Social Network Analysis

#5 PageRank centrality

PageRank centrality

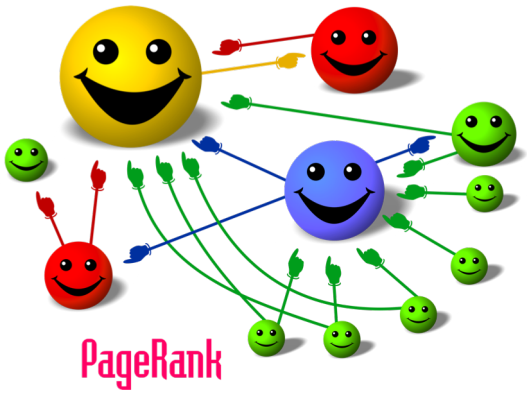
What is PageRank?

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 organize the web?

Idea : 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

Two approaches

Conceptually similar

PageRank

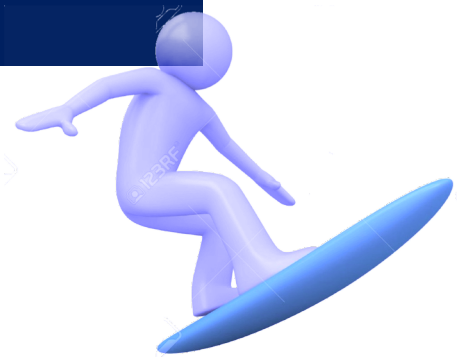
Page, Brin, Motwani, Winograd
1999
«The PageRank citation ranking: bringing order to the web»
Stanford InfoLab

HITS – hubs and authorities

Kleinberg, J.M.
1999
«Authoritative sources in a hyperlinked environment»
Journal of the ACM

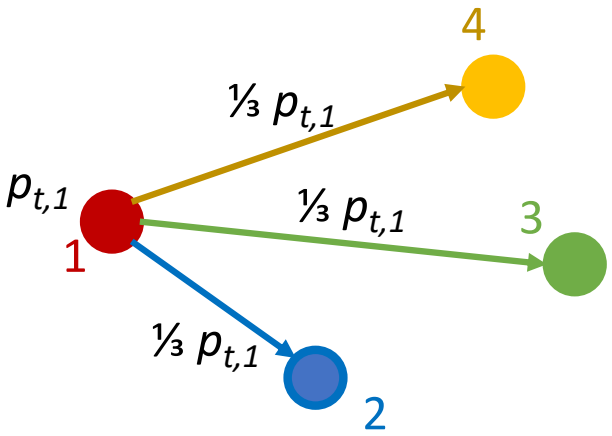
- it is authority directed
- can inspect hubs by transposing *A*

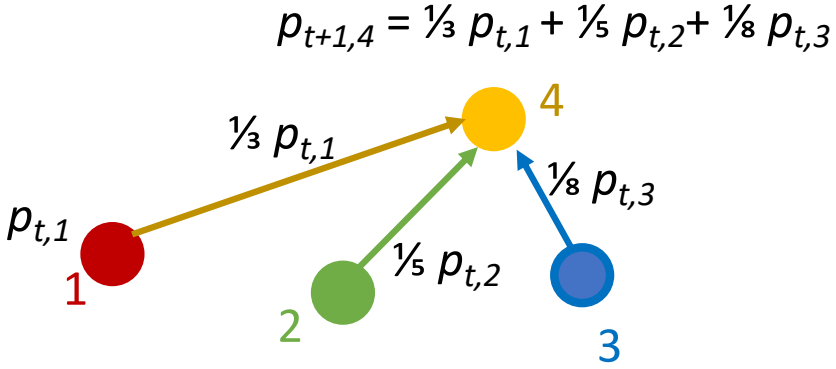
Rationale (Markov chain)



At time *t* a web surfer

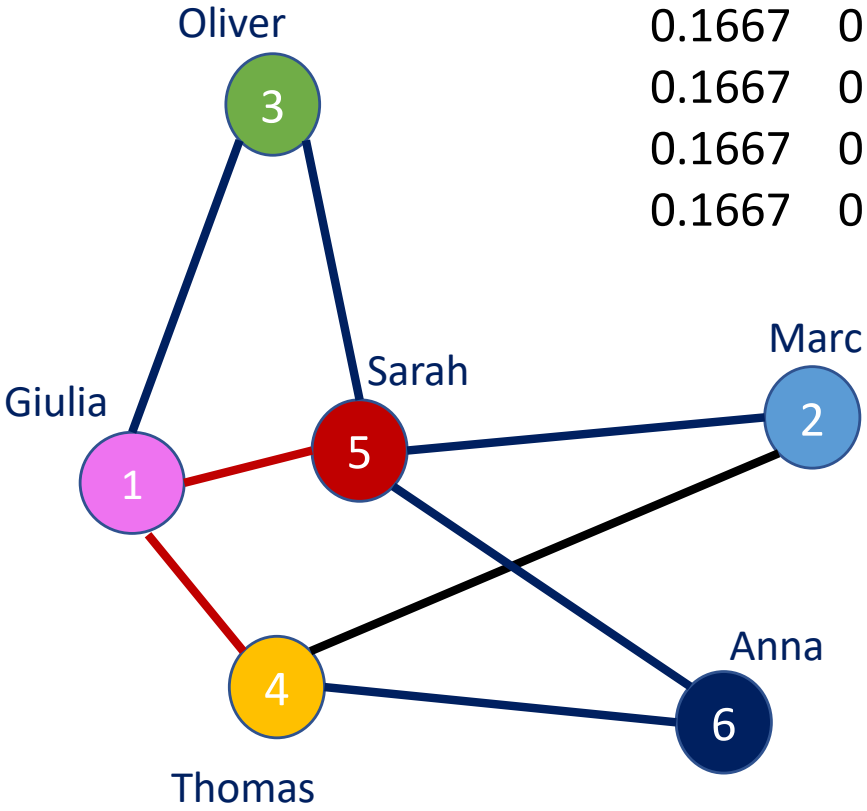
- is at site *i* with probability $p_{t,i}$
- chooses with equal probability one of the sites linked by site *i*





- after a while probabilities settle to a steady state = the PageRank vector (authority score)

Example



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

Known problems

With high probability the surfer ends in:

- Dead ends: some nodes do not have a way out = zero valued columns of ***A***
- Spider traps: some set of nodes do not have a way out, and further induce a periodic behaviour

Teleportation

Idea:

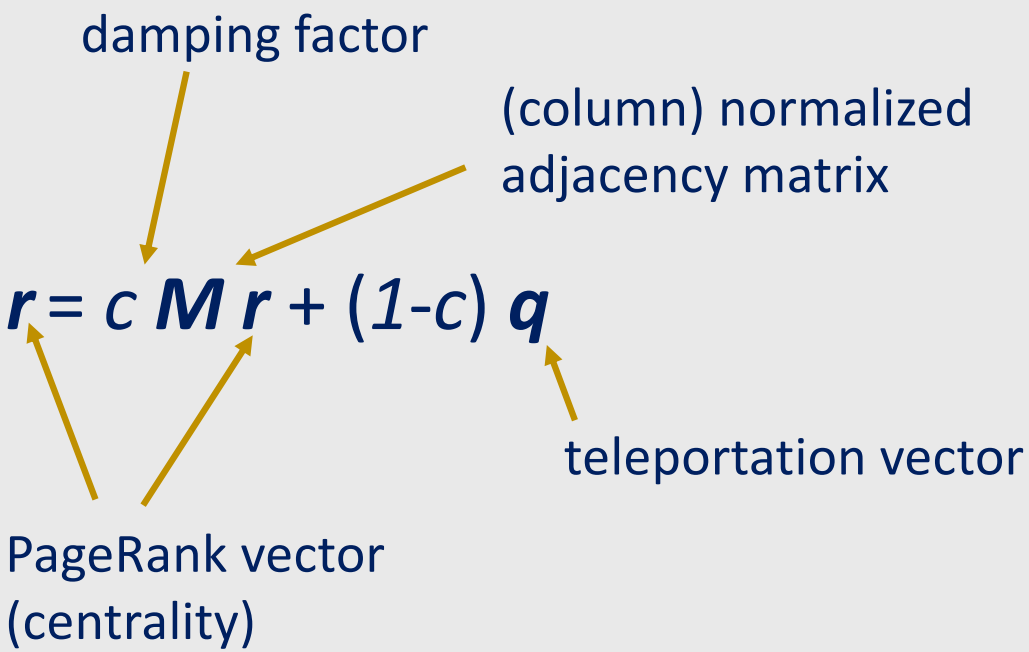
- the surfer does not necessarily move to one of the links of the page she/he is viewing
- with a certain probability, might jump to a random page

damping factor, typically $c = 0.85$, meaning that 85% of the times the surfer moves to one of the links of the page

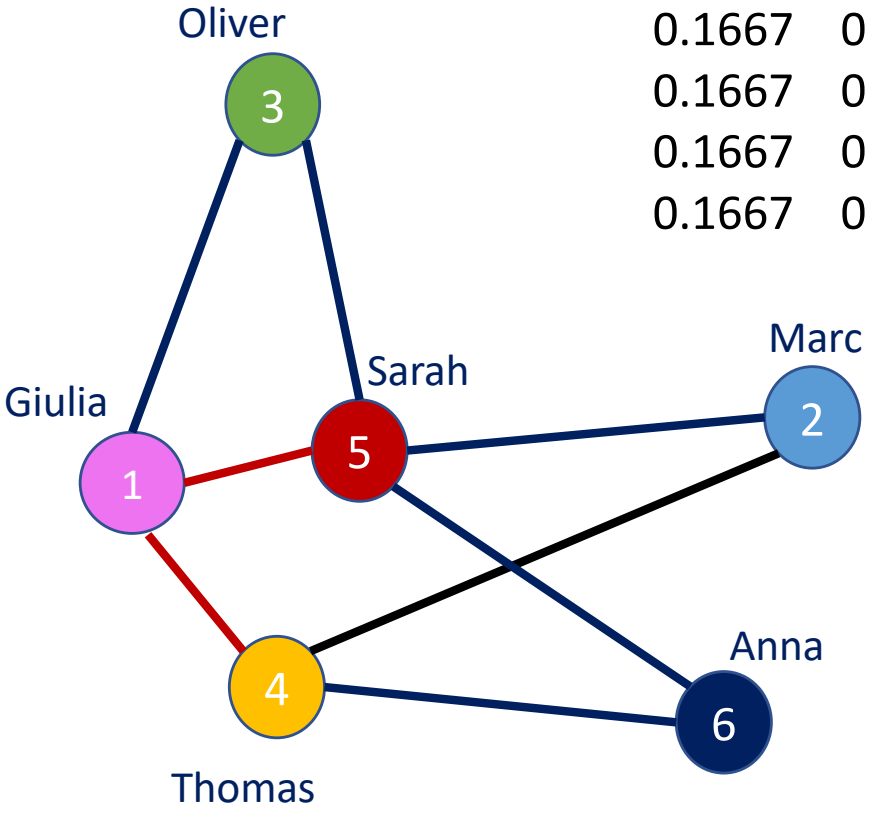
the remaining $1 - c$ = 15% of the times the surfer moves to a random page according to a probability vector $\boldsymbol{q}$, e.g., $\boldsymbol{q}=\boldsymbol{1}/N$ for uniform probability

PageRank with restart

PageRank equation



Example (cont'd)



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

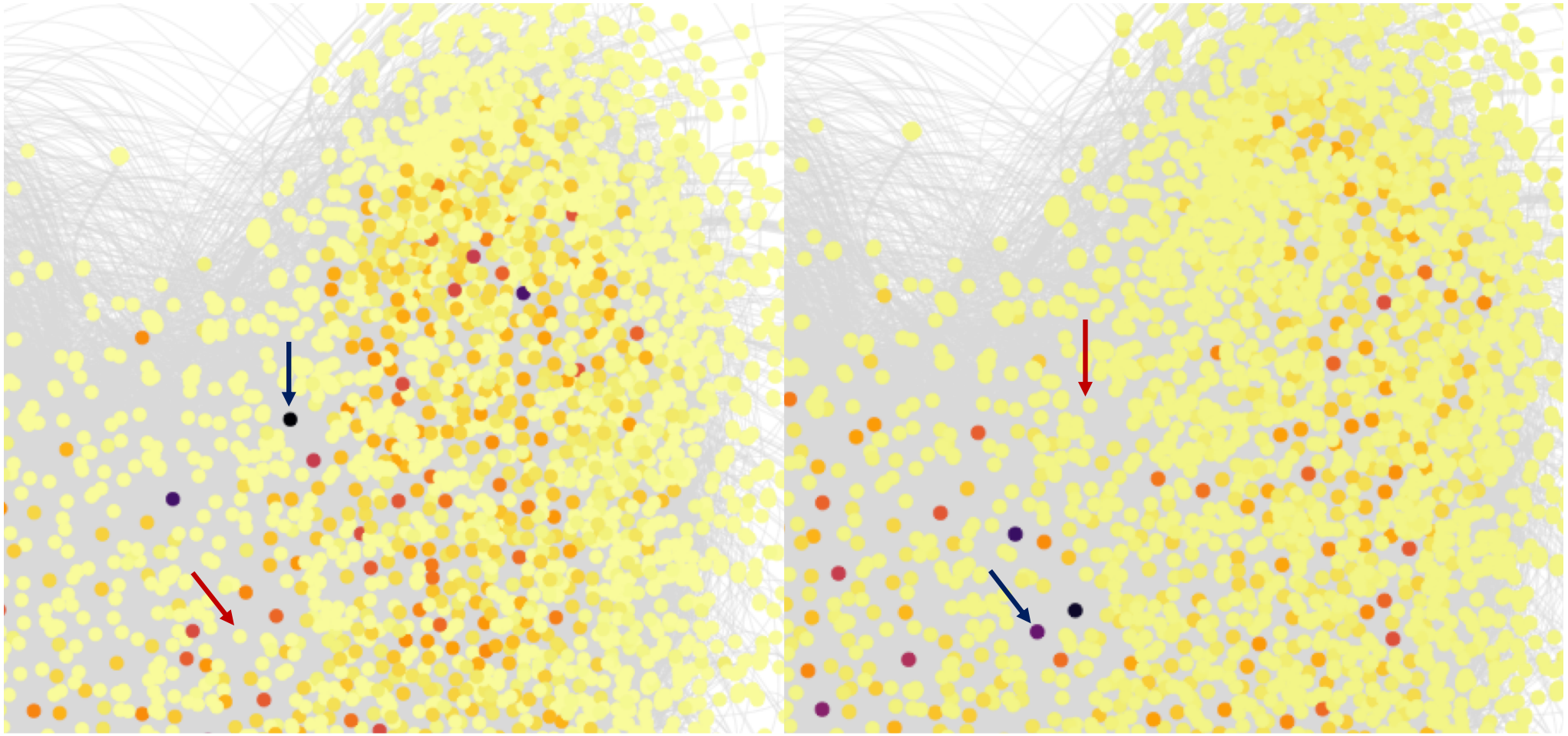
Degree vs PageRank

Wikipedia administrator elections and vote history data
@ Stanford Network Analysis Project
https://snap.stanford.edu/data/wiki-Vote.html

PageRank centrality

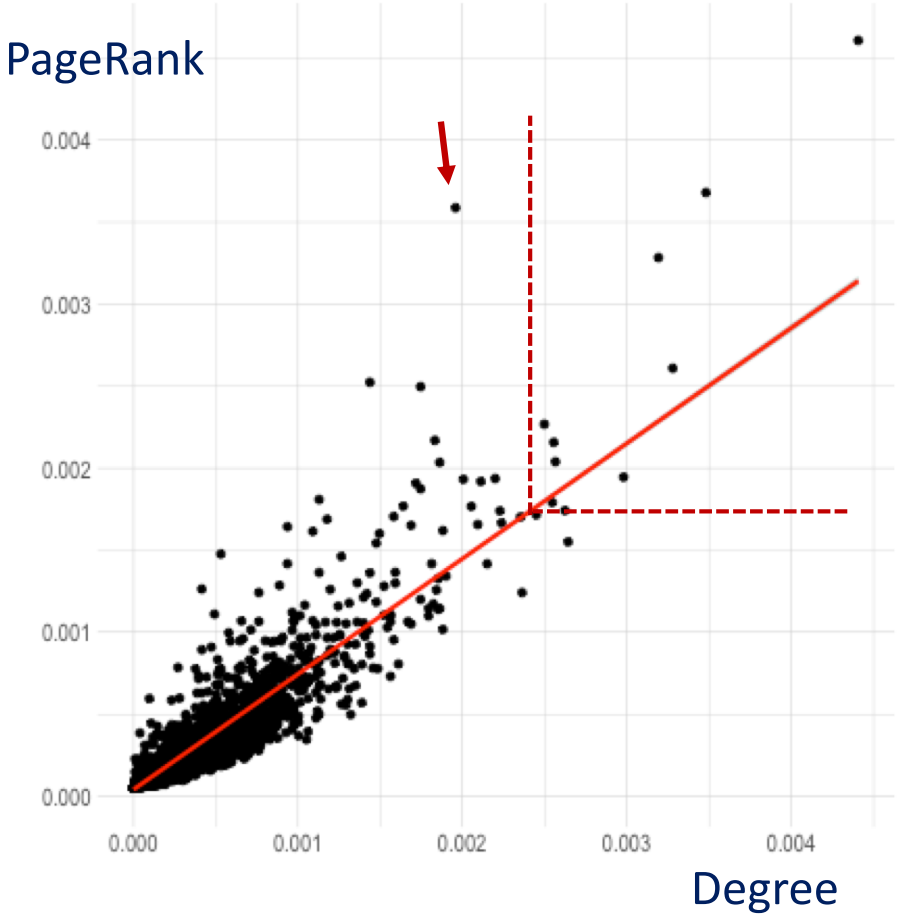
Authorities

Hubs

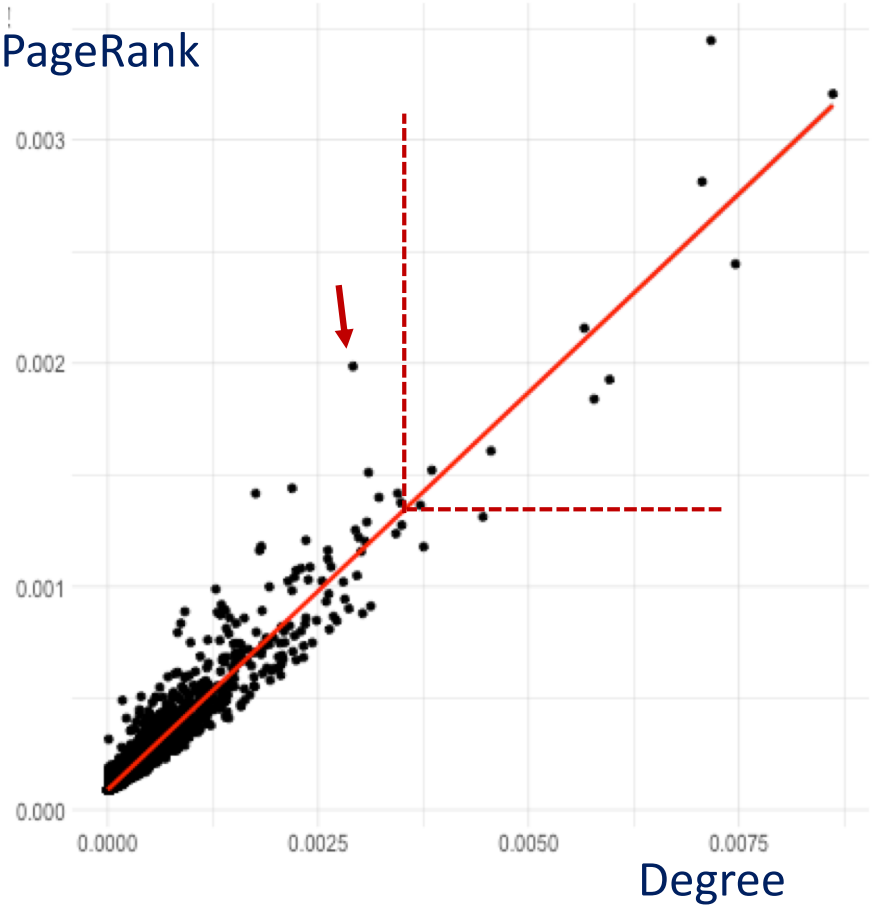


Degree vs PageRank

Authorities



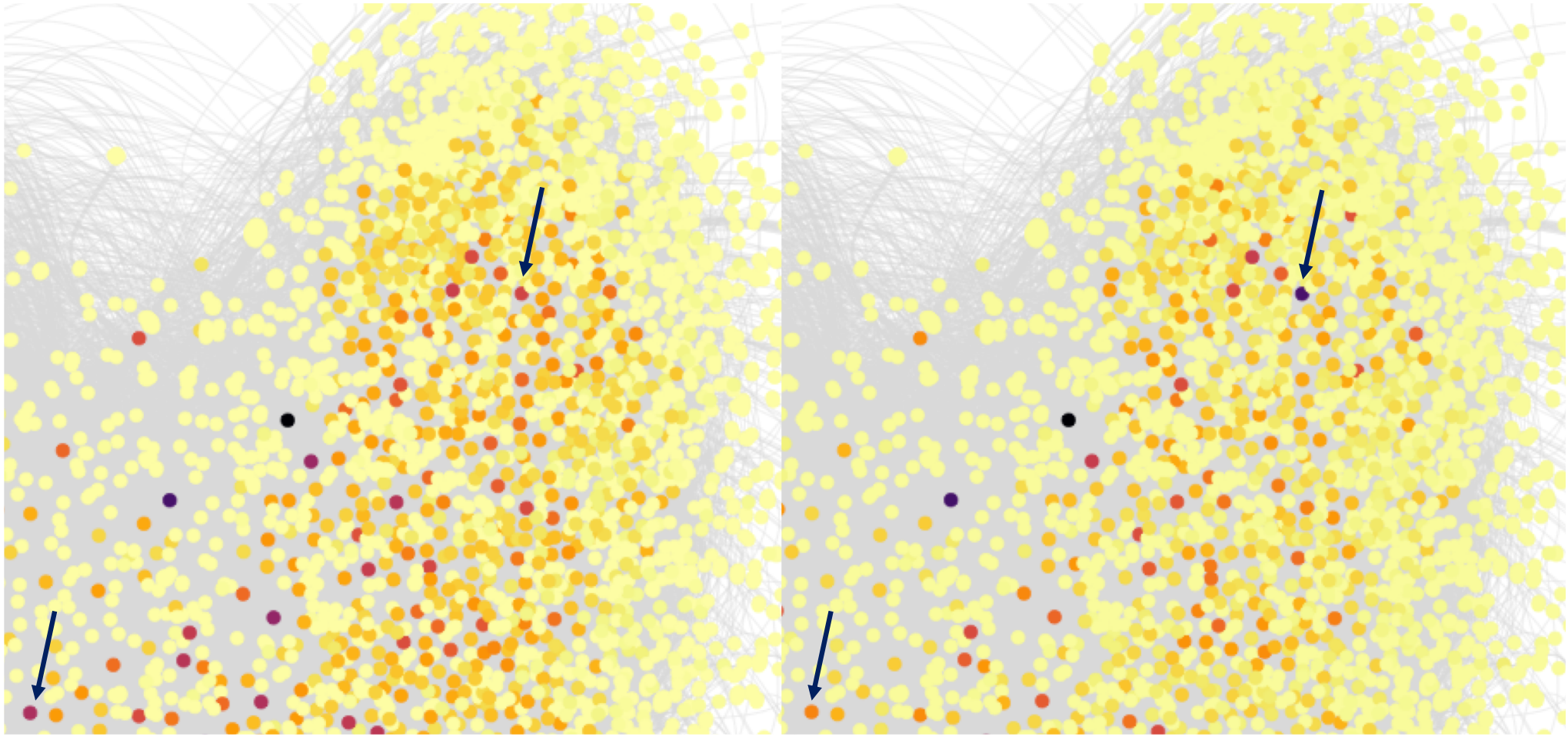
Hubs



Authorities

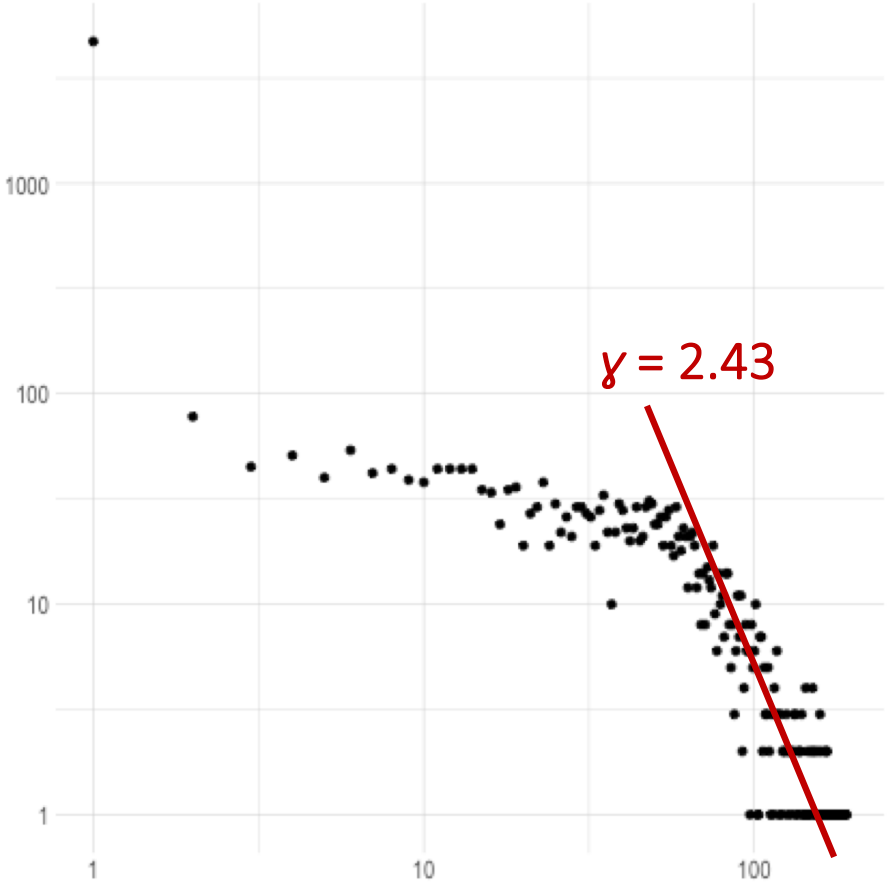
Degree

PageRank

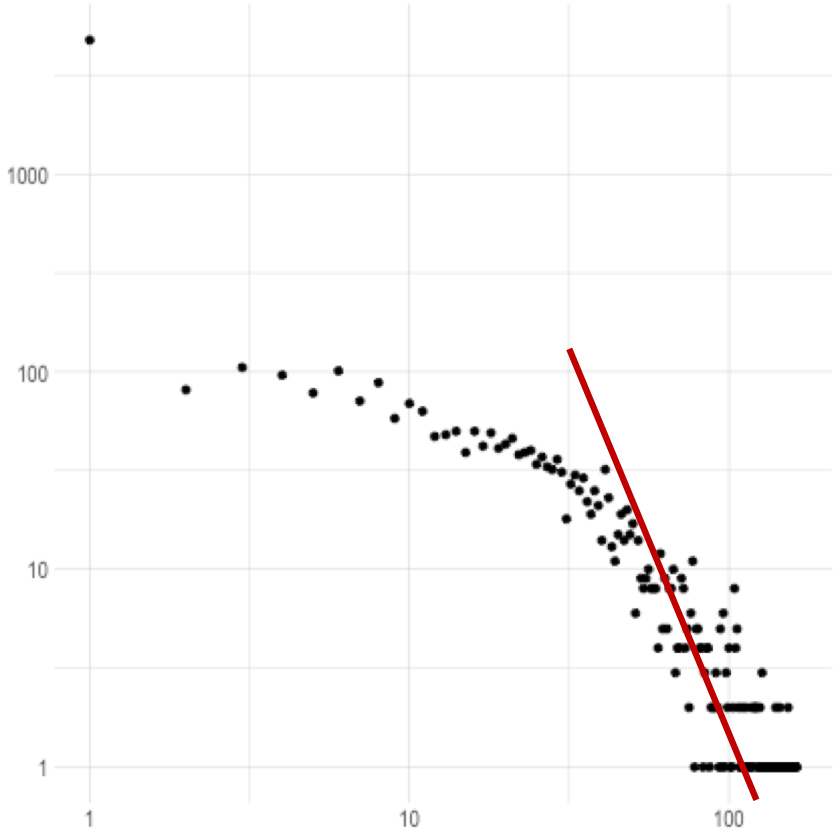


Authorities

Degree



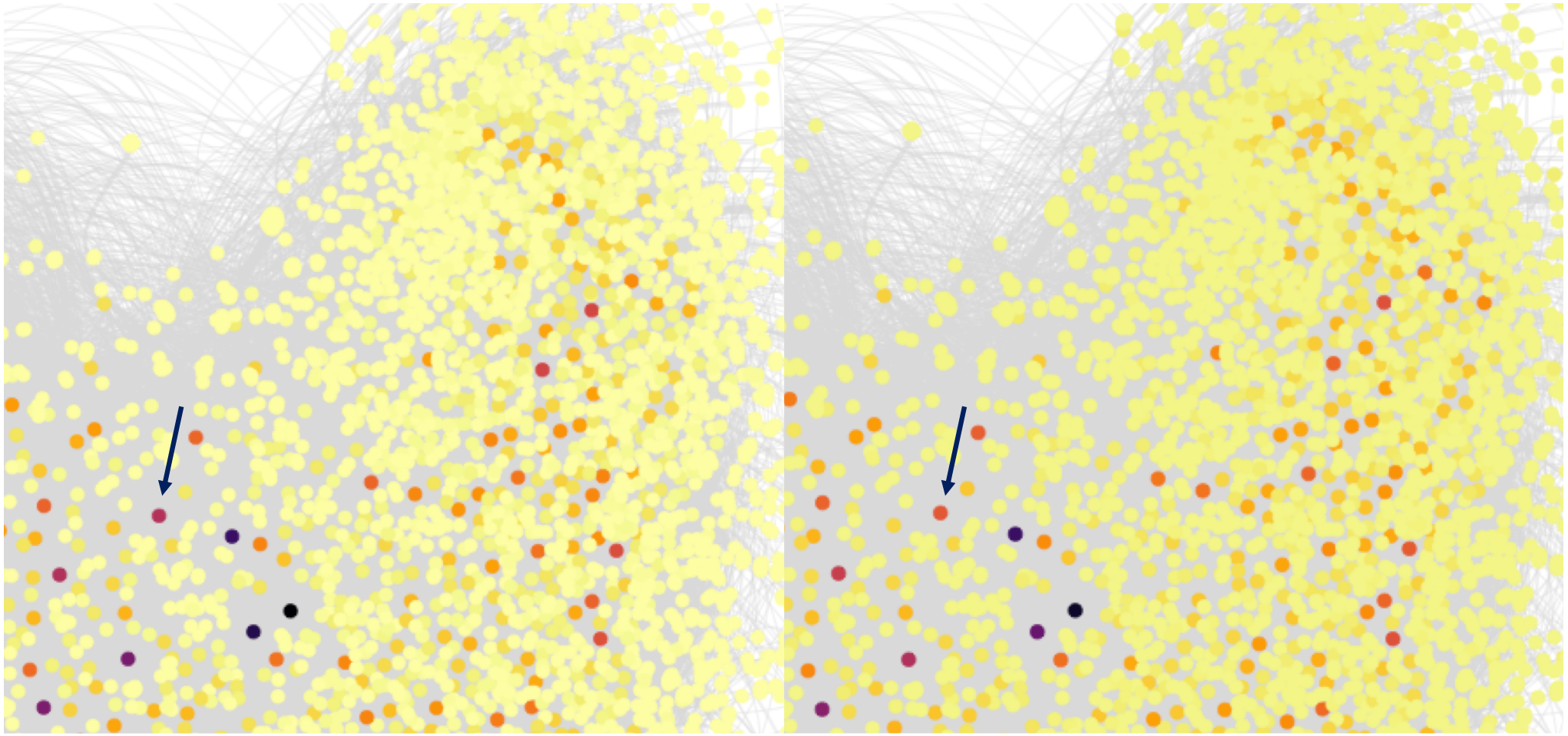
PageRank



Hubs

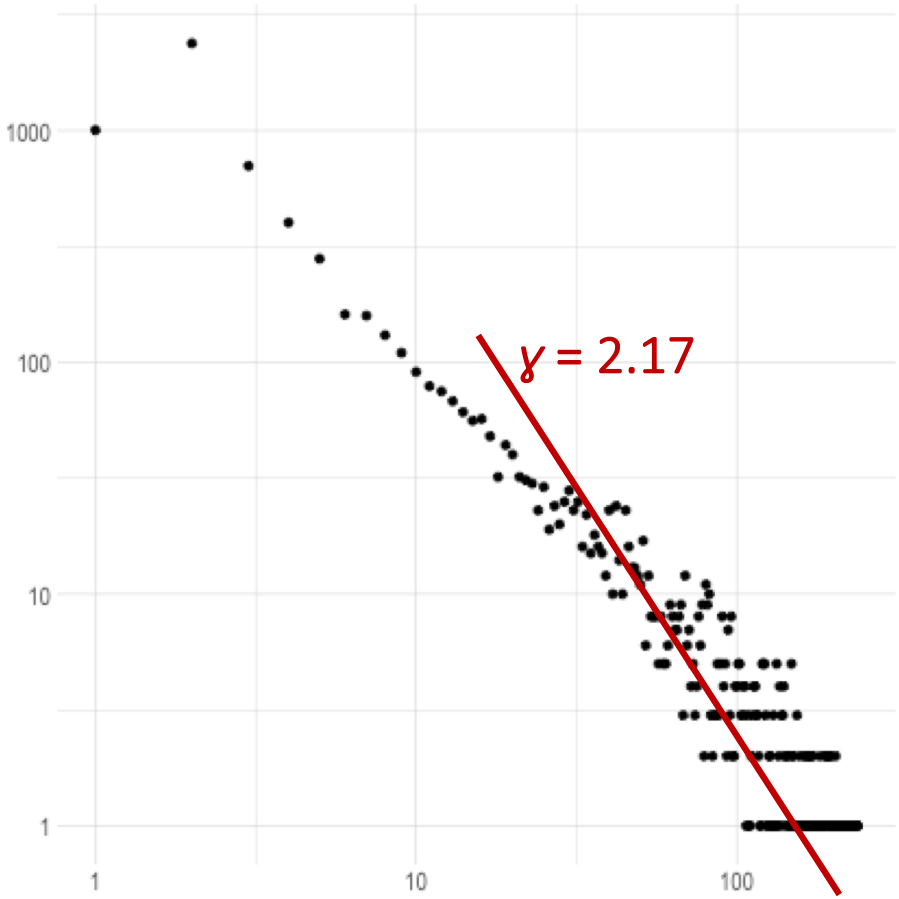
Degree

PageRank

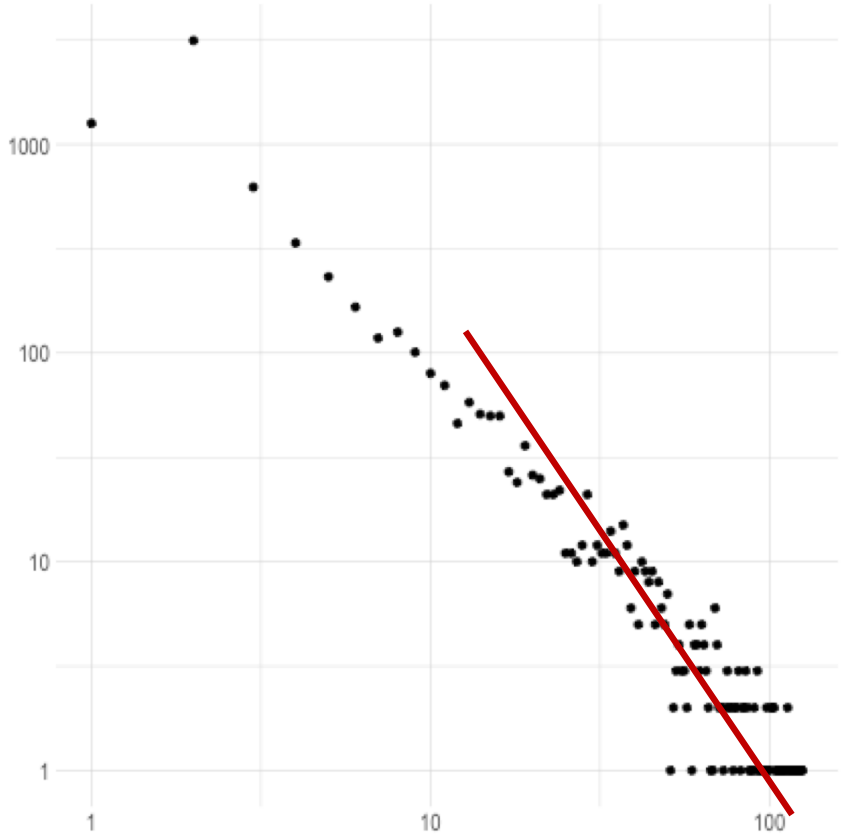


Hubs

Degree



PageRank



Local PageRank

How can we use PageRank?

Want to know about a specific topic? TopicSpecific PageRank

Want to measure proximity/similarity to a node? Local PageRank

… appropriately select your teleport vector $\boldsymbol{q}$!

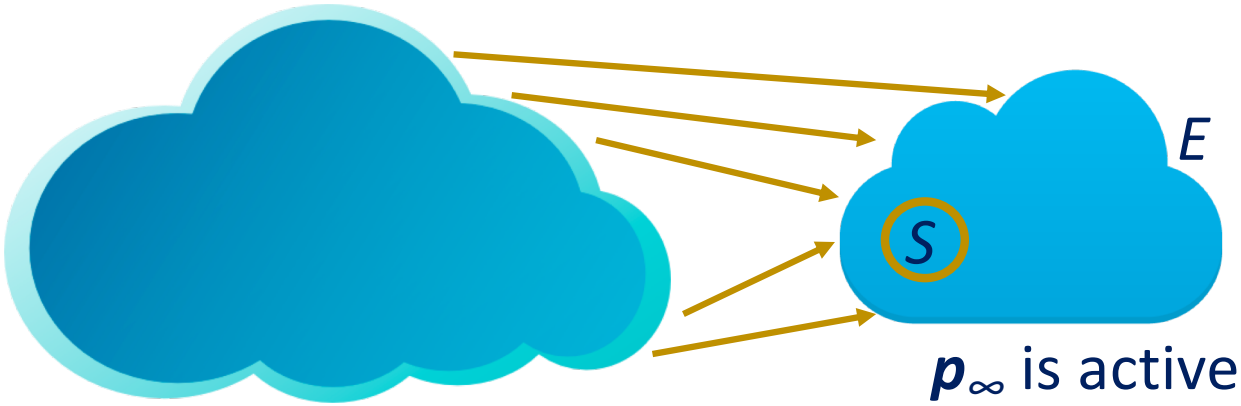
Topic specific PageRank

Idea

- Bias the random walk towards a topic specific teleport set *S* of nodes, i.e., make sure that $\boldsymbol{q}$ is active in *S* only
- *S* should contain only pages that are relevant to the topic

Result

- The random walk deterministically ends in a small set *E*, containing *S*, and being in some sense close to it



$\boldsymbol{p}_\infty$ is active in E only

Measuring closeness: Local PageRank

Idea

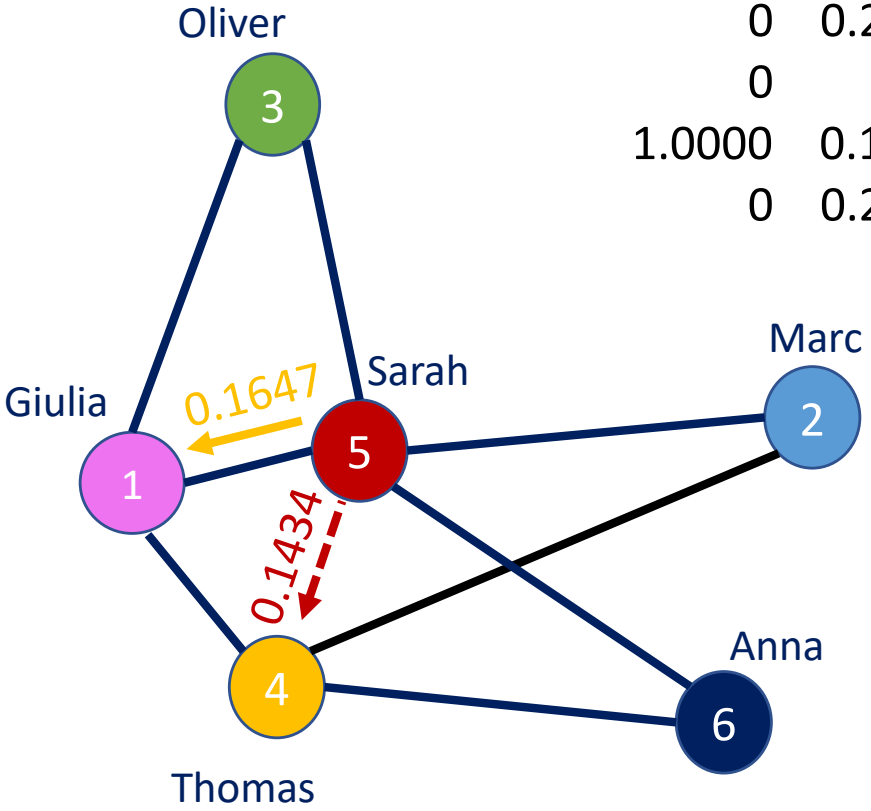
- Measure similarity to node *i* by applying TopicSpecific PageRank with a teleport set with a unique element $S = \{i\}$ and $q = [0 \dots 0\ 1\ 0 \dots 0]$

Result

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

MiME.

Example: who's Sarah's best friend?

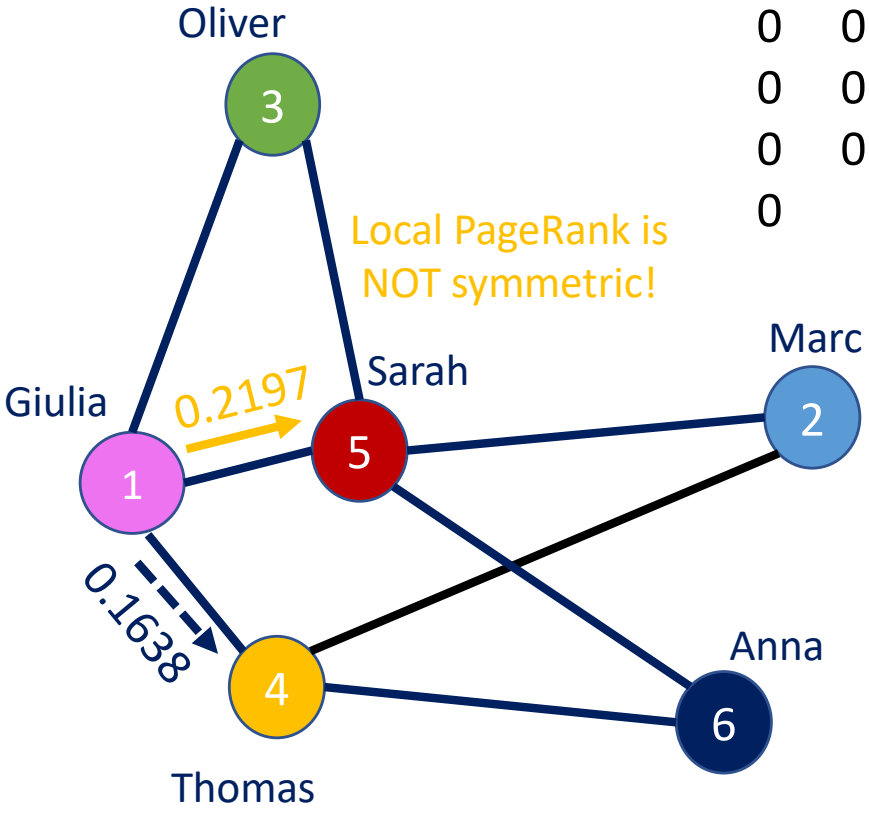


t=1	*2*	*3*	*4*	*5*
0	0.2125	0.1222	0.2096	0.1290
0	0.2125	0.0319	0.1705	0.0708
0	0.2125	0.0921	0.1369	0.1127
0	0	0.2408	0.0617	0.2043
1.0000	0.1500	0.4811	0.2508	0.4125
0	0.2125	0.0319	0.1705	0.0708

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

MiME.

Example: who's Giulia's best friend?



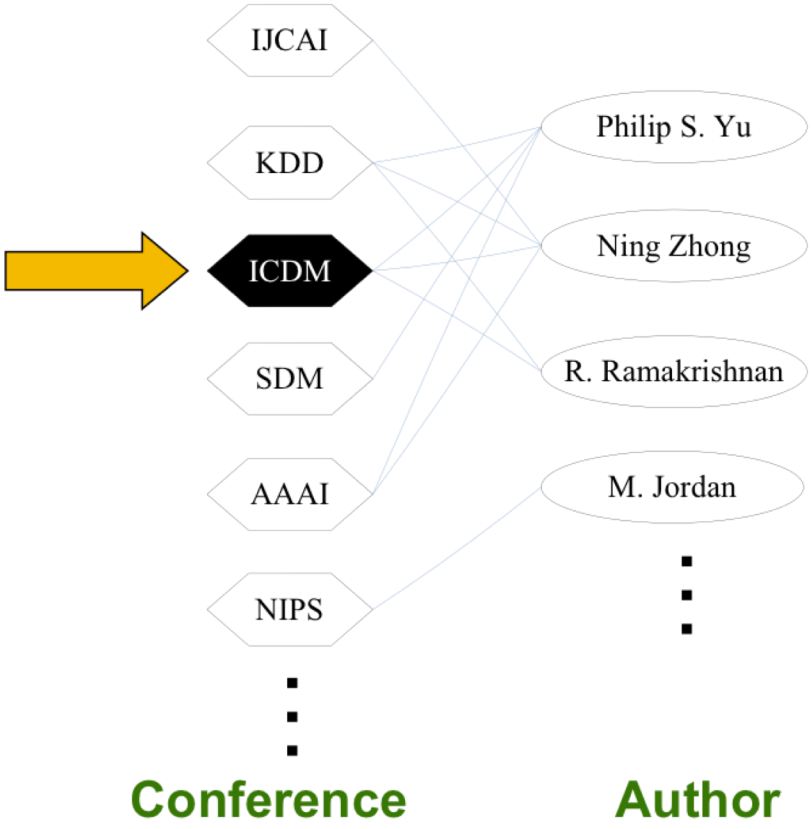
t=1	*2*	*3*	*4*	*5*
1.0000	0.1500	0.4109	0.2403	0.3404
0	0	0.1405	0.0467	0.1262
0	0.2833	0.1027	0.1510	0.1275
0	0.2833	0.0425	0.2358	0.1078
0	0.2833	0.1629	0.2795	0.1719
0	0	0.1405	0.0467	0.1262

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

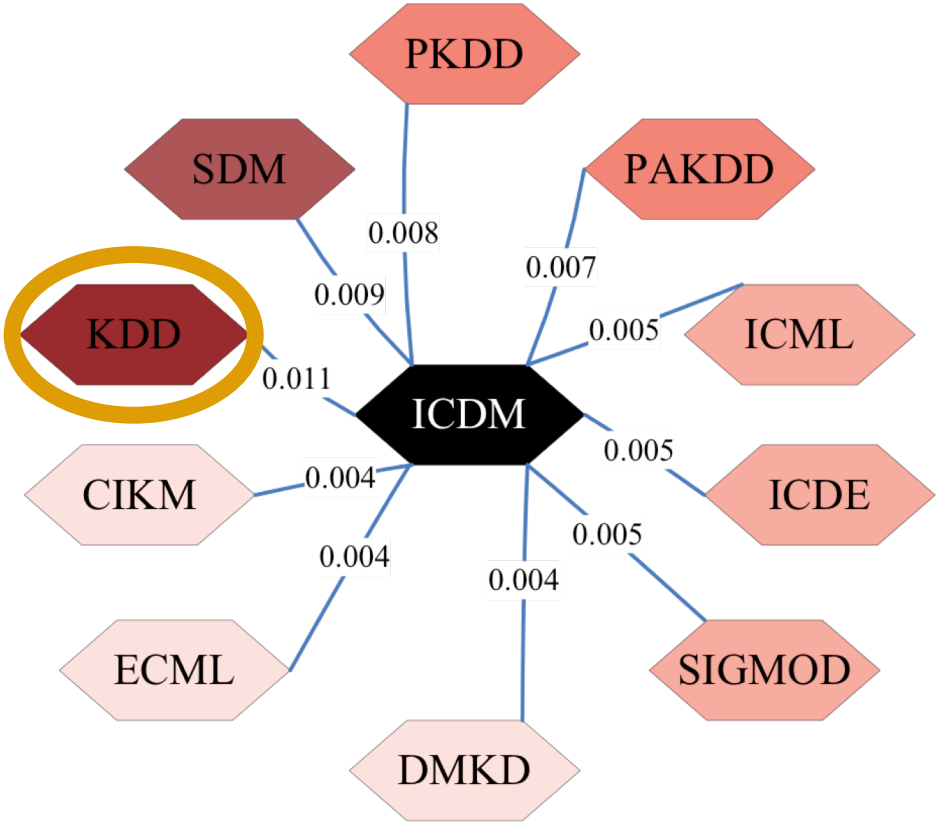
MiME.

Example

What is the most related conference to ICDM?



Top 10 ranking results

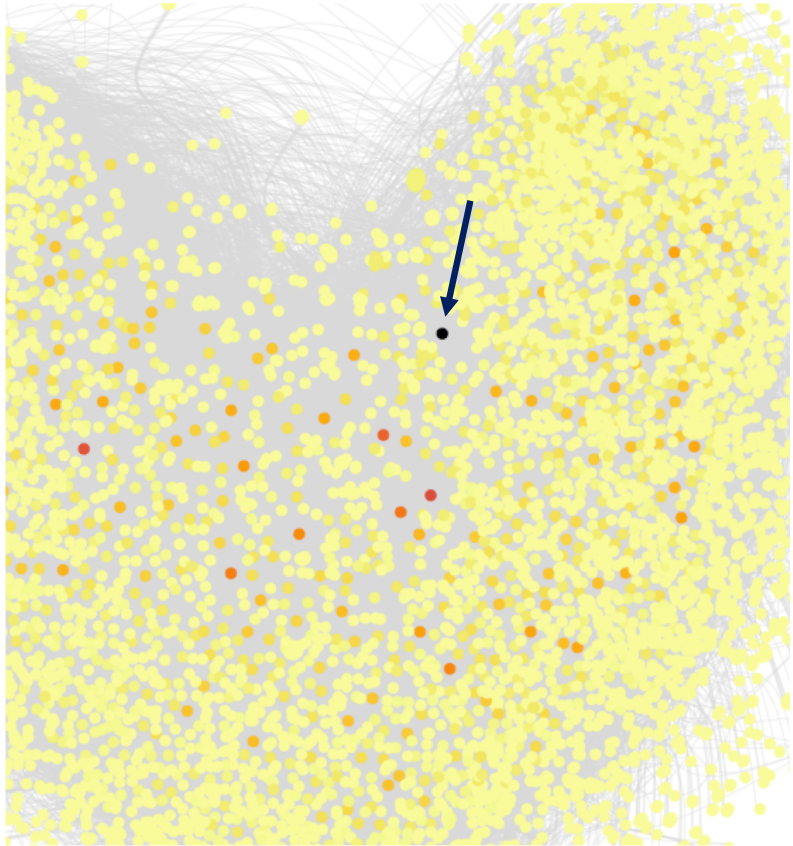


ICDM = international conf. on data mining
KDD = knowledge discovery and data mining

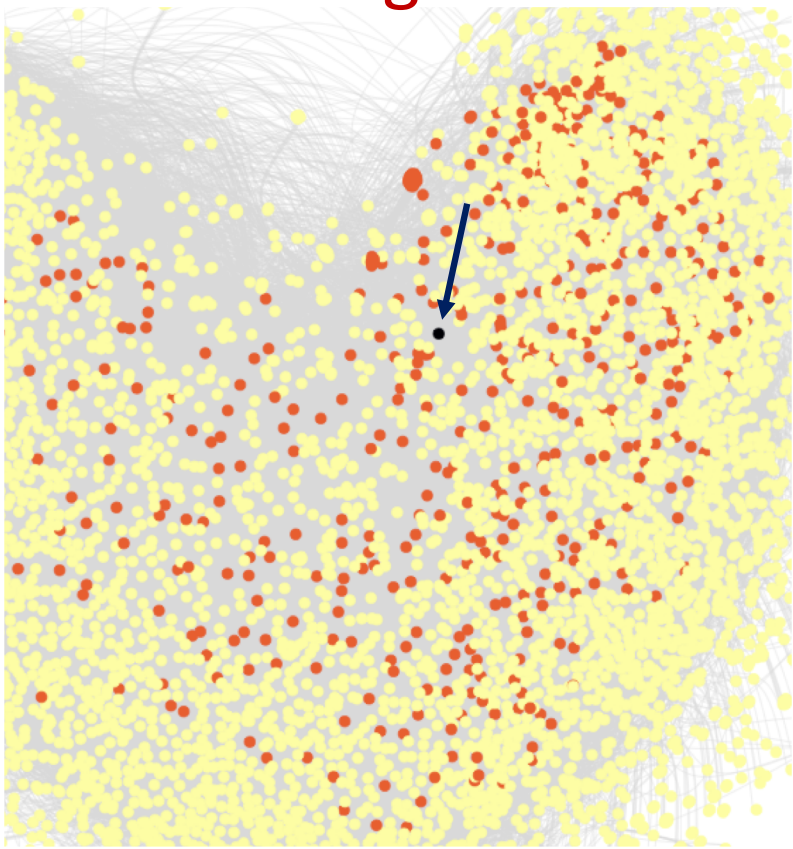
MiME

Local PageRank (authorities , *A*)

Local PageRank



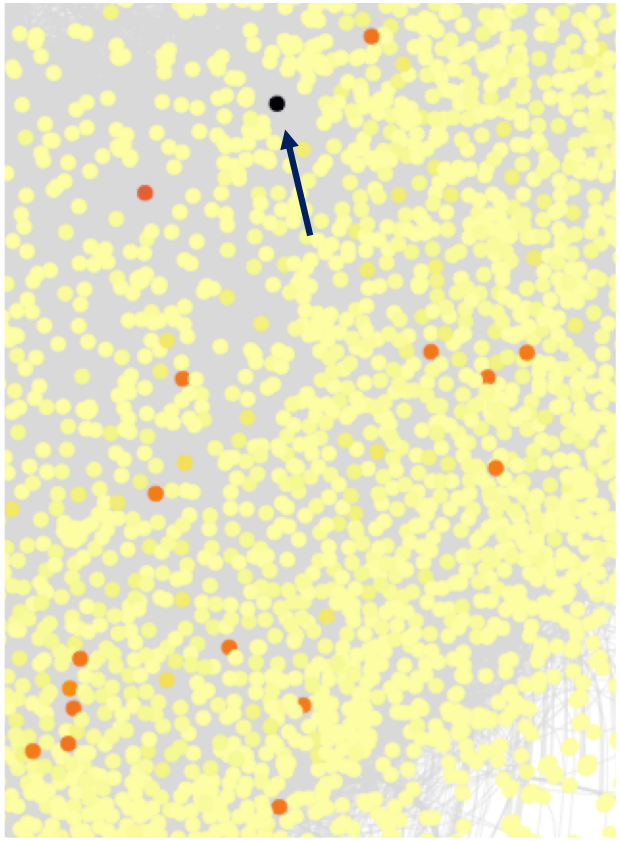
1-hop
out-neighbours



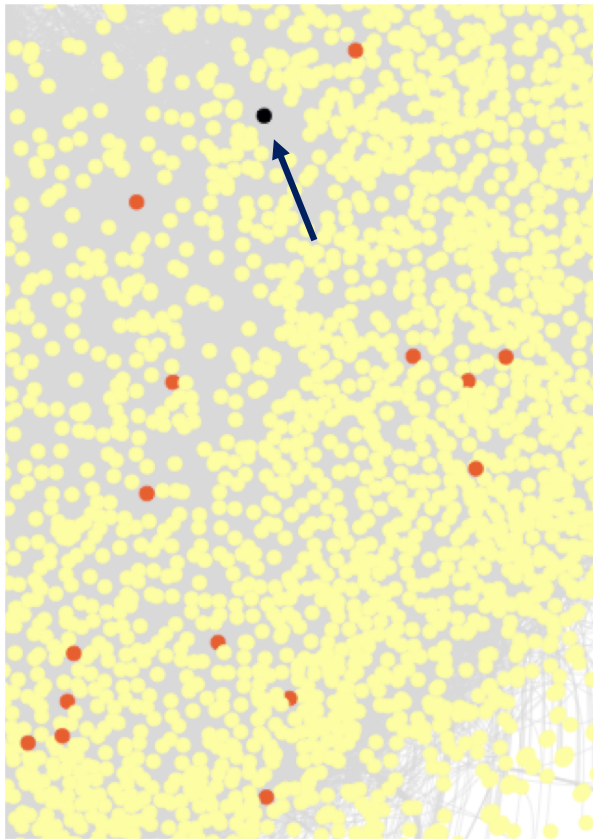
neighbours authority score =
local node → neighbours

Local PageRank (hubs, A^T)

Local PageRank



1-hop
in-neighbours



neighbours hub score =
neighbours → local node

Closeness centrality

What is Closeness?

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:

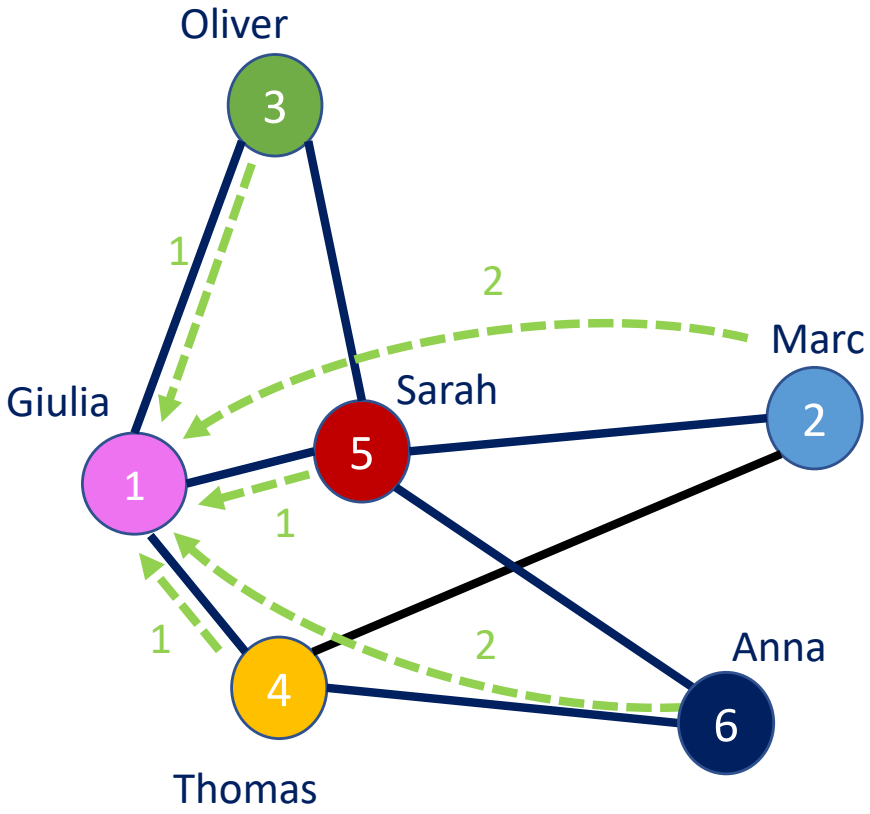
$$C(x) = \frac{1}{\sum_y d(y,x)}.$$

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

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

Example

count the lengths of the shortest paths leading to Giulia
1 + 2 + 1 + 2 + 1 = 7



Closeness

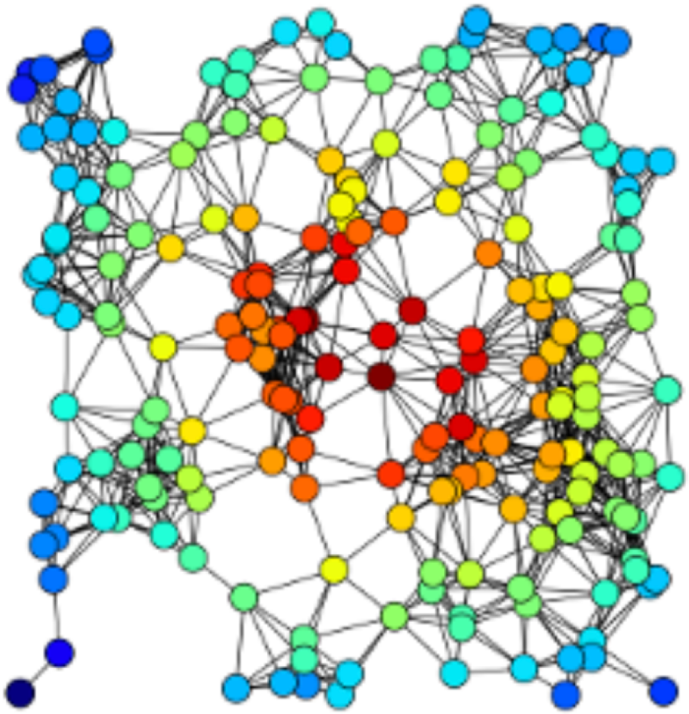
0.1429	Giulia
0.1250	Marc
0.1250	Oliver
0.1429	Thomas
0.1667	Sarah
0.1250	Anna

Sarah is the preferred node for spreading information

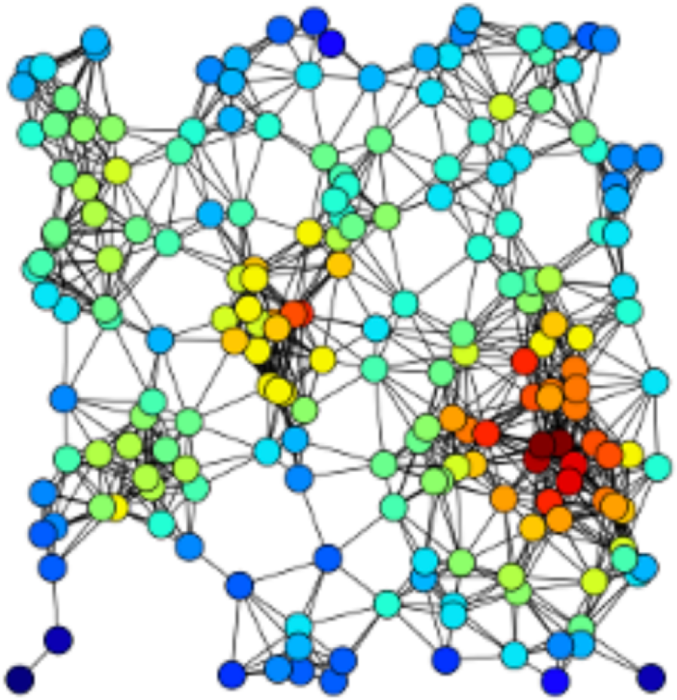
C(Giulia) = 1/7
= 0.1429

Closeness versus Degree centrality

Closeness



Degree



Betweenness centrality

What is Betweenness?

Betweenness centrality

From Wikipedia, the free encyclopedia

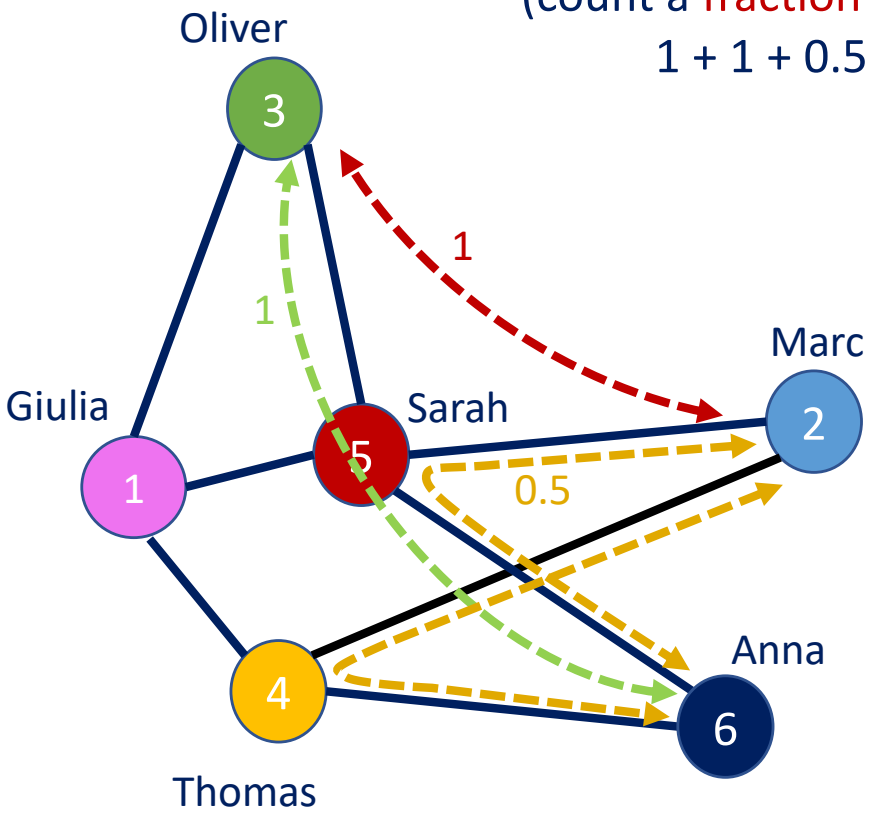
In graph theory, **betweenness centrality** is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs)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

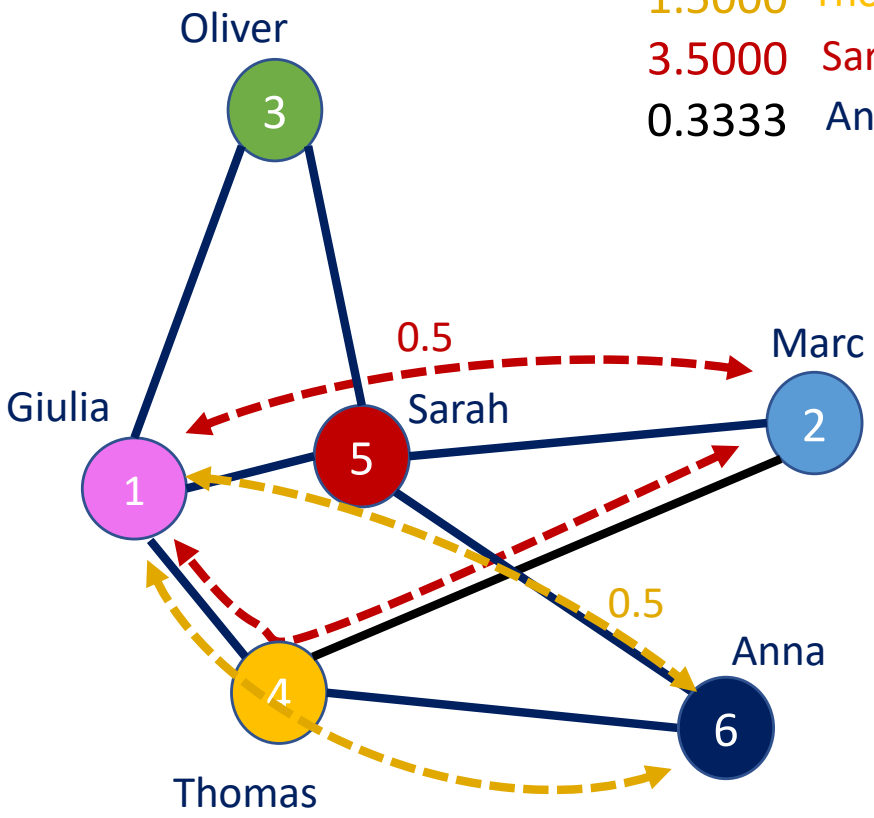
Betweenness centrality was devised as a general measure of centrality:[1] it applies to a wide range of problems in network theory, including problems related to social networks, biology, transport and scientific cooperation. Although earlier authors have intuitively described centrality as based on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

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

Example

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



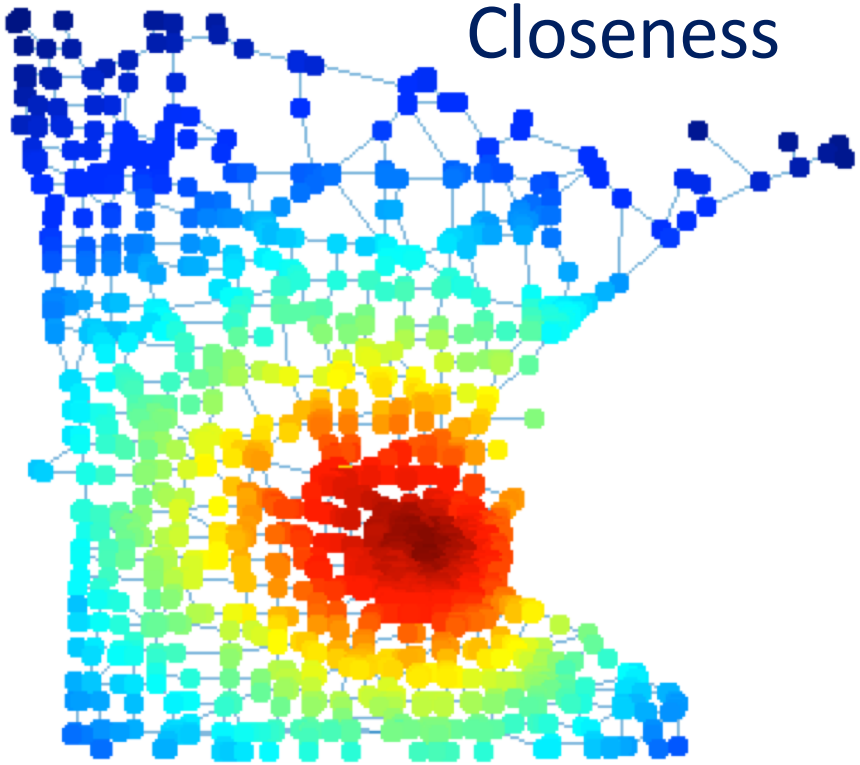


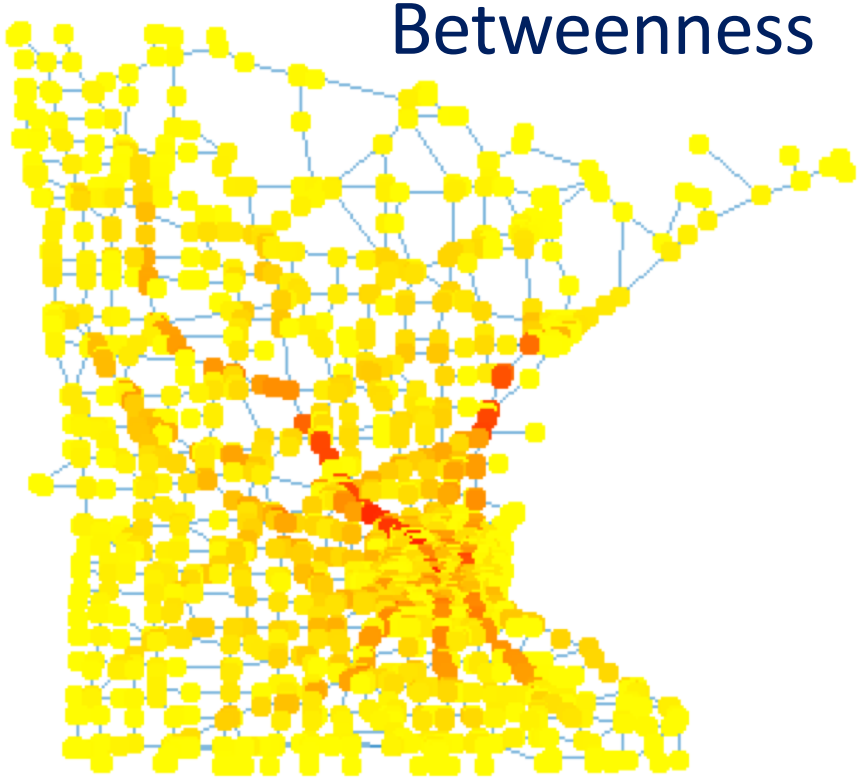
Betweenness

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

Closeness versus Betweenness centrality

Minnesota road network





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

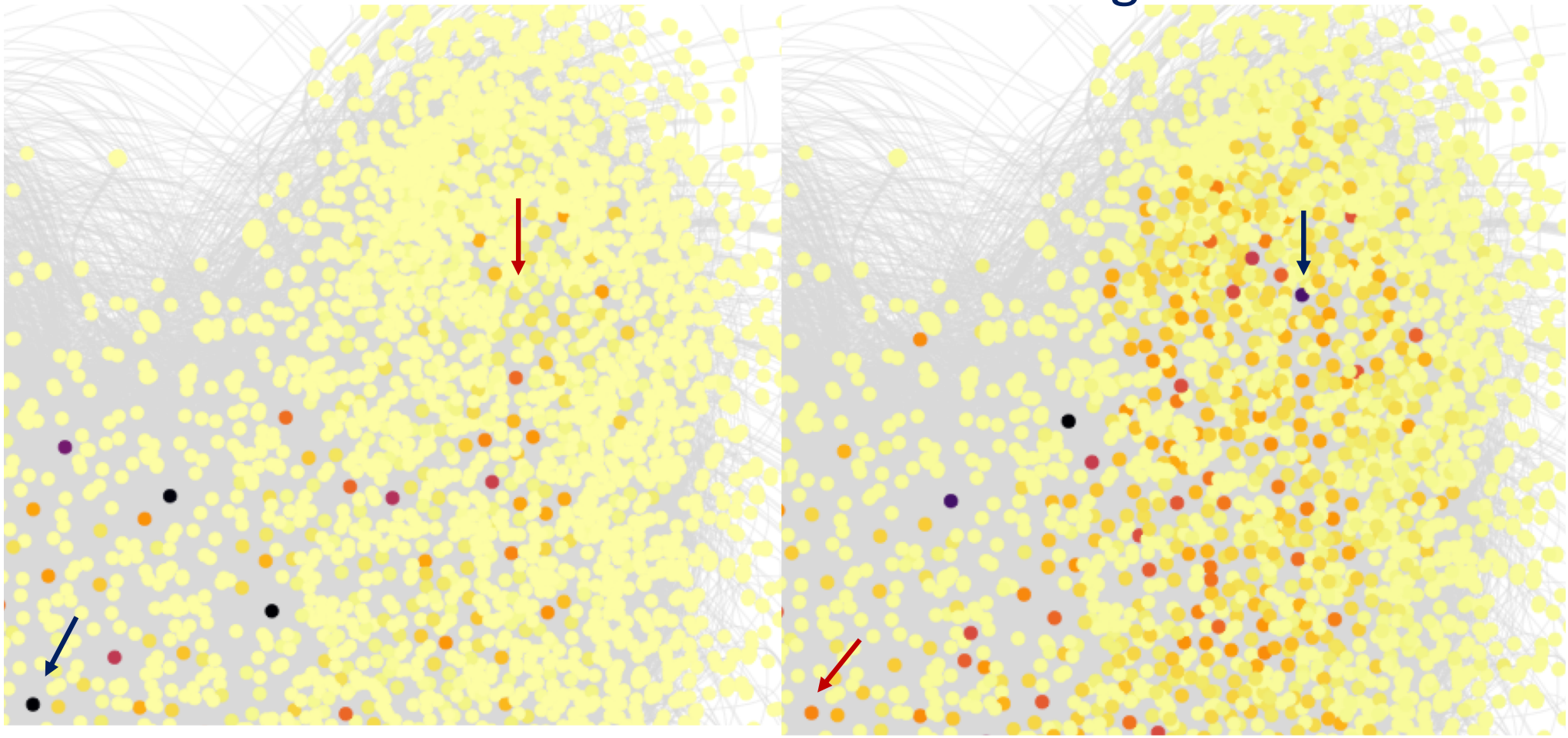
Betweenness is a measure of brokerage (i.e., being a bridge)

MiME.

Betweenness versus PageRank centrality

Betweenness

PageRank



Take-aways

Centrality measure	Technical property	Meaning
Degree (in/out)	Measures number (and quality) of connections	Cohesion Entrepreneurship Extraversion
PageRank (authorithies/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

Take-aways

https://reticular.hypotheses.org/1745