Network Science

#5 PageRank centrality

What is centrality?

Centrality

From Wikipedia, the free encyclopedia

In graph theory and network analysis, indicators of **centrality** identify the most important vertices within a graph. Applications include identifying the most influential person(s) in a social network, key infrastructure nodes in the Internet or urban networks, and super-spreaders of disease. Centrality concepts were first developed in social network analysis, and many of the terms used to measure centrality reflect their sociological origin.[1]

Degree centrality [edit]

Main article: Degree (graph theory)

PageRank centrality [edit]

Main article: PageRank

Betweenness centrality [edit]

Main article: Betweenness centrality

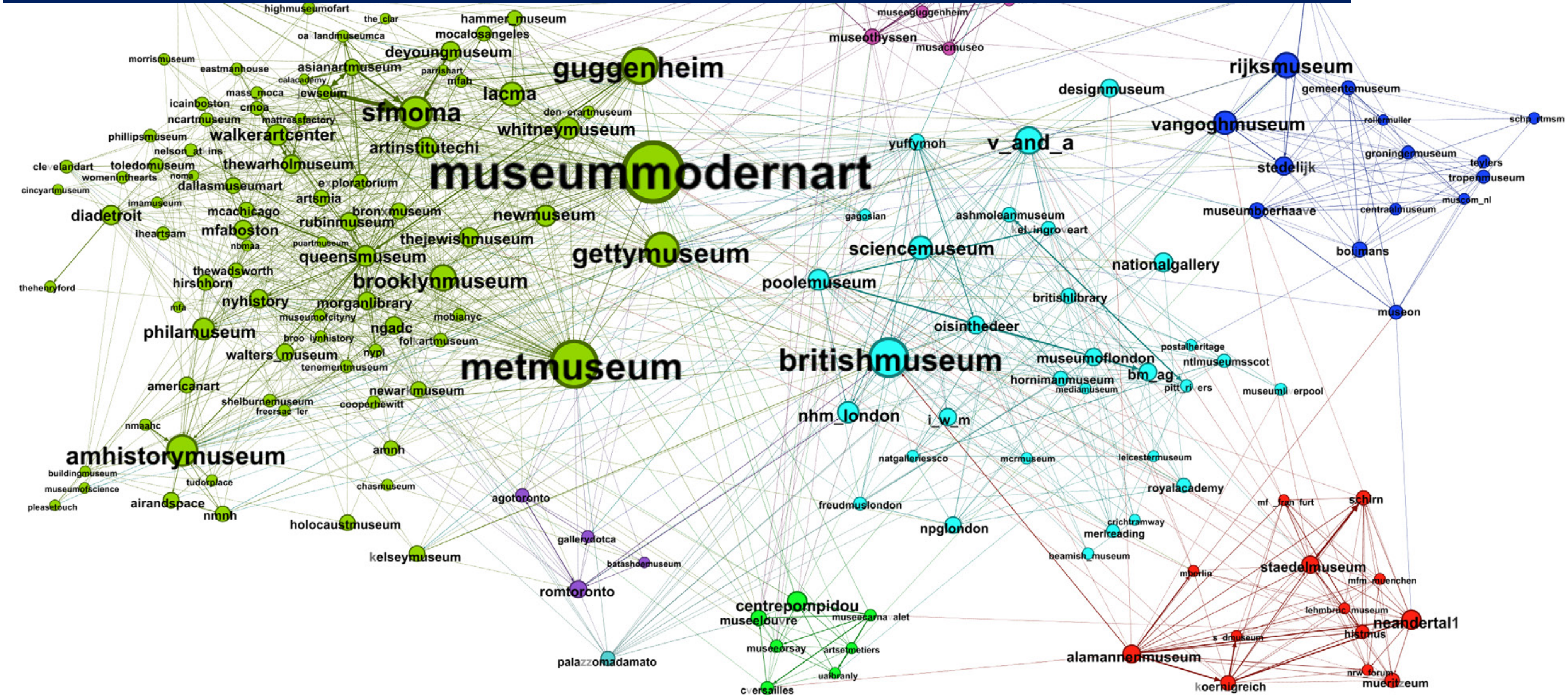
Eigenvector centrality [edit]

Main article: Eigenvector centrality

Closeness centrality [edit]

Main article: Closeness centrality

How to rank nodes in a network?



Can we do this efficiently, i.e., by using an automatic, reliable, and fast method?

The solution comes from the web

How to organize the web?

Idea: links as votes

- the higher (and stronger) the number of incoming links, the more important a node
- the more important a node, the more valuable the output links

MiME.

Two approaches

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

Conceptually similar

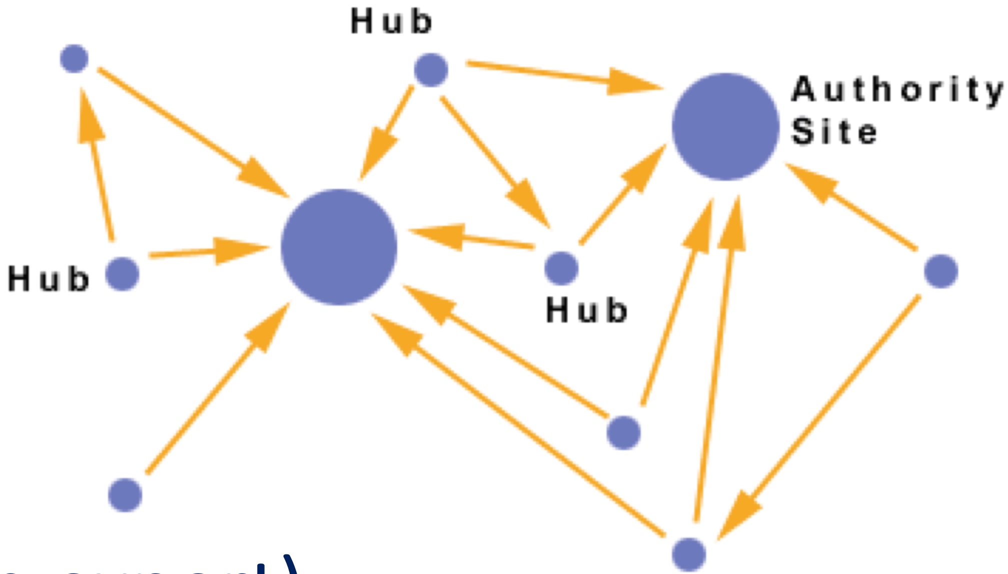
HITS centrality

Hiperlink induced topic search (HITS)

Two classes of nodes:

❑ Authorities (quality as a content provider)

nodes that contain useful information, or having a high number of edges pointing to them (e.g., course homepages)



❑ Hubs (quality as an expert)

trustworthy nodes, or nodes that link to many authorities (e.g., course bulletin)

authority or hub?

MIME

HITS equations – authority score

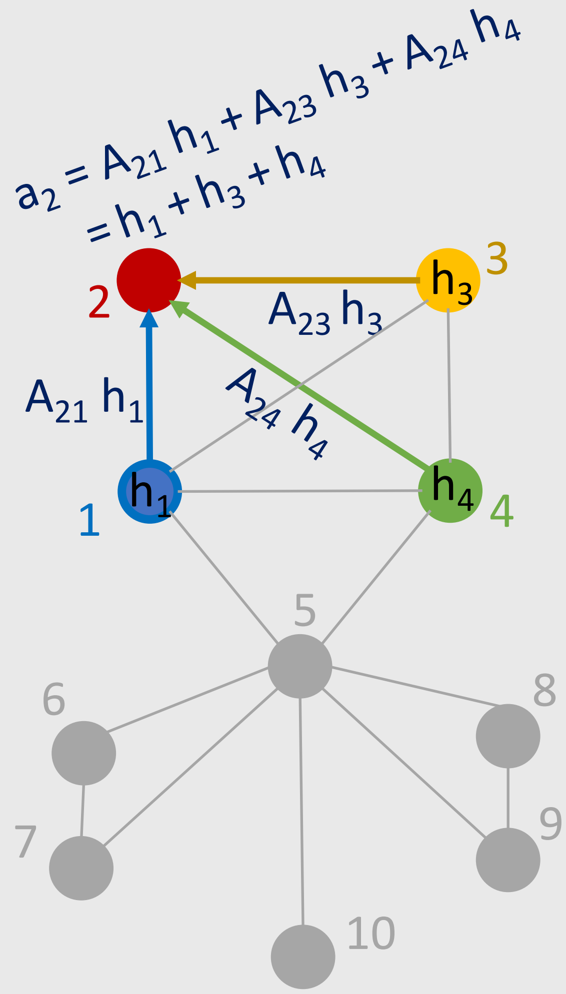
$A_{2,4}$ = weight of connection 4 → 2

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \\ a_9 \\ a_{10} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \\ h_9 \\ h_{10} \end{bmatrix}$$

$$\boldsymbol{a} = \boldsymbol{Ah}$$

authority scores

hub scores



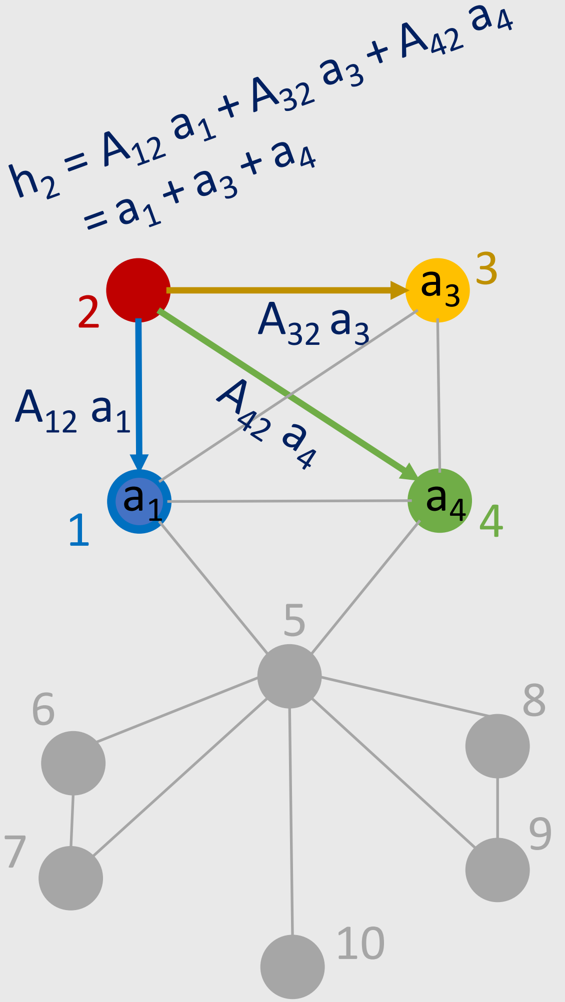
MIME.

HITS equations – hub score

$A_{3,2}$ = weight of connection 2 → 3

$$
\begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \\ h_9 \\ h_{10} \end{bmatrix}
=
\begin{bmatrix}
0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}^T
\cdot
\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \\ a_9 \\ a_{10} \end{bmatrix}
$$

$$\boldsymbol{h} = \boldsymbol{A}^T \boldsymbol{a}$$



MIME

HITS equations

$$\boldsymbol{a} = c_a \cdot \boldsymbol{A}\boldsymbol{h}$$

$$\boldsymbol{h} = c_h \cdot \boldsymbol{A}^T\boldsymbol{a}$$

hubs

$$\boxed{\boldsymbol{h} = c\,\boldsymbol{M}\boldsymbol{h}}$$

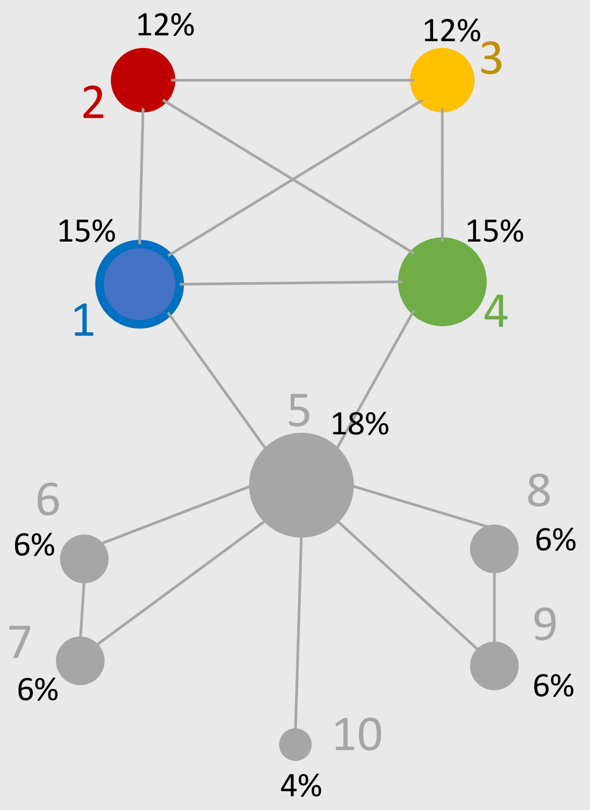
$$M = \boldsymbol{A}^T\boldsymbol{A}$$

$$c = c_a c_h$$

authorities

$$\boxed{\boldsymbol{a} = c_a \cdot \boldsymbol{A}\boldsymbol{h}}$$

- The formula says we are interested in the (principal) eigenvector of matrix $M = A^{\mathrm{T}} \cdot A$
- Can be obtained by standard linear algebra algorithms



Power iteration method

0. Start from an initial guess $\boxed{\boldsymbol{a}_0}$

1. Let the time go by

$$\boxed{\boldsymbol{a}_{t+1} = \boldsymbol{M}\,\boldsymbol{a}_t}$$

product by a sparse matrix (twice) $\boldsymbol{M} = \boldsymbol{A}\,\boldsymbol{A}^T$

2. Keep normalizing (divide $\boldsymbol{a}_{t+1}$ by the sum of elements)

3. Stop when $\boldsymbol{a}$ converges (few iterations)

principle of repeated improvement

Power iteration method

Main convergence property:

- $||\boldsymbol{a}_t - \boldsymbol{a}_\infty||_2 \leq \sqrt{N} \cdot (\lambda_2/\lambda_1)^t$
- λ_1 largest eigenvalue of $\boldsymbol{M}$
- λ_2 second largest eigenvalue of $\boldsymbol{M}$
- Triang. inequality ensures $||\boldsymbol{a}_t - \boldsymbol{a}_{t+1}||_2 \leq 2\sqrt{N} \cdot (\lambda_2/\lambda_1)^t$

Worst case result:

- Precision ε implies: $||\boldsymbol{a}_t - \boldsymbol{a}_{t+1}||_2 < \varepsilon$
- Iterations required: $\boxed{t = [\ln(2/\varepsilon) + ½\ln(N)] \,/\, \ln(\lambda_1/\lambda_2)}$

$N = 10^9 \rightarrow 10.3$

10^{-3} precision $\rightarrow 7.6$

slow if λ_2 close to λ_1

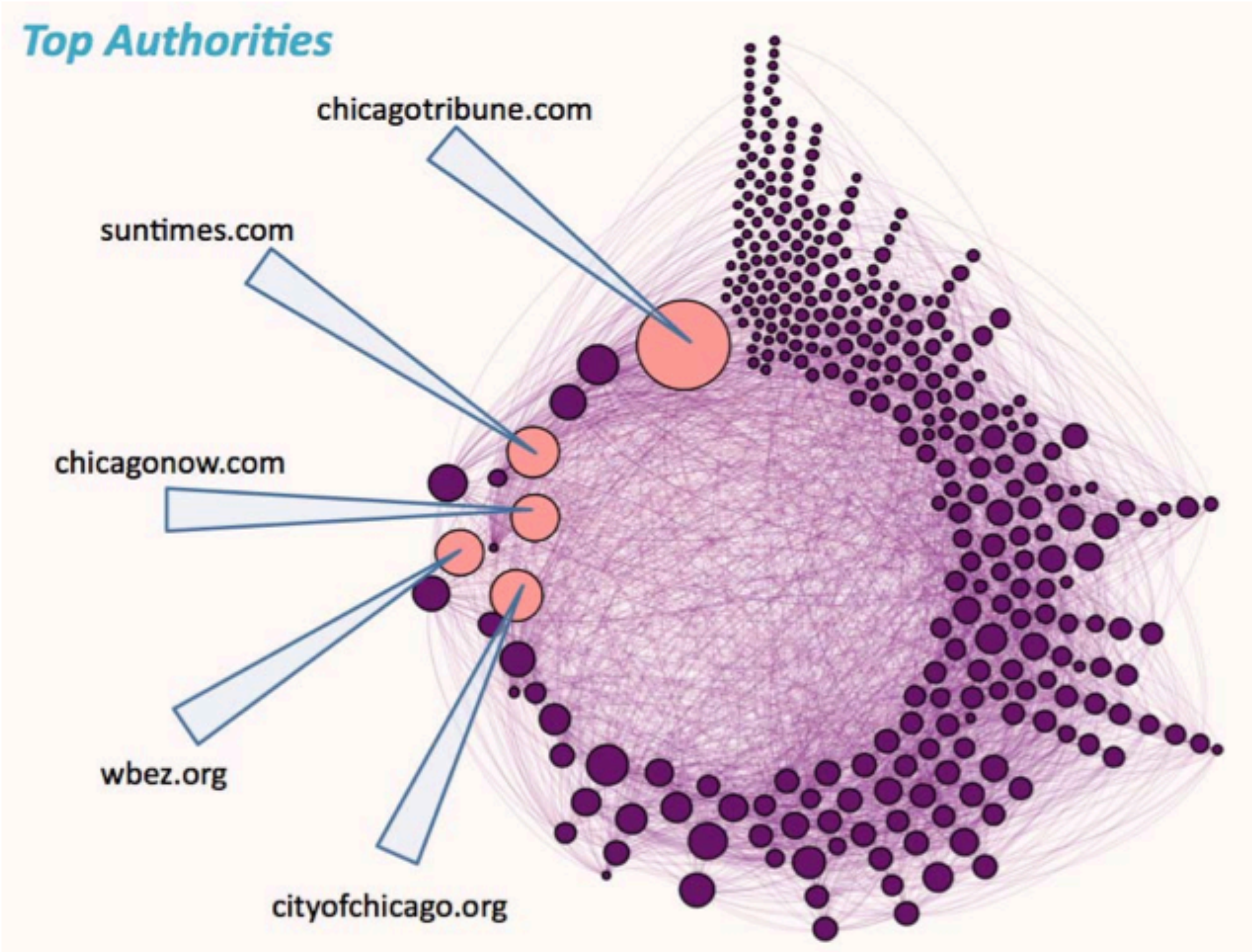
Application example – The news

Sonderman [2012]

«Study: Smaller news websites depend more on social media for traffic than larger sites,»

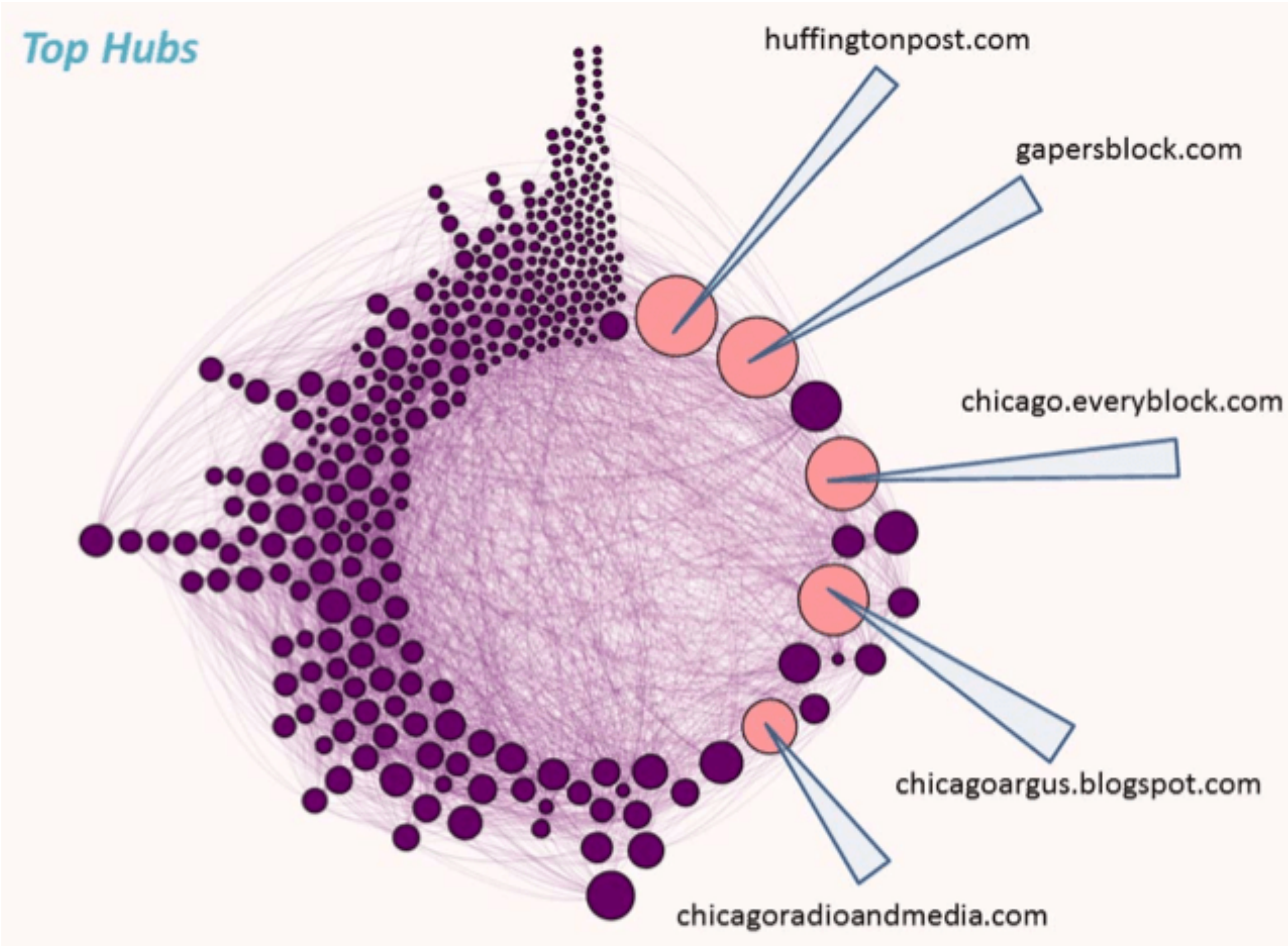
- Examined links between 301 news websites, for a two-week period
- 23 percent of all their referrals were from social media
- Small websites got more than half their referrals from social media, while the large sites got only about 19 percent from social

Authorities



- ❑ The legacy media brands

Hubs



- ❑ A different set of leading websites

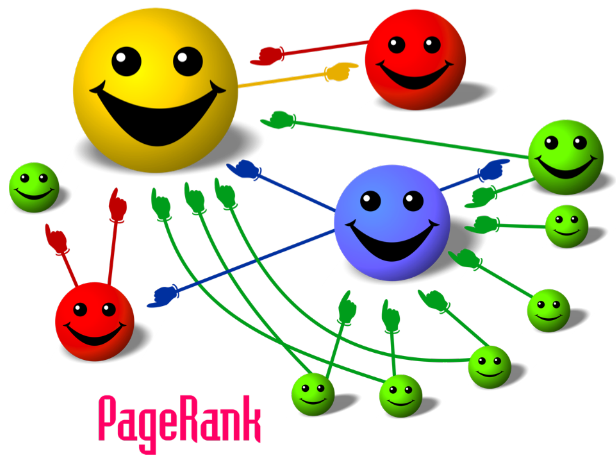
MIME.

PageRank centrality

PageRank

Quoting 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

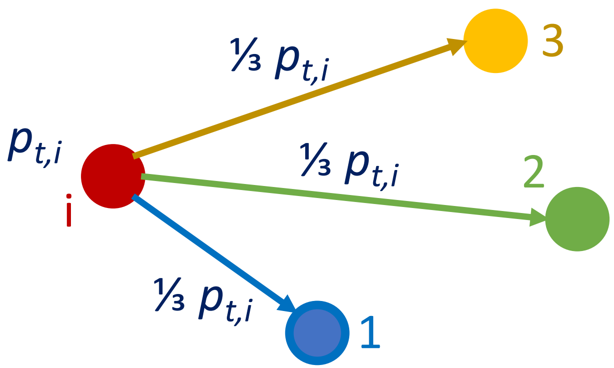


- Same ideas as HITS authorities
- Can be extended to hubs by using A^T

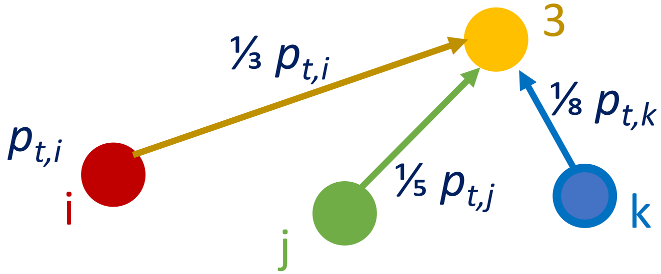
PageRank

The rationale behind PageRank

- Random walk
- at time t, a web surfer is at page *i* with probability $p_{t,i}$
- let the surfer choose with equal probability one of the sites linked by site *i*
- this identifies a Markov chain
- after a while probabilities settle to a steady state = the PageRank vector (authority score)



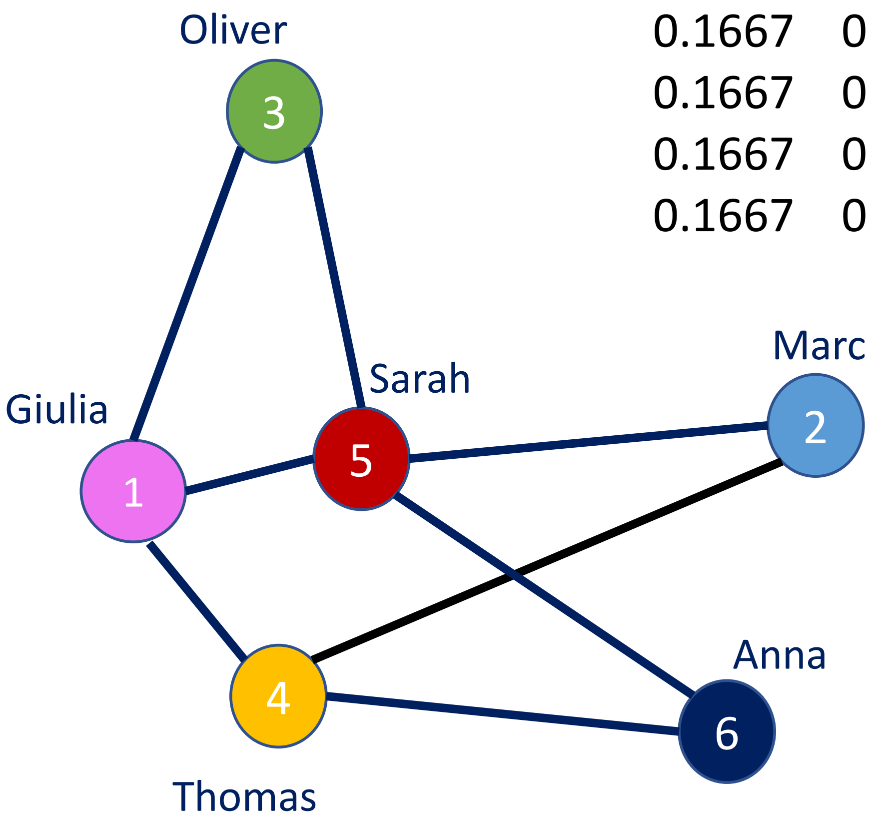
$$p_{t+1,3} = ⅓\, p_{t,i} + ⅕\, p_{t,j} + ⅛\, p_{t,k}$$



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 !!!

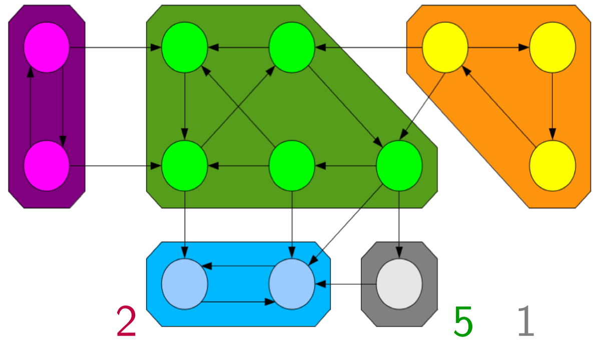


10	20	50	75	100	
0.1783	0.1848	0.1874	0.1875	0.1875	Giulia
0.1153	0.1222	0.1249	0.1250	0.1250	Marc
0.1242	0.1248	0.1250	0.1250	0.1250	Oliver
0.2020	0.1917	0.1876	0.1875	0.1875	Thomas
0.2649	0.2543	0.2501	0.2500	0.2500	Sarah
0.1153	0.1222	0.1249	0.1250	0.1250	Anna

PageRank

Markov chain interpretation

- $\boldsymbol{p}_{t+1} = \boldsymbol{M}\,\boldsymbol{p}_t$
- $\boldsymbol{p}_t$ stochastic vector (positive entries which sum up to 1)
- $\boldsymbol{M}$ normalized adjacency matrix (column stochastic)
- $\boldsymbol{M} = \boldsymbol{A}\,\mathrm{diag}^{-1}(\boldsymbol{d})$
- $\boldsymbol{d} = \boldsymbol{A}^{\mathsf{T}}\,\boldsymbol{1}$ output degree vector
- $\boldsymbol{p}_\infty = \boldsymbol{M}\,\boldsymbol{p}_\infty$ converges to an eigenvector of $\boldsymbol{M}$ (with eigenvalue 1)



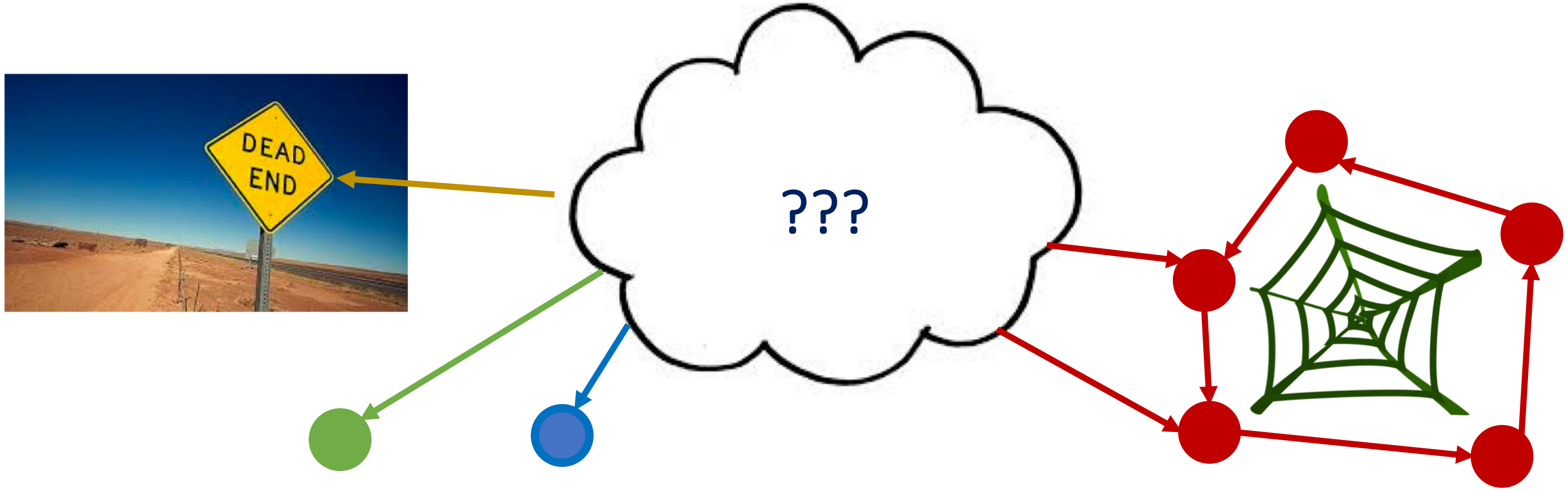
$$
\begin{array}{cc}
 & \begin{array}{ccccc} 3 & 2 & 5 & 1 & 2 \end{array} \\
\boldsymbol{M} = & \left[\begin{array}{ccc|cc|ccccc|c|cc}
 & \frac{1}{3} & & & & & & & & & & & \\
 & & 1 & & & & & & & & & & \\
1 & & & & & & & & & & & & \\
\hline
 & & & & \frac{1}{2} & & & & & & & & \\
 & & & \frac{1}{2} & & & & & & & & & \\
\hline
 & \frac{1}{3} & & & & & & & & \frac{1}{2} & & & \\
 & & & \frac{1}{2} & & \frac{1}{2} & & & \frac{1}{3} & & & & \\
 & \frac{1}{3} & & & & \frac{1}{2} & & & & & & & \\
 & & & & & & & \frac{1}{3} & & & & & \\
 & & & & \frac{1}{2} & & 1 & & \frac{1}{3} & & & & \\
 & & & & & & & \frac{1}{3} & & & 0 & & \\
\hline
 & & & & & & & \frac{1}{3} & \frac{1}{3} & & 1 & & 1 \\
 & & & & & & & & & \frac{1}{2} & & 1 &
\end{array}\right]
\end{array}
$$

columns sum to 1

PageRank

With high probability the surfer ends in:

- ❑ Dead ends: some nodes do not have a way out = zero valued columns of ***M***
- ❑ 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
- $\boldsymbol{p}_{t+1} = \mathrm{c}\, \boldsymbol{M}\, \boldsymbol{p}_t + (1\text{-}c)\, \boldsymbol{q}$

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 at random according to a probability vector $\boldsymbol{q}$ independent of the node she/he is in, e.g., $\boldsymbol{q}=\boldsymbol{1}/\mathrm{N}$ for uniform probability

PageRank with restart

original adjacency matrix
(can be fractional)

dead ends $\boldsymbol{A}_0$

teleportation vector

no dead ends $\boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{b}\,\boldsymbol{e}^{\mathrm{T}}$

indicating vector
of dead ends

normalization $\boldsymbol{M} = \boldsymbol{A}\ \mathrm{diag}^{-1}(\boldsymbol{d}),\qquad \boldsymbol{d} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{1}$

no spider traps $\boldsymbol{M}_1 = c\,\boldsymbol{M} + (1\text{-}c)\ \boldsymbol{q}\,\boldsymbol{1}^{\mathrm{T}}$

equivalent formulation
matrix is no more sparse

$$\boxed{\boldsymbol{p}_{t+1} = \boldsymbol{M}_1\,\boldsymbol{p}_t}$$

MIME.

PageRank with restart

damping factor

(column) normalized
adjacency matrix

PageRank equation

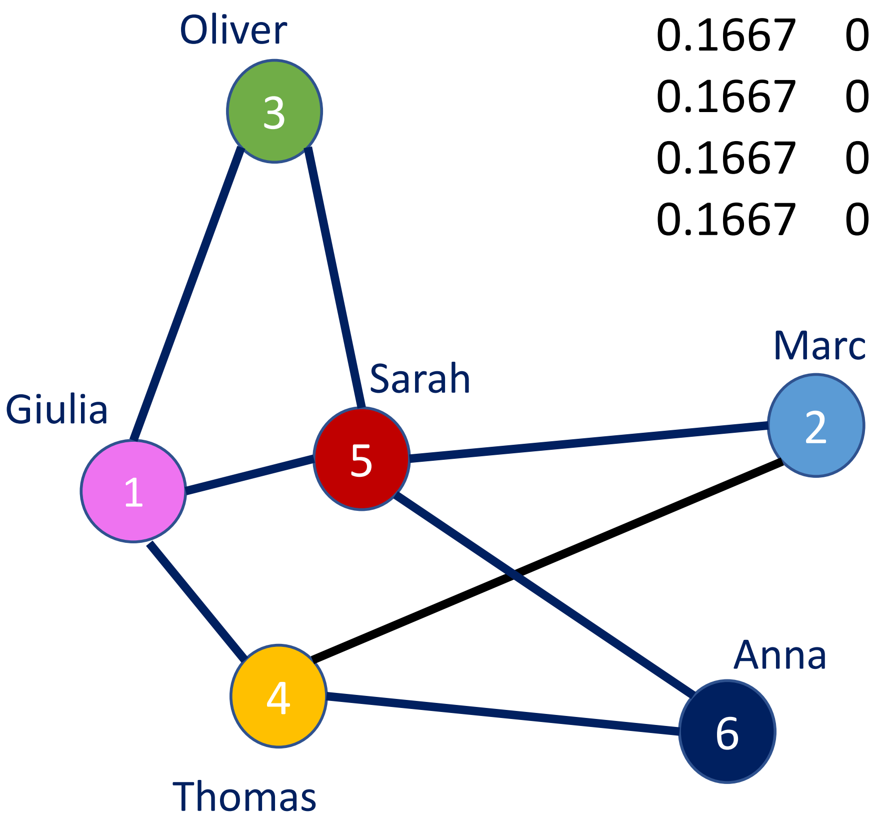
$$\boldsymbol{r} = c\,\boldsymbol{M}\,\boldsymbol{r} + (1-c)\,\boldsymbol{q}$$

teleportation vector

PageRank vector
(centrality)

MIME.

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 !!!

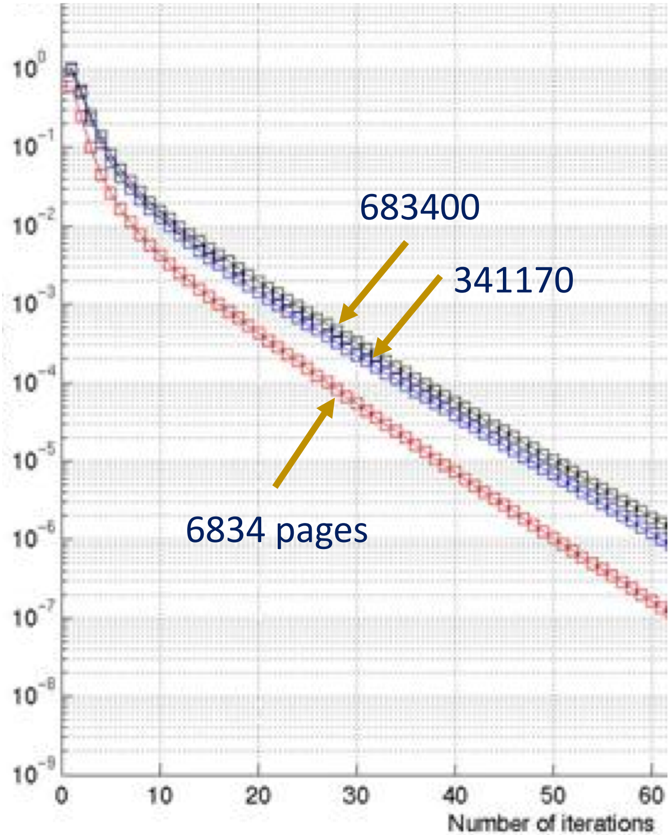
10	20	50	75	100	
0.1820	0.1839	0.1840	0.1840	0.1840	Giulia
0.1273	0.1293	0.1294	0.1294	0.1294	Marc
0.1283	0.1285	0.1285	0.1285	0.1285	Oliver
0.1902	0.1873	0.1871	0.1871	0.1871	Thomas
0.2449	0.2419	0.2417	0.2417	0.2417	Sarah
0.1273	0.1293	0.1294	0.1294	0.1294	Anna

MIME.

Properties

- The 

vector is the probability $\boldsymbol{p}_t$ for large t
- It corresponds to the stationary behaviour of the Markov chain
- $\boldsymbol{p}_\infty$ is unique
- $\boldsymbol{p}_\infty$ is a stochastic vector (i.e., with positive entries summing to *1*)
- $\boldsymbol{p}_\infty$ depends on the choice of the teleportation vector $\boldsymbol{q}$ (and of c)
- $\boldsymbol{p}_\infty$ converges in few iterations, typically $p_{40} \simeq p_\infty$



Power iteration method

Main convergence property:

- $||\boldsymbol{p}_t - \boldsymbol{a}_\infty||_2 \lesssim K\, c^t\, t^{m-1} \sim K\, c^t$
- Triangular inequality: $||\boldsymbol{p}_{t+1} - \boldsymbol{p}_t||_2 \lesssim 2K\, c^t$

Complexity considerations:

- Precision ε at: $||\boldsymbol{p}_{t+1} - \boldsymbol{p}_t||_2 < \varepsilon$
- Iterations: $\boxed{t = [\ln(2/\varepsilon) + \ln(K)] / \ln(1/c)}$

precision 10^{-3} → 7.6

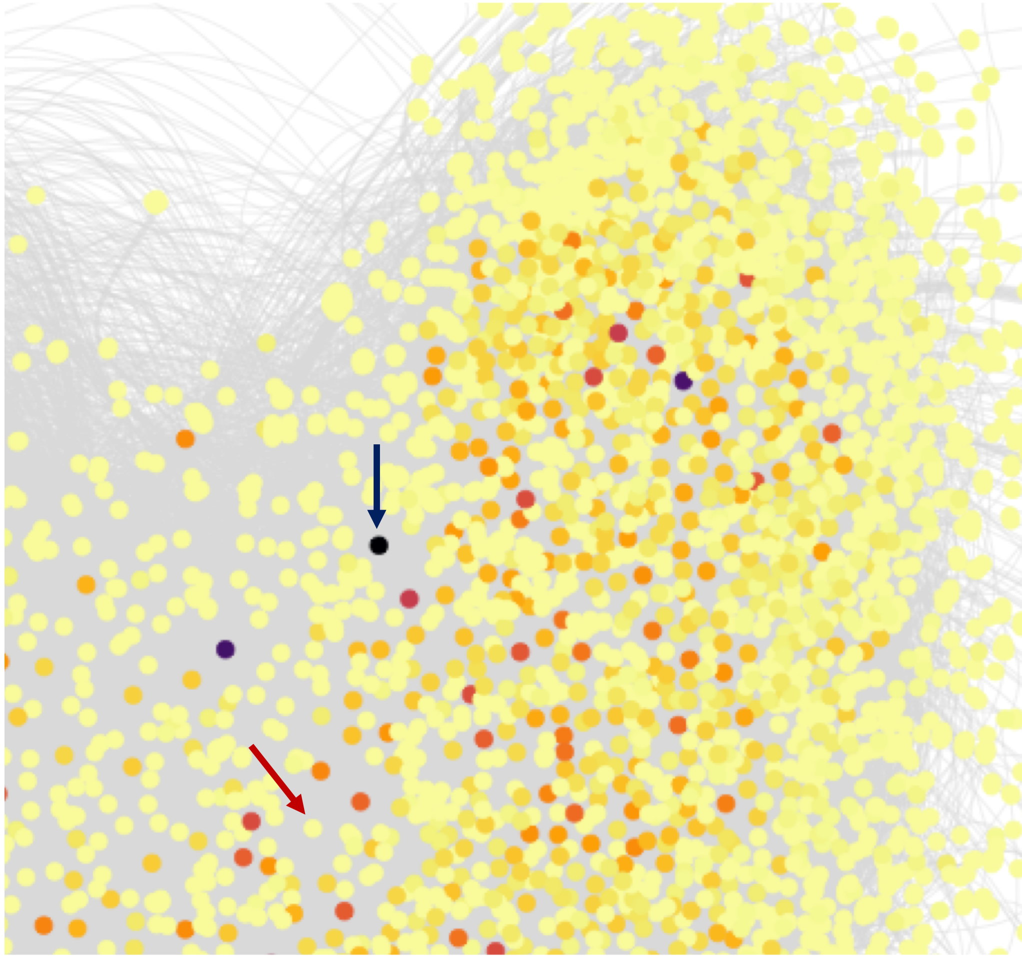
can be proportional to ln(N) → slow algorithm

c=0.85 → 1/ln(1/c) = 6

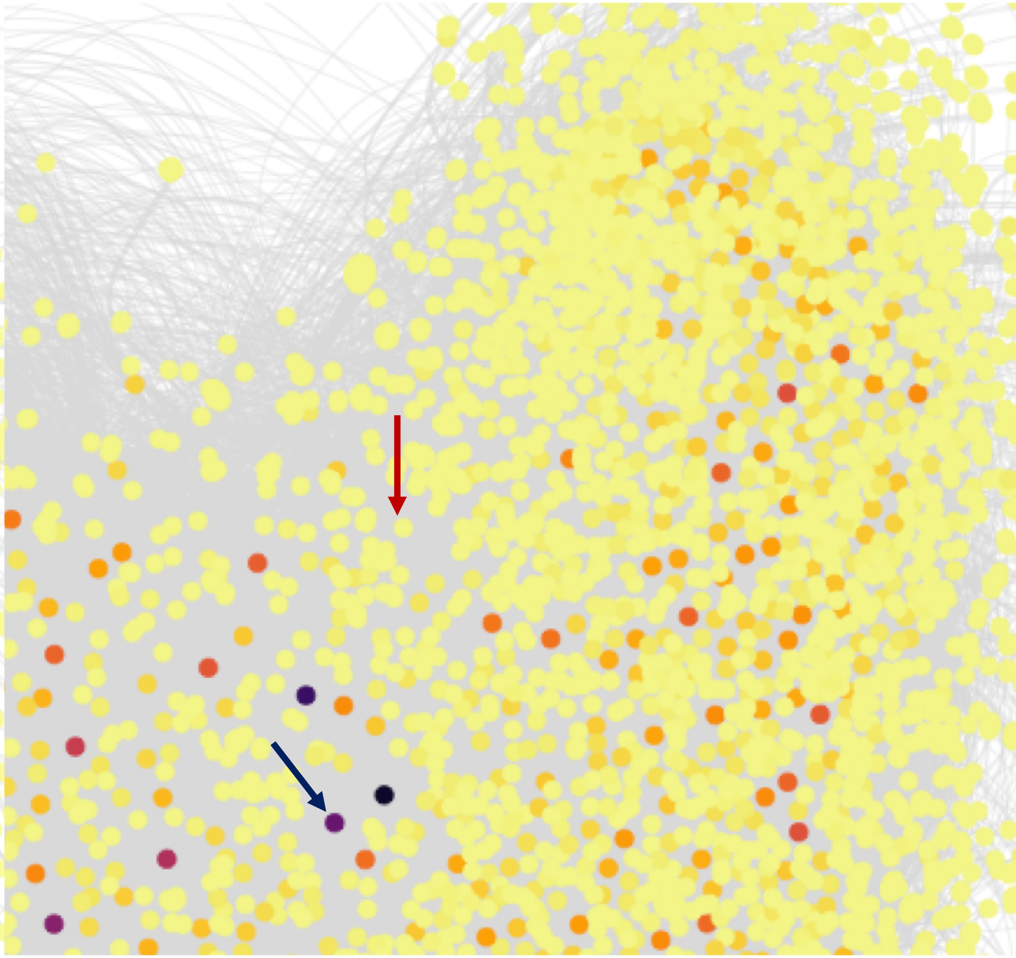
Wiki-vote network example

PageRank centrality

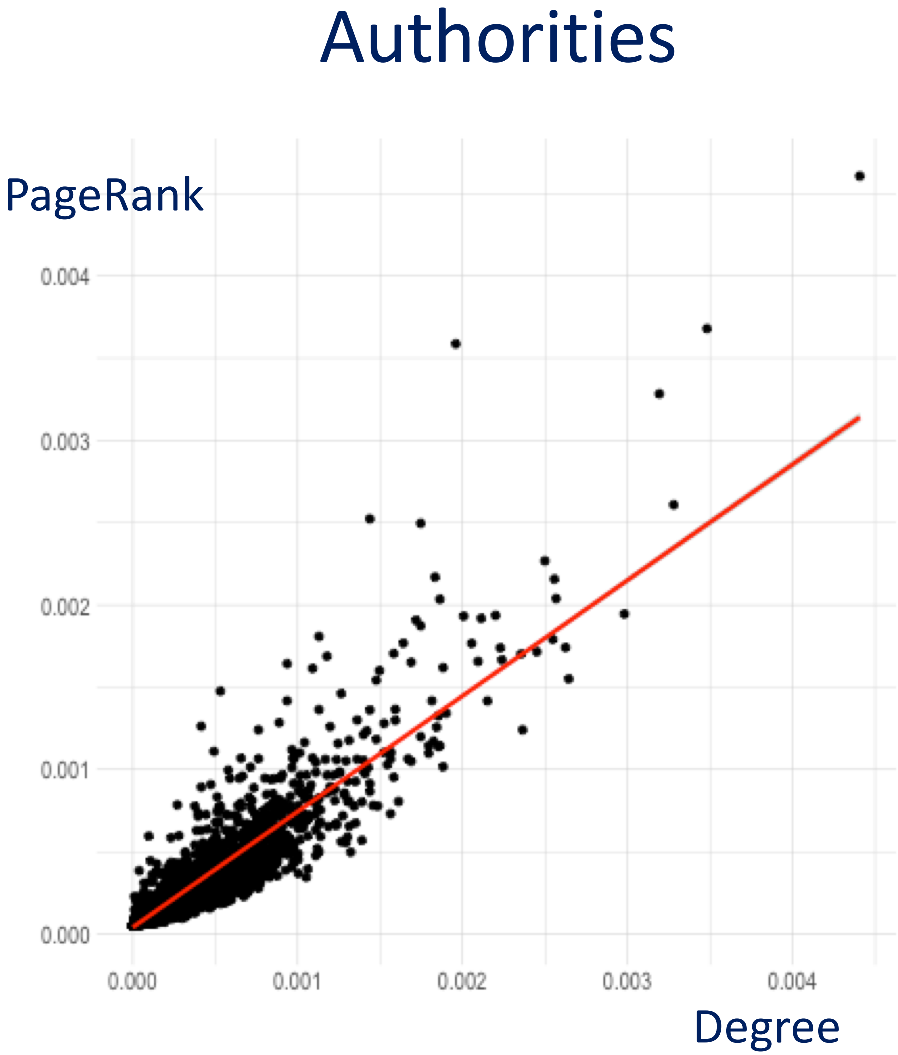
Authorities

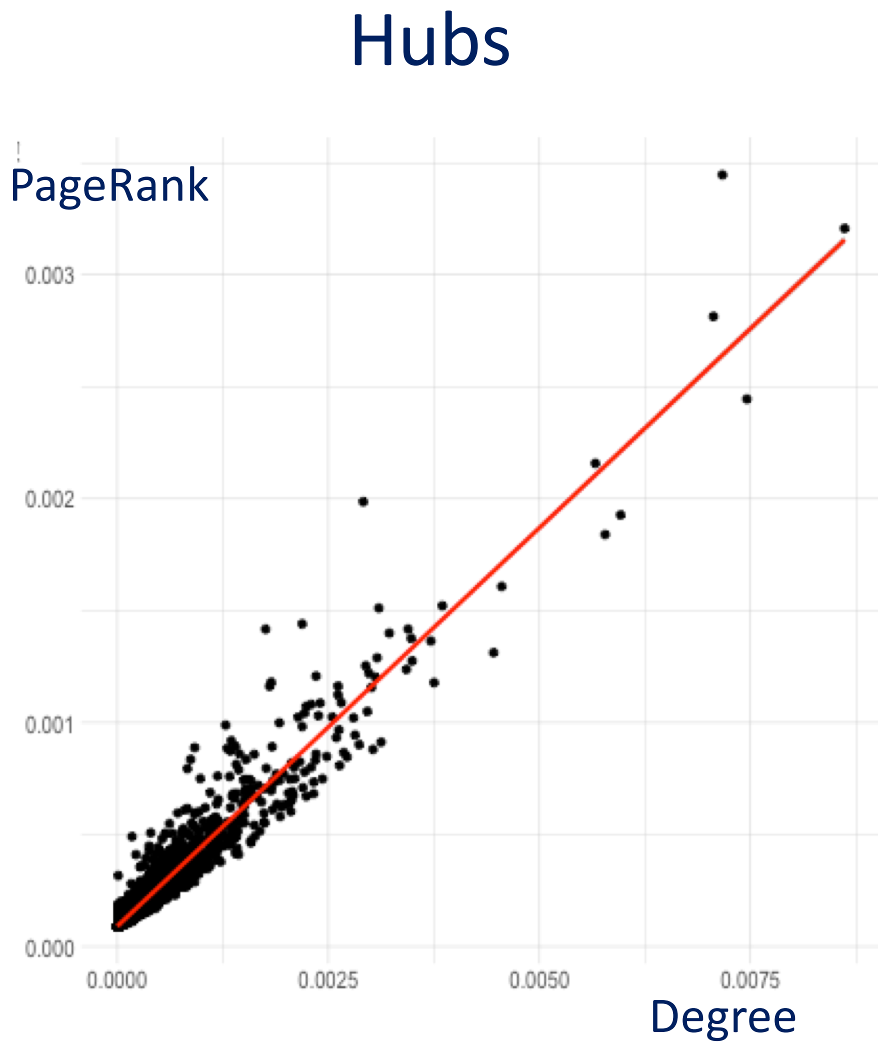


Hubs



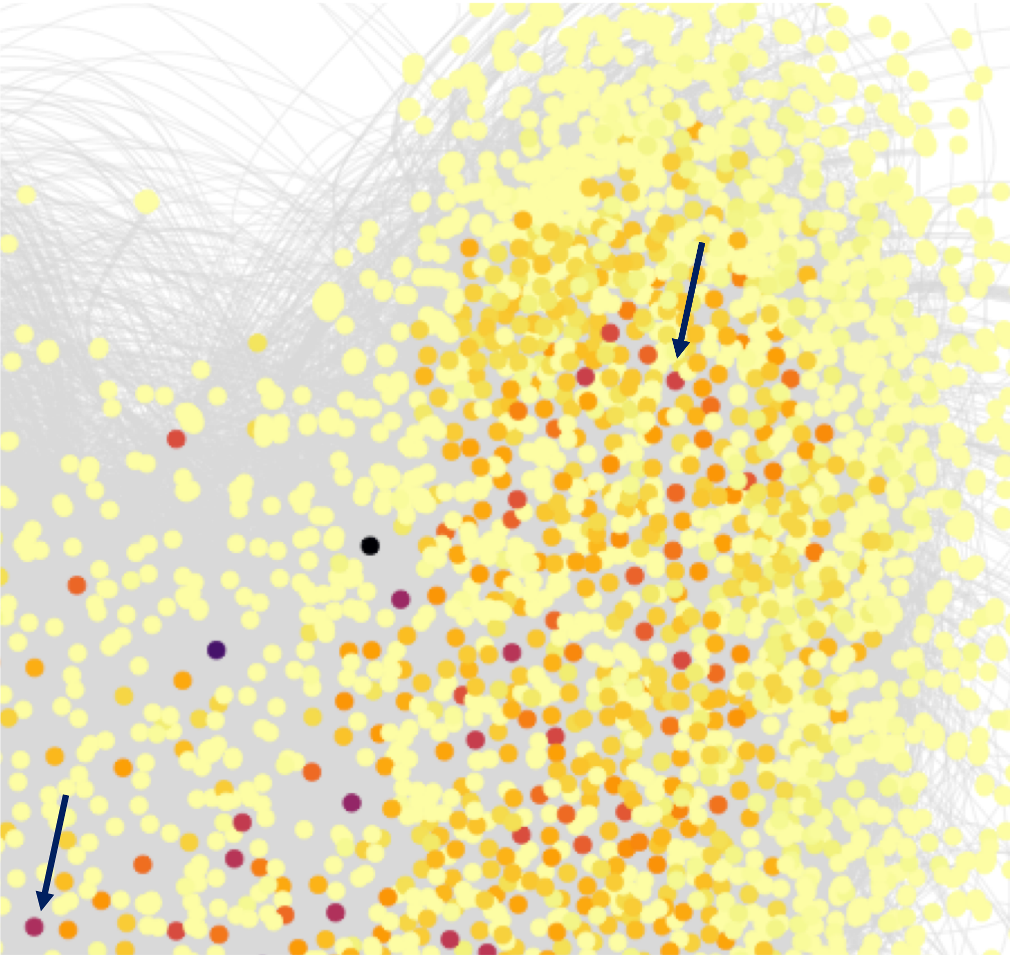
Degree vs PageRank



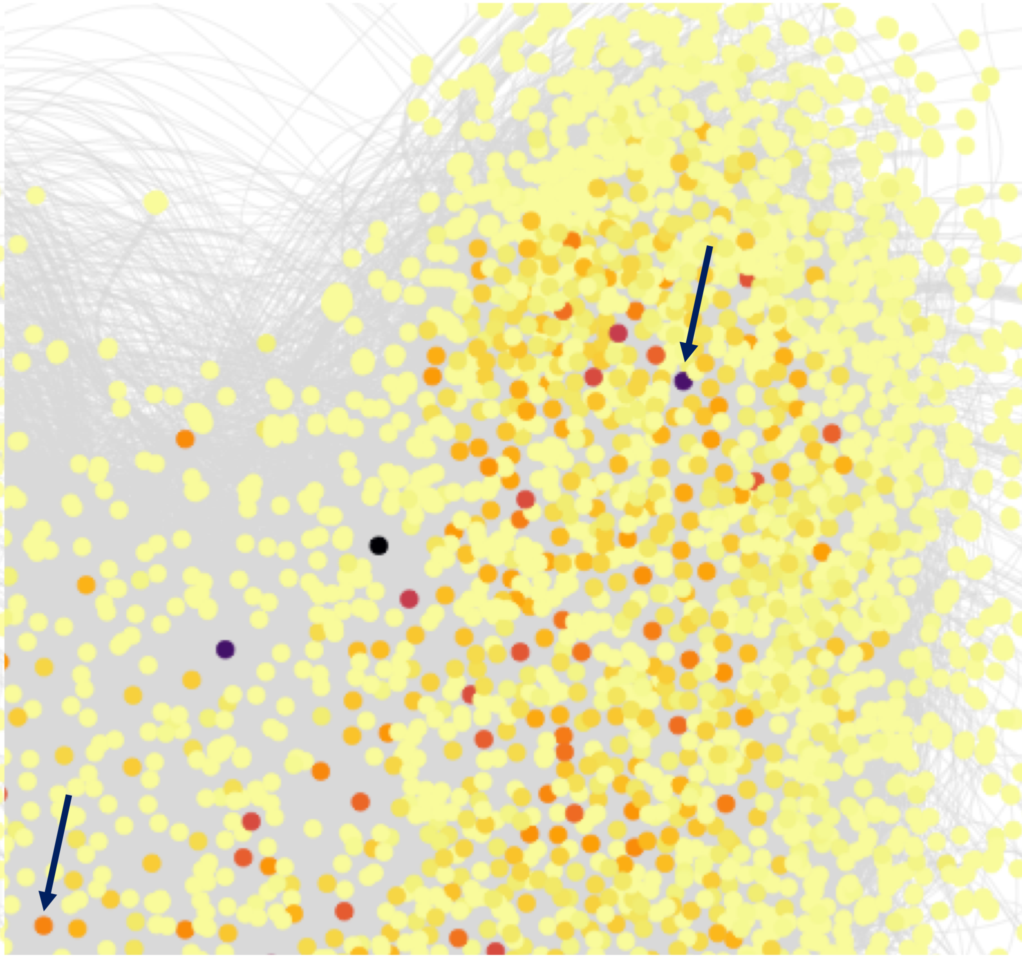


Authorities

Degree

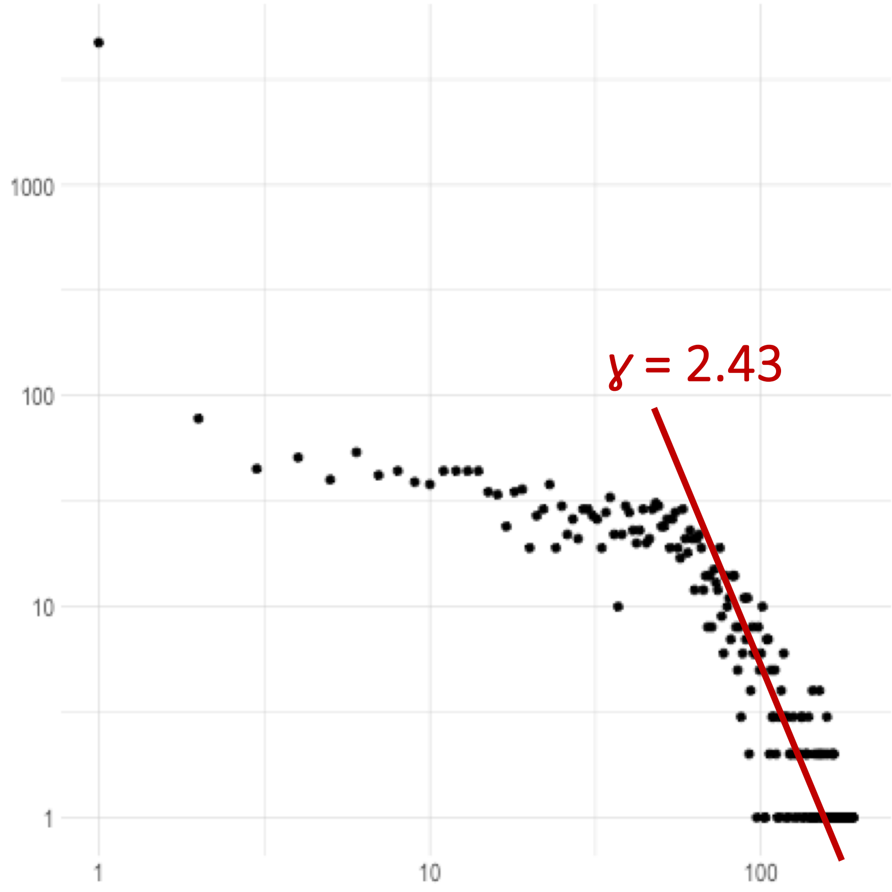


PageRank

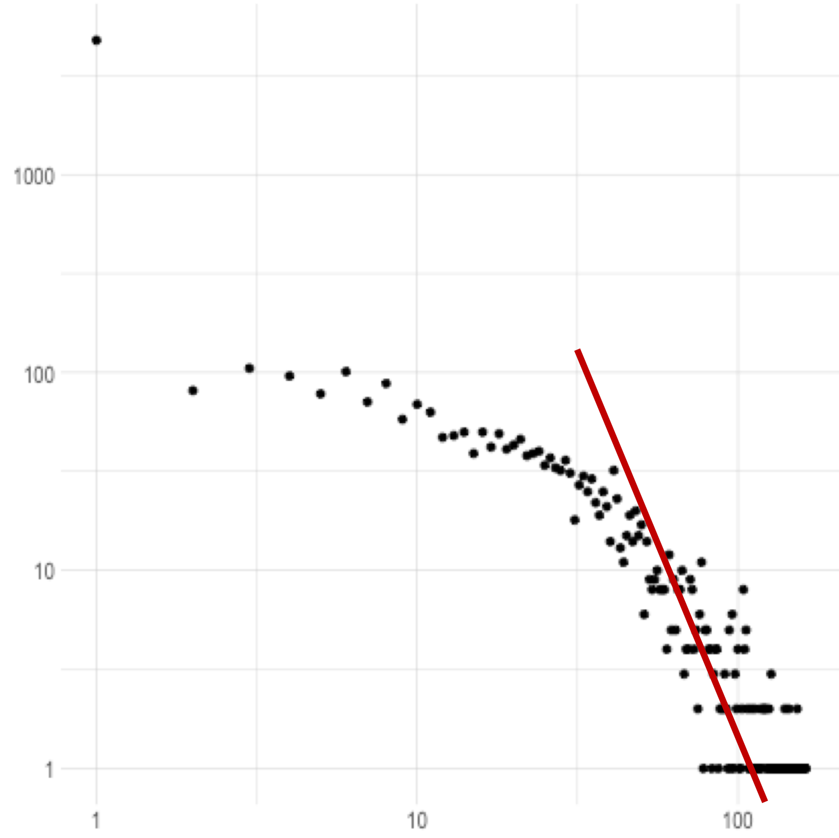


Authorities

Degree

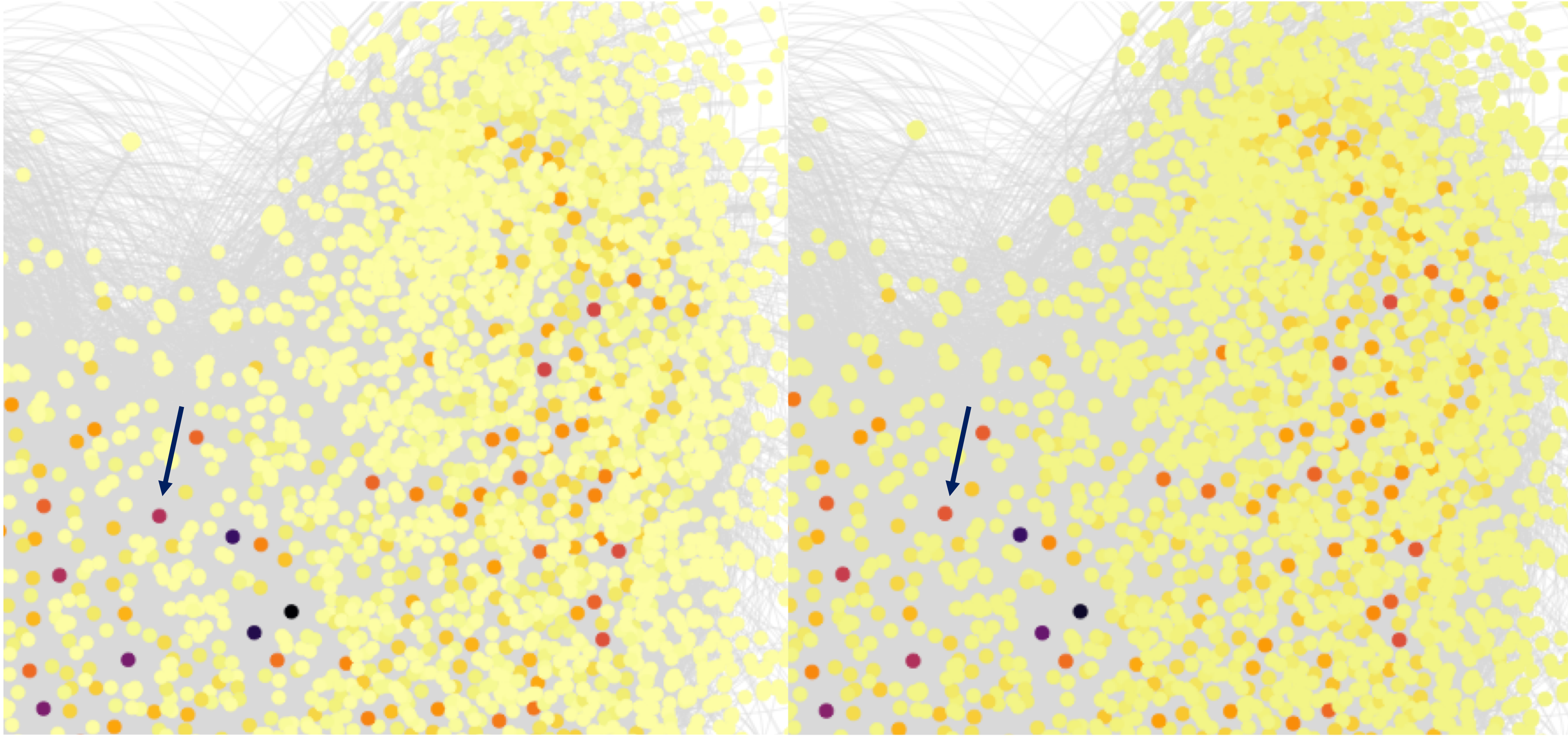


PageRank



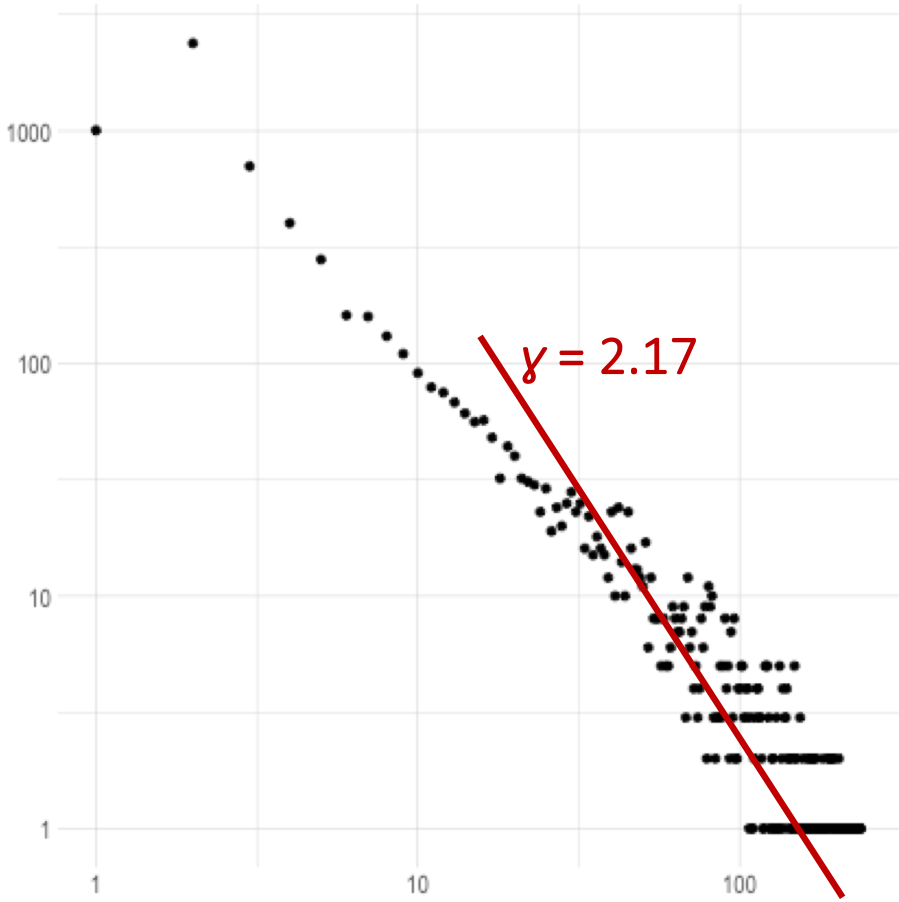
Hubs

Degree PageRank

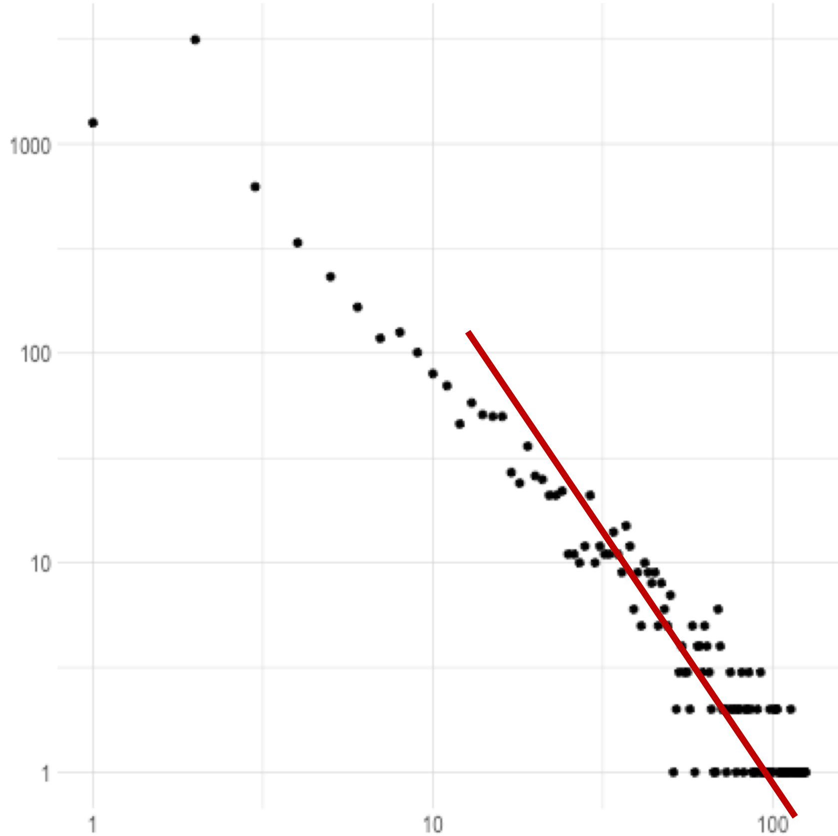


Hubs

Degree



PageRank



Lessons learned

- Two different approaches
- Based on simple linear algebra concepts
- Scalable
- Implementable via simple message exchange algorithms

HITS

$\mathbf{A}^T\mathbf{A}$

$$\boldsymbol{h} = c\, \boldsymbol{M}\, \boldsymbol{h}$$

$$\boldsymbol{a} = c_a\, \boldsymbol{A}\, \boldsymbol{h}$$

PageRank

$$\boldsymbol{r} = c\, \boldsymbol{M}\, \boldsymbol{r} + (1 - c)\, \boldsymbol{q}$$

$\boldsymbol{A}\ \mathrm{diag}^{-1}(\boldsymbol{d})$

MiME.

Readings

HITS

- Kleinberg, "Authoritative sources in a hyperlinked environment," 1999

https://www.cs.cornell.edu/home/kleinber/auth.pdf

PageRank

- Brin and Page, "The anatomy of a large-scale hypertextual web search engine," 1998
- Page, Brin, Motwani, Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," 1999

http://ilpubs.stanford.edu/422/1/1999-66.pdf

Power iteration

- Wikipedia, "Power iteration"

Google Scholar
https://scholar.google.com/