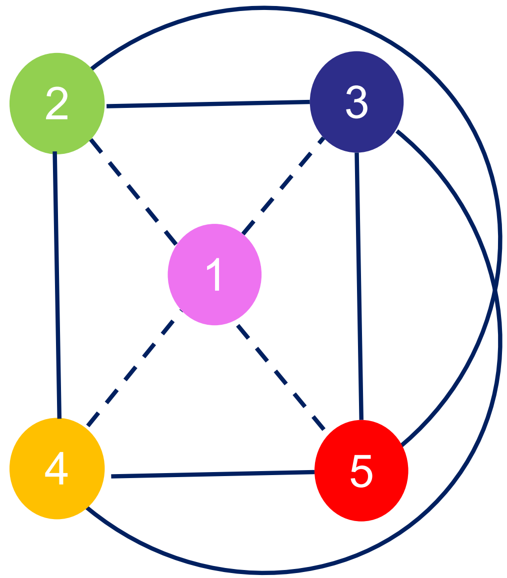
$C_1 = ½ = 3 / (4x3/2)$

$C_2 = C_3 = ⅔$

$C_4 = C_5 = 1$

strongly connected neighbourhood

$<C> = 1$



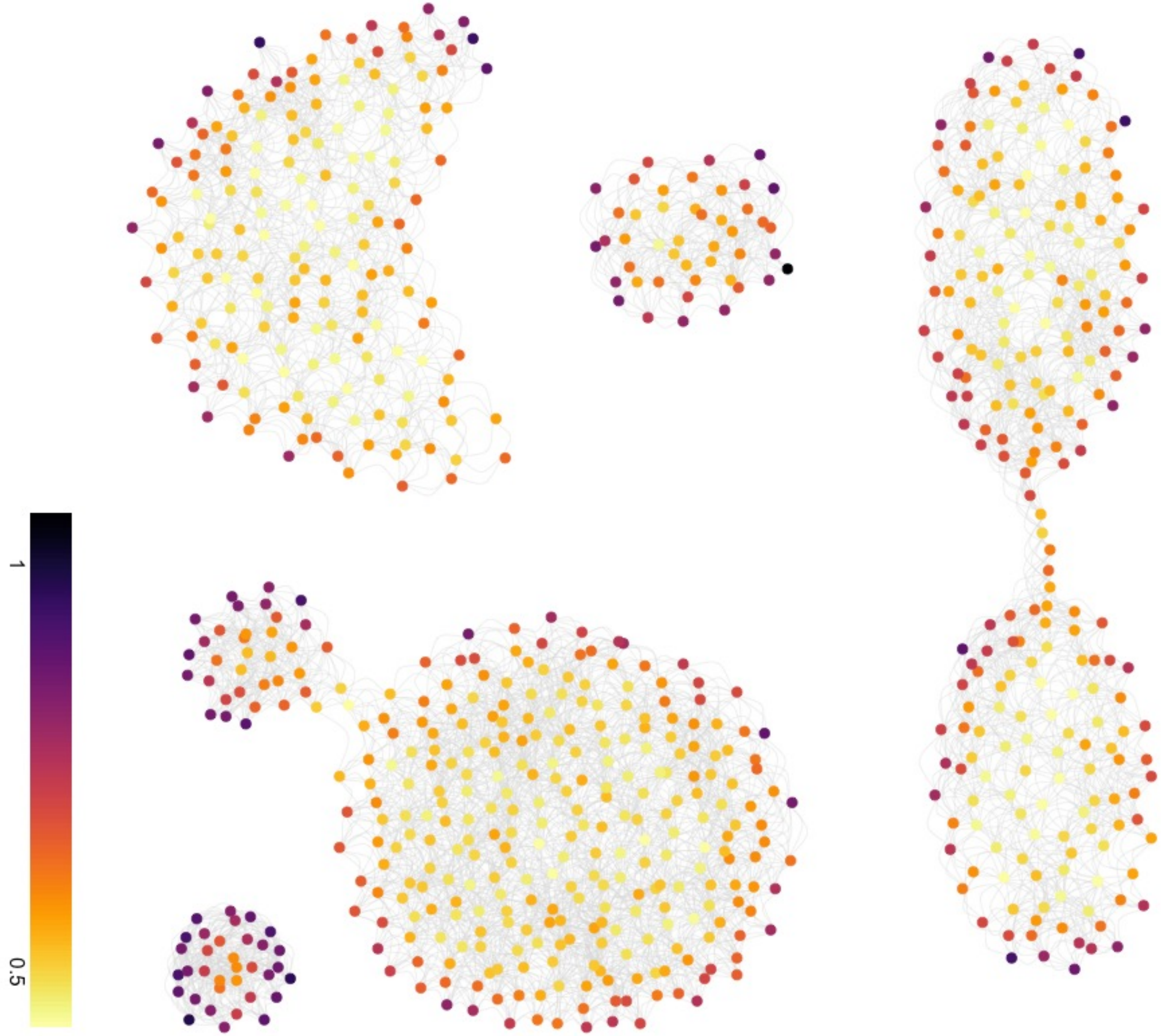
$C_1 = 1 = 6 / (4x3/2)$

Warning

But clustering coefficient is generally hard to see and visual interpretation is considered unreliable

UNIVERSITÀ DEGLI STUDI DI PADOVA

Visual example



Wrap-up

on centrality measures

Takeaways

for Closeness, Betwenness and Clustering coefficient

- Closeness, betweenness and clustering coefficient are alternative centrality measures that have a different view wrt PageRank
- They provide useful insights especially in social networks, as they are linked to sociology concepts
- Closeness and betweenness are based on distances, that require algorithms that are less scalable than PageRank
- Exploit their potential at your best

Takeaways
on centrality measures

Centrality measure	Technical property	Meaning
Degree (in/out)	Measures number (and quality) of connections	Cohesion Entrepreneurship
PageRank (authorities/hubs)	Measures number (and quality) of direct and indirect connections	Cohesion Entrepreneurship Closeness/Similarity/Friendship (with a direction) Dependence
Closeness	Measures length of min paths	Visual centrality Significant spreading points Outliers
Betweenness	Measures number of min paths	Brokerage Structural holes Ostracism
Clustering coeff.	Measures number of triadic closures	Centrality in a community Cohesion of the neighbourhood

UNIVERSITÀ DEGLI STUDI DI PADOVA

More on the meaning

https://reticular.hypotheses.org/1745