Università degli Studi di Padova

Network Science

A.Y. 23/24

ICT for Internet & multimedia, Data science, Physics of data

Centrality

Importance of nodes in a network

Centrality

From Wikipedia, the free encyclopedia

For the statistical concept, see Central tendency.

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] They should not be confused with node influence metrics, which seek to quantify the influence of every node in the network.

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

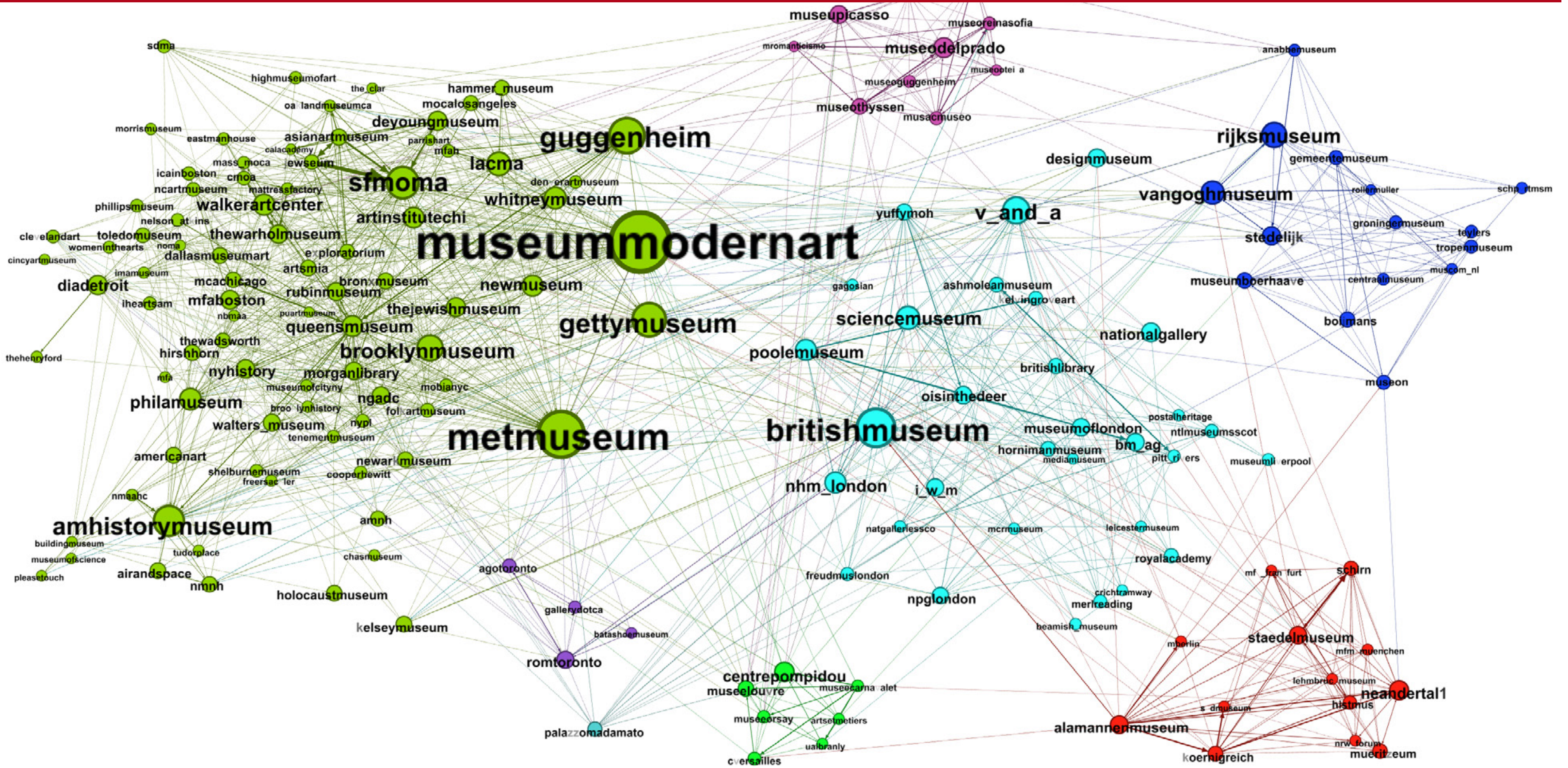
Closeness centrality [edit]

Main article: Closeness centrality

UNIVERSITÀ DEGLI STUDI DI PADOVA

An example of node centrality

museums network



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

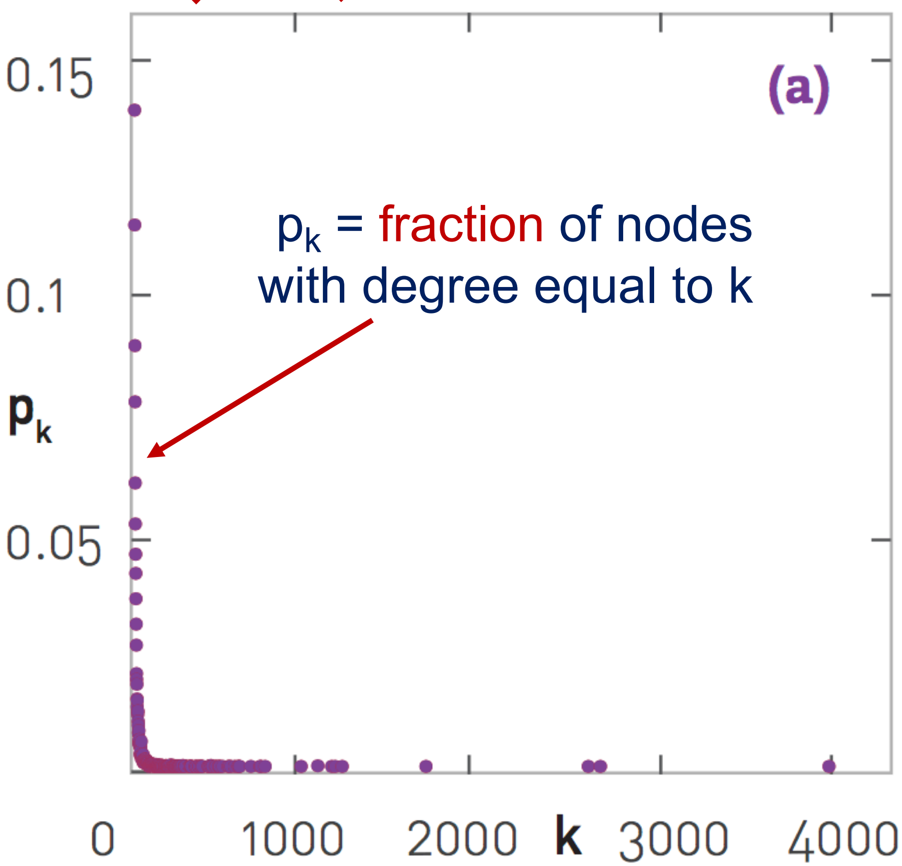
Degree centrality

Counting the in/out degrees of nodes

Università degli Studi di Padova

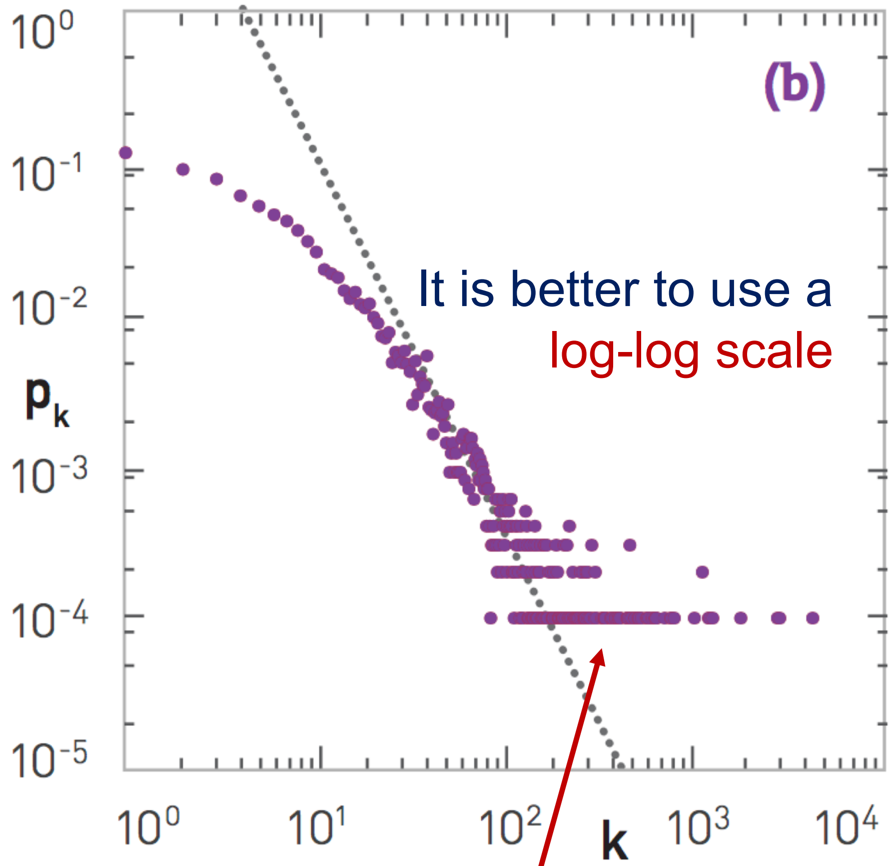
The degree distribution
for an undirected network

~~LINEAR~~ SCALE



Wide range for the degree k !
Wide range for the probability p_k !

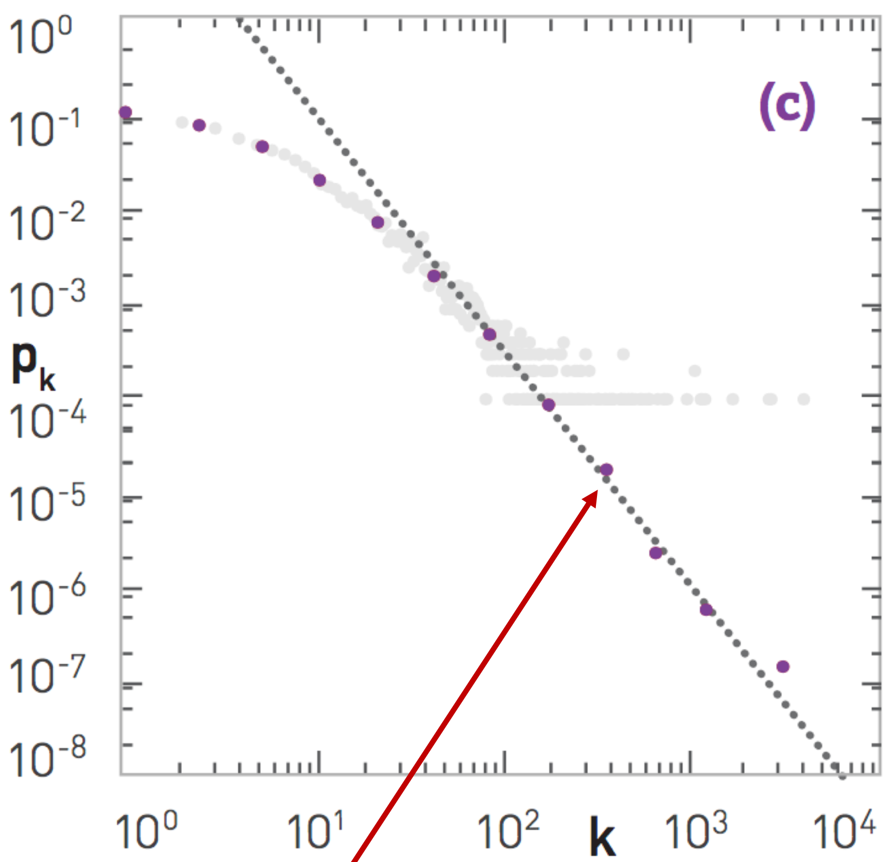
LINEAR BINNING



plateau

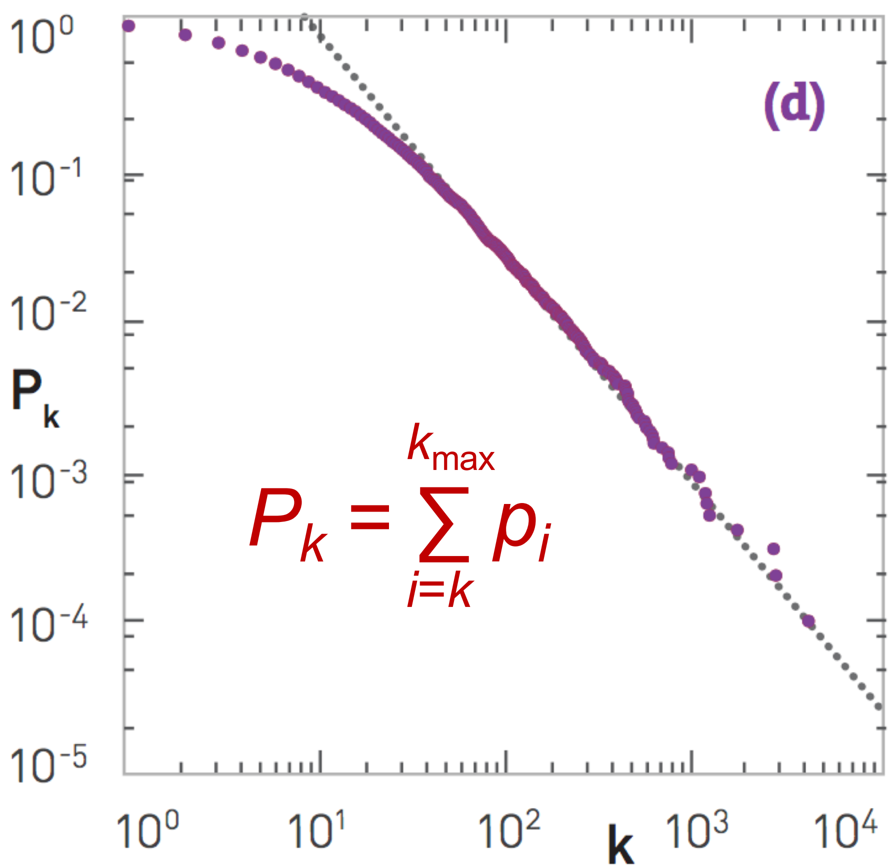
Alternative log representations
for an undirected network

LOG-BINNING



p_{k_i} = fraction of nodes with degree in the range $[k_i,k_{i+1})$ where k_i are uniformly distributed in the log-domain, $k_{i+1}=k_i \cdot \Delta$

CUMULATIVE

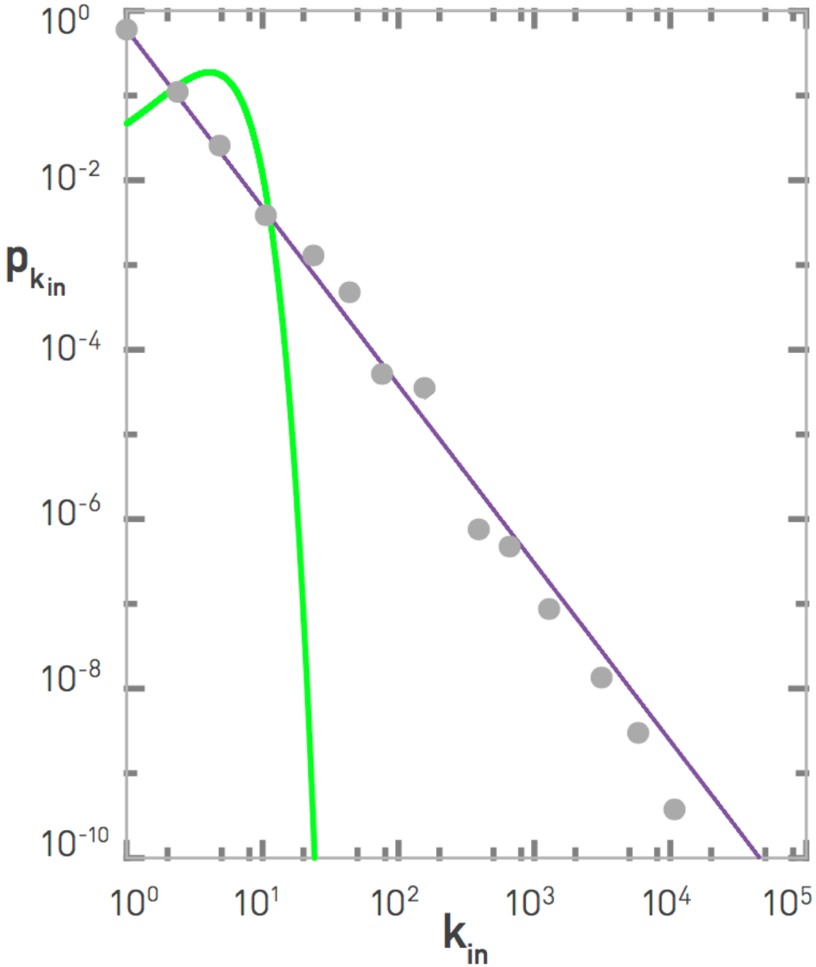


$$P_k = \sum_{i=k}^{k_{max}} p_i$$

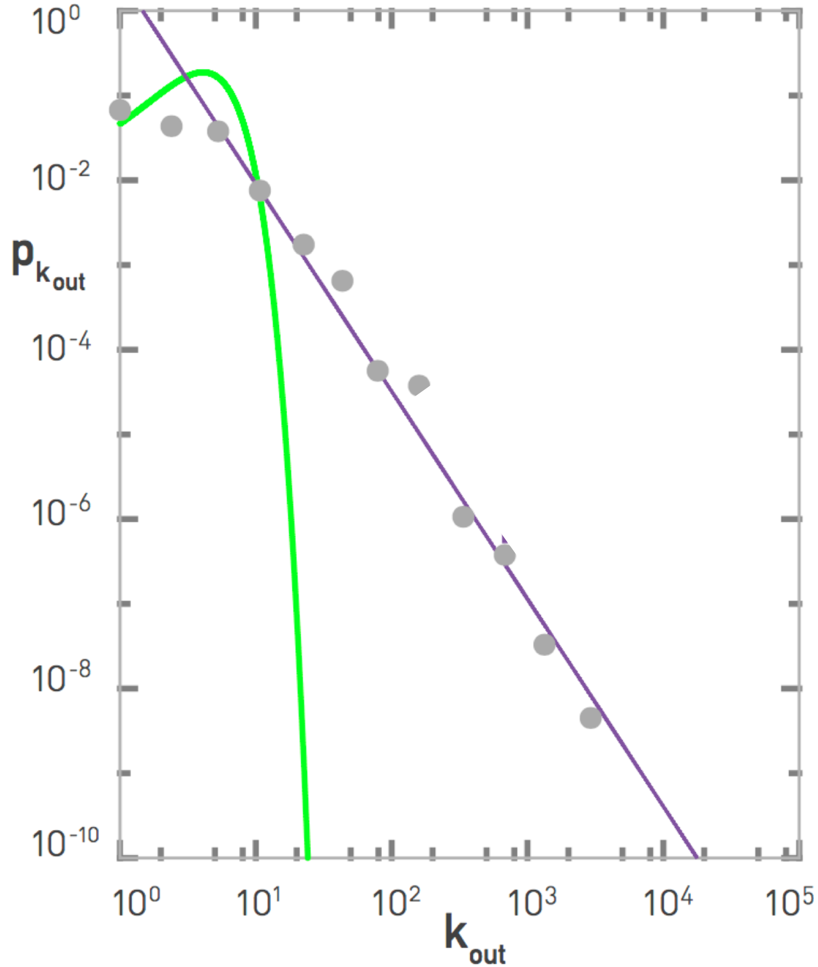
complementary cumulative distr. function (CCDF)

Two degree distributions

for directed networks



p_{kin} = fraction of nodes with input degree equal to k_{in}

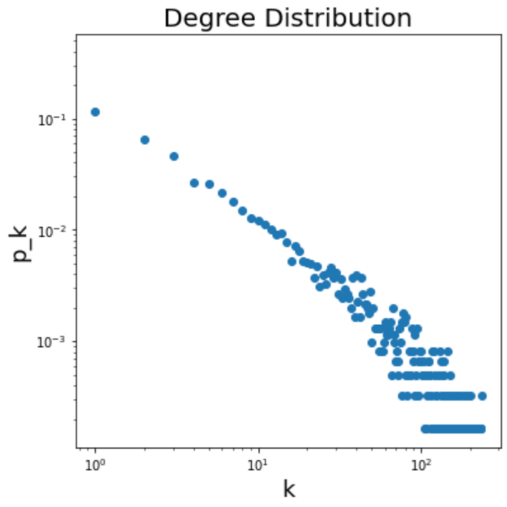


p_{kout} = fraction of nodes with output degree equal to k_{out}

Pseudocode example

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

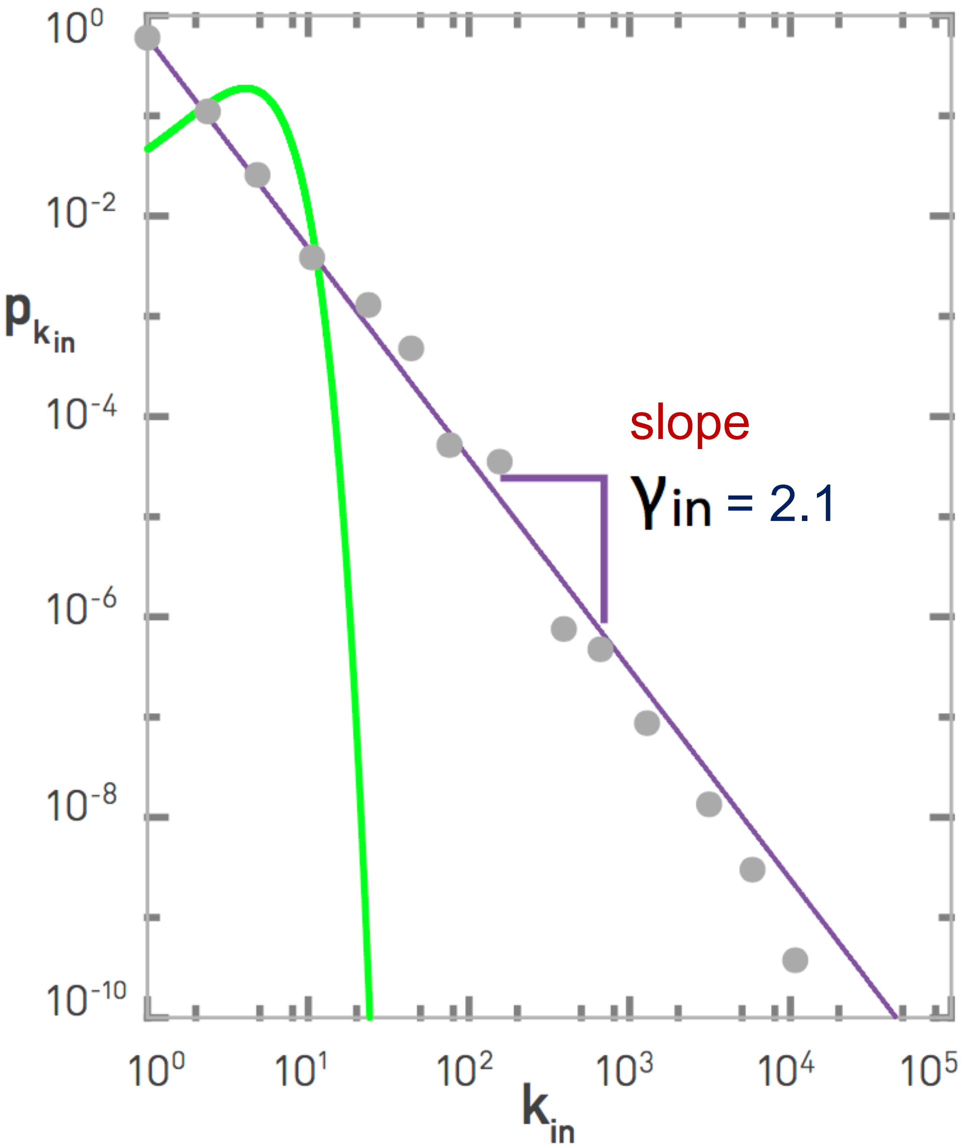
```
G = np.loadtxt('Wiki-Vote.txt').astype(int)

# adjacency matrix
N = np.max(G)
A = csr_matrix((np.ones(len(G)), (G[:, 1], G[:, 0])))


#distribution
which_deg = 0 # 0=out degree, 1=in degree
d = np.sum(A, which_deg) # out degree for each node
d = np.squeeze(np.asarray(d)) # from matrix to array
d = d[d>0] # avoid zero degree
k = np.unique(d) # degree samples
pk = np.histogram(d, k)[0] # occurrence of each degree
pk = pk/np.sum(pk) # normalize to 1
Pk = 1 - np.cumsum(pk) # complementary cumulative
```




```
fig = plt.figure()
plt.loglog(pk, 'o')
plt.title("Degree Distribution", size = 20)
plt.xlabel("k", size = 18)
plt.ylabel("p_k", size = 18)
plt.show()
```

The power-law

typical behaviour of social networks



many networks follow a power-law

$$\ln(p_k) = c - \gamma \cdot \ln(k)$$

$$\boxed{p_k = C \cdot k^{-\gamma}}$$

how to correctly estimate the slope γ ?

Degree distribution $p_k = C\, k^{-\gamma}$

Constant C is determined by the (approx.) normalization condition

$$\int_{k_{min}}^{\infty} p_k\, dk = C \cdot k_{min}^{-(\gamma-1)} / (\gamma-1) = 1$$

Target PDF $\boxed{p(k|\gamma) = (\gamma-1)/k_{min} \cdot (k/k_{min})^{-\gamma}}$

ML estimate for the exponent γ

the most reliable approach

ML criterion: find the γ that best fits the data

$$\max_\gamma \sum_i \ln p(k_i|\gamma)$$

where k_i is the measured degree of node i

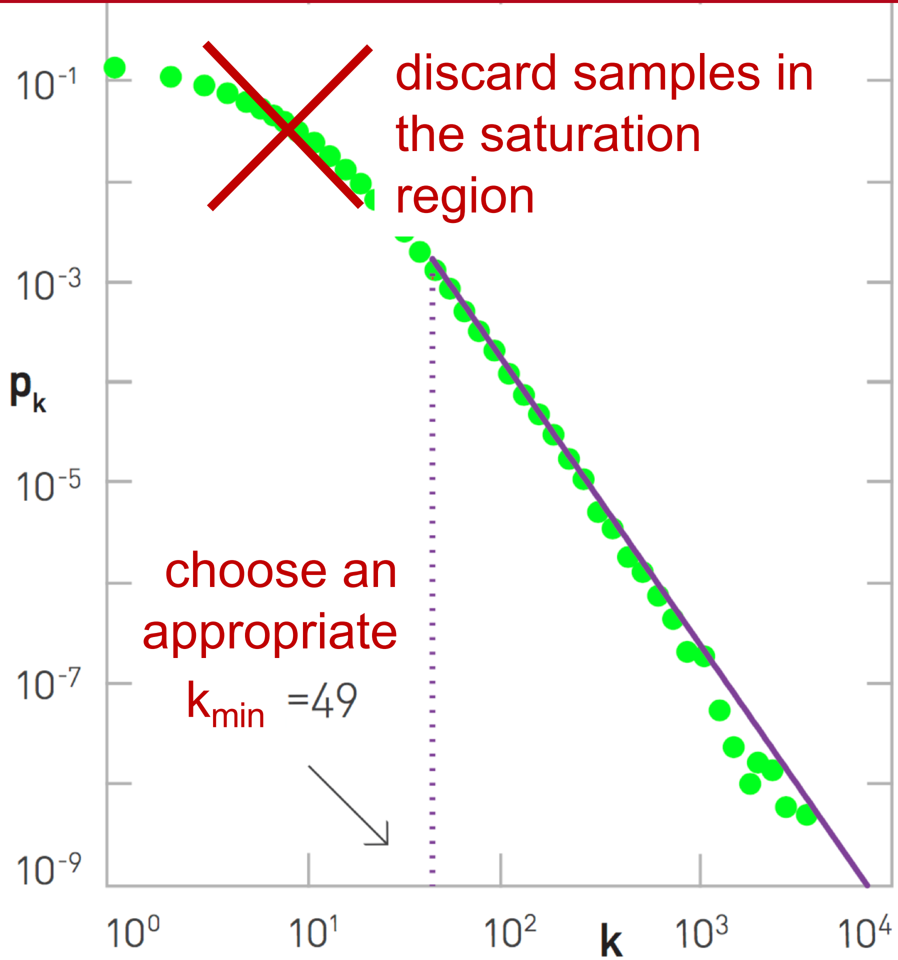
$$f(\gamma) = \sum \ln((\gamma-1)/k_{\text{min}}) - \gamma \ln(k_i/k_{\text{min}})$$

$$f'(\gamma) = \sum 1/(\gamma-1) - \ln(k_i/k_{\text{min}}) = 0$$

$$\boxed{\gamma = 1 + \sum_i 1 / \sum_i \ln(k_i/k_{\text{min}})}$$

Pseudocode example

to estimate the exponent



MATLAB

```
which_deg = 1; % 1 = out degree
d = full(sum(A,which_deg));
d2 = d(d>=kmin); % restrict range
ga = 1+1/mean(log(d2/kmin)); % estimate the exponent
```

The value of the exponent ɣ
in real networks ɣ ∊ [2,5]

NETWORK	N	L	$\langle k \rangle$	γ_{in}	γ_{out}	γ
Internet	192,244	609,066	6.34	-	-	3.42*
WWW	325,729	1,497,134	4.60	2.00	2.31	-
Power Grid	4,941	6,594	2.67	-	-	Exp.
Mobile Phone Calls	36,595	91,826	2.51	4.69*	5.01*	-
Email	57,194	103,731	1.81	3.43*	2.03*	-
Science Collaboration	23,133	93,439	8.08	-	-	3.35*
Actor Network	702,388	29,397,908	83.71	-	-	2.12*
Citation Network	449,673	4,689,479	10.43	3.03**	4.00*	-
E. Coli Metabolism	1,039	5,802	5.58	2.43*	2.9 0*	-
Protein Interactions	2,018	2,930	2.9 0	-	-	2.89*

* = good statistical fit with a power-law

** = good fit for a power-law with an exponential cutoff

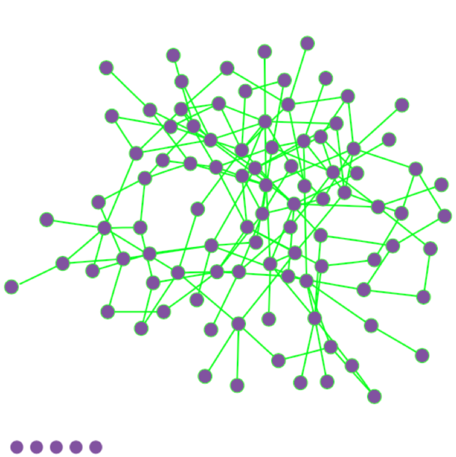
Exp = good fit with an exponential distribution e^{-ak}

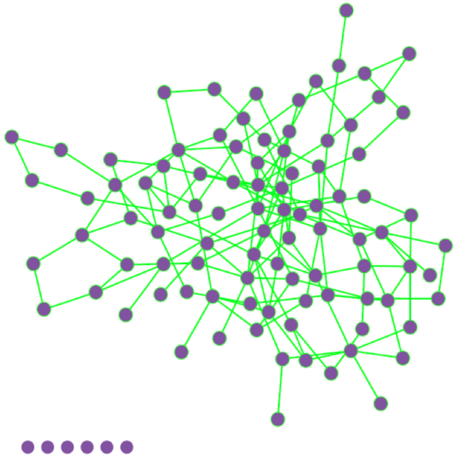
Explaining the power-law

Preferential attachment

Random networks

Erdös-Rényi model 1959/60





- The random network is the simplest model:

  pick a probability *p*, with *0<p<1*

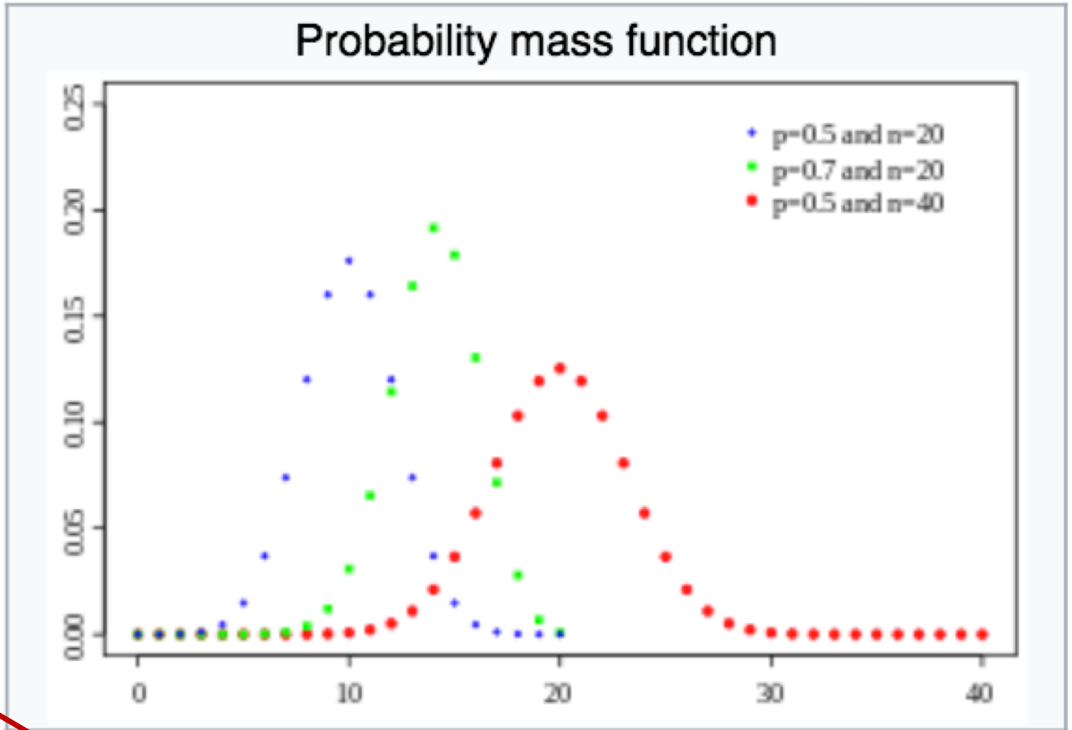
  activate each link (*i,j*) with probability p
- The number of links is variable
- There might be isolates
- Easy to calculate fundamental parameters

UNIVERSITÀ DEGLI STUDI DI PADOVA

Binomial distribution

explains the degree distribution for random networks

Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial $q = 1 - p$
Support	$k \in \{0,1,\ldots,n\}$ – number of successes
PMF	$\binom{n}{k} p^k q^{n-k}$
CDF	$I_q(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	npq
Skewness	$\frac{q-p}{\sqrt{npq}}$
Ex. kurtosis	$\frac{1-6pq}{npq}$



P(k;n,p) = probability that *k* out of *n* trials are positive, where each is positive with probability *p*

WIKIPEDIA
L'enciclopedia libera

Degree distribution
in random networks

- The number of neighbours is binomially distributed

  P($k;n,p$) = probability that a node has exactly k neighbours, with number of possible neighbours $n = N\text{-}1$

- Average # of neighbours

  this defines p

  $$\langle k \rangle = (N-1)p \;\rightarrow\; \boxed{p = \langle k \rangle/(N-1)}$$

- Variance

  p is usually very small (since $\langle k \rangle \ll$ N)

  $$\sigma_x^2 = (N-1)p(1-p) \simeq \langle k \rangle$$

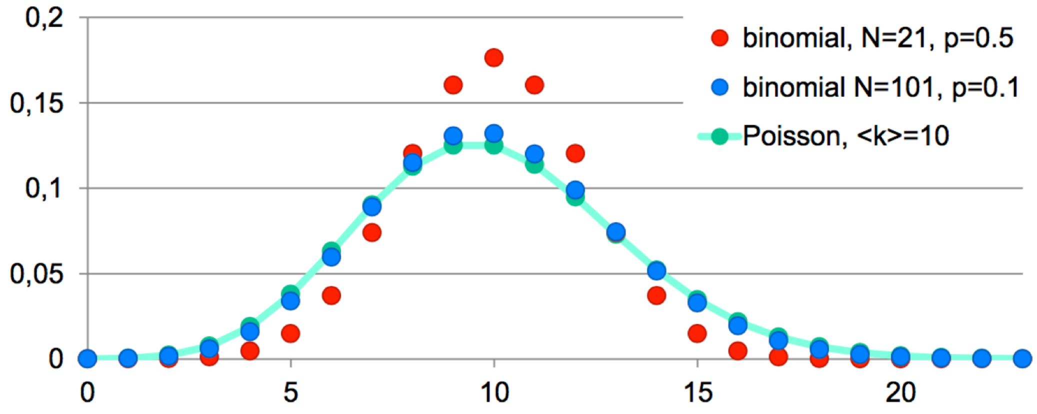
  tight around the mean

Poisson approximation

why random networks are called Poisson networks

- Poisson distribution (easier to use)

$$P\left[x=k\right]=\frac{m_x^k}{k!}\cdot e^{-m_x}$$



- Very good approximation of binomial for small *p* (and at small *k*)

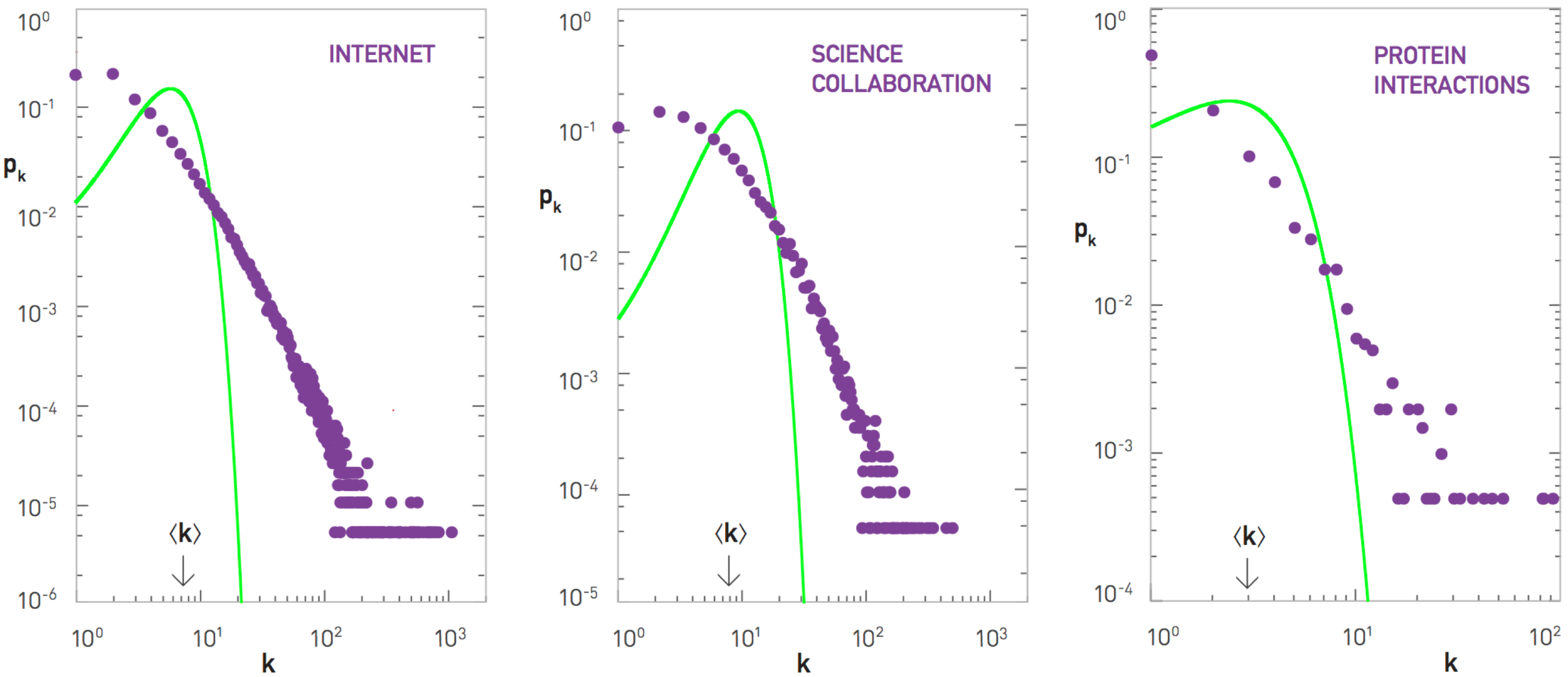
active part

p

$$P\left[x=k\right]=\underbrace{\frac{(n-k+1)\ldots(n-1)n}{n^k}}_{\simeq 1}\cdot\frac{m_x^k}{k!}\cdot\underbrace{\left(1-\frac{m_x}{n}\right)^{n-k}}_{\simeq \text{const}}$$

Are real networks Poisson?

no, they aren't

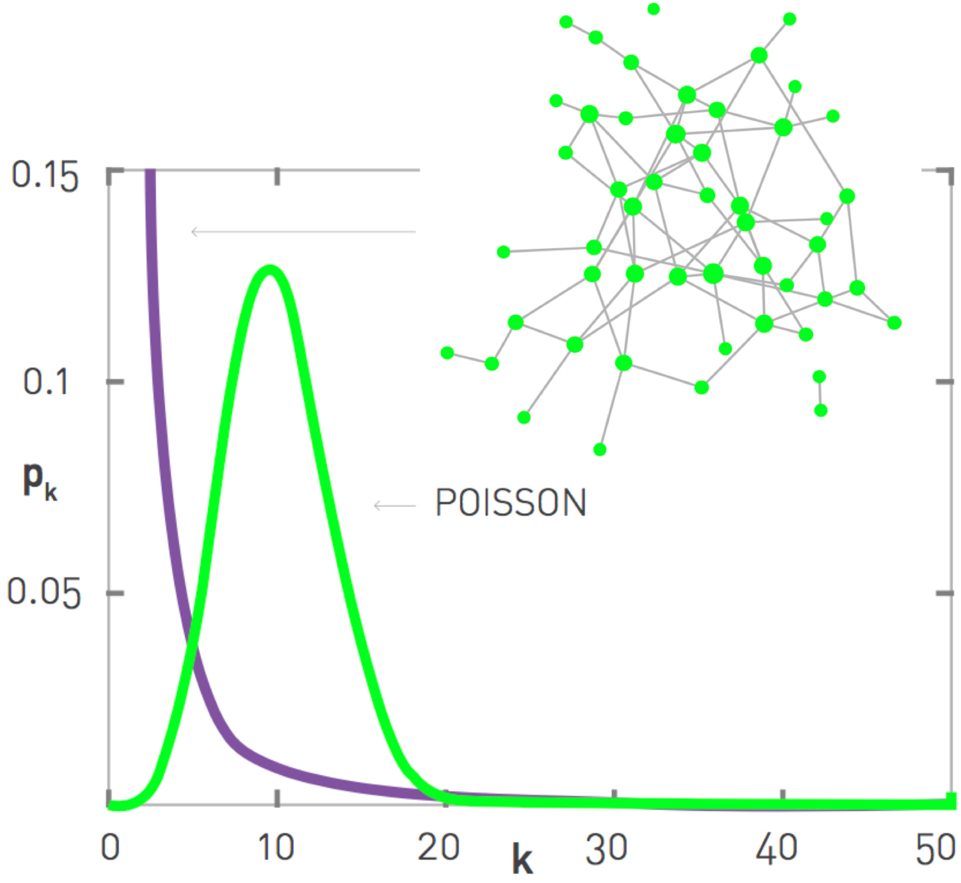


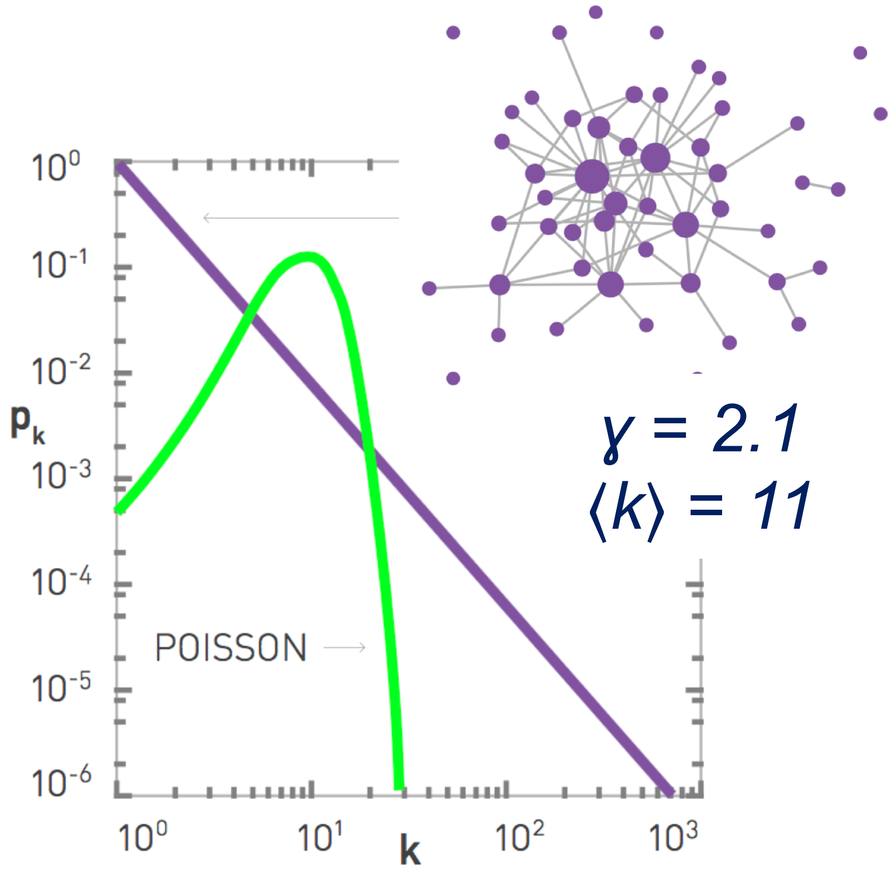
No! Poisson networks are deprived of hubs

… but, nevertheless, Poisson networks capture some aspects

Poisson versus power law

a comparison





Power-law is heavy tailed (presence of hubs) - like Weibull, lognormal, Lévy

Università degli Studi di Padova

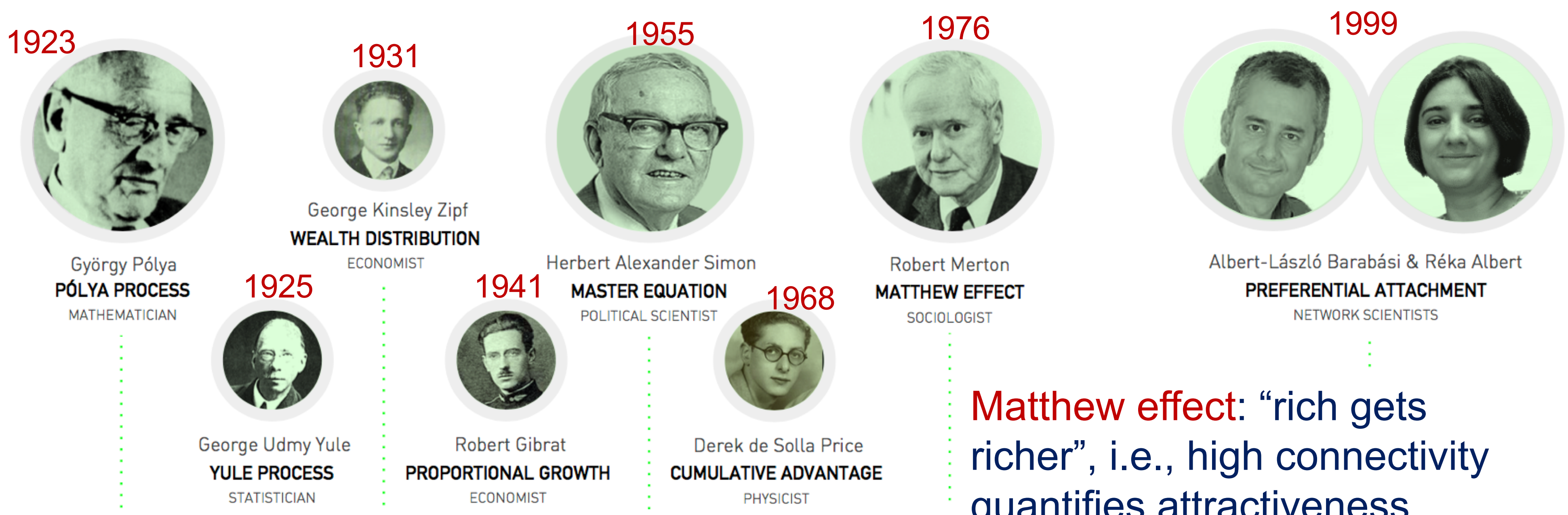
Preferential attachment

a simple concept that (partially) explains the power-law

Nodes link to the more connected nodes

e.g., think of www

This idea has a long history



Matthew effect: “rich gets richer”, i.e., high connectivity quantifies attractiveness

The Barabasi-Albert model

Barabási, Albert. "Emergence of scaling in random networks" (1999)

Start with m_0 nodes arbitrarily connected, with $\langle k \rangle = m$

- Growth

  add a node (the Nth) with m links that connect the node to nodes in the network

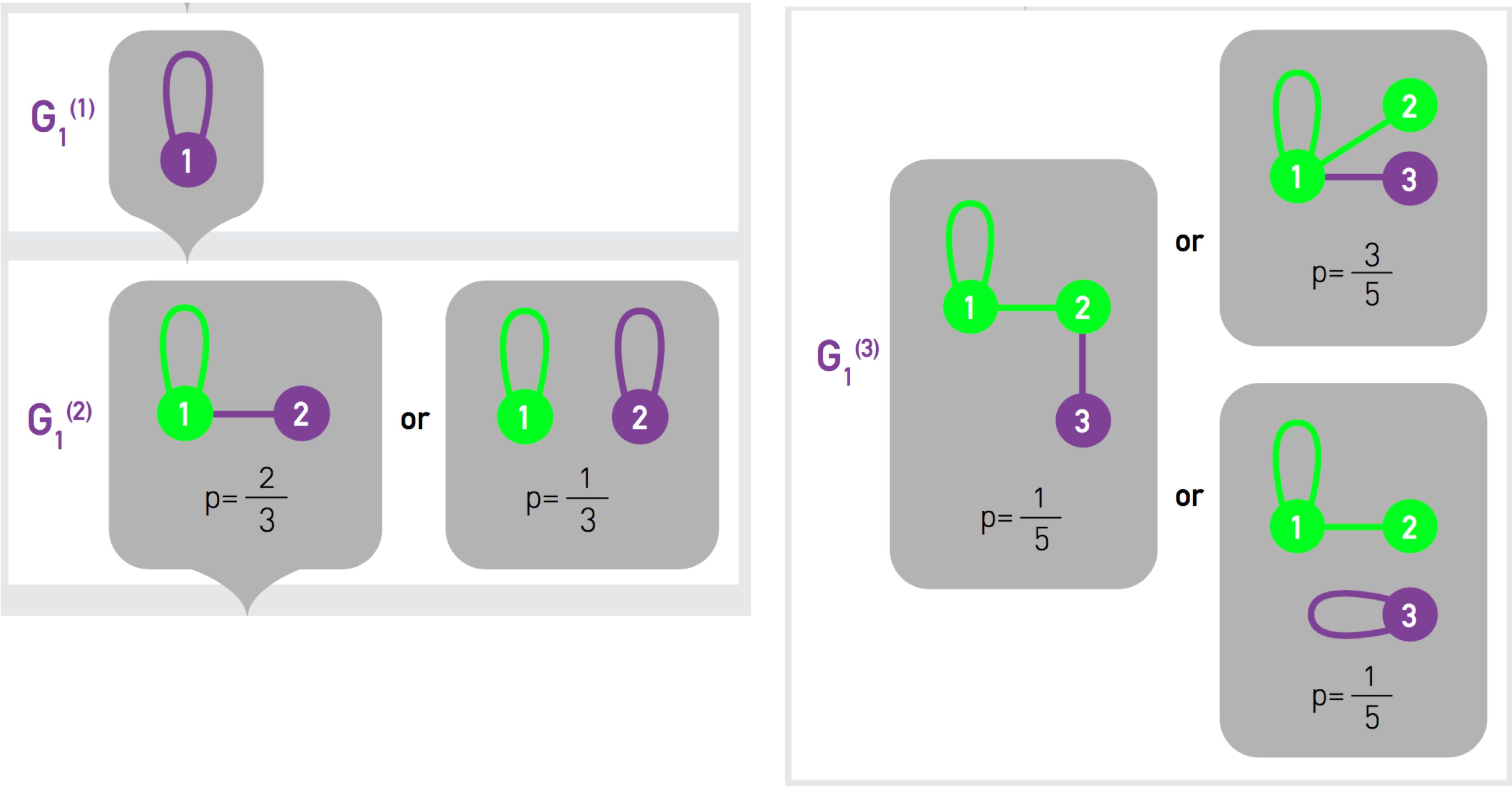
- Preferential attachment

  $p_i = k_i / C$ probability of connecting to node i

  $p_i = 1/C$ for self-loops

  $$C = 1 + \sum k_i = 1 + 2(N-1)m$$

An example
with m=1



Approximate analysis
evolution of nodes degree

- Increase in the degree (at each step)

$$\Delta k_i \simeq \underset{\uparrow \text{ trials}}{m} \cdot \underset{\uparrow \text{ probability per trial}}{k_i / (1+2m(N\text{-}1))} \simeq k_i / 2N$$

- Approximation in the continuous domain

$$\Delta k_i \simeq \mathrm{d}k_i/\mathrm{d}N \;\rightarrow\; \boxed{\mathrm{d}k_i/k_i \simeq ½\, \mathrm{d}N/N}$$

- Integration

$$\ln(k_i) = ½ \ln(N) + \text{cost.} \rightarrow\; k_i = c\, N^{½}$$

- Recalling that node *i* joins the network at time $N = i$

$$k_i\,(N{=}i) = m \;\rightarrow\; \boxed{k_i\,(N) = m\,(\mathrm{N/i})^{½}}$$

½ is the dynamic exponent

Approximate analysis

degree distribution

- Recall $k_i = m\,(N/i)^{½}$
- The number of nodes with degree smaller than k is

$$k_i < k \rightarrow m\,(N/i)^{½} < k$$

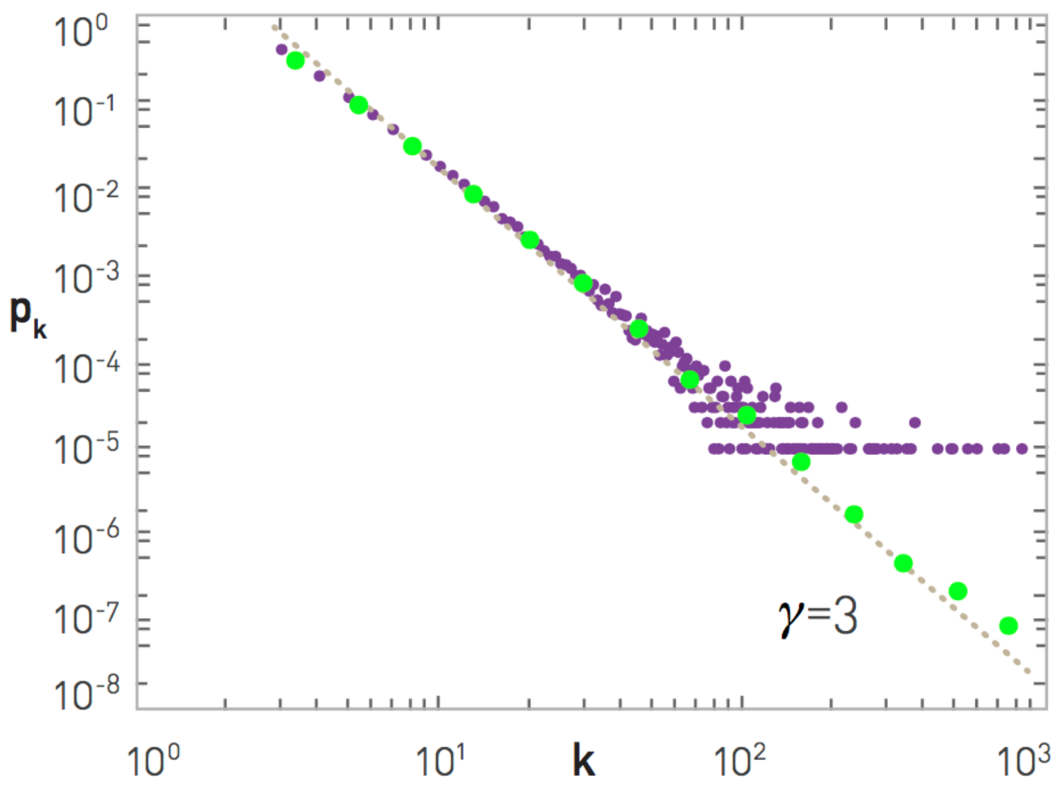
$$\rightarrow i > N\,(m/k)^2 \rightarrow \boxed{N - N\,(m/k)^2}$$

- CDF is $P_k = P[k_i \leq k] = 1 - (m/k)^2$
- The degree distribution is

$$dP_k/dk = \boxed{p_k = 2\,m^2 / k^3}$$

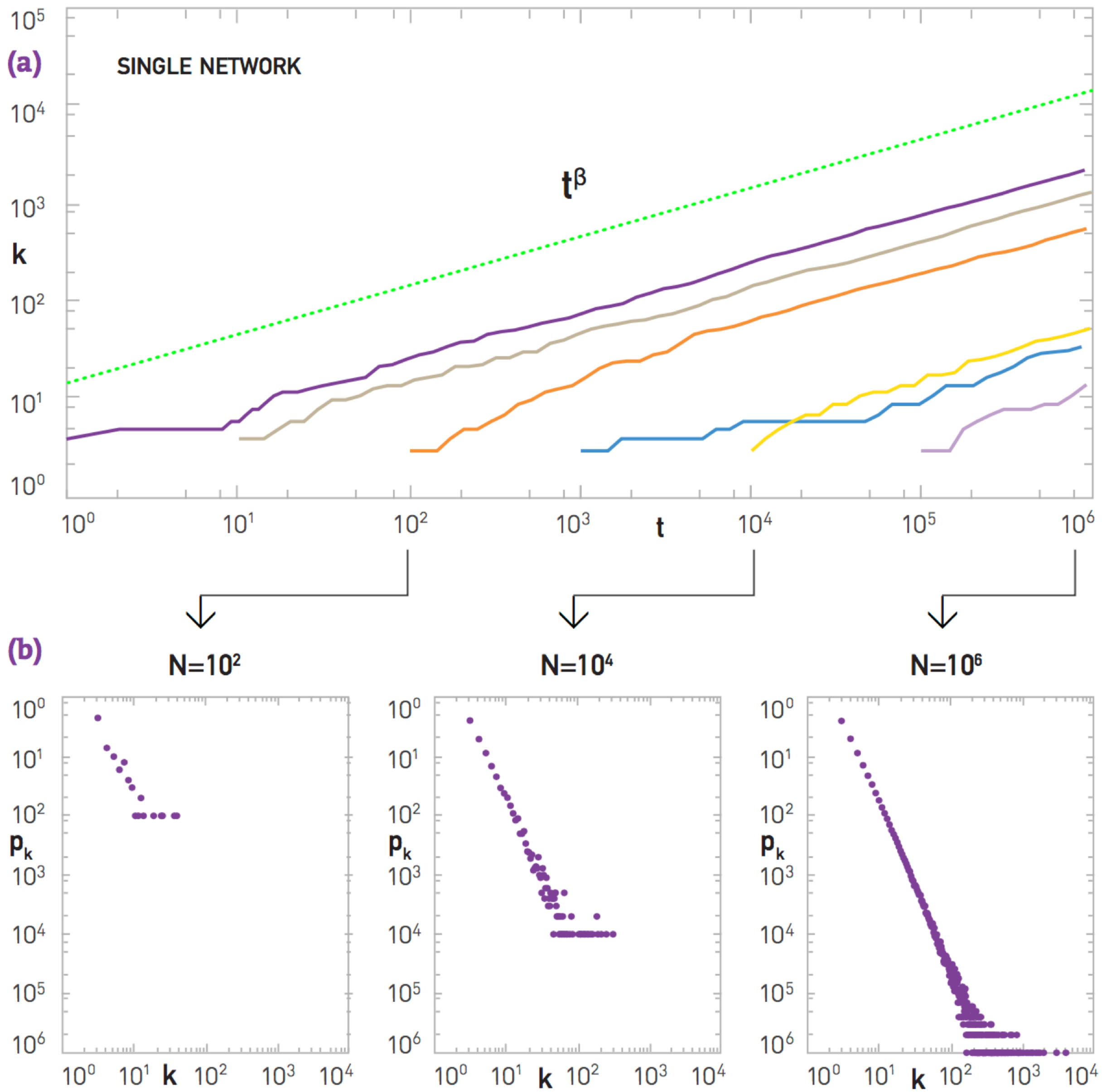
The Barabasi-Albert model
wrap up

- Depending on the implementation there might be self/multiple links
- Most nodes have a small degree (exactly *m* for the youngest ones)
- Hubs appear
- The average degree is $\langle k \rangle = 2m$, and in fact $L = Nm = \frac{1}{2}\langle k \rangle N$
- The resulting degree distribution is always a power-law with exponent $\gamma = 3$



The Barabasi-Albert model

consequence of $k_i = m\ (N/i)^{½}$



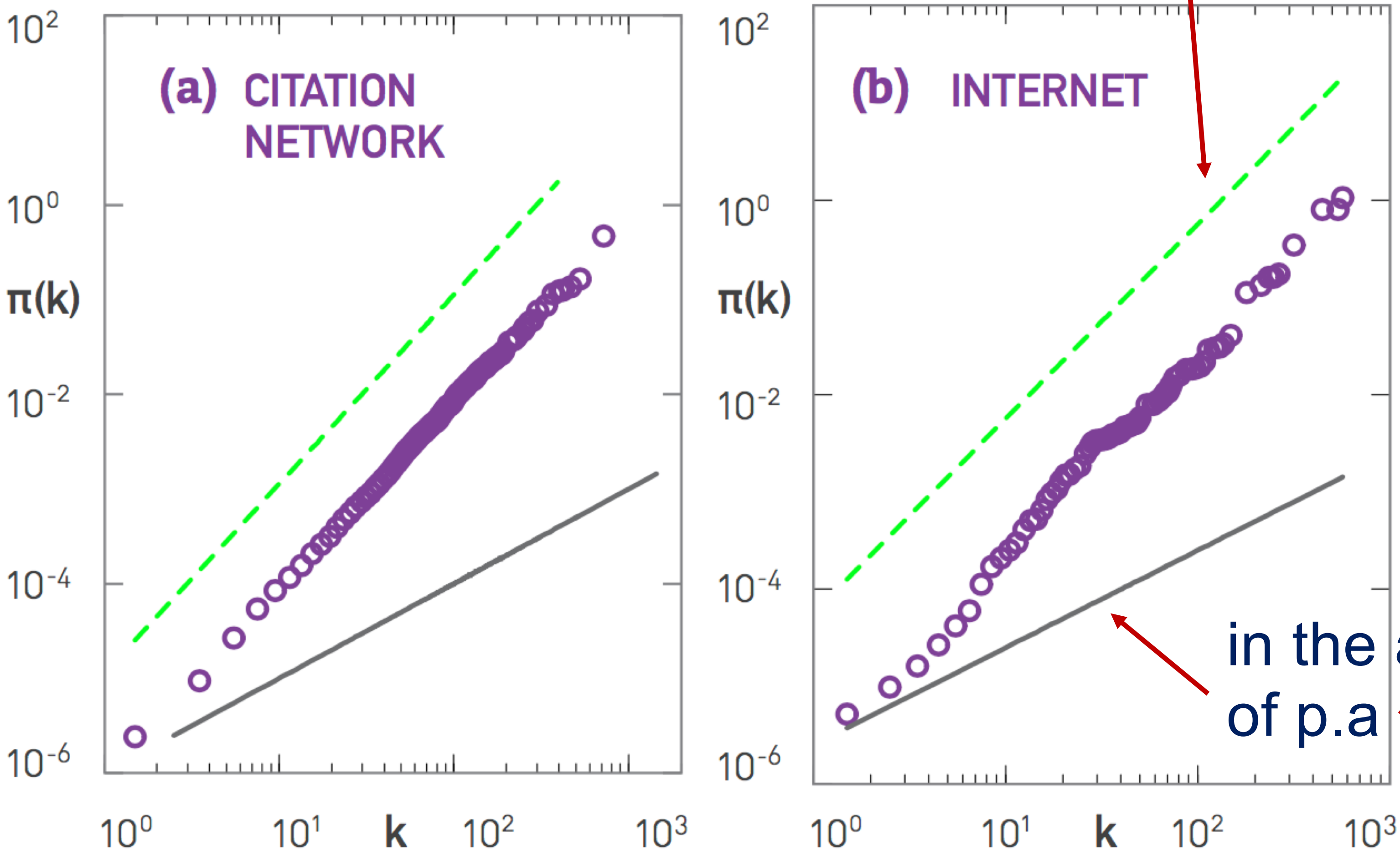
- all nodes follow the same dynamics
- the growth is sub-linear: nodes are competing with the others
- the earlier the node is added, the higher the degree – “first-mover advantage”
- older nodes acquire more links
- this explains the hub formation

Measuring preferential attachment

in real networks

$$\pi(k) = \sum_{k_i}^{k} \Delta k_i / \Delta N$$

under preferential attachment $\pi(k) \sim k^2$



in the absence of p.a $\pi(k) \sim k$

The Bianconi-Barabasi model

Bianconi, Barabási. "Competition and multiscaling in evolving networks" (2001)

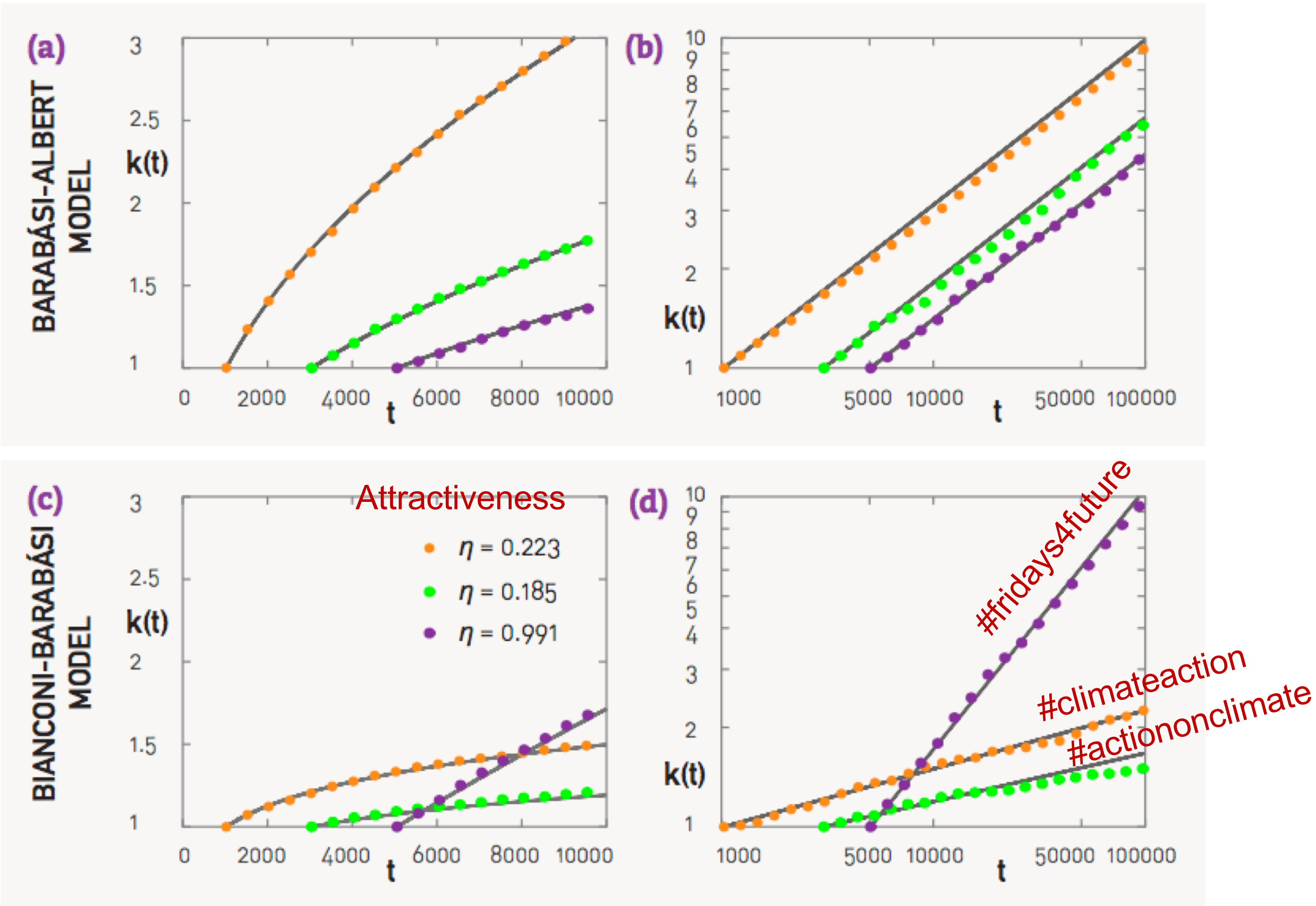
The model:

- Growth – at time step N a new node i=N is added with m links and fitness η_i
- Attractiveness (or fitness) is a random number drawn from a given distribution $\rho(\eta)$ - a quality of the individual to attract links
- Preferential attachment - probability of linking to node i is proportional to both the degree and the attractiveness, i.e., $p_i = k_i\eta_i / \sum k_j\eta_j$

An example

properties of the Bianconi-Barabasi model

we guess $k_i \simeq m\ (N/i)^{\beta(\eta_i)}$ for some $\beta(\eta)$



Approximate analysis
starting point

- We guess $\boxed{k_i \simeq m\,(N/i)^{\beta(\eta_i)}}$

- Increase in the degree $\Delta k_i \simeq \underset{\text{trials}}{m} \cdot \underset{\text{probability per trial}}{k_i\,\eta_i / \sum k_j\,\eta_j}$
- We show that $\sum k_j\,\eta_j \simeq m\,N \cdot C$ (see proof)

Approximate analysis

the denominator

- Analysis of denominator $\sum k_i \eta_i$
  - → average value wrt η
  - → hypothesis $k_i \simeq m\,(N/i)^{\beta(\eta_i)}$
- $A = E[\sum_i k_i \eta_i] = \sum E[k_i \eta_i] \simeq \int_1^N E[k_i \eta_i]\, di$
- $E[k_i \eta_i] = \int m(N/i)^{\beta(\eta)}\, \eta \cdot \rho(\eta)\, d\eta$
- Swap integrals

$$A \simeq \int m\, N^{\beta(\eta)} \left[\int_1^N i^{-\beta(\eta)}\, di \right] \eta \cdot \rho(\eta)\, d\eta$$

- Integrate

constant C

$$A \simeq m\, N \cdot \boxed{\int \frac{(1 - \cancel{N^{\beta(\eta)-1}})}{1-\beta(\eta)}\, \eta\, \rho(\eta)\, d\eta}$$

negligible for large N if $0 < \beta < 1$

Approximate analysis

evolution of nodes degrees

- We guess $\boxed{k_i \simeq m\,(N/i)^{\beta(\eta_i)}}$
- Increase in the degree $\Delta k_i \simeq m \cdot k_i\,\eta_i / \sum k_j\,\eta_j$
- It is $\sum k_j\,\eta_j \simeq m\,N \cdot C$

Hence:

1. By inspection of the above

$$\Delta k_i \simeq m\,(N/i)^{\beta(\eta_i)}\,\eta_i / N\,C$$

2. By continuum theory

$$\Delta k_i \simeq \mathrm{d}k_i/\mathrm{d}N \simeq m\,\beta(\eta_i)\,N^{\beta(\eta_i)-1}\,i^{-\beta(\eta_i)}$$

3. By combining the results $\boxed{\beta(\eta_i) \simeq \eta_i/C}$

We conclude $\boxed{k_i \simeq m\,(N/i)^{\eta_i/C}}$

Approximate analysis
constant C

$$\beta(\eta) \simeq \eta / C$$

$$C = \int \frac{\eta\, \rho(\eta)\, d\eta}{1-\beta(\eta)} \quad \rightarrow \quad \boxed{1 = \int_0^{\eta_{max}} (C-\eta)^{-1}\, \eta\, \rho(\eta)\, d\eta}$$

this identifies C for a given $\rho(\eta)$

it is $\boxed{C > \eta_{max}}$, i.e., $\beta < 1$, → the integral makes sense

growth with exponent <1

it also is $C \leq 2\eta_{max}$

Approximate analysis
degree distribution

Want to identify P_k = P[$k_i \leq k$] = 1 – P[$k_i > k$]

- $k_i > k$ and $k_i = m\,(N/i)^{\eta i/C}$ $\rightarrow$ $i < N\,(m/k)^{C/\eta i}$
- Hence P[$k_i > k|\eta_\mathrm{i}$] = $(m/k)^{C/\eta i}$
- and P[$k_i \leq k|\eta_\mathrm{i}$] = 1 - $(m/k)^{C/\eta i}$
- We have $P_k = 1 - \int (m/k)^{C/\eta}\, \rho(\eta)\, \mathrm{d}\eta$

The degree distribution is

$$p_k = P_k' = \boxed{C \int_0^{\eta_{\max}} k^{-(C/\eta+1)}\, m^{C/\eta}\, \eta^{-1}\, \rho(\eta)\, \mathrm{d}\eta}$$

weighted combination of power laws with exponent in [2,∞) since $\eta_{\max} < C$

What if $\rho(\eta) = \delta(\eta\text{-}1)$?

- Coefficient $C = 2$ since

$$\int_0^{\eta_{max}} (C/\eta - 1)^{-1}\, \delta(\eta\text{-}1)\, d\eta = (C - 1)^{-1} = 1$$

- Exponential degree $k_i \simeq m\, (N/i)^{½}$

Degree distribution

$$p_k = C \int_0^{\eta_{max}} \eta^{-1}\, m^{C/\eta}\, k^{-(C/\eta+1)}\, \delta(\eta\text{-}1)\, d\eta = 2\, m^2\, k^{-3}$$

Uniform fitness

the model

What if $\rho(\eta) = 1$ and $\eta_{max} = 1$?

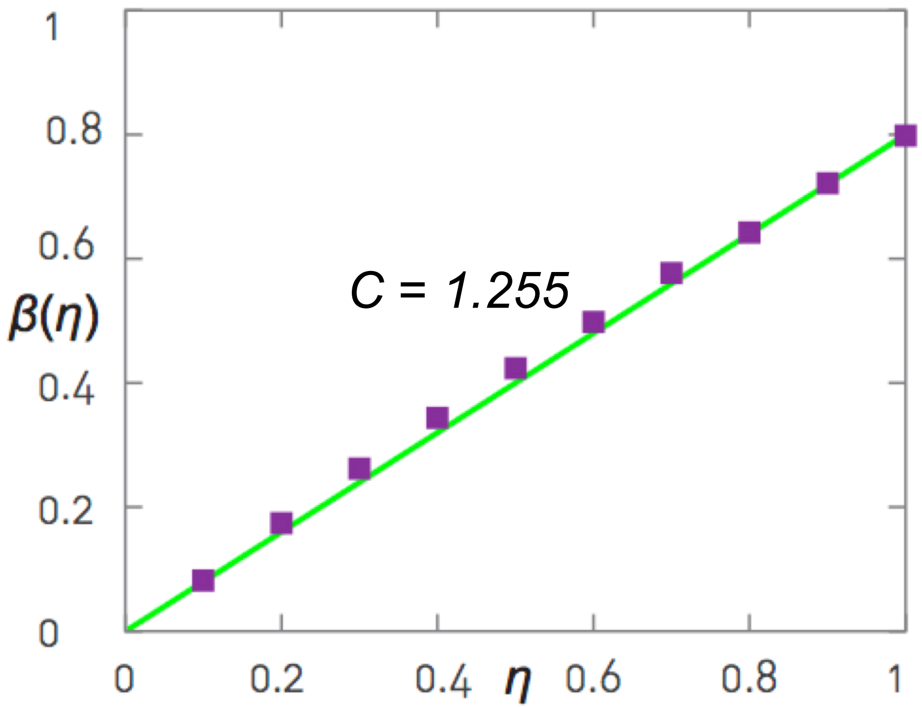
- Coefficient $C = 1.255$ since
$$\int_0^1 (C/\eta - 1)^{-1}\, d\eta = 1 \;\rightarrow\; e^{-2/C} = 1-1/C$$
- Exponential degree $k_i \simeq m\,(N/i)^{\eta_i/C}$
- Each node has its own dynamic exponent !!!

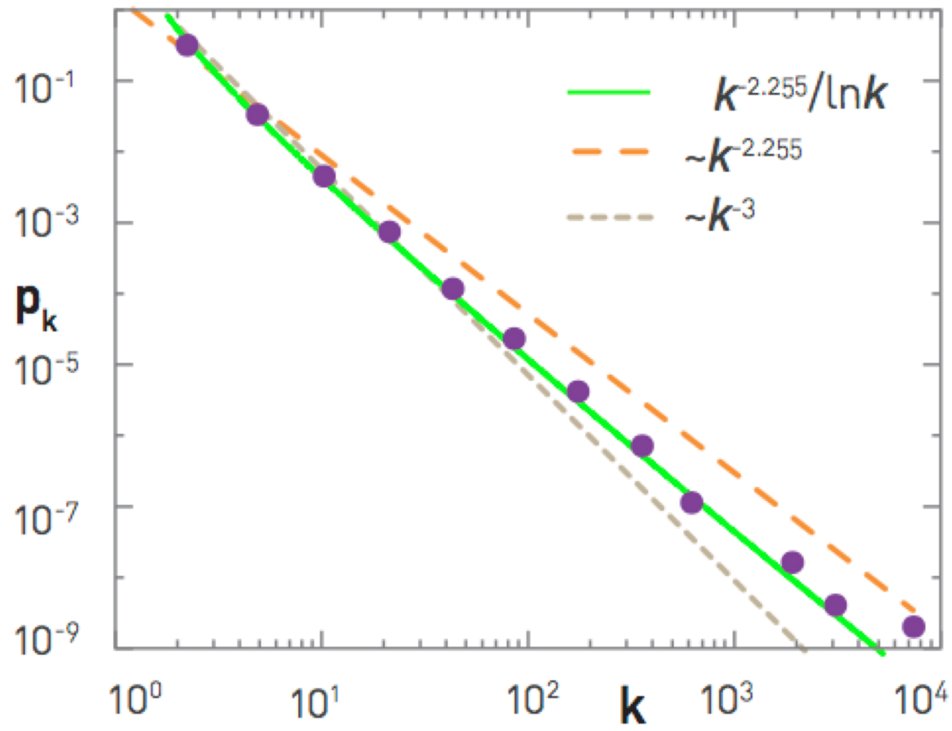
Degree distribution

$$p_k = C/k \int_0^1 \eta^{-1}\, e^{-C\ln(k/m)/\eta}\, d\eta \sim \boxed{k^{-(1+C)} / \ln(k)}$$

$e^{-b} - b\,E_1(b)$, $b = C\ln(k/m)$
exponential integral E_1

Uniform fitness
the measured data





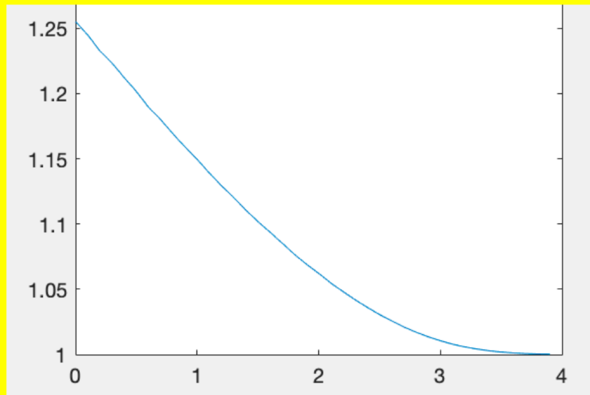
degree distribution $p_k \sim k^{-2.255} / \ln(k)$

corrective term

What if $\rho(\eta) = a\, e^{-a\eta} / (1-e^{-a})$ and $\eta_{max} = 1$?

- C rapidly converges to C=1

$$\int_0^1 (C/\eta - 1)^{-1}\, \rho(\eta)\, d\eta = 1$$



- Exponential degree $k_i \simeq m\, (N/i)^{\eta_i/C}$
- Each node has its own dynamic exponent !!!

Degree distribution

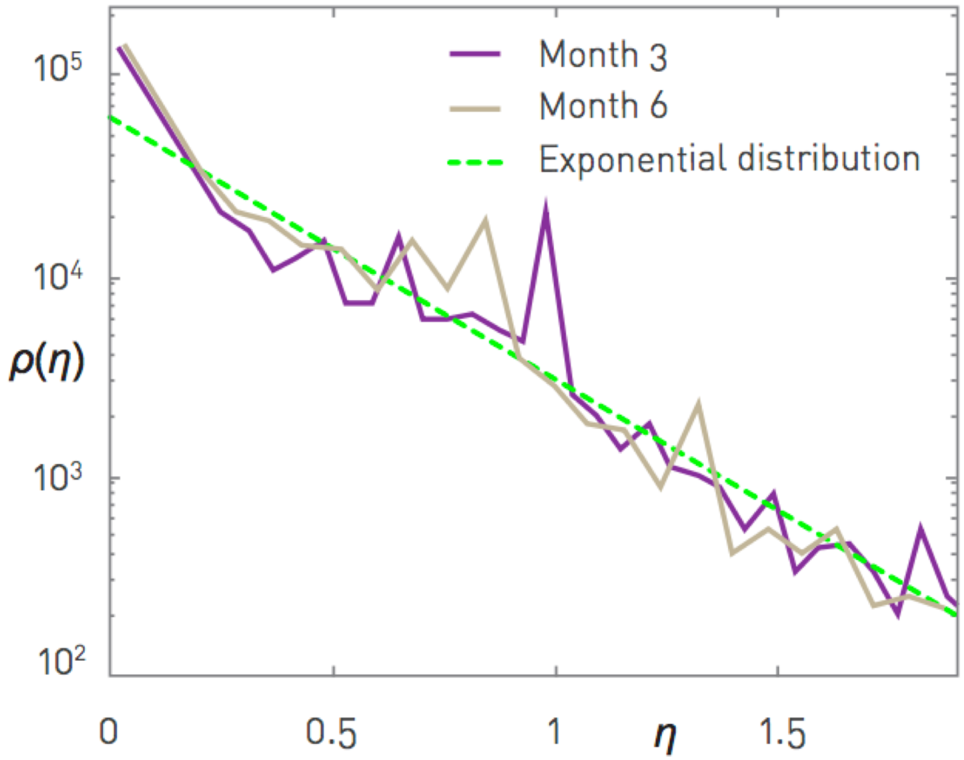
$$p_k = C/k \int_0^1 \eta^{-1}\, e^{-C \ln(k/m)/\eta}\, \rho(\eta)\, d\eta \sim \boxed{k^{-(1+C)} / \ln(k)}$$

exponential integral E_1

Exponential fitness

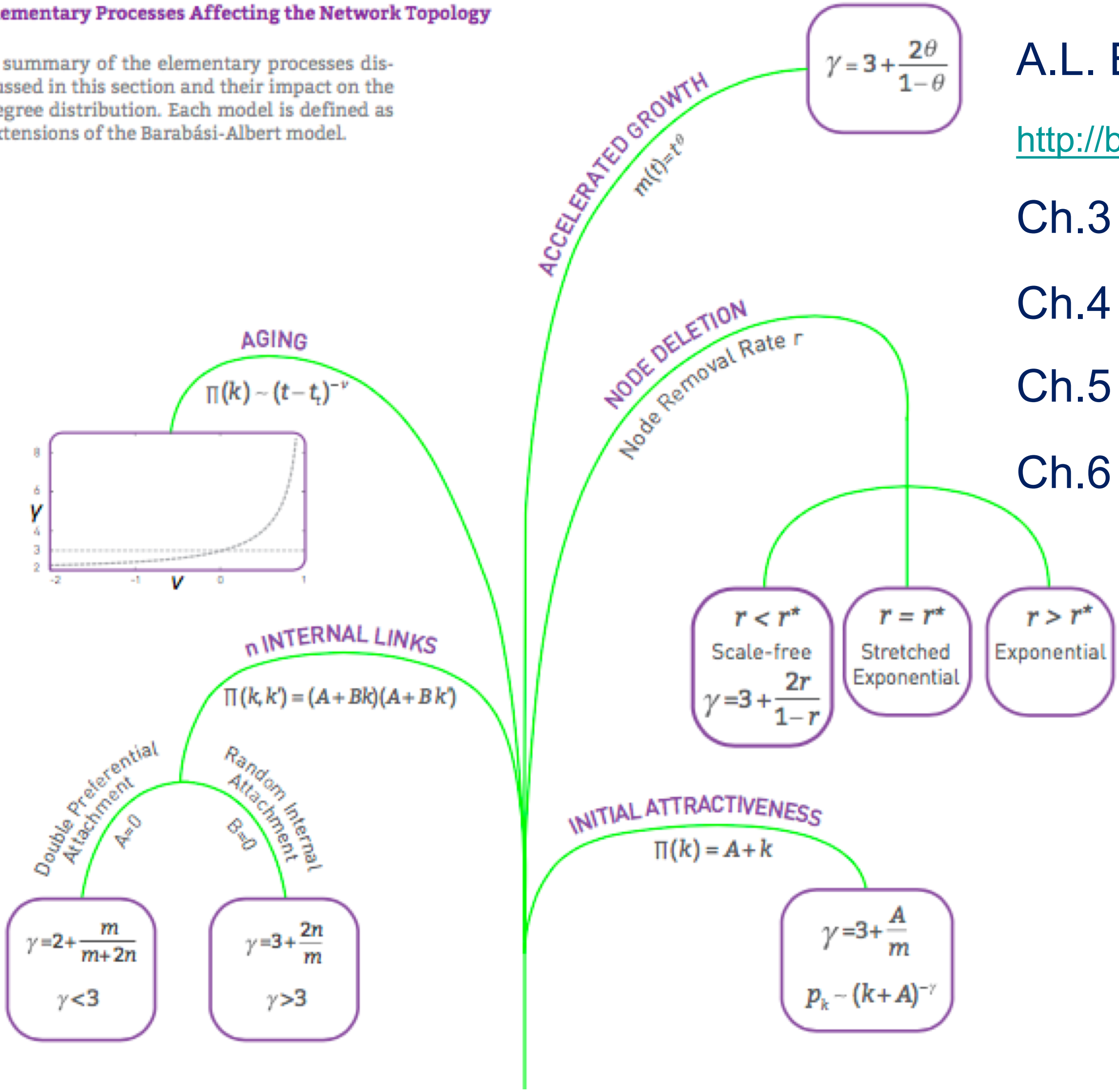
the www

$a = 4.6$
$C = 1$



degree distribution $p_k \sim \boxed{k^{-2} / \ln(k)}$

Other ideas for extension
of the Albert-Barabasi model



Elementary Processes Affecting the Network Topology

A summary of the elementary processes discussed in this section and their impact on the degree distribution. Each model is defined as extensions of the Barabási-Albert model.

A.L. Barabási, Network science

http://barabasi.com/networksciencebook

Ch.3 “Random networks”

Ch.4 “The scale-free property”

Ch.5 “The Barabási-Albert model”

Ch.6 “Evolving networks”

Properties of the power-law

scale-free and random networks

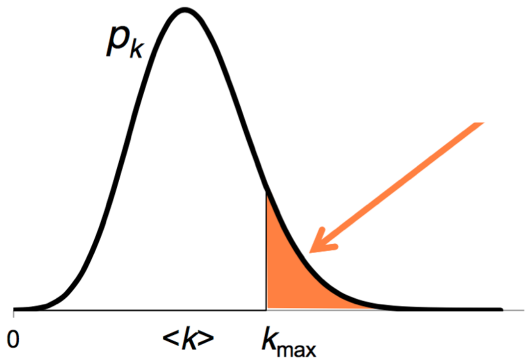
The largest hub

natural cutoff under the power-law

Degree distribution $p_k = C\, k^{-\gamma}$ with $C = (\gamma-1)\, k_{min}^{\gamma-1}$

The size of the largest hub is captured by

$$\int_{k_{max}}^{\infty} p_k\, dk = C \cdot k_{max}^{-(\gamma-1)} / (\gamma-1) = 1/N$$



$$\boxed{k_{max} = k_{min}\, N^{\,1/(\gamma-1)}}$$

is the natural cutoff
it explains large hubs

Moments
of the power-law

- $\langle k^n \rangle = \int_{k_{min}}^{k_{max}} k^n \, p_k \, dk$ with $p_k = C\, k^{-\gamma}$

$$= C \cdot (k_{max}^{n-\gamma+1} - k_{min}^{n-\gamma+1})/(n-\gamma+1)$$

$$= C\, k_{min}^{n-\gamma+1} \cdot (N^{-1+n/(\gamma-1)} - 1) / (n-\gamma+1)$$

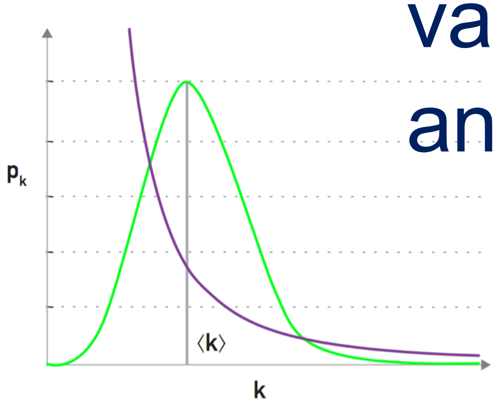
- They diverge with N if $\gamma < n+1$

  mean ($n=1$) doesn't diverge for $\gamma \geq 2$

  variance ($n=2$) diverges for $\gamma < 3$

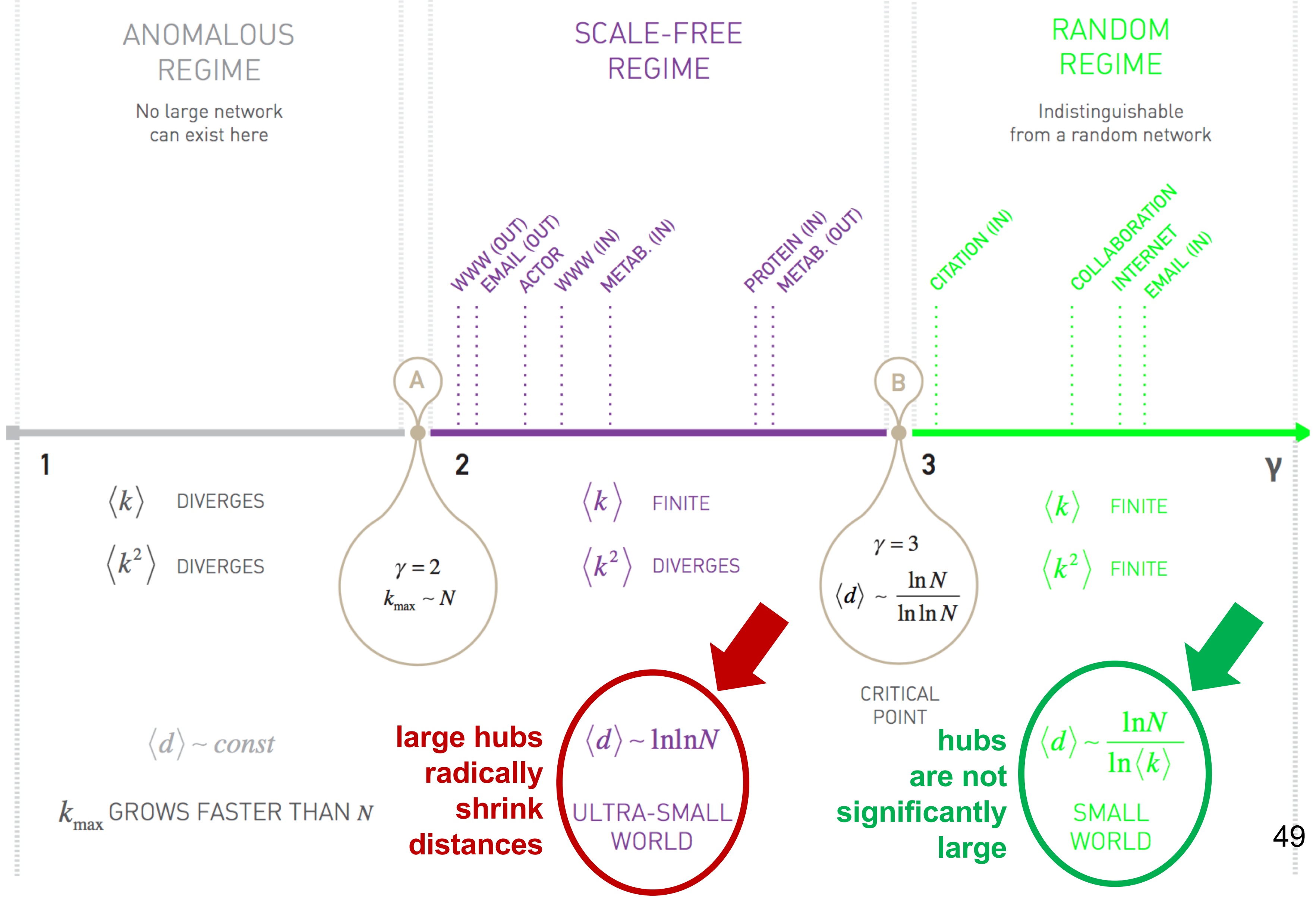
  and the network <u>does not have a scale</u>

  (scale-free regime)



The scale-free regime

for 2<γ<3

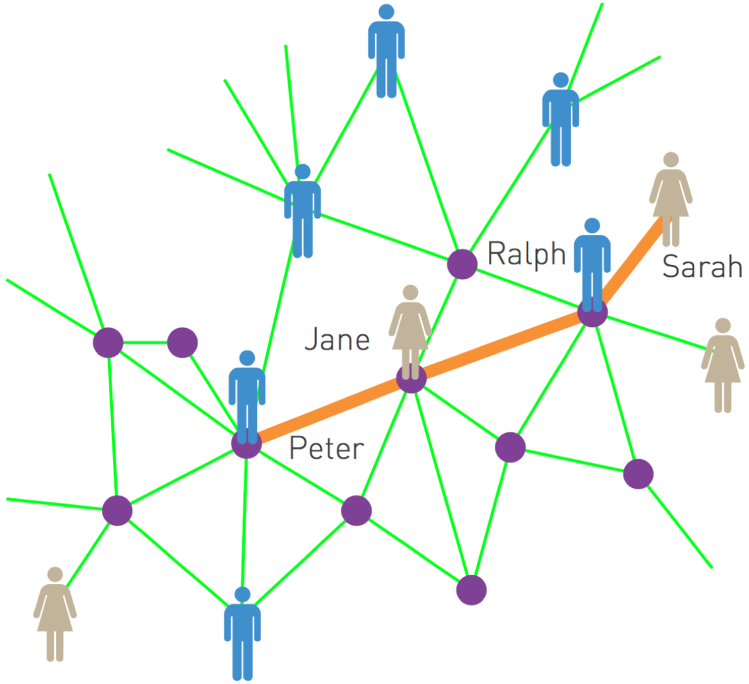


Small world property

Watts, Strogatz, «*Collective dynamics of small-world networks*», (1998)

In real networks distance between two randomly chosen nodes is generally short

Milgram [1967]: *6 degrees of separation*



What does this mean?

We are more connected than we think

Distances in random graphs

theoretical result

- we reach $\langle k \rangle$ nodes in one hop, $\langle k \rangle^2$ in two, $\langle k \rangle^3$ in three, etc.
- an estimate of the average distance $\langle d \rangle$ is found by solving for $N = \langle k \rangle^{\langle d \rangle}$ to have

$$\langle d \rangle = \ln(N) / \ln(\langle k \rangle)$$

- $\langle d \rangle$ is often taken as an estimate of the network diameter $d_{\max}$

e.g.: on earth we are $N=7 \cdot 10^9$ individuals,
with $\langle k \rangle$=1000 acquaintances each → $\langle d \rangle = 3.28$

Distances in random graphs

fitting with real data

NETWORK	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\ln N}{\ln \langle k \rangle}$
Internet	192,244	609,066	6.34	6.98	26	6.58 ✓
WWW	325,729	1,497,134	4.60	11.27	93	8.31 ✓
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.72	39	11.42 ✓
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	93,439	8.08	5.35	15	4.81 ✓
Actor Network	702,388	29,397,908	83,71	3,91	14	3,04 ✓
Citation Network	449,673	4,707,958	10.43	11,21	42	5.55
E. Coli Metabolism	1,039	5,802	5.58	2.98	8	4.04
Protein Interactions	2,018	2,930	2.90	5.61	14	7.14 ✓

Very good fit ! Correct at least as order of magnitude

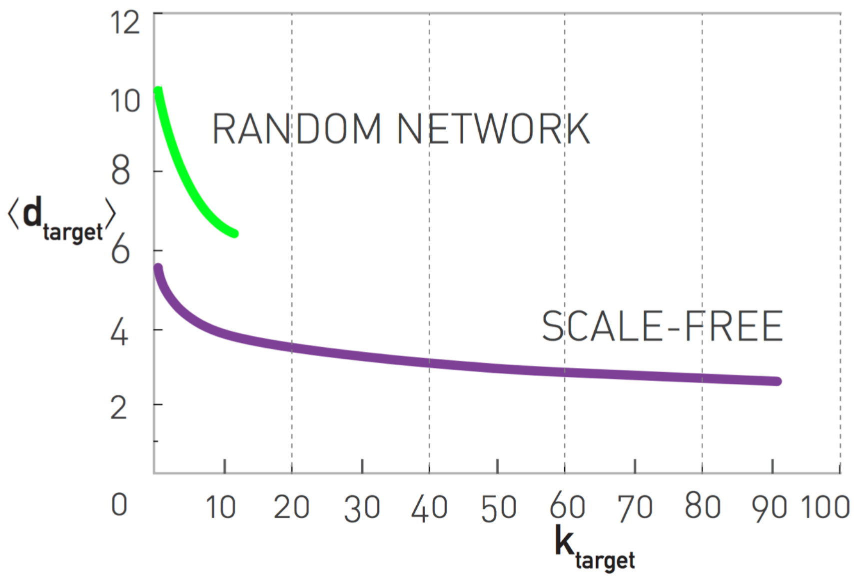
Distances in scale-free networks

the ultra-small-world

- The average distance increases as ln(ln(N)), much slower than N or ln(N)

e.g. in www $N=7\cdot 10^9$, ln(N)=22.7, ln(ln(N))=3.12 (very small)

distance of a node of degree $\langle k \rangle$ from a node of degree k_{target}



- The large hubs radically shrink the distance between nodes → ultra small world

In many social experiments people avoided hubs for entirely perceptual reasons (e.g., they assumed they are busy, better use them only if really needed)

We live in a ultra-small-world, but we perceive that we are more distant from others than we really are!

Friendship paradox

my friends are more popular than me (Feld 1991)

- Can be observed in the ultra-small-world under the presence of big hubs
- Rationale: a node is very likely to be connected to a big hub, having a very large number of connections
- # of friends (in the average) = $\langle k \rangle$
- # of friends of friends $\simeq N$

Takeaways

for degree centrality

- Do not use it for resizing nodes according to their importance (will use PageRank for this)
- Provide useful information in the form of a degree distribution
- Always plot degree distributions in the log scale
- Always evaluate their slope ɣ, but please use the ML approach: ɣ provides useful insights on the network
- Preferential attachment and attractiveness can be measured if you have temporal info on the network

PageRank centrality

Google’s approach to centrality

How to organise the web?

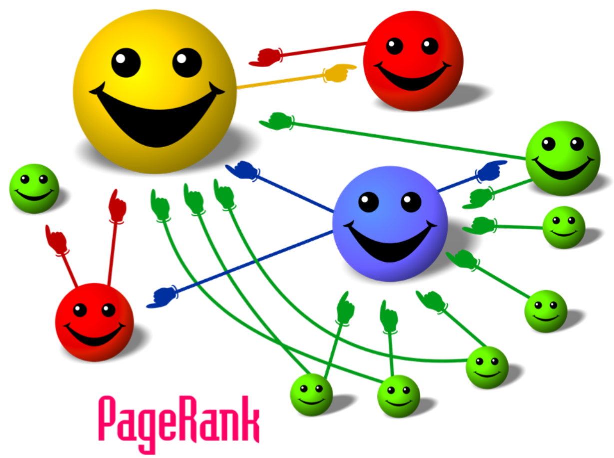
links as votes

- 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

The Google's view
quoting Google

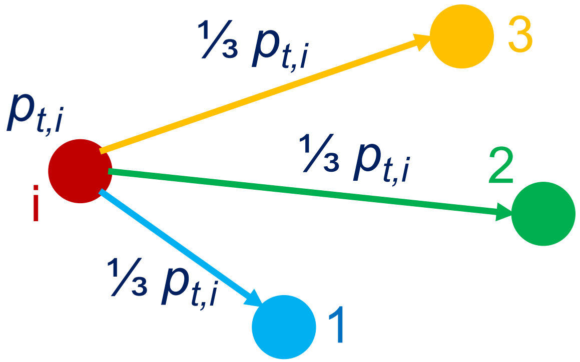
- PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is
- The underlying assumption is that more important websites are likely to receive more links from other websites



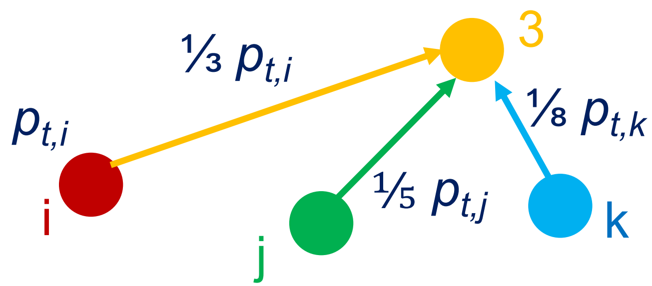
A random walk on www

the rationale behind PageRank

- 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*



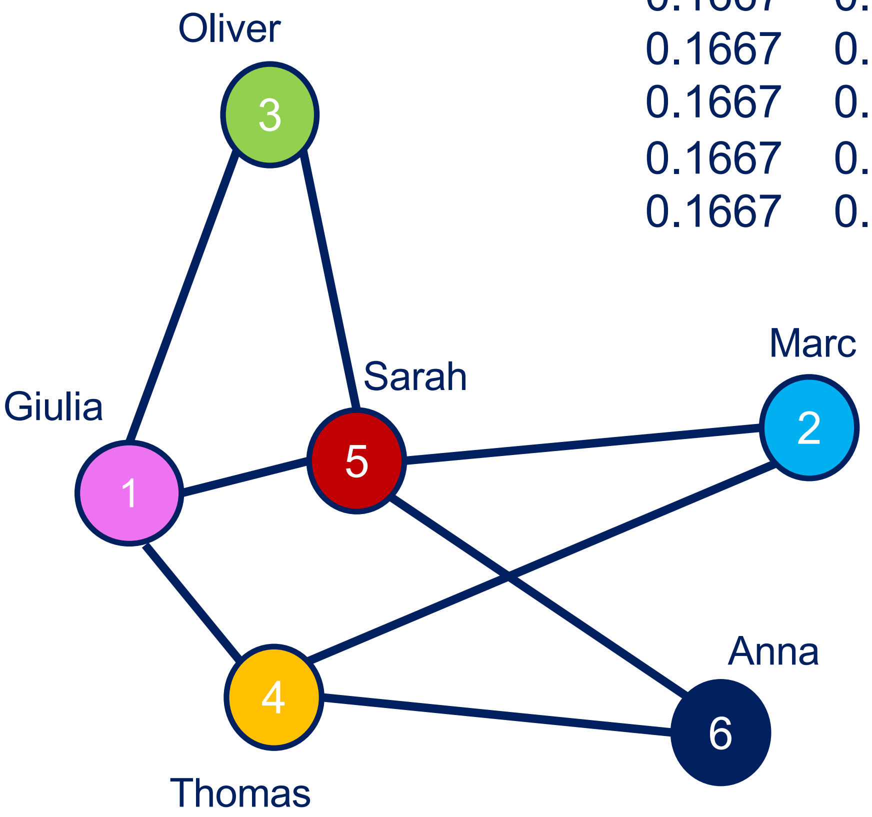
$$p_{t+1,3} = \tfrac{1}{3}\, p_{t,i} + \tfrac{1}{5}\, p_{t,j} + \tfrac{1}{8}\, p_{t,k}$$



- this identifies a Markov chain
- after a while probabilities settle to a steady state = the PageRank vector

Example

of the random walk effect on a friends' network



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

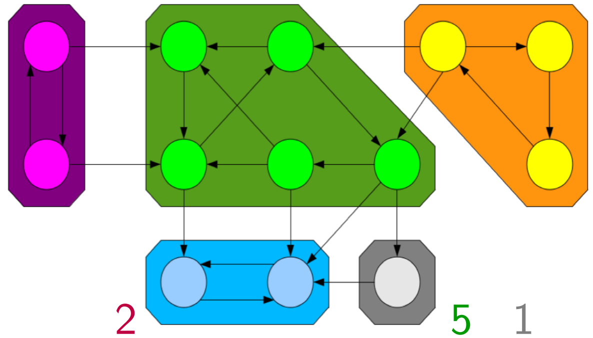
Equal to (normalized) degree centrality in undirected networks !!!

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

Matrix formalization

of the random walk

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



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

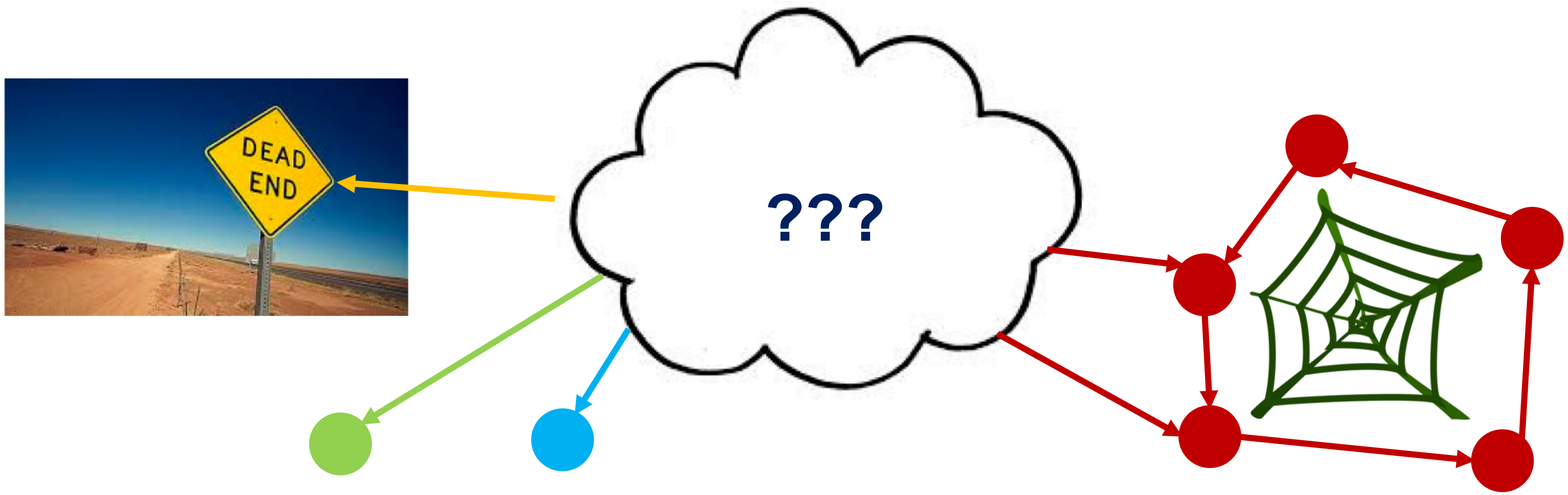
columns sum to 1

Problems in the random walk

dead ends and spider traps

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

as a method to overcome problems

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
- $\boldsymbol{p}_{t+1}$ = c $\boldsymbol{M}$ $\boldsymbol{p}_t$ + (1-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., $\boldsymbol{q}$ = 1/N for uniform probability

PageRank with restart

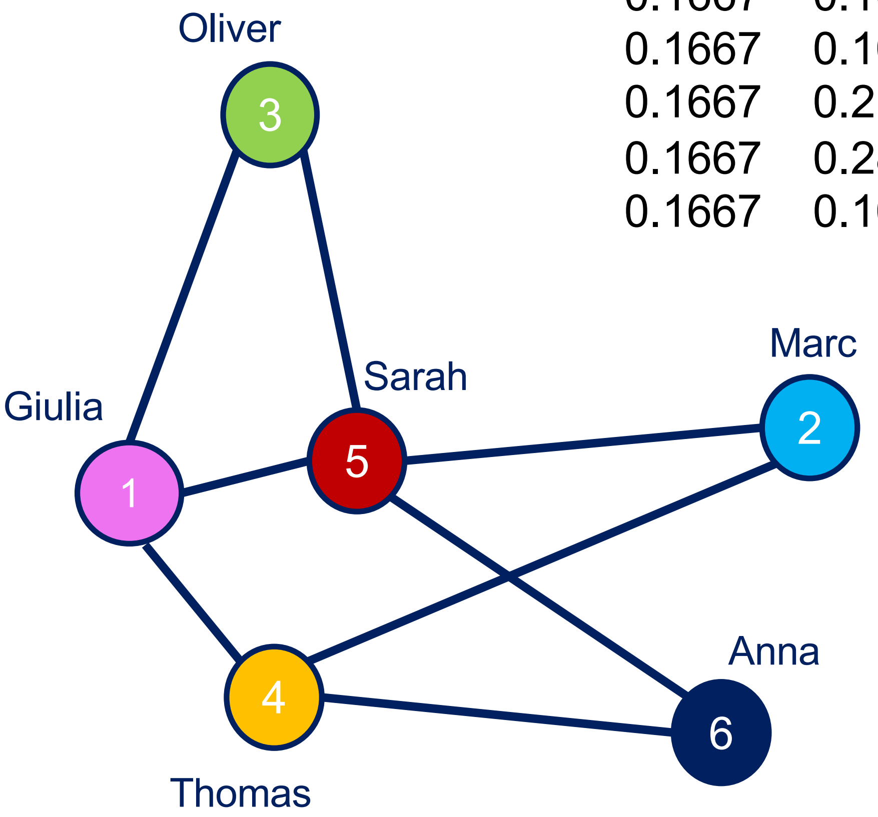
or simply PageRank

dead ends	$\boldsymbol{A}_0$	original adjacency matrix (can be fractional)
no dead ends	$\boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{b}\,\boldsymbol{e}^{\mathrm{T}}$	$\boldsymbol{b}$: teleportation; $\boldsymbol{e}$: 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	$\boldsymbol{M}_1 = c\,\boldsymbol{M} + (1-c)\,\boldsymbol{q}\,\boldsymbol{1}^{\mathrm{T}}$	equivalent formulation matrix is no more sparse
Markov chain	$\boldsymbol{p}_{t+1} = \boldsymbol{M}_1\,\boldsymbol{p}_t$	
PageRank equation	$\boldsymbol{r} = c\,\boldsymbol{M}\,\boldsymbol{r} + (1-c)\,\boldsymbol{q}, \qquad \boldsymbol{r} = \boldsymbol{p}_\infty$	PageRank centrality vector

UNIVERSITÀ DEGLI STUDI DI PADOVA

Example

of PageRank with restart on a friends' network



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

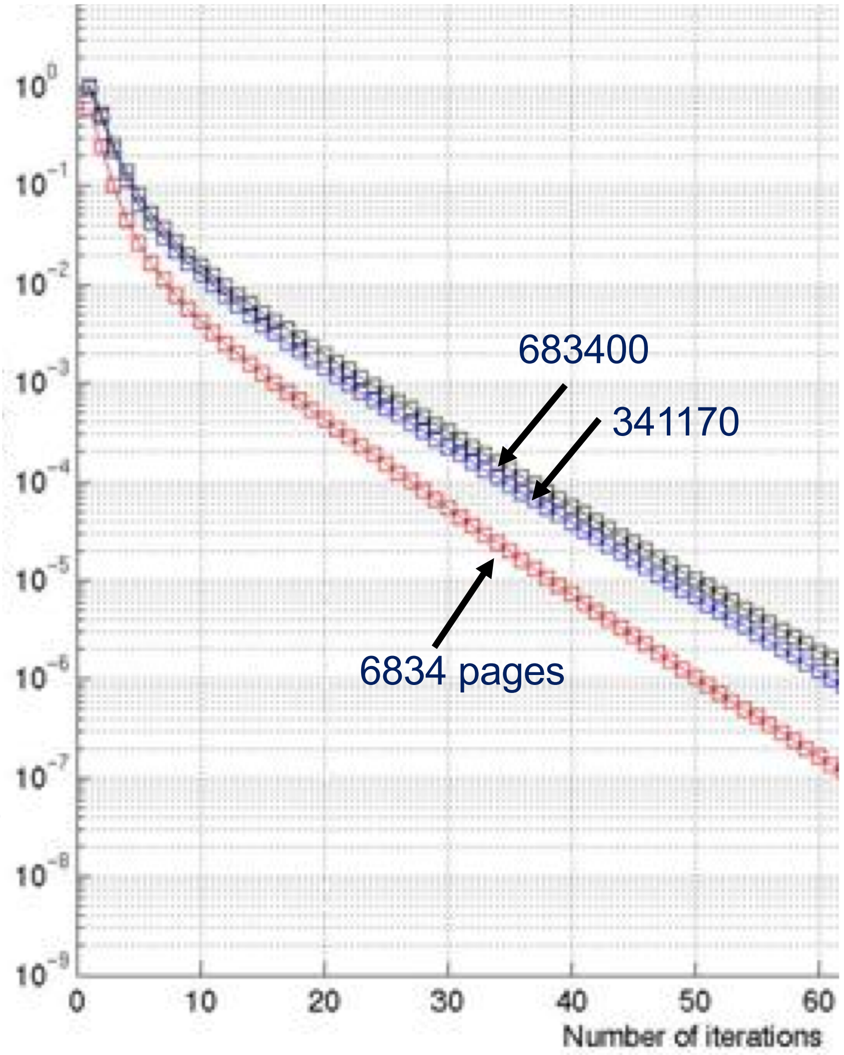
Convergence properies of PageRank

an overview

- The 

vector is the probability $\boldsymbol{p}_t$ for large t
- It corresponds to the stationary behaviour of the Markov chain

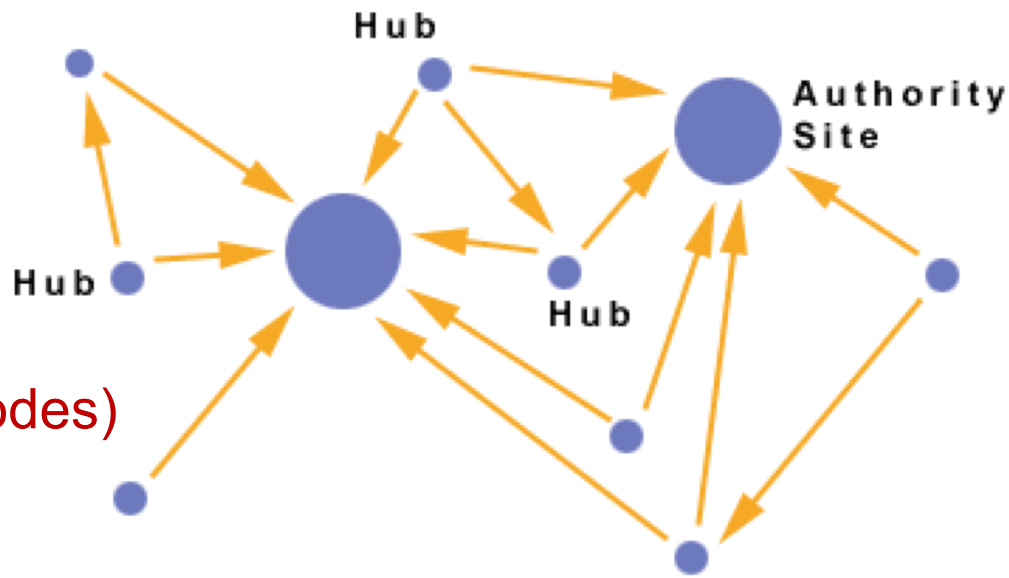
- $\boldsymbol{p}_\infty$ is unique
- $\boldsymbol{p}_\infty$ is a stochastic vector (with positive entries summing to *1*)
- $\boldsymbol{p}_\infty$ depends on the choice of the teleportation vector $\boldsymbol{q}$ (and of c)
- $\boldsymbol{p}_\infty$ converges in few iterations, typically $\boldsymbol{p}_{40} \simeq \boldsymbol{p}_\infty$



Hubs and Authorities
what we can get from PageRank

❑ Authority (quality as a content provider)

nodes that contain useful information, or having a high number of edges pointing to them (e.g., course homepages)
= PageRank vector
(related to the in-degree of nodes)



❑ Hub (quality as an expert)

trustworthy nodes, or nodes that link to many authorities (e.g., course bulletin)
= PageRank vector starting from $\mathbf{A}_0^T$
(related to the out-degree of nodes)

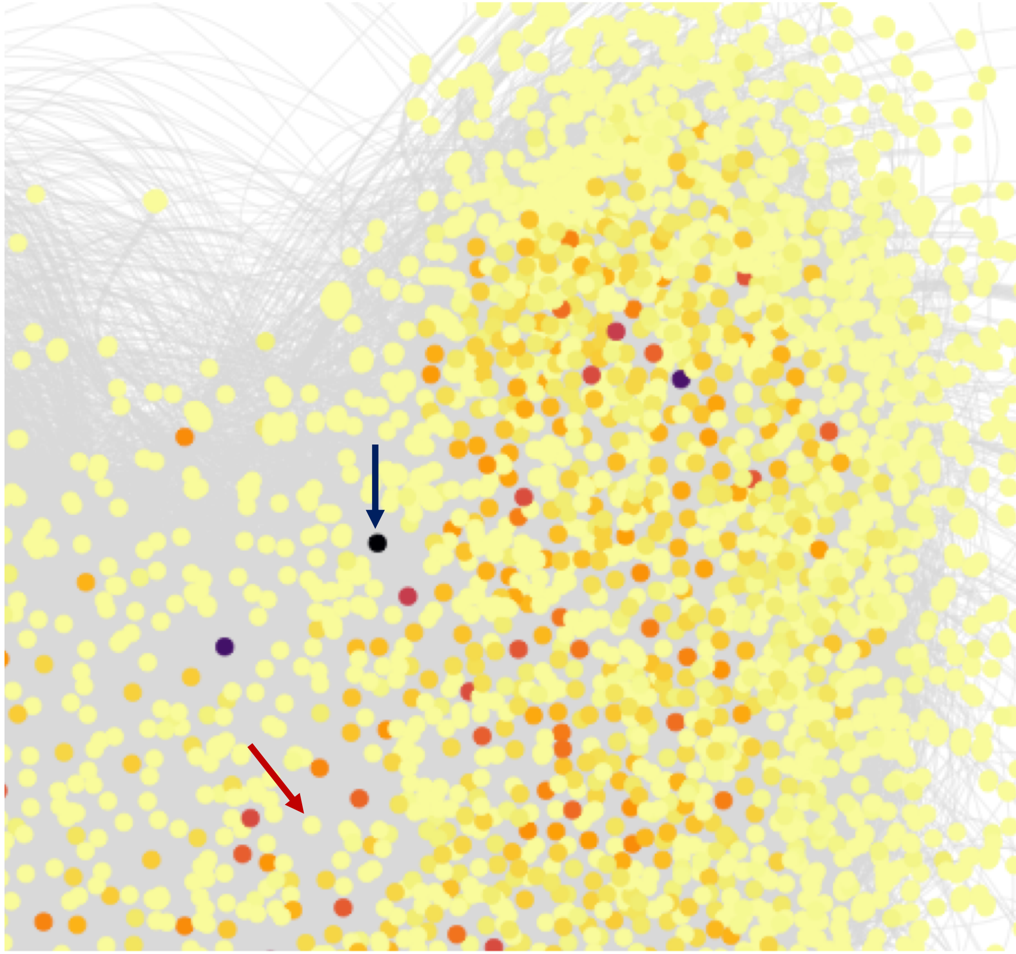
authority or hub?

Example of PageRank centrality

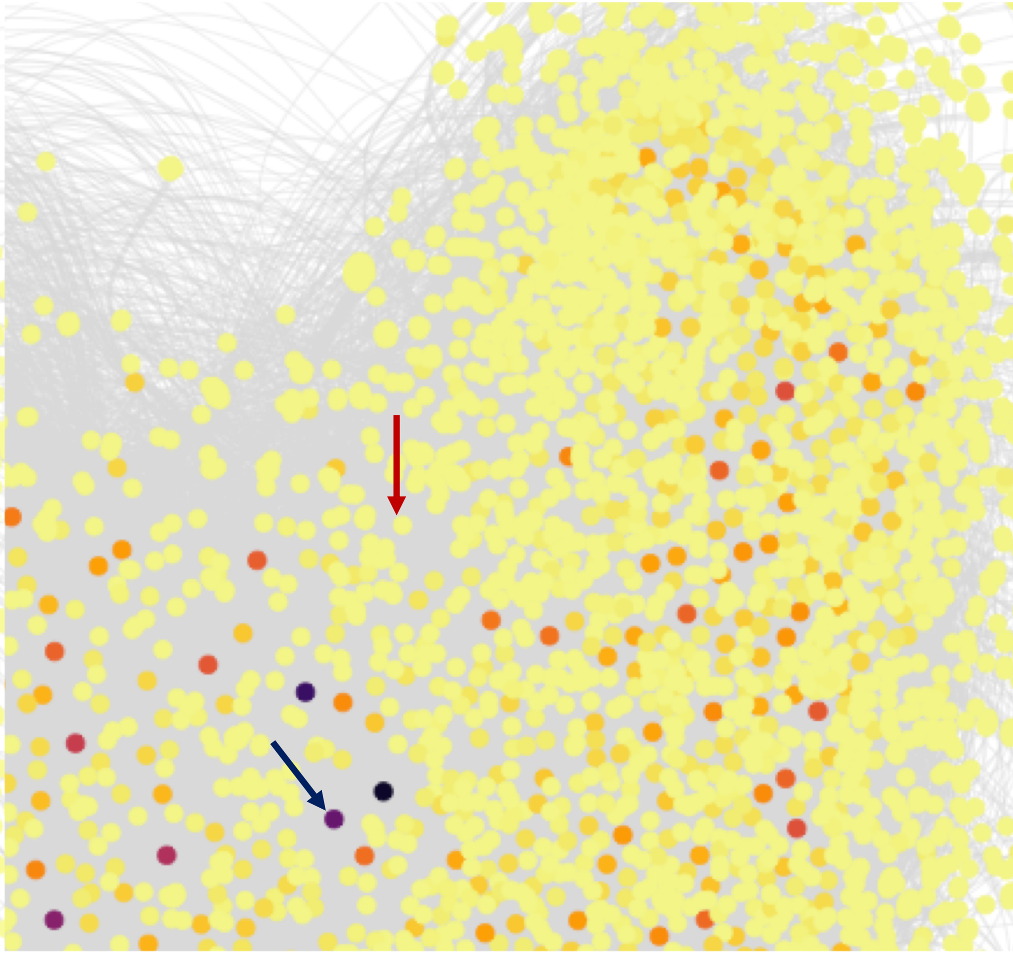
wikipedia administrator elections and vote history data

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

Authorities



Hubs

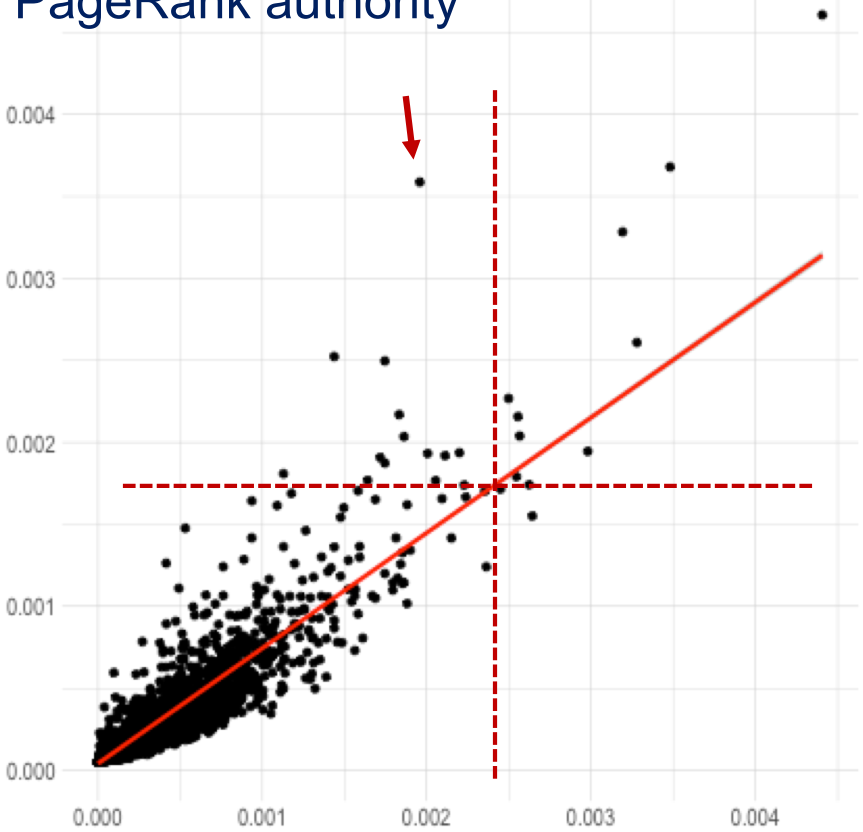


PageRank versus degree centrality

wikipedia administrator elections and vote history data

Authorities

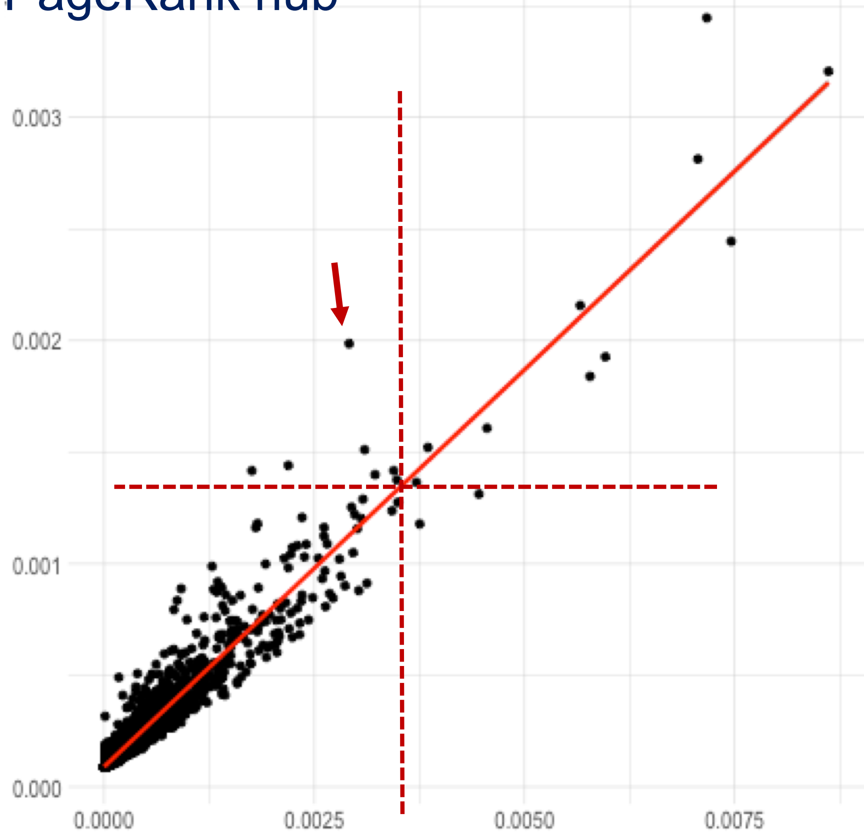
PageRank authority



in-degree

Hubs

PageRank hub

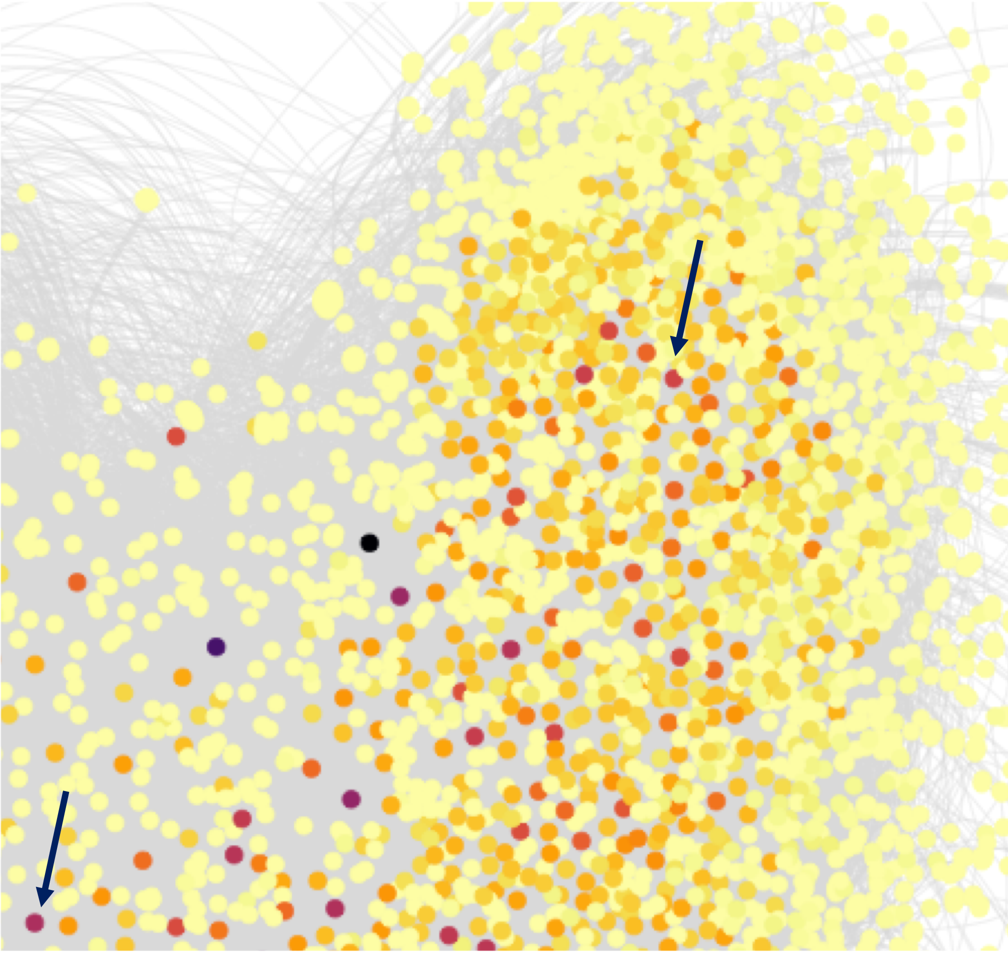


out-degree

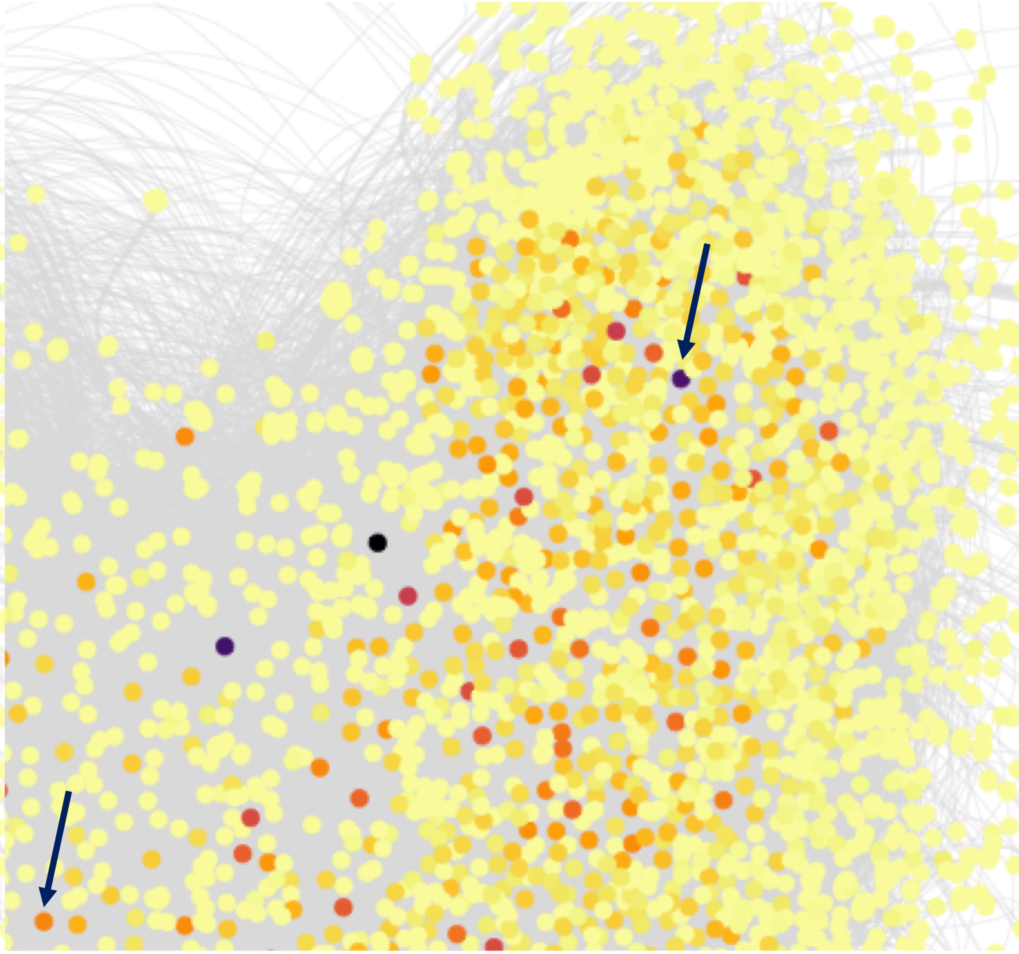
PageRank versus degree authorities

wikipedia administrator elections and vote history data

Degree



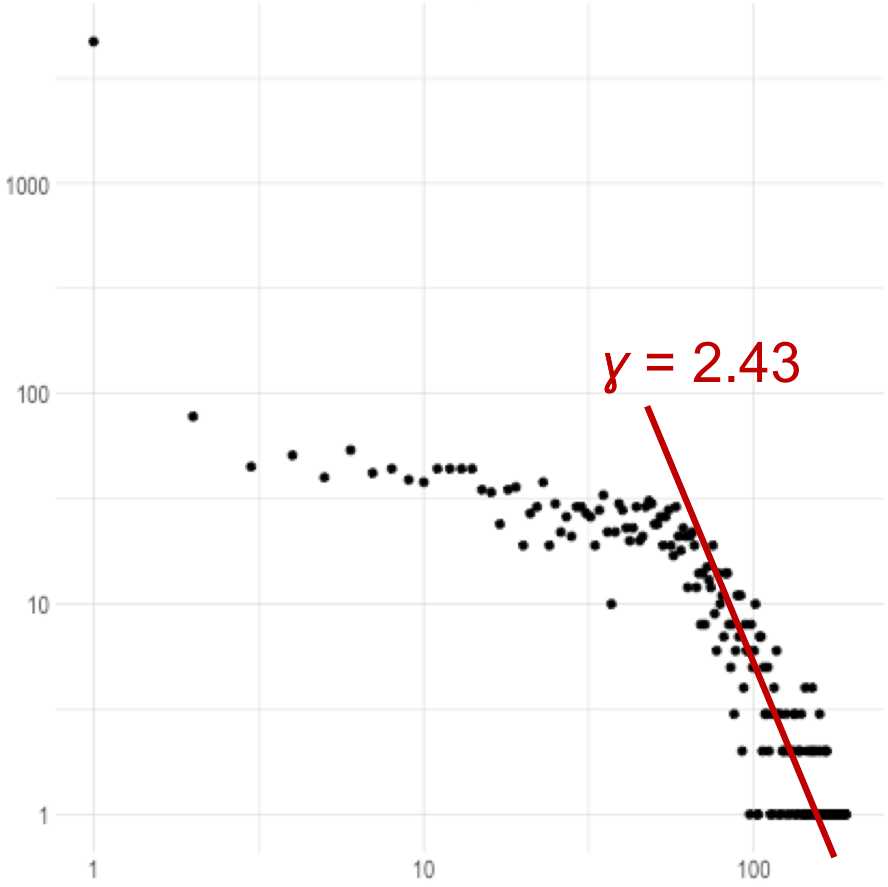
PageRank



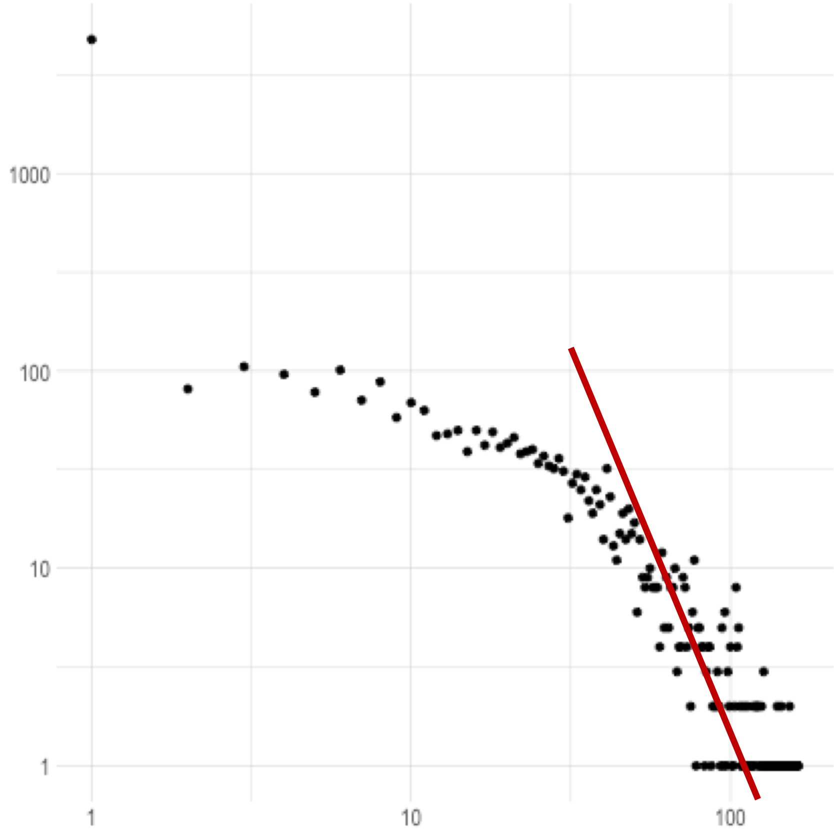
PageRank versus degree authorities

wikipedia administrator elections and vote history data

Degree



PageRank

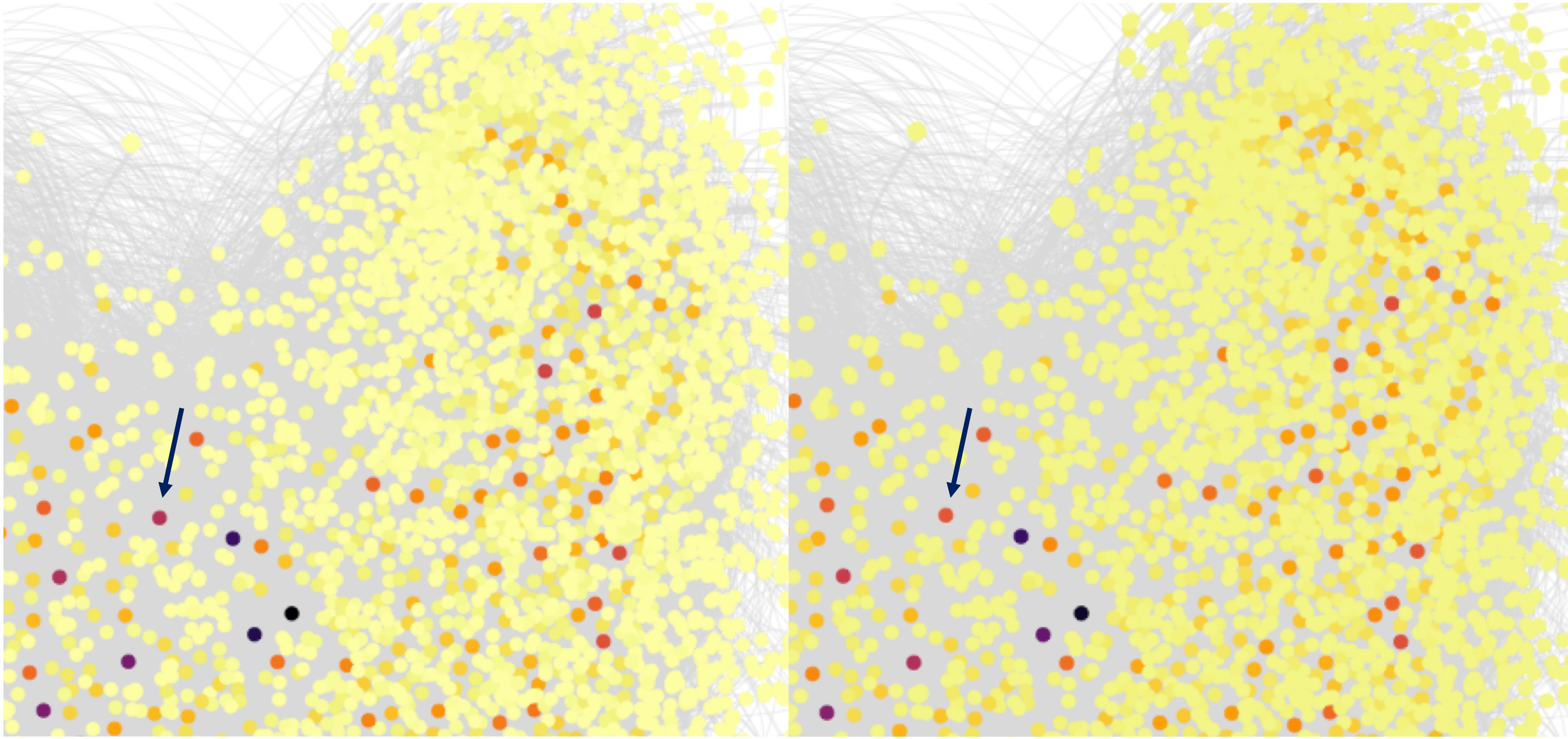


PageRank versus degree hubs

wikipedia administrator elections and vote history data

Degree

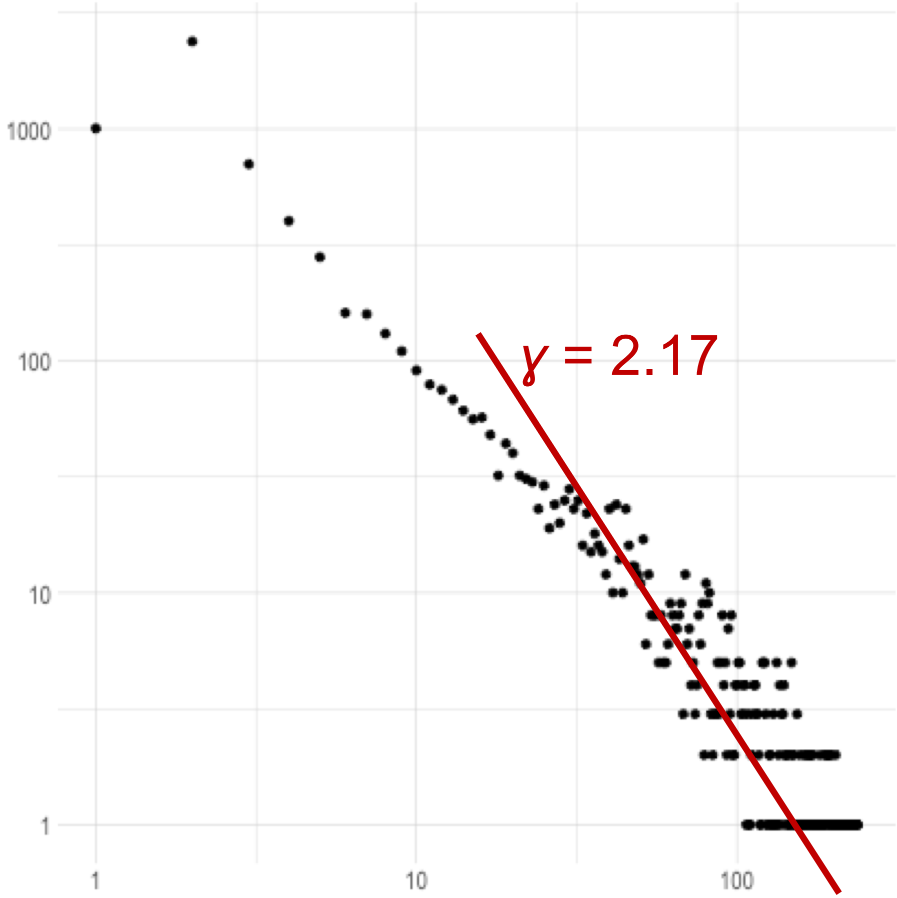
PageRank



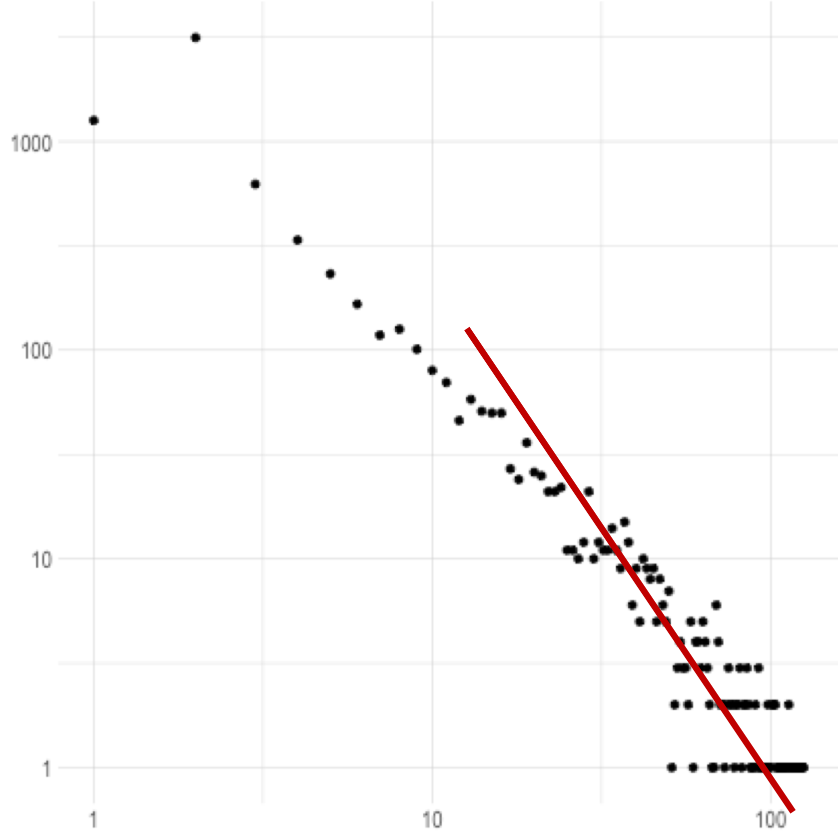
PageRank versus degree hubs

wikipedia administrator elections and vote history data

Degree



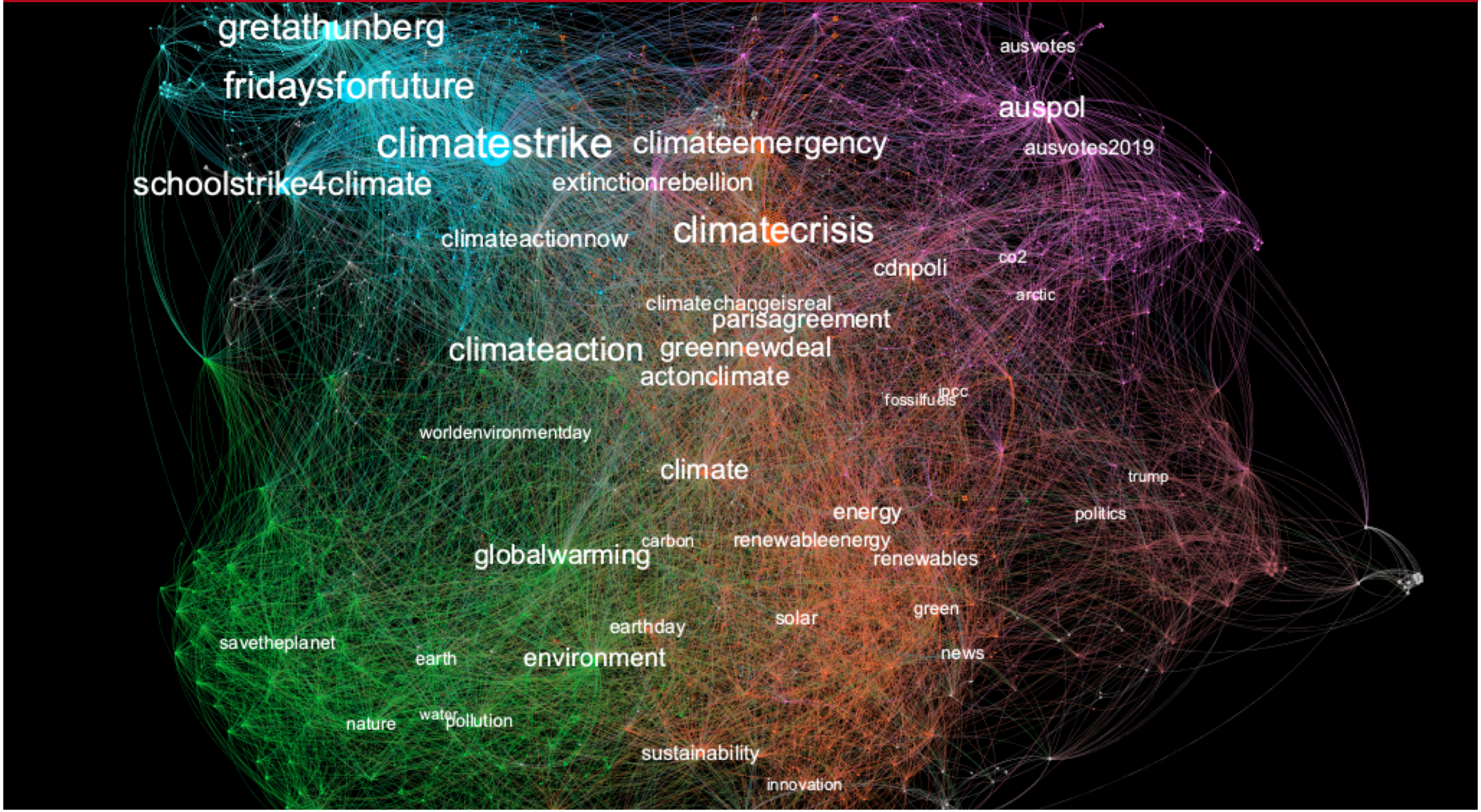
PageRank



UNIVERSITÀ DEGLI STUDI DI PADOVA

PageRank on a semantic network

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



Readings
on 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

Google Scholar

https://scholar.google.com/

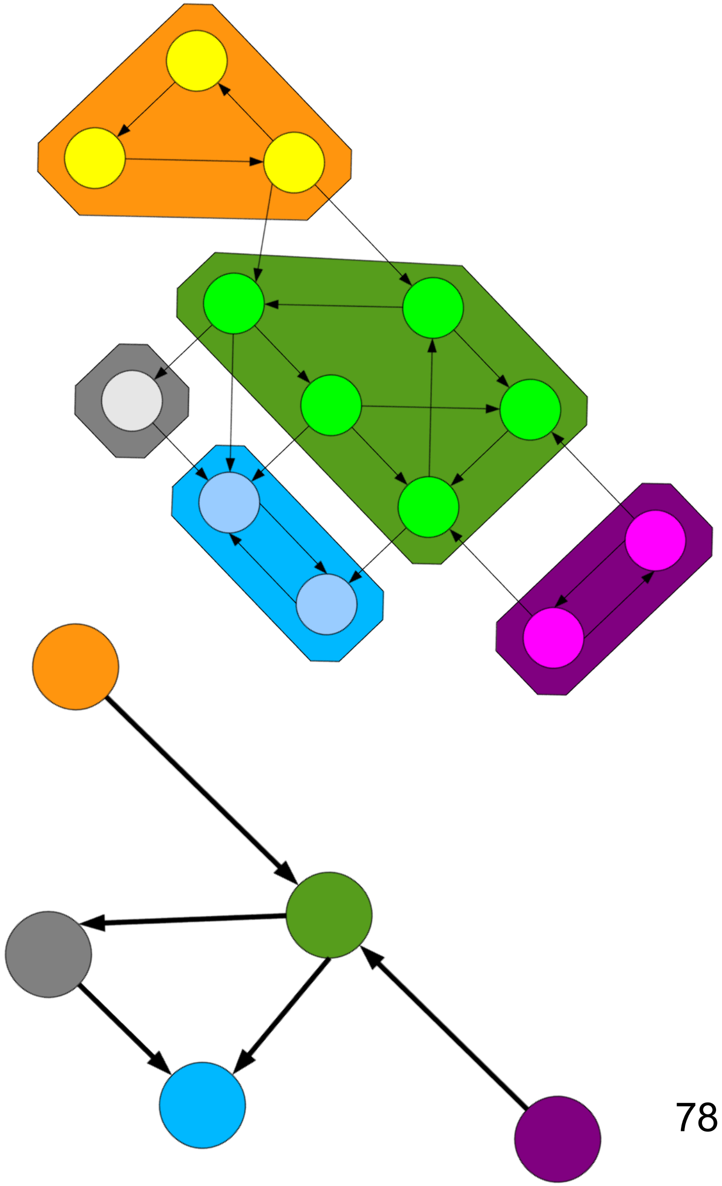
Convergence properties

of PageRank power iterations

The condensation graph

ordering an adjacency matrix

- Strong connectivity induces a partition in disjoint strongly connected sets $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_K$
- By reinterpreting the sets as nodes we obtain a condensation graph $\mathcal{G}^*$ where $i \rightarrow j$ is an edge if a connection exists between sets $\mathcal{V}_i \rightarrow \mathcal{V}_j$



Properties of the condensation graph

ordering an adjacency matrix

- $\mathcal{G}^*$ does not contain cycles

  otherwise the sets in the cycle would be strongly connected

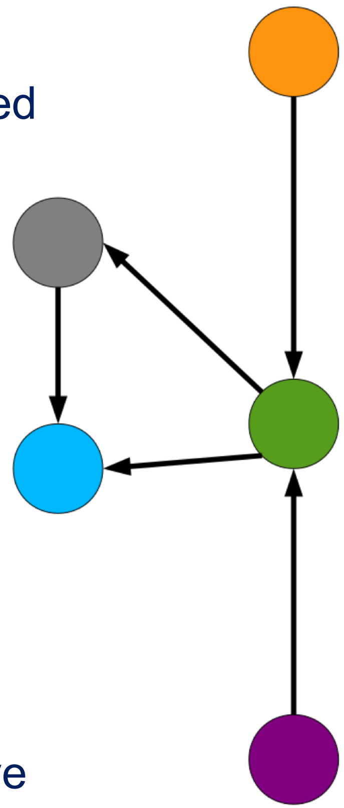
- $\mathcal{G}^*$ has at least one root and one leaf

  and every node in the graph can be reached from one of the roots

- $\mathcal{G}^*$ allows a particular reordering

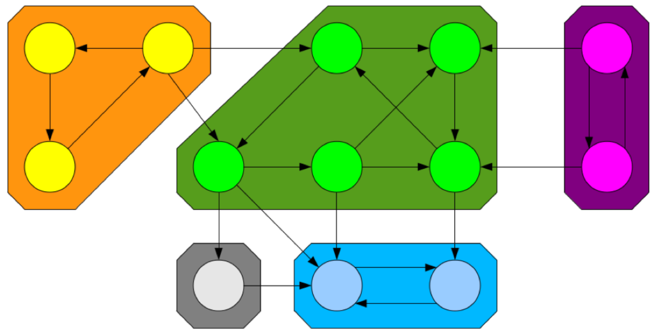
  where node n_i does not reach any of the nodes n_j with $j<i$

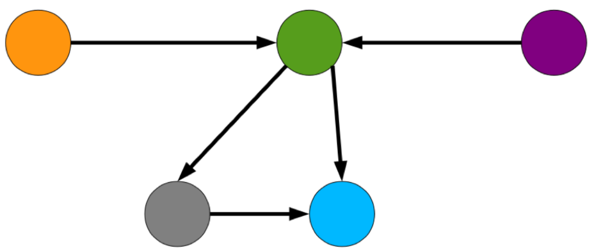
  procedure: identify a root n_1 and remove it from the network, then identify a new root; cycle until all nodes have been selected



Matrix representation
of the condensation graph

The condensation graph ordering induces a block-lower-triangular matrix structure on the adjacency matrix





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

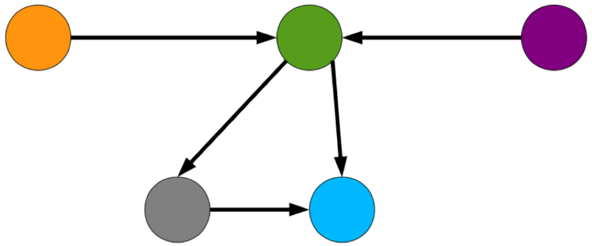
blocks in the diagonal are irreducible = no block-diagonal form !

Perron-Frobenius theorem

of the condensation graph

UNIVERSITÀ DEGLI STUDI DI PADOVA

the eigenvalues of the diagonal blocks, except for the leaves, lie inside the unit circle, i.e., $|\lambda|<1$

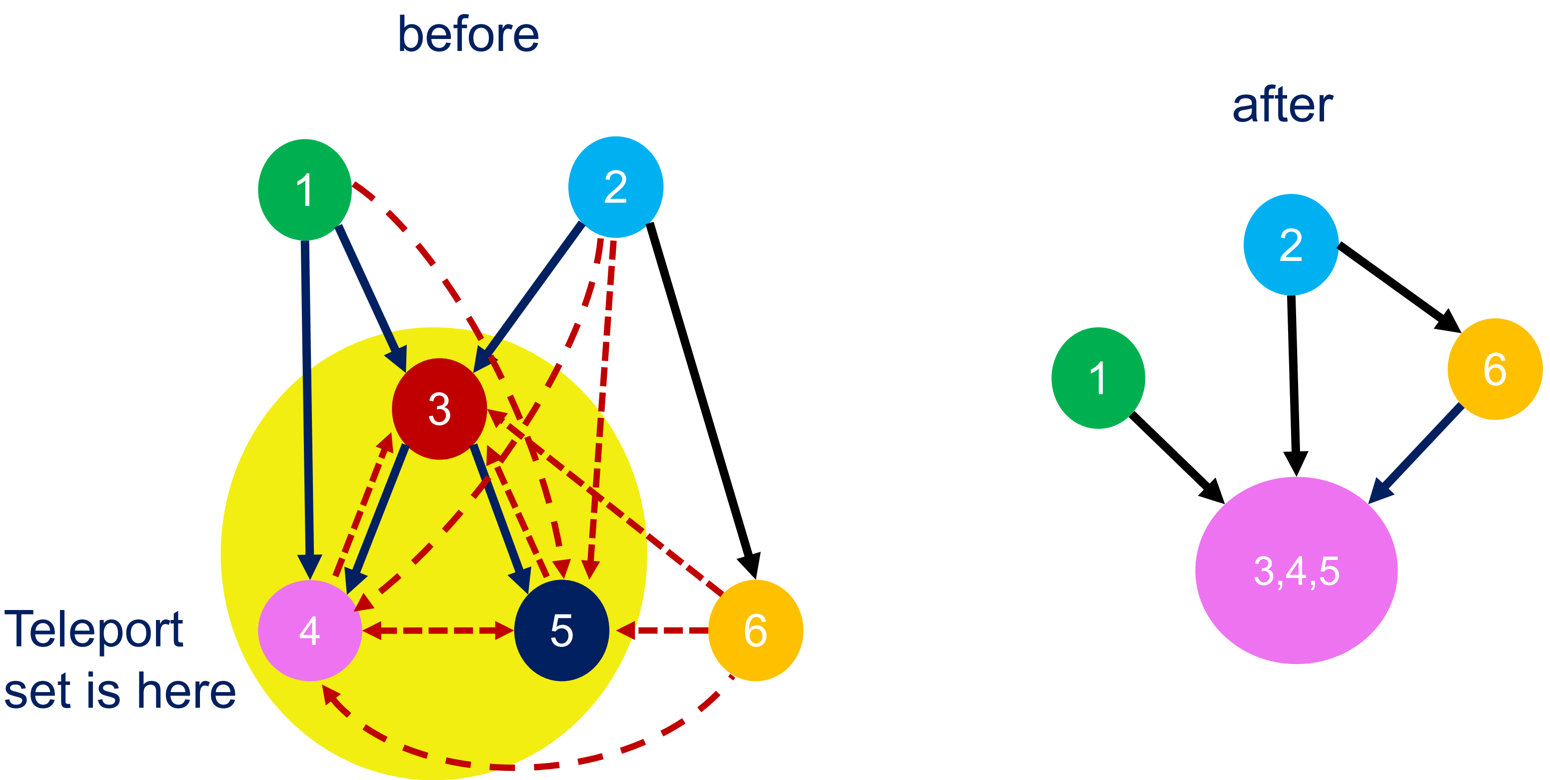


$$M=\begin{array}{c} \begin{array}{ccccc} 3 & 2 & 5 & 1 & 2 \end{array} \\ \left[\begin{array}{ccc|cc|ccccc|c|cc}
 & \frac{1}{3} & & & & & & & & & & & \\
 & & 1 & & & & & & & & & & \\
1 & & & & & & & & & & & & \\ \hline
 & & & & \frac{1}{2} & & & & & & & & \\
 & & & \frac{1}{2} & & & & & & & & & \\ \hline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

each leaf-block has at least one eigenvalue in the unit circle; $\lambda=1$ is always available, the others are distinct

The teleportation effect

it implies only one leaf



Hence M_1 carries only one eigenvector associated with the eigenvalue λ=1

Lemma

on generalized eigenvectors

- PageRank matrix $\boldsymbol{M_1} = c\,\boldsymbol{M} + (1\text{-}c)\,\boldsymbol{q}\,\boldsymbol{1}^T$
- Normalization property $\boldsymbol{1}^T \boldsymbol{M_1} = \boldsymbol{1}^T$
- Jordan form $\boldsymbol{M_1} = \boldsymbol{V}\,\boldsymbol{J}\,\boldsymbol{V}^{-1}$

carries the right (generalized) eigenvectors $\mathbf{e}_i$ of $\mathbf{M}_1$

carries the eigenvalues of $\mathbf{M}_1$

$$\boldsymbol{J} = \begin{bmatrix} \boxed{\begin{matrix}\lambda_1 & 1 & \\ & \lambda_1 & 1 \\ & & \lambda_1\end{matrix}} & & & & \\ & \boxed{\begin{matrix}\lambda_2 & 1 \\ & \lambda_2\end{matrix}} & & & \\ & & \boxed{\lambda_3} & & \\ & & & \ddots & \\ & & & & \boxed{\begin{matrix}\lambda_n & 1 \\ & \lambda_n\end{matrix}} \end{bmatrix}$$

$$\boldsymbol{1}^T \boldsymbol{M_1}\,\boldsymbol{V} = \boldsymbol{1}^T \boldsymbol{V} = \boldsymbol{1}^T \boldsymbol{V}\,\boldsymbol{J} \;\Rightarrow\; \underbrace{\boldsymbol{1}^T \boldsymbol{V}}_{\rho}\;\underbrace{(\boldsymbol{J} - \boldsymbol{I})}_{\text{only one value is 0}} = \boldsymbol{0}$$

Hence $\boldsymbol{1}^T \mathbf{e}_i = 0$ for $i>1$, i.e., except for the eigenvector associated with eigenvalue 1

Main result

for the eigenstructure of the PageRank matrix

$$\underbrace{\boldsymbol{M}_1\, \boldsymbol{e}_i = c\, \boldsymbol{M}\, \boldsymbol{e}_i}_{\text{same eigenvalues of } \mathbf{M}\text{, but multiplied by c !!!}} + (1\text{-}c)\, \boldsymbol{q}\, \cancel{\boldsymbol{1}^{\mathrm{T}}}\, \boldsymbol{e}_i \qquad \text{for } i>1$$

- ❑ $\boldsymbol{M}_1$ has one eigenvalue equal to 1
- ❑ The remaining eigenvalues satisfy $|\lambda| \leq c$

Haveliwala and Kamvar, “The second eigenvalue of the Google matrix,” 2003

http://ilpubs.stanford.edu:8090/582/1/2003-20.pdf

Convergence properties

of the PageRank power iteration

$$\boldsymbol{p}_t = \boldsymbol{M}_1\, \boldsymbol{p}_{t-1} = \boldsymbol{M}_1^t\, \boldsymbol{p}_0 = \boldsymbol{V}\, \boldsymbol{J}^t\, \boldsymbol{V}^{-1}\, \boldsymbol{p}_0$$

gets large for high multiplicity   max eigenvalue multiplicity

- $\|\boldsymbol{p}_t - \boldsymbol{p}_\infty\|_2 \lesssim K\, c^t\, t^{m-1} \sim K\, c^t$
- Triangular inequality: $\|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\|_2 \lesssim 2K\, c^t$
- Precision ε: $\|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\|_2 < \varepsilon$
- Iterations required: $\boxed{t = [\ln(2/\varepsilon) + \ln(K)] \,/\, \ln(1/c)}$

precision 10^{-3} → 7.6

c=0.85 → 1/ln(1/c) = 6

Is usually small
→ fast algorithm

Local PageRank

measuring similarity/closeness among nodes

Measuring closeness: LocalPageRank

for the eigenstructure of the PageRank matrix

Idea

- Measure similarity / closeness to node *i* by applying PageRank with teleport set $S=\{i\}$, i.e., with $\mathbf{q} = \boldsymbol{\delta}_i$

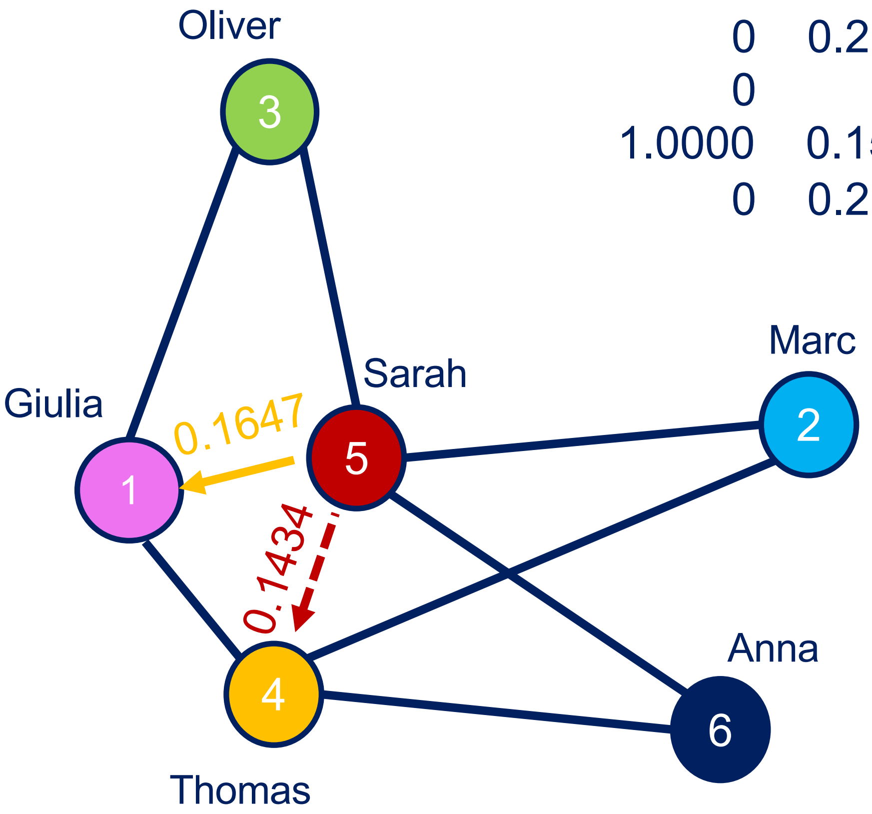
Result

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

Example

who's Sara'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

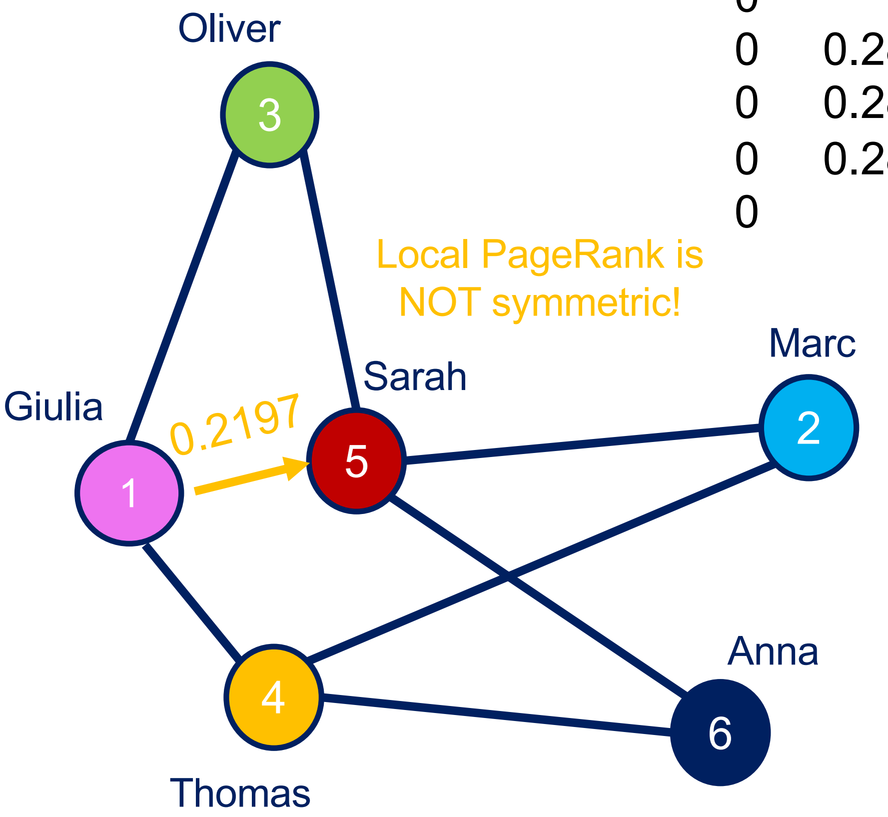


10	20	50	75	100	
0.1743	0.1653	0.1647	0.1647	0.1647	Giulia
0.1238	0.1144	0.1138	0.1138	0.1138	Marc
0.1206	0.1199	0.1199	0.1199	0.1199	Oliver
0.1285	0.1426	0.1434	0.1434	0.1434	Thomas
0.3290	0.3435	0.3444	0.3444	0.3444	Sarah
0.1238	0.1144	0.1138	0.1138	0.1138	Anna

Example

who's Giulia's best friend?

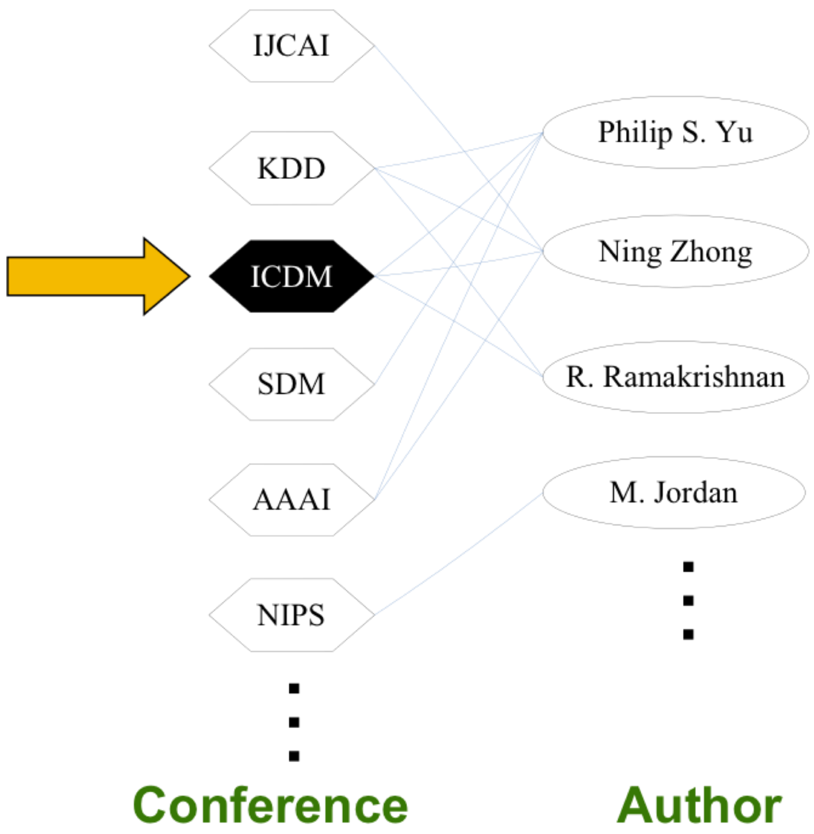
t=1	2	3	4	5
1.0000	0.1500	0.4109	0.2403	0.3404
0	0	0.1405	0.0467	0.1262
0	0.2833	0.1027	0.1510	0.1275
0	0.2833	0.0425	0.2358	0.1078
0	0.2833	0.1629	0.2795	0.1719
0	0	0.1405	0.0467	0.1262



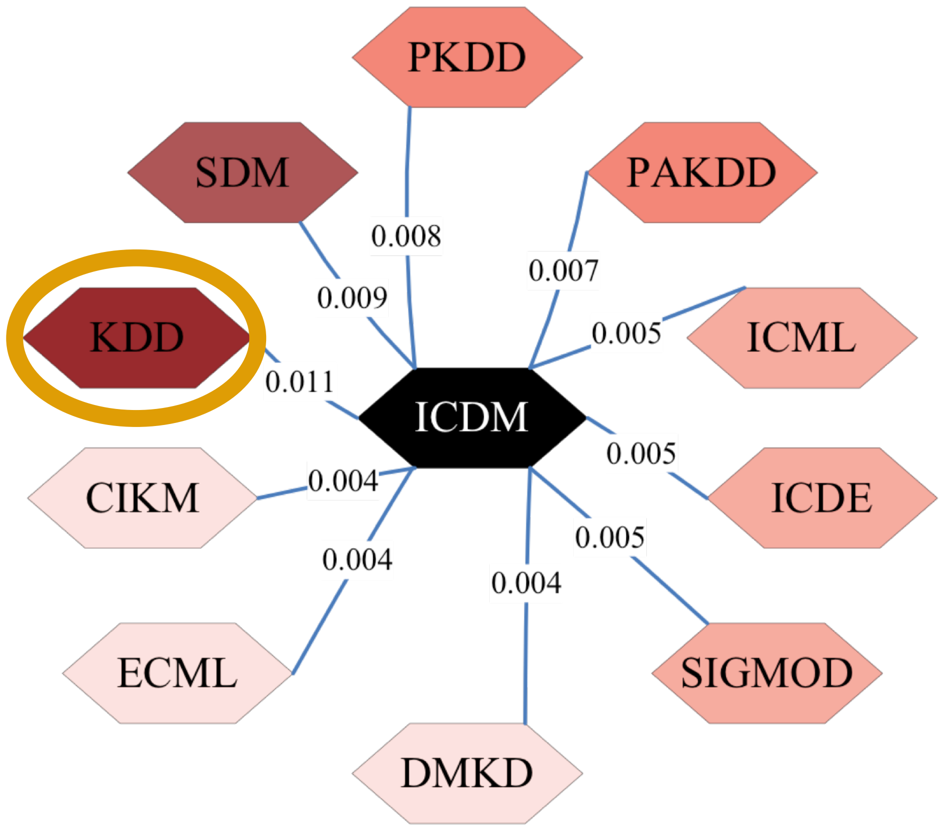
10	20	50	75	100	
0.2909	0.2985	0.2989	0.2989	0.2989	Giulia
0.0848	0.0926	0.0931	0.0931	0.0931	Marc
0.1309	0.1313	0.1314	0.1314	0.1314	Oliver
0.1763	0.1645	0.1638	0.1638	0.1638	Thomas
0.2324	0.2204	0.2197	0.2197	0.2197	Sarah
0.0848	0.0926	0.0931	0.0931	0.0931	Anna

Example

what is the most related conference to ICDM?



Top 10 ranking results

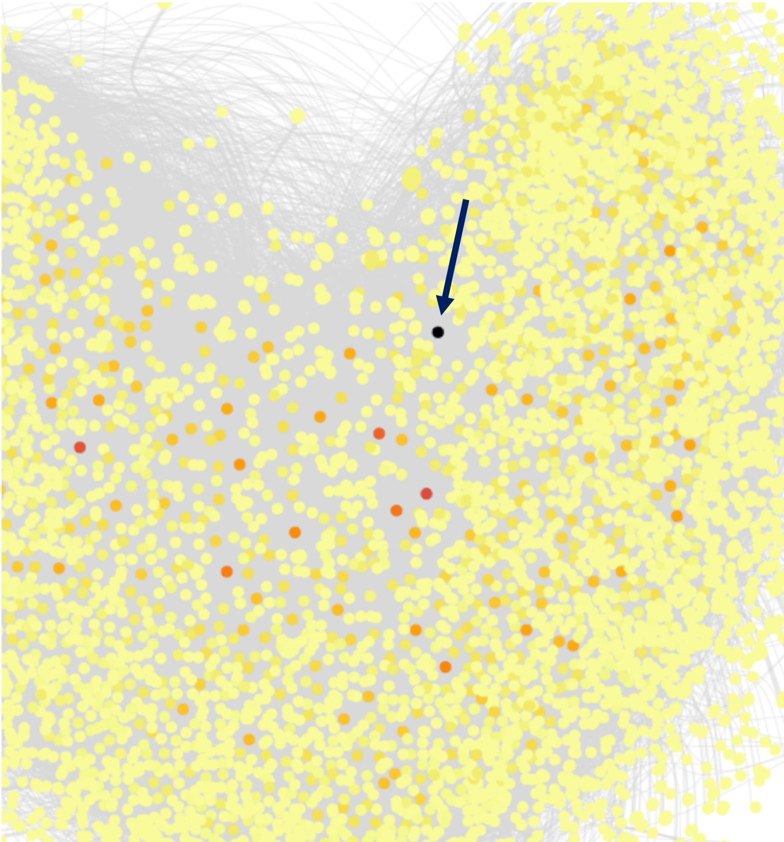


ICDM = international conf. on data mining
KDD = knowledge discovery and data mining

Local PageRank versus degree

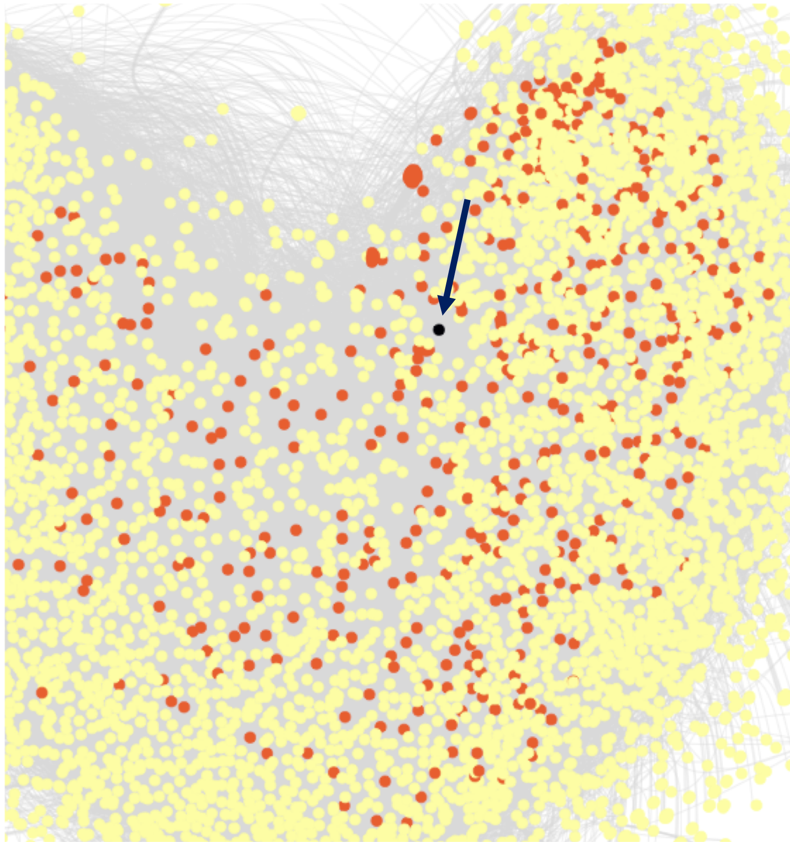
authorities

Local PageRank



neighbours authority score = local node → neighbours

1-hop out-neighbours



On the complexity of Local PageRank

approximate PageRank

Andersen, Chung, Lang, “Local graph partitioning using PageRank vectors,” 2006

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4031383

use institutional Sign In with your unipd credentials

Approximate PageRank algorithm

the push operation

- Start from $\boldsymbol{u} = \boldsymbol{0}$ and $\boldsymbol{v} = \boldsymbol{q}$

precision

degree of node i

sum of the degrees

- To all the nodes *i* satisfying $v_i > \varepsilon\, d_i/D$ apply the **push** operation

u constantly increases

$$\boldsymbol{u}^{+} = \boldsymbol{u} + (1\text{-}c)\,\boldsymbol{\delta}$$

$$\boldsymbol{v}^{+} = \boldsymbol{v} - \boldsymbol{\delta} + c\,\boldsymbol{M}\,\boldsymbol{\delta}$$

only one active element in position i with value v_i

v always positive

Returns $\boldsymbol{u} \simeq \boldsymbol{r}$ with precision $|\boldsymbol{r} - \boldsymbol{u}|_1 < \varepsilon$

It is simple

Linearity of PageRank

to build a lemma for the proof

column stochastic matrix $\boldsymbol{1}^{\mathrm{T}} \boldsymbol{M} = \boldsymbol{1}^{\mathrm{T}}$

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

stochastic ranking vector
$\boldsymbol{1}^{\mathrm{T}} \boldsymbol{r}_q = 1,\ \boldsymbol{r}_q \geq 0$

stochastic Teleport vector
$\boldsymbol{1}^{\mathrm{T}} \boldsymbol{q} = 1$

- Alternative equation $\boldsymbol{r}_q = (\boldsymbol{I} - c\, \boldsymbol{M})^{-1}\, (1-c)\, \boldsymbol{q}$

linear in q

$$\boldsymbol{r}_{au+bv} = a\, \boldsymbol{r}_u + b\, \boldsymbol{r}_v$$

Modifying the PageRank equation

the lemma for the proof

one-step random walk

- PageRank equation $\boldsymbol{r}_q = c\ \boldsymbol{r}_{Mq} + (1\text{-}c)\ \boldsymbol{q}$

$\Downarrow$

- $\boldsymbol{r}_q = (\boldsymbol{I} - c\ \boldsymbol{M})^{-1}\ (1\text{-}c)\ \boldsymbol{q}$
- $\boldsymbol{r}_q = (1\text{-}c)\ \Sigma\ (c\ \boldsymbol{M})^k\ q$
- $\boldsymbol{M}\ r_q = (1\text{-}c)\ \Sigma\ (c\ \boldsymbol{M})^k\ \boldsymbol{M}\ \boldsymbol{q}$
- $\boldsymbol{M}\ \boldsymbol{r}_q = \boldsymbol{r}_{Mq}$

Main property of push: $\boldsymbol{r}_q = \boldsymbol{u} + \boldsymbol{r}_v$

almost there

- At starting point $\boldsymbol{u} = \boldsymbol{0}$ and $\boldsymbol{v} = \boldsymbol{q}$ imply $\boldsymbol{r}_q = \boldsymbol{0} + \boldsymbol{r}_q$
- The following steps are proved by induction

$$\boldsymbol{u}^+ = \boldsymbol{u} + (1\text{-}c)\,\delta$$
$$\boldsymbol{v}^+ = \boldsymbol{v} - \delta + c\,\boldsymbol{M}\,\delta$$

$$\Downarrow$$

$$\boldsymbol{u}^+ + \boldsymbol{r}_{v+} = \boldsymbol{u} + (1\text{-}c)\,\delta + \overbrace{\boldsymbol{r}_v - \boldsymbol{r}_\delta + \underbrace{c\,\boldsymbol{r}_{M\delta}}_{\boldsymbol{r}_\delta - (1\text{-}c)\,\delta}}^{\text{by linearity}}$$

$$\Downarrow$$

$$\boldsymbol{u}^+ + \boldsymbol{r}_{v+} = \boldsymbol{u} + \boldsymbol{r}_v = \boldsymbol{r}_q$$

Precision guarantee: $|\boldsymbol{r}_q - \boldsymbol{u}|_1 < \varepsilon$

and the result is proved

- The push property implies $\boldsymbol{r}_q = \boldsymbol{u} + \boldsymbol{r}_v$
- Hence $|\boldsymbol{r}_q - \boldsymbol{u}|_1 = |\boldsymbol{r}_v|_1 = \boldsymbol{1}^T \boldsymbol{r}_v$

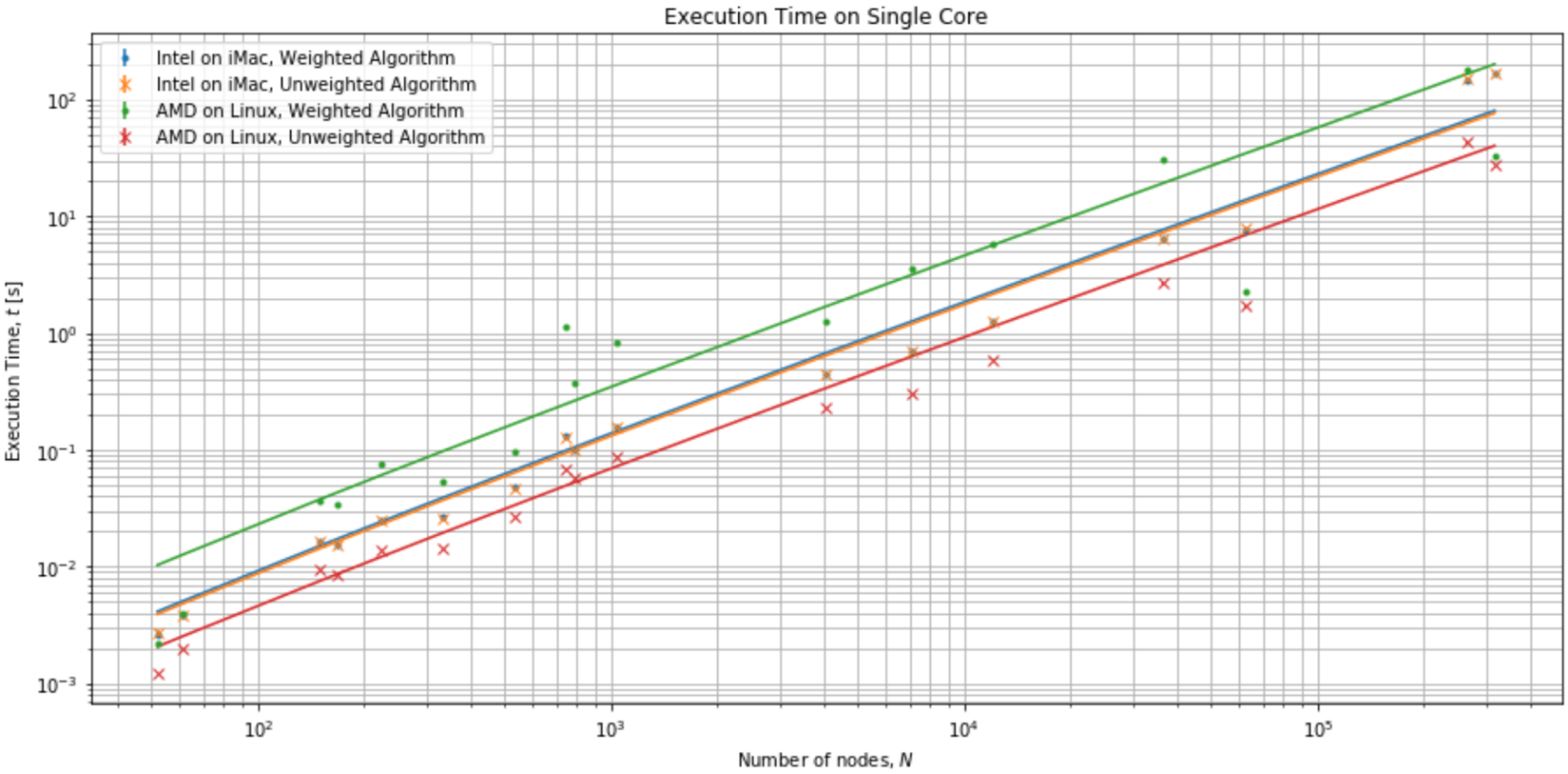
- The PageRank equation is $\boldsymbol{r}_v = c\, \boldsymbol{M}\, \boldsymbol{r}_v + (1-c)\, \boldsymbol{v}$
- Hence $\boldsymbol{1}^T \boldsymbol{r}_v = c\, \underbrace{\boldsymbol{1}^T \boldsymbol{M}}_{\boldsymbol{1}^T}\, \boldsymbol{r}_v + (1-c)\, \boldsymbol{1}^T \boldsymbol{v}$ so that $\boldsymbol{1}^T \boldsymbol{r}_v = \boldsymbol{1}^T \boldsymbol{v}$

As a result $|\boldsymbol{r}_q - \boldsymbol{u}|_1 = \boldsymbol{1}^T \boldsymbol{v} < \Sigma\, \varepsilon\, d_i / D = \varepsilon$

Scalability properties

of Local PageRank using Approximate PageRank

(Francesco Barbato & Tommaso Boccato, 2020)



Quasi-linear behaviour = scalability of Local PageRank

Beware of the Lazy PageRank

which is suggested in the paper

- Lazy PageRank $\boldsymbol{r} = a\, \boldsymbol{M}_2\, r + (1\text{-}a)\, \boldsymbol{q}$

  $$\boldsymbol{M}_2 = b\, \boldsymbol{I} + (1\text{-}b)\, \boldsymbol{M}$$

- Lazy because a fraction b of the times the surfer stays where she/he is

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

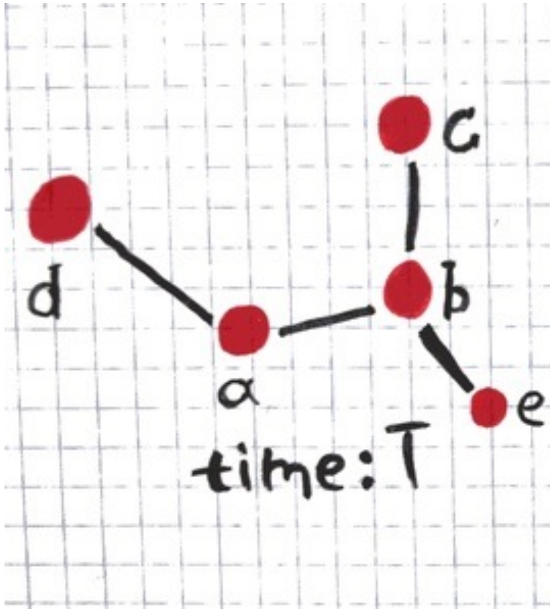
  $$c = a(1\text{-}b)/(1\text{-}ab) < a$$

  slower algorithm, as its convergence speed depends on a>c, better use c directly!

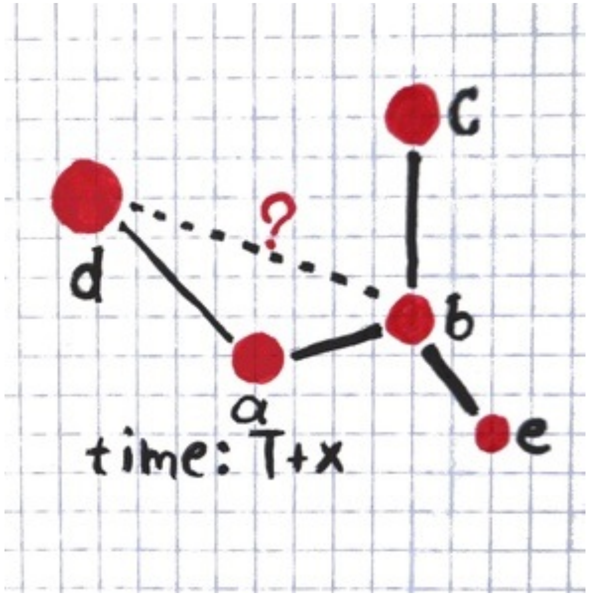
Application #1

the link prediction task

Recommendation in social networks

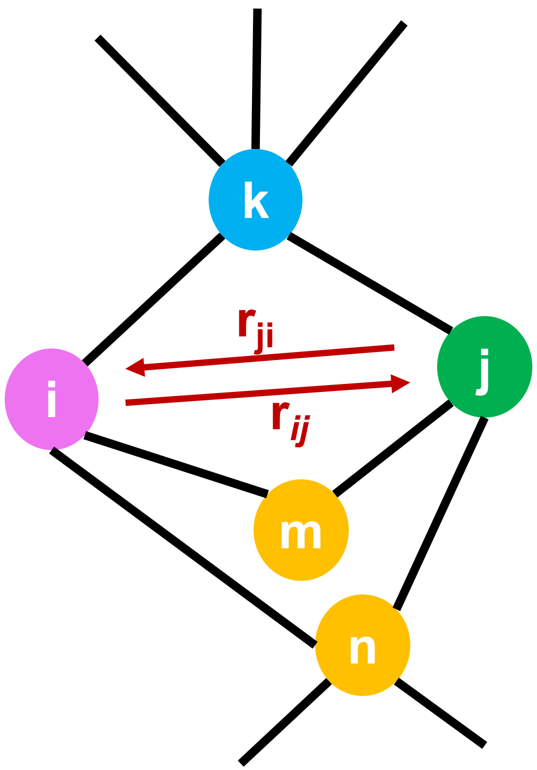


Given a graph at time T, can we output a ranked list of links that are predicted to appear in the graph at time T+x ?



Application #1

random walk with restart (RWR) method



Local PageRank vector

teleportation to node *i*

$$\mathbf{r}_i = c\ \mathbf{M}\ \mathbf{r}_i + (1\text{-}c)\ \boldsymbol{\delta}_i$$

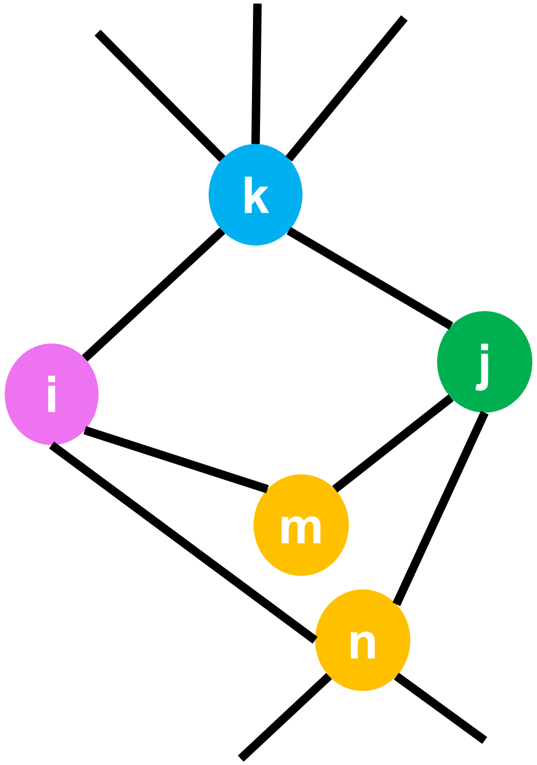
Likelihood of activating the link (i,j)

$$\boxed{L_{RWR}(i,j) = r_{ij} + r_{ji}}$$

Select the highest values of L_{RWR} for recommendation pourposes

Application #1

the resorse allocation (RA) counterpart



$$L_{RA}(i,j) = \sum_{k \in N_i \cap N_j} 1/d_k$$

common neighbours

related to a two-hop RWR

$$\mathbf{r}_i \simeq (1\text{-c}) \sum_{n=0}^{2} (\text{c}\, \mathbf{M})^n\, \boldsymbol{\delta}_i$$

to have

$$r_{ij} \simeq (1\text{-c})\, c^2 / d_i\, L_{RA}(i,j)$$

$$L_{RWR}(i,j) \simeq (1\text{-c})\, c^2\, (1/d_i + 1/d_j)\, L_{RA}(i,j)$$

Application #1
performance metrics

fraction of links correctly guessed
(out of 100 recomendations)

Precision	CN	RA	LP	ACT
USAir	0.59	0.64	0.61	0.49
NetScience	0.26	0.54	0.30	0.19
Power	0.11	0.08	**0.13**	0.08
Yeast	0.67	0.49	0.68	0.57
C.elegans	0.12	0.13	**0.14**	0.07

	RWR	HSM	LRW	SRW
USAir	0.65	0.28	0.64(3)	**0.67**(3)
NetScience	**0.55**	0.25	0.54(2)	0.54(2)
Power	0.09	0.00	0.08(2)	0.11(3)
Yeast	0.52	0.84	**0.86**(3)	0.73(9)
C.elegans	0.13	0.08	**0.14**(3)	**0.14**(3)

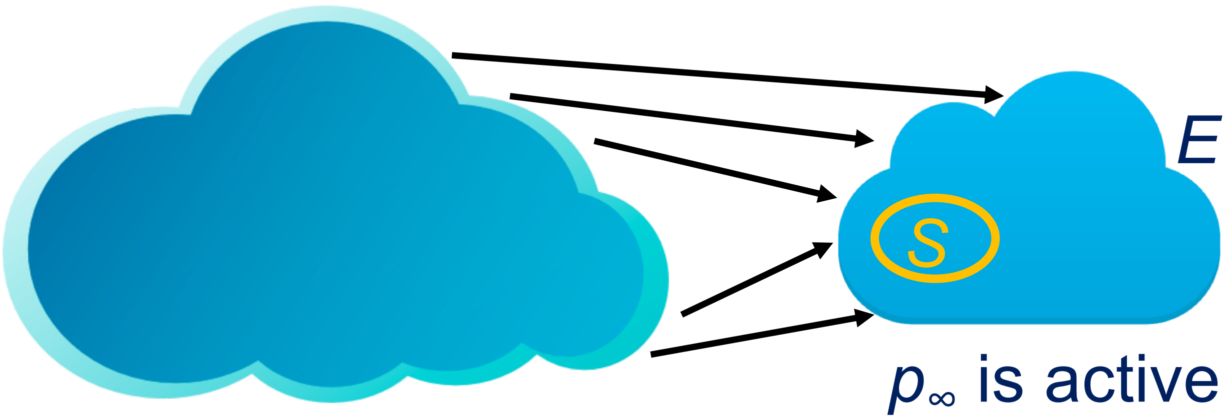
Among the best performance in social networks

But not strikingly good compared to simpler methods (e.g., RA = resource allocation)

- Bias the random walk towards a topic specific teleport set *S* of nodes, i.e., make sure that *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



p_∞ is active in E only

Application #2

assigning documents to topics in semantic networks

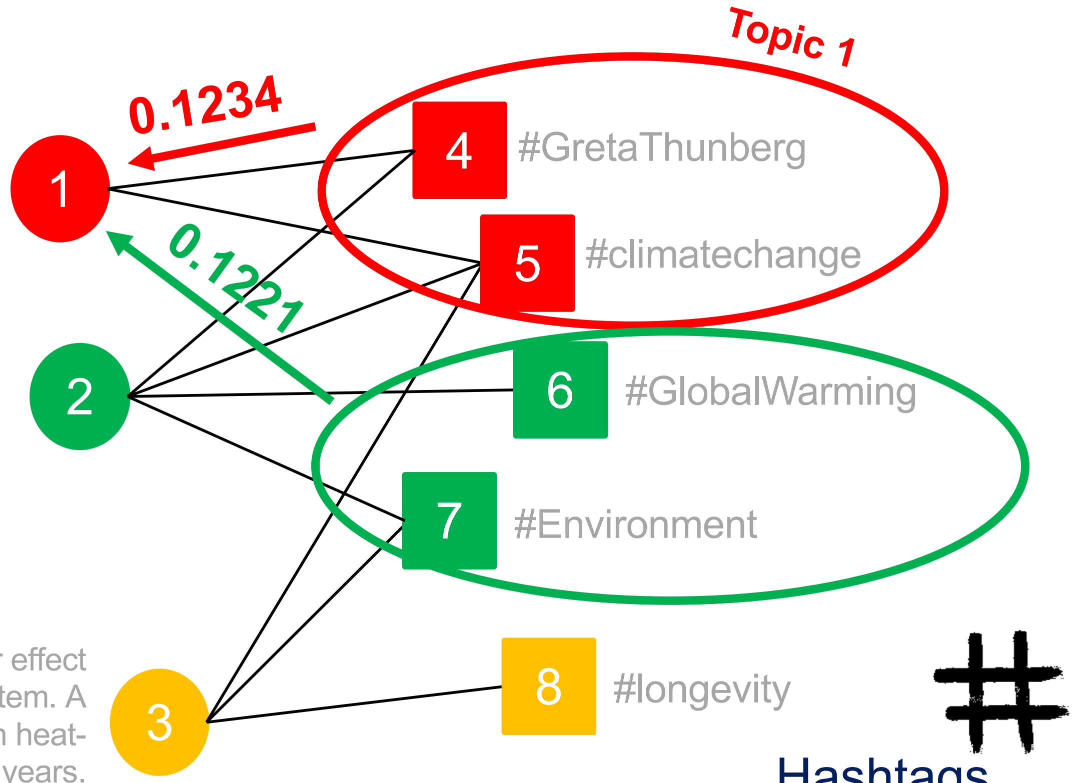
Tweet 1 is assigned
to **Topic 1** !!!

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



Hashtags

Signed PageRank

modifications for signed networks

PageRank in signed networks

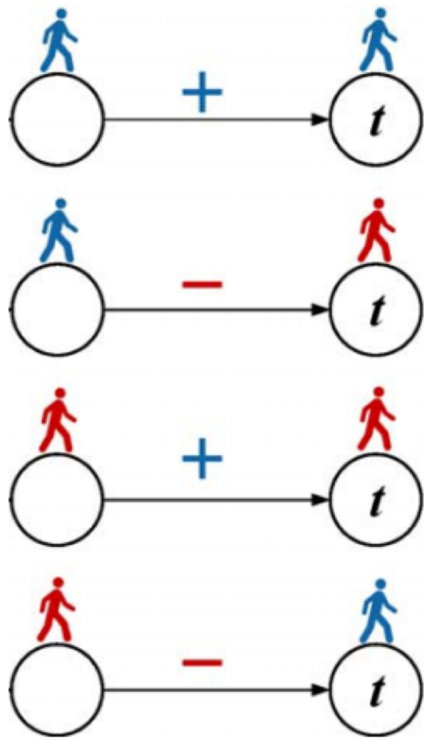
Jung, Jim, Sael, Kang, "Personalized ranking in signed networks using signed random walk with restart," 2016

https://ieeexplore.ieee.org/iel7/7837023/7837813/07837935.pdf

- Identify + (favourable) and – (adversarial) paths, i.e., ranking values $\boldsymbol{r}_+$ and $\boldsymbol{r}_-$ for positive and negative surfers
- Extract positive $\boldsymbol{A}_+$ and negative $\boldsymbol{A}_-$ contributions to $\boldsymbol{A} = \boldsymbol{A}_+ - \boldsymbol{A}_-$
- Normalize the absolute value, to get $\boldsymbol{M}_+$ and $\boldsymbol{M}_-$ (with normalized $\boldsymbol{M}_+ + \boldsymbol{M}_-$)
- Run a signed random walk

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

$$\boldsymbol{r}_- = c\,\boldsymbol{M}_-\,\boldsymbol{r}_+ + c\,\boldsymbol{M}_+\,\boldsymbol{r}_-$$



Signed PageRank

power iteration

damping factor

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_+ & \boldsymbol{M}_- \\ \boldsymbol{M}_- & \boldsymbol{M}_+ \end{bmatrix}$$ (column) normalized adjacency matrix

$$\boldsymbol{r} = c\,\boldsymbol{M}\,\boldsymbol{r} + (1\text{-}c)\,\boldsymbol{q}_0$$

teleportation vector

PageRank vector (centrality) $\boldsymbol{r} = \begin{bmatrix} \boldsymbol{r}_+ \\ \boldsymbol{r}_- \end{bmatrix}$

$$\boldsymbol{q}_0 = \begin{bmatrix} \boldsymbol{q} \\ \boldsymbol{0} \end{bmatrix}$$

signed centrality outcome $\boldsymbol{r}_{+-} = \boldsymbol{r}_+ - \boldsymbol{r}_-$

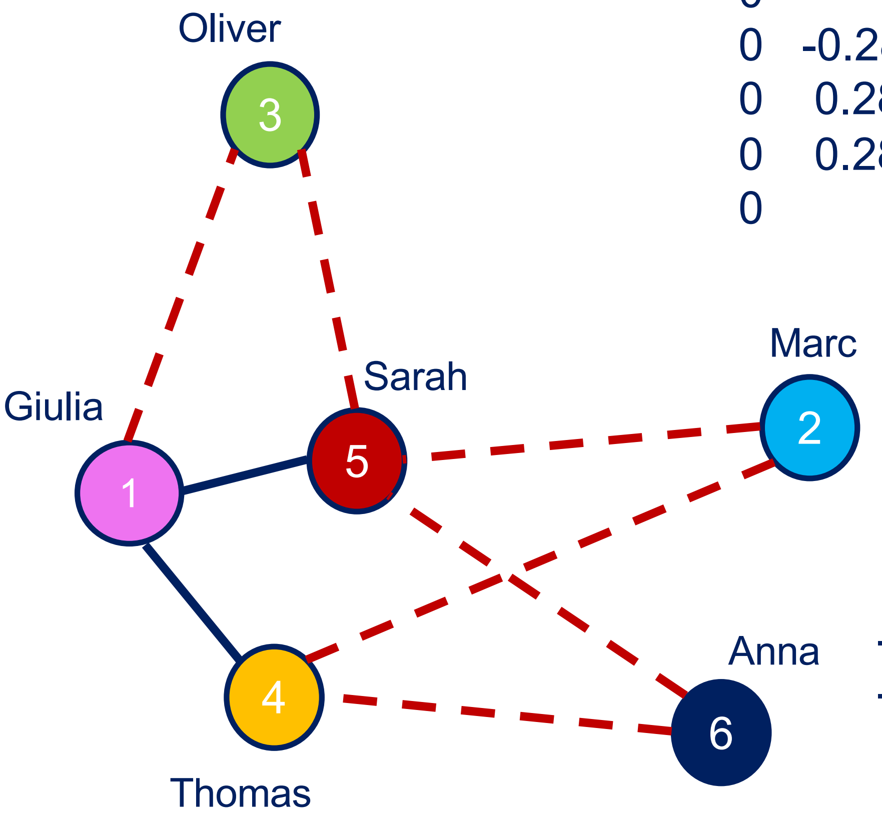
$$\boldsymbol{r}_{+-} = c\,\boldsymbol{M}_{+-}\,\boldsymbol{r}_{+-} + (1\text{-}c)\,\boldsymbol{q}$$

can be signed

$$\boldsymbol{M}_{+-} = \boldsymbol{A}\,\mathrm{diag}^{-1}(|\boldsymbol{A}|^{\mathrm{T}}\boldsymbol{1})$$

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

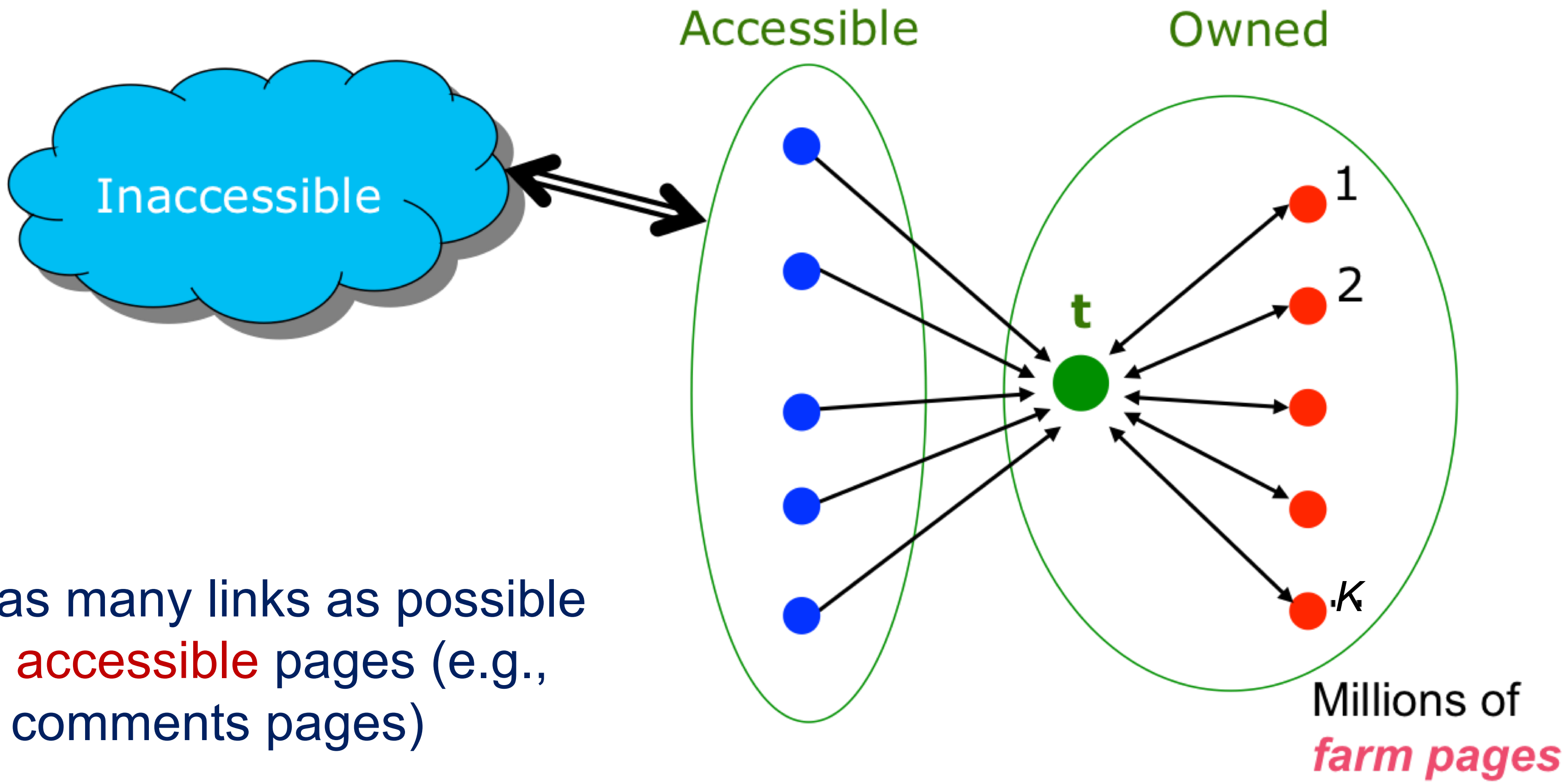
10	20	50	75	100	
0.2909	0.2985	0.2989	0.2989	0.2989	Giulia
-0.0848	-0.0926	-0.0931	-0.0931	-0.0931	Marc
-0.1309	-0.1313	-0.1314	-0.1314	-0.1314	Oliver
0.1763	0.1645	0.1638	0.1638	0.1638	Thomas
0.2324	0.2204	0.2197	0.2197	0.2197	Sarah
-0.0848	-0.0926	-0.0931	-0.0931	-0.0931	Anna

Preventing spamming

on the role of the teleport vector

Spam farm

how to boost PageRank for a web page



1. Get as many links as possible from accessible pages (e.g., blog comments pages)

2. Construct link farm to get a PageRank multiplier effect

Google bombs
in 2004 US elections

Web Images Groups News Froogle Local **more »**

miserable failure | Search | Advanced Search | Preferences

Web Results **1** - **10** of about **969,000** for **miserable failure**. (**0.06** seconds)

Biography of President George W. Bush
Biography of the president from the official White House web site.
www.whitehouse.gov/president/gwbbio.html - 29k - Cached - Similar pages
Past Presidents - Kids Only - Current News - President
More results from www.whitehouse.gov »

Welcome to MichaelMoore.com!
Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, **...**
www.michaelmoore.com/ - 35k - Sep 1, 2005 - Cached - Similar pages

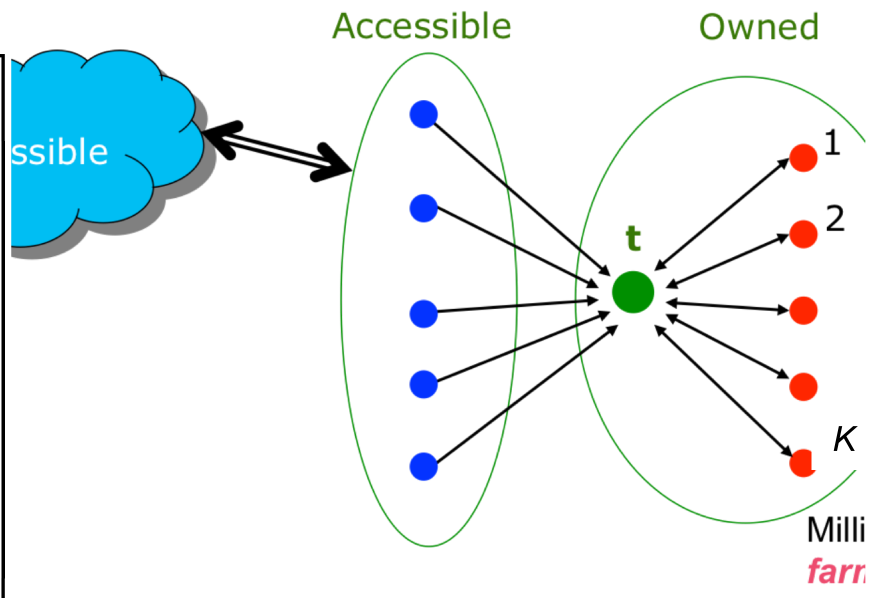
BBC NEWS | Americas | '**Miserable failure**' links to Bush
Web users manipulate a popular search engine so an unflattering description leads to the president's page.
news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - Cached - Similar pages

Google's (and Inktomi's) **Miserable Failure**
A search for **miserable failure** on Google brings up the official George W. Bush biography from the US White House web site. Dismissed by Google as not a **...**
searchenginewatch.com/sereport/article.php/3296101 - 45k - Sep 1, 2005 - Cached - Similar pages

UNIVERSITÀ DEGLI STUDI DI PADOVA

PageRank analysis
of spam farms

$$
\begin{array}{ll}
\begin{array}{r} \text{inacc.} \\ \\ \text{acc.} \\ M = \\ \text{t} \\ \\ \text{owned} \end{array} &
\left[\begin{array}{c|c|c|c}
? & ? & \begin{matrix}0\\ \vdots \\ 0\end{matrix} & \mathbf{0} \\
\hline
? & ? & \begin{matrix}0\\ \vdots \\ 0\end{matrix} & \mathbf{0} \\
\hline
0 \ \cdots \ 0 & ? \ \cdots \ ? & 0 & 1 \ \cdots \ 1 \\
\hline
\mathbf{0} & \mathbf{0} & \begin{matrix}\frac{1}{K}\\ \vdots \\ \frac{1}{K}\end{matrix} & \mathbf{0}
\end{array}\right] \\
& \begin{array}{cccc} \text{inaccessible} & \text{accessible} & \text{t} & \text{owned} \end{array}
\end{array}
$$



ranking due to accessible pages

teleportation value to pages owned by the spammer

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

$$r_t = a + cKr_o + (1-c)q_o$$

$$r_o = c\frac{1}{K}r_t + (1-c)q_o$$

PageRank outcome
of spam farms

teleportation value to pages owned by the spammer

ranking due to accessible pages

$$r_t = \frac{a}{1-c^2} + \frac{cK+1}{1+c} q_o$$

scaling factor (≃3.6)

spam factor (can be made as large as desired)

solution teleport only to trusted pages (i.e., set $q_o = 0$)
can also be used as a method to identify spam farms

Row-normalized PageRank

For spreading information over the network

Row-normalized PageRank

an overview

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

row-normalized

$\boldsymbol{M}\,\boldsymbol{1} = \boldsymbol{1}$

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

Markov chain
$$\boldsymbol{p}_{t+1} = c\,\boldsymbol{M}\,\boldsymbol{p}_t + (1\text{-}c)\,\boldsymbol{q}$$
$$\boldsymbol{p}_0 = \boldsymbol{q}$$

$$\boldsymbol{M}_1 = c\,\boldsymbol{M} + (1\text{-}c)\,\boldsymbol{q}\,\boldsymbol{v}^T$$
$$\boldsymbol{v}^T\boldsymbol{M} = \boldsymbol{v}^T$$
$$\boldsymbol{v}^T\boldsymbol{q} = 1$$

same properties of column-normalized PageRank:

- $\boldsymbol{M}_1$ has one eigenvalue equal to 1
- The remaining eigenvalues satisfy $|\lambda| \le c$

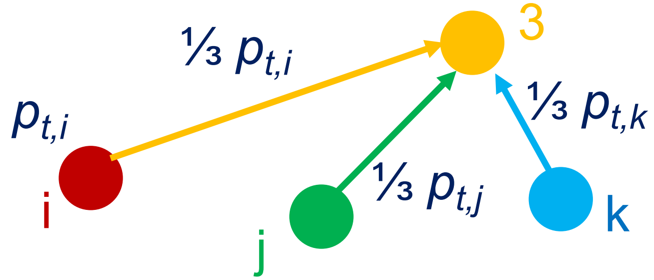
Università degli Studi di Padova

Row-normalized PageRank

interpreting its action

A node gathers the average value of the neighbour nodes pointing to it

$$p_{t+1,3} = c\ \tfrac{1}{3}\ (\ p_{t,i} + p_{t,j} + p_{t,k}\ ) + (1-c)\ q_3$$



It is a way of spreading the original information **q** over the network

Università degli Studi di Padova

Semantic network example

agency = action and goal orientation, sense of which is necessary for people to attempt social change

q values of agency (in colour) | r values after spreading

q values of agency (in colour)

r values after spreading

Takeaways
for PageRank centrality

- This is the metric to be used it for resizing nodes according to their importance
- Provides elaborate information on the relevance of nodes in the network
- For directed networks, it can be used in both its authority and hub forms
- Can also be put in the form of a PageRank distribution
- Can be used in different useful ways, e.g., to evaluate similarity or closeness, to spread information
- Exploit its potential at your best

HITS centrality

a (less interesting) alternative to PageRank

HITS – hubs and authorities

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

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

Conceptually similar to PageRank

Provides scores for authorities and hubs, separately, as PageRank can do

We deprecate its use

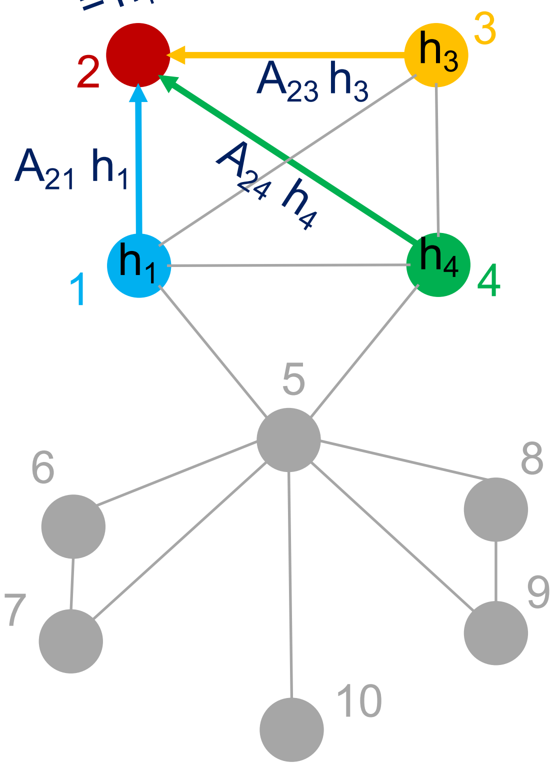
HITS equations

authorities score

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

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

$$a_2 = A_{21}\, h_1 + A_{23}\, h_3 + A_{24}\, h_4$$
$$= h_1 + h_3 + h_4$$



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

authority scores

hub scores

HITS equations

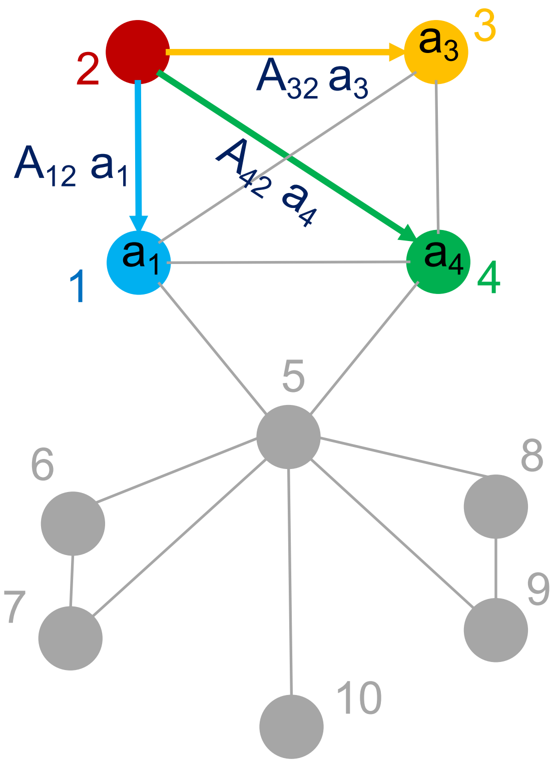
hubs score

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

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

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

$h_2 = A_{12}\, a_1 + A_{32}\, a_3 + A_{42}\, a_4$
$= a_1 + a_3 + a_4$



HITS equations

hubs and authorities

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

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

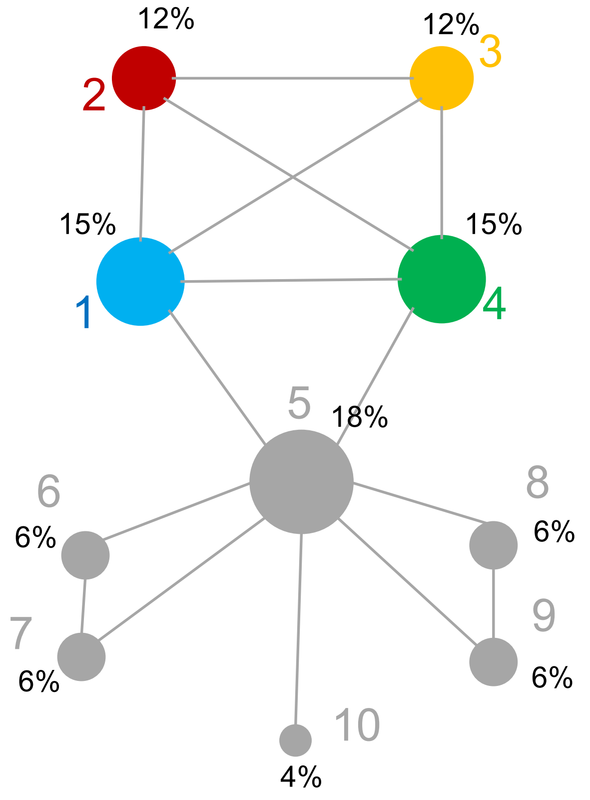
hubs

$$\boxed{\boldsymbol{h} = c\,\boldsymbol{M}\boldsymbol{h}} \quad \boldsymbol{M} = \boldsymbol{A}^T\boldsymbol{A}$$

$$c = c_a c_h$$

authorities

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



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

Power iteration method

for HITS

0. Start from an initial guess $\boxed{a_0}$

1. Let the time go by $\boxed{a_{t+1} = M\, a_t}$

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

2. Keep normalizing (divide a_{t+1} by the sum of elements)

3. Stop when a converges (few iterations?)

Convergence properties

for HITS

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

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 → 10.3

slow if λ_2 close to λ_1

10^{-3} precision → 7.6

Eigenvector and Kats centralities

other (less interesting) alternatives to PageRank

UNIVERSITÀ DEGLI STUDI DI PADOVA

Eigenvector and Kats centralities

an overview

	with constant term	without constant term
normalized	PageRank $\mathbf{r} = c\,\mathbf{M}\,\mathbf{r} + (1\text{-}c)\,\mathbf{q}$	Degree $\mathbf{r} = \mathbf{M}\,\mathbf{r}$
unnormalized	Katz $\mathbf{r} = c\,\mathbf{A}\,\mathbf{r} + \mathbf{1}$	Eigenvector $\mathbf{r} = c\,\mathbf{A}\,\mathbf{r}$

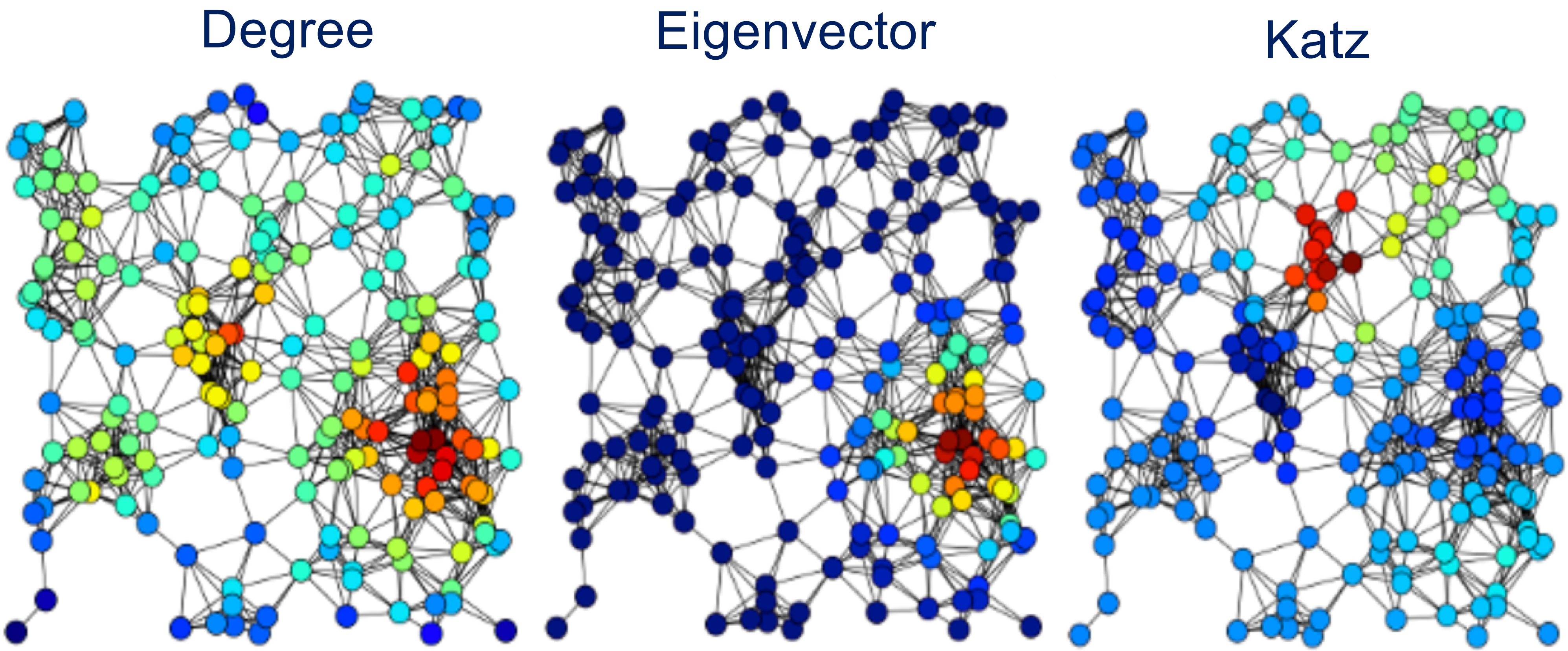
$$\mathbf{r} = (\mathbf{I} - c\,\mathbf{A})^{-1}\,\mathbf{1}$$
$$= \sum (c\,\mathbf{A})^k\,\mathbf{1}$$

The absence of normalization makes them less robust and meaningful compared to PageRank

They are deprecated

Eigenvector and Kats centralities

their graphical interpretation



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

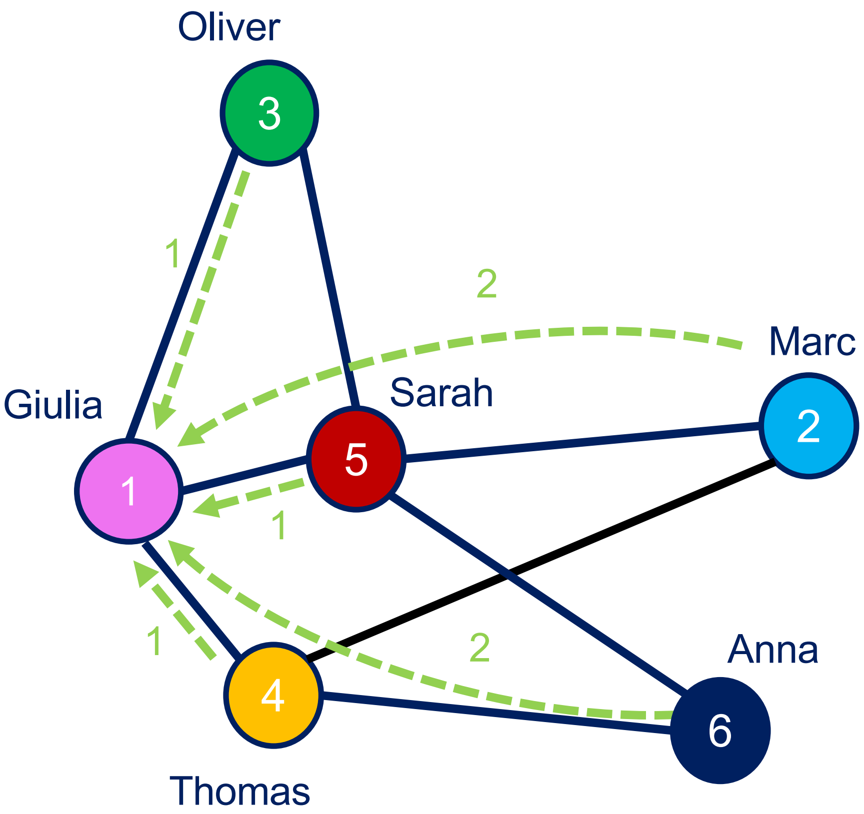
Rationale: the node which is the easiest to reach, the one which is the best for spreading information

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

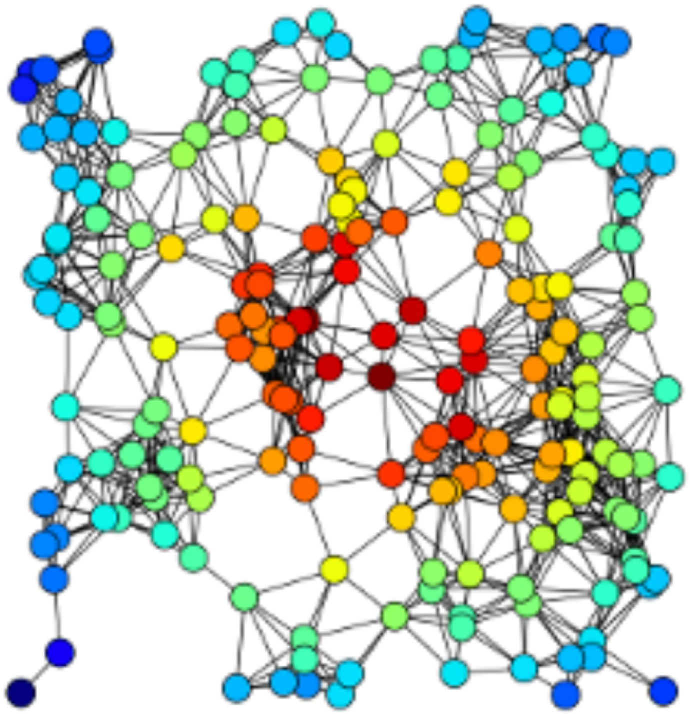
Sarah is the preferred node for spreading information

$$C(\text{Giulia}) = 1/7 = 0.1429$$

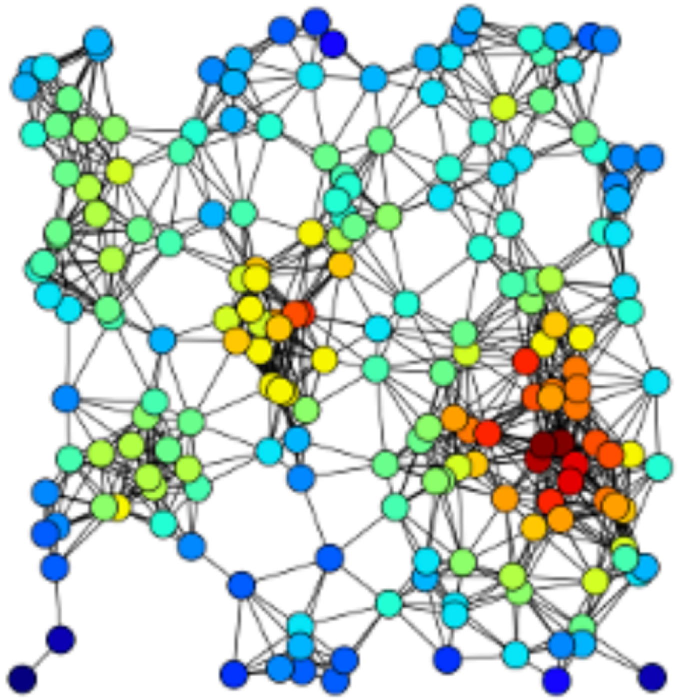
Closeness versus degree centrality

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 $1/\infty = 0$:

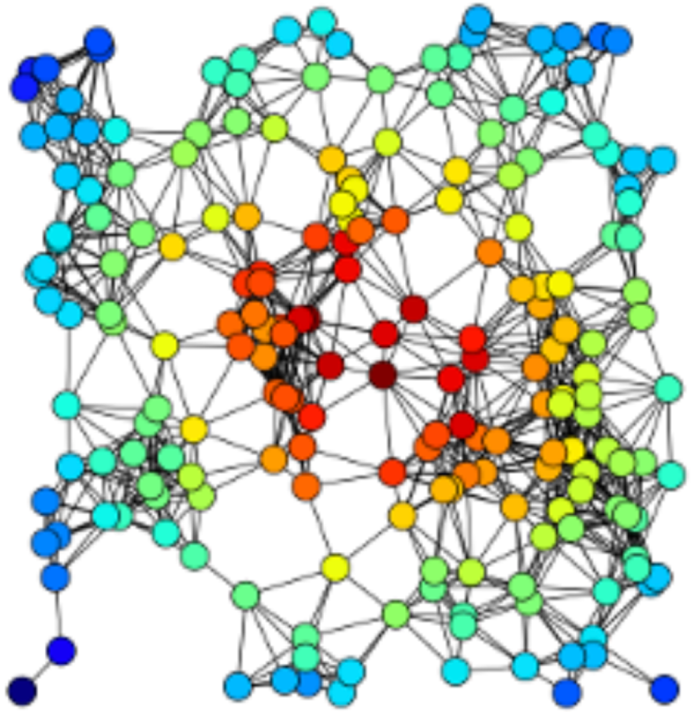
$$H(x) = \sum_{y \neq x} \frac{1}{d(y, x)}.$$

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.

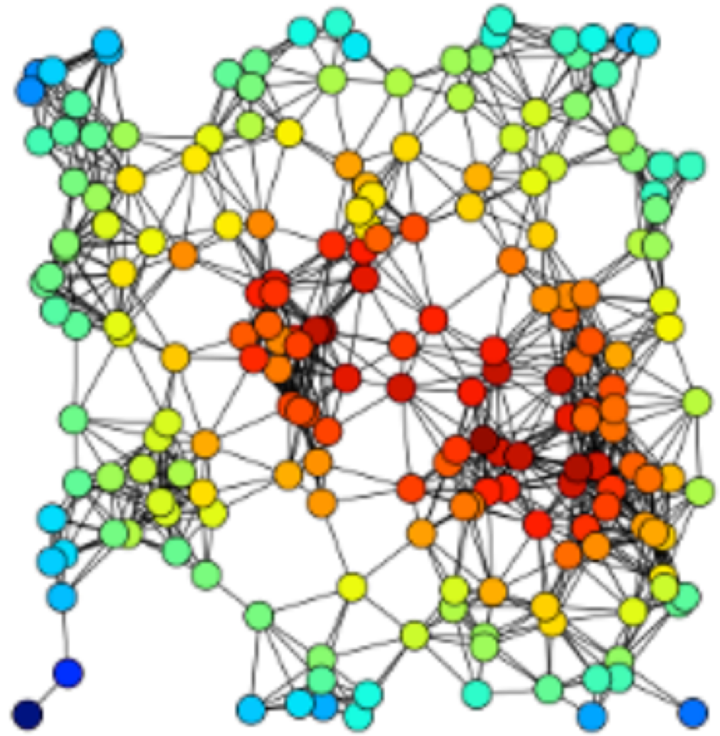
Closeness versus harmonic centrality

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.

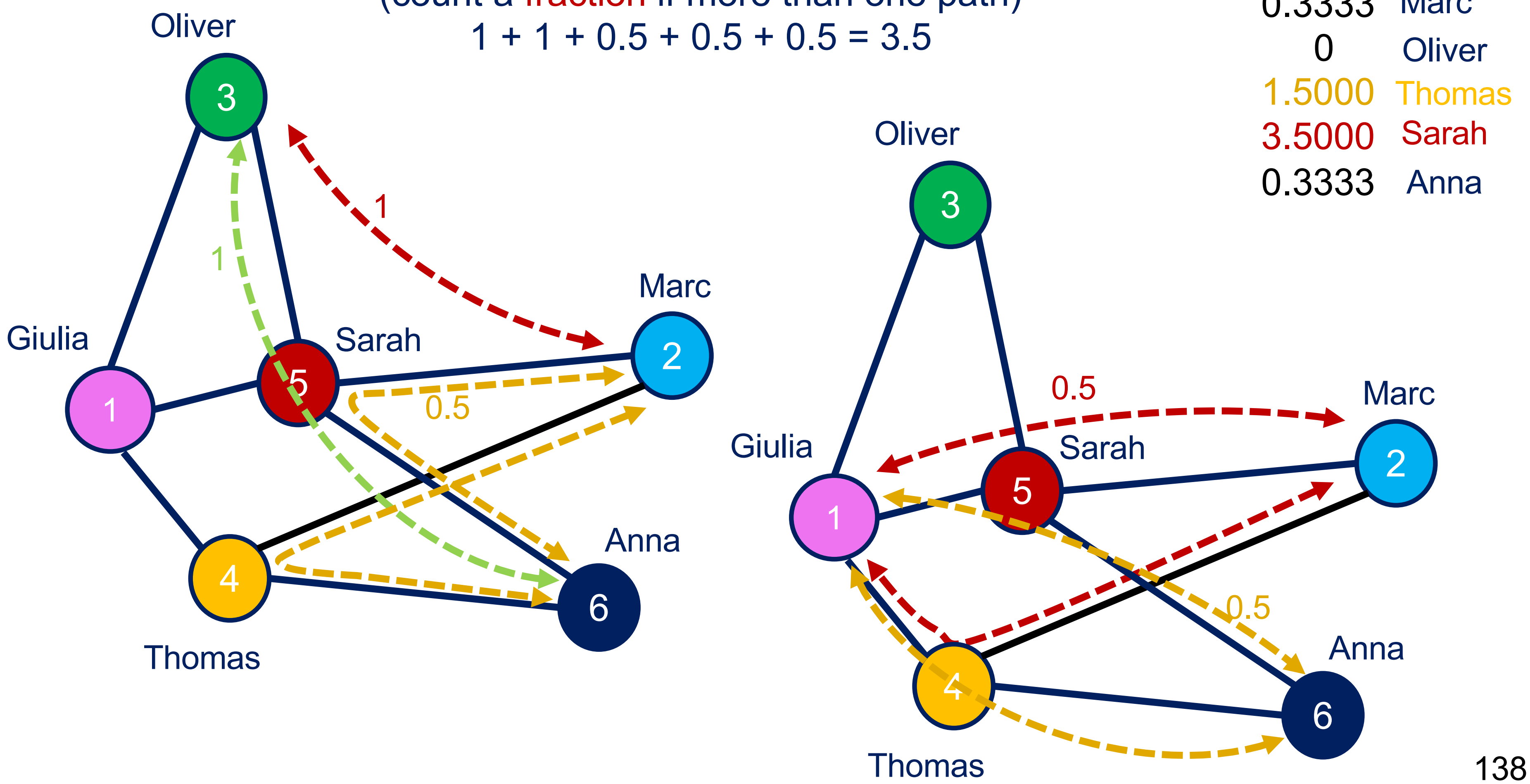
Betweenness centrality was devised as a general measure of centrality:[1] it applies to a wide range of problems in network theory, including problems related to social networks, biology, transport and scientific cooperation. Although earlier authors have intuitively described centrality as based on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



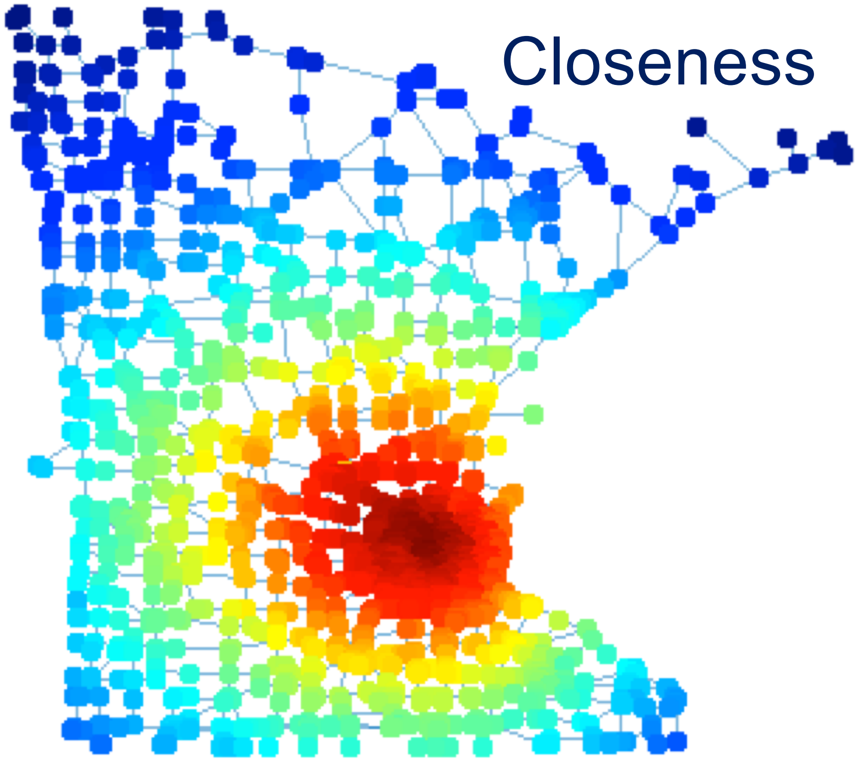
Betweenness

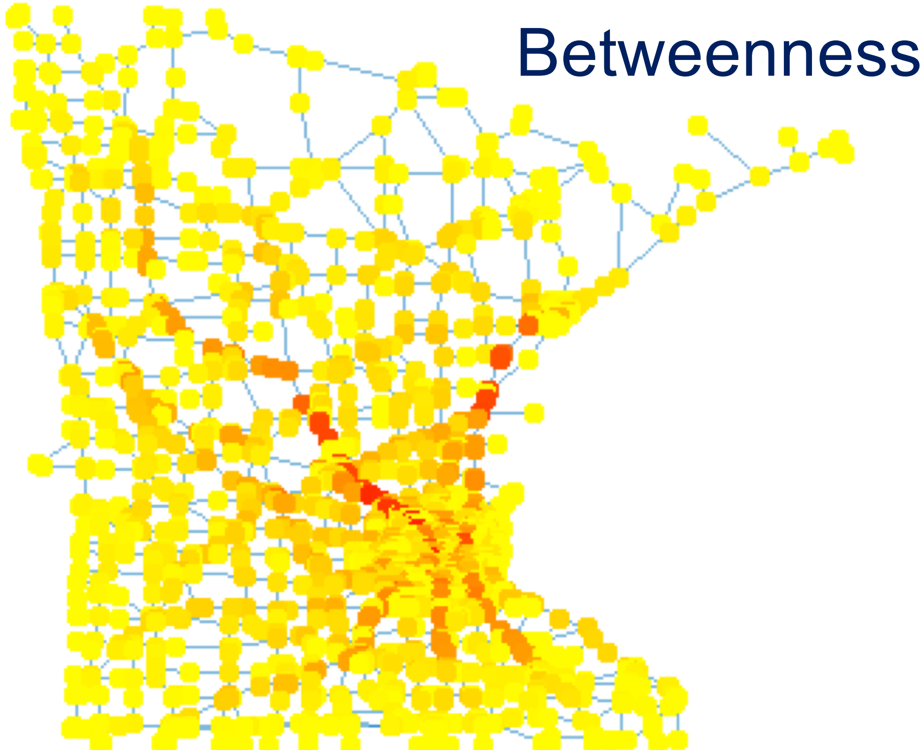
1.3333	Giulia
0.3333	Marc
0	Oliver
1.5000	Thomas
3.5000	Sarah
0.3333	Anna

Closeness vs betweenness centrality

a graphical interpretation

Minnesota road network





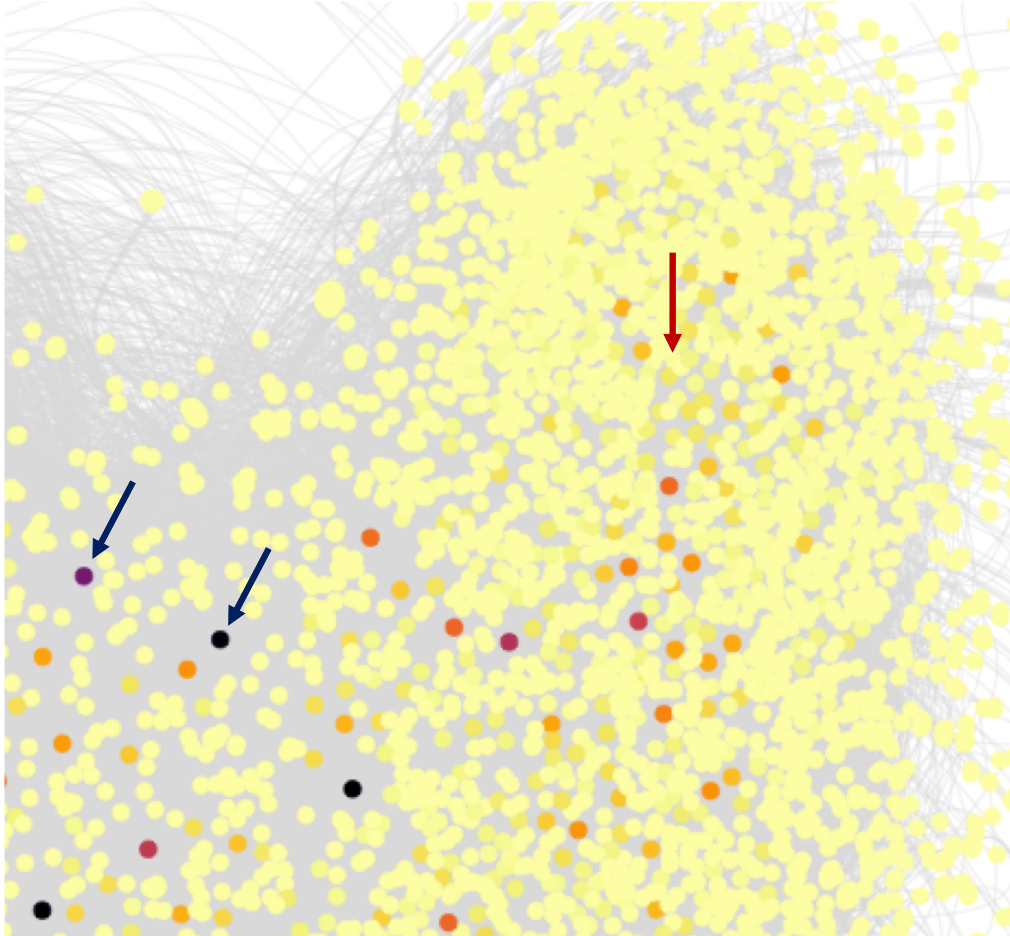
Closeness is a measure of center of gravity (best node to spread info)

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

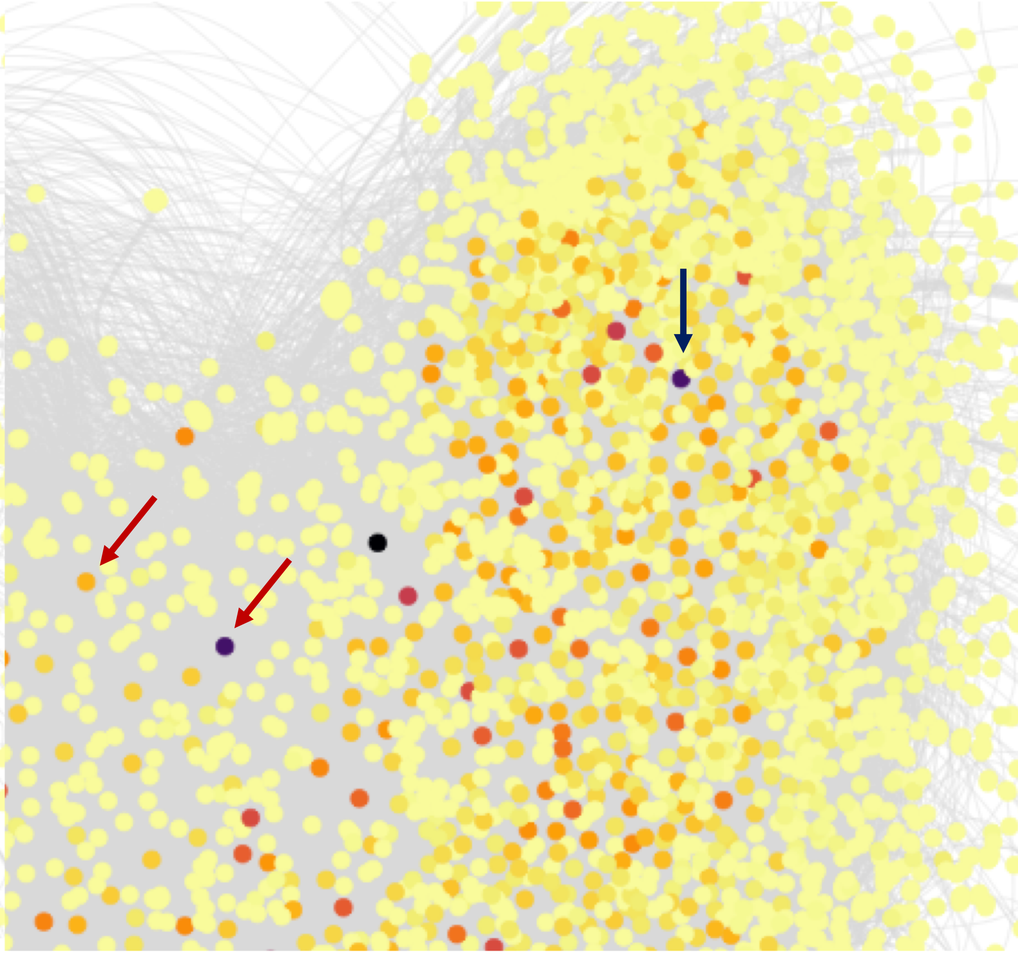
Betweenness vs PageRank centrality

wiki vote network

Betweenness

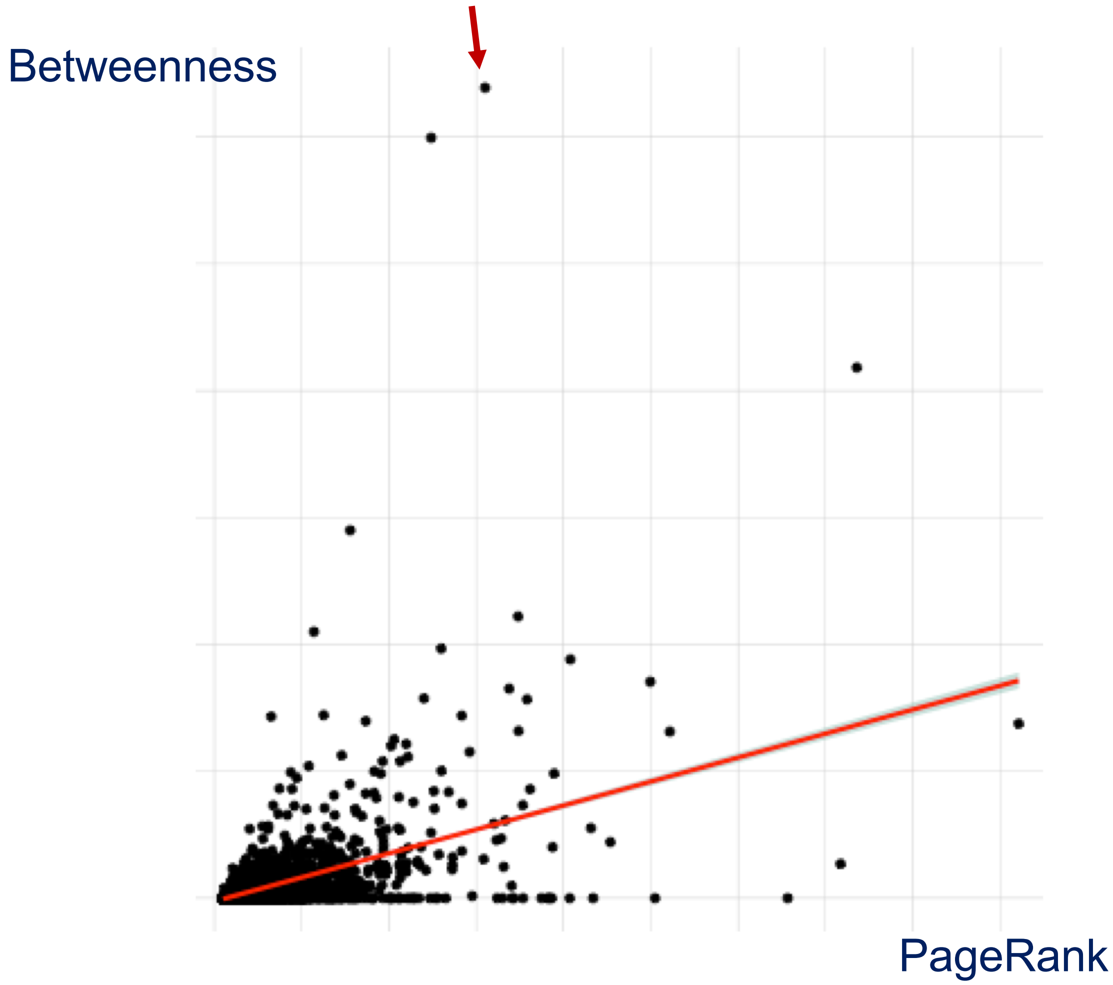


PageRank



Betweenness vs PageRank centrality

a correlation view



Clustering coefficient

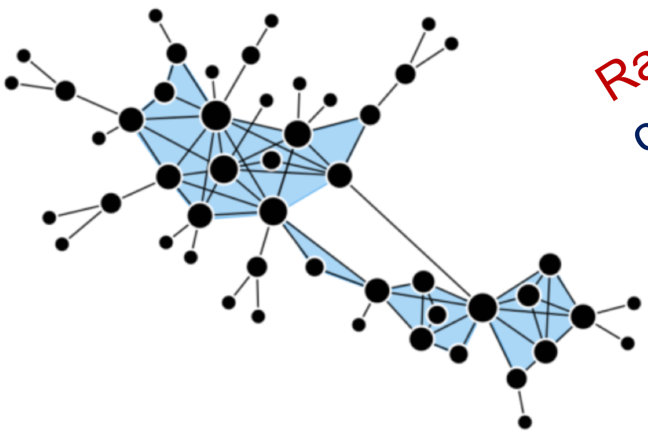
how tightly linked is the network locally

Clustering coefficient

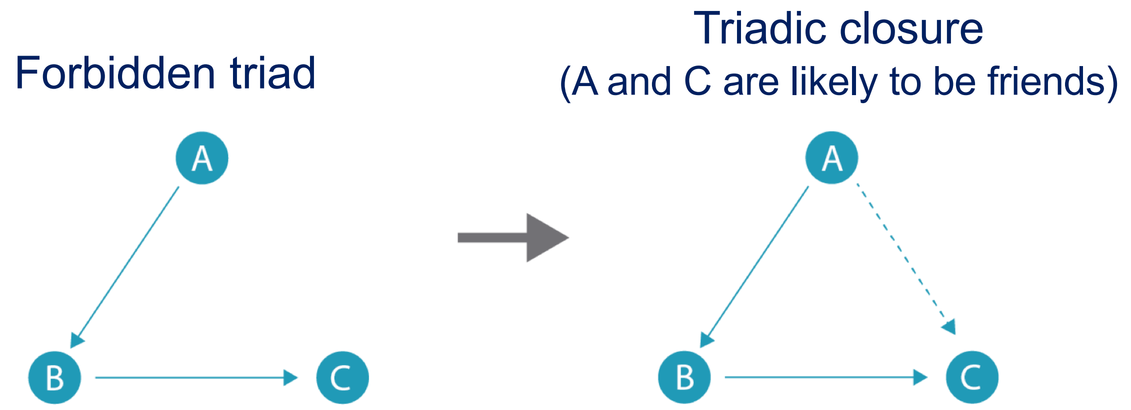
a definition

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



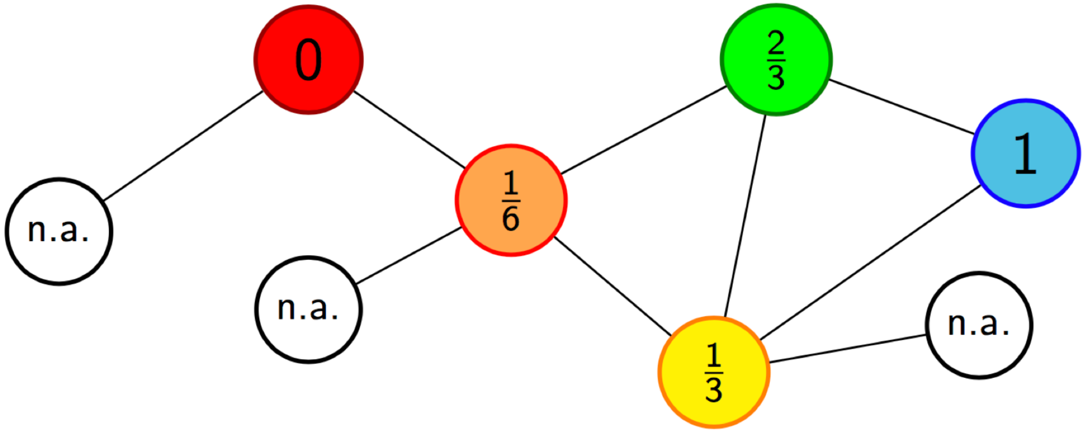
Triadic closure

- 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($\mathbf{A}^3$)

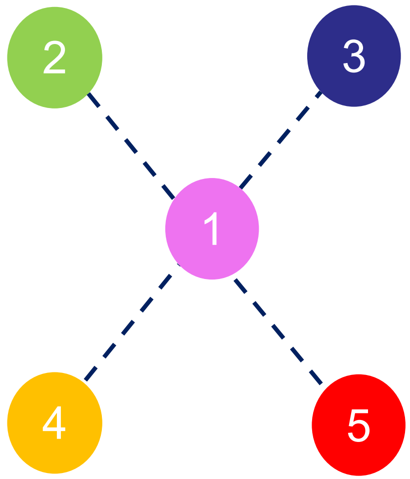
where $tc_{ijk} = 1$ if the triplet (i,j,k) forms a triadic closure, and zero otherwise

Local clustering coefficient

examples

not connected neighbourhood

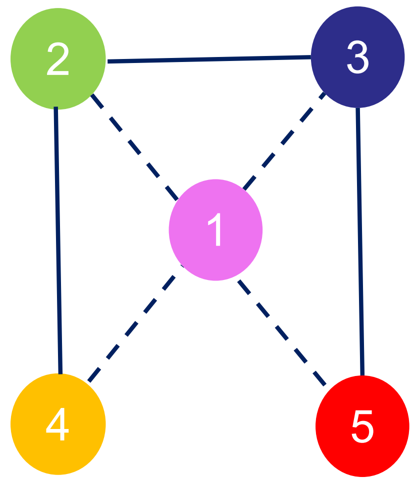
$\langle C \rangle = 0$



$C_1 = 0$

weakly connected neighbourhood

$\langle C \rangle = 0.766$



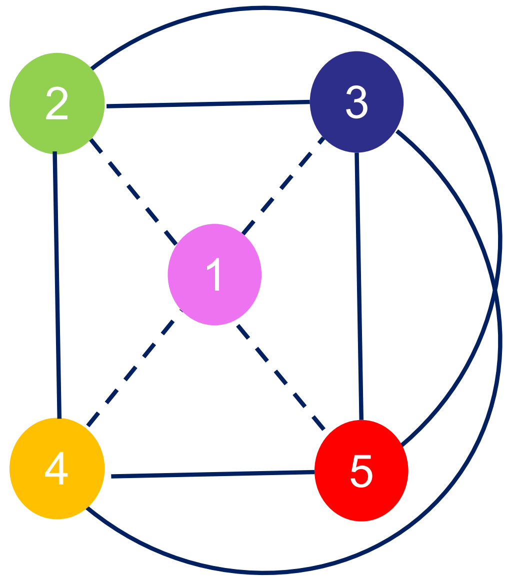
$C_1 = ½ = 3 / (4x3/2)$

$C_2 = C_3 = ⅔$

$C_4 = C_5 = 1$

strongly connected neighbourhood

$\langle C \rangle = 1$

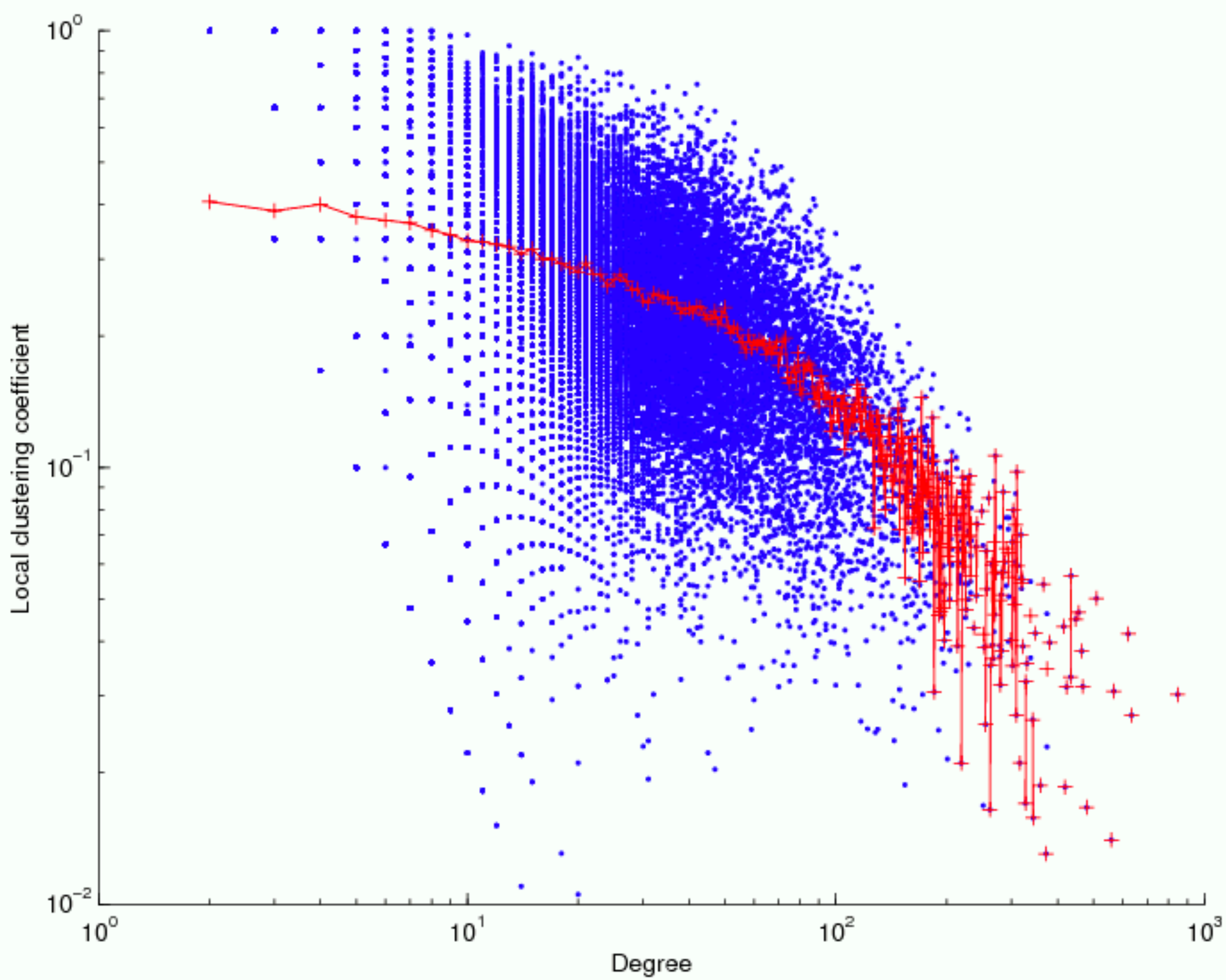


$C_1 = 1 = 6 / (4x3/2)$

Clustering coeff. vs degree centrality

a correlation view

citation network from arXiv's High Energy Physics / Phenomenology section

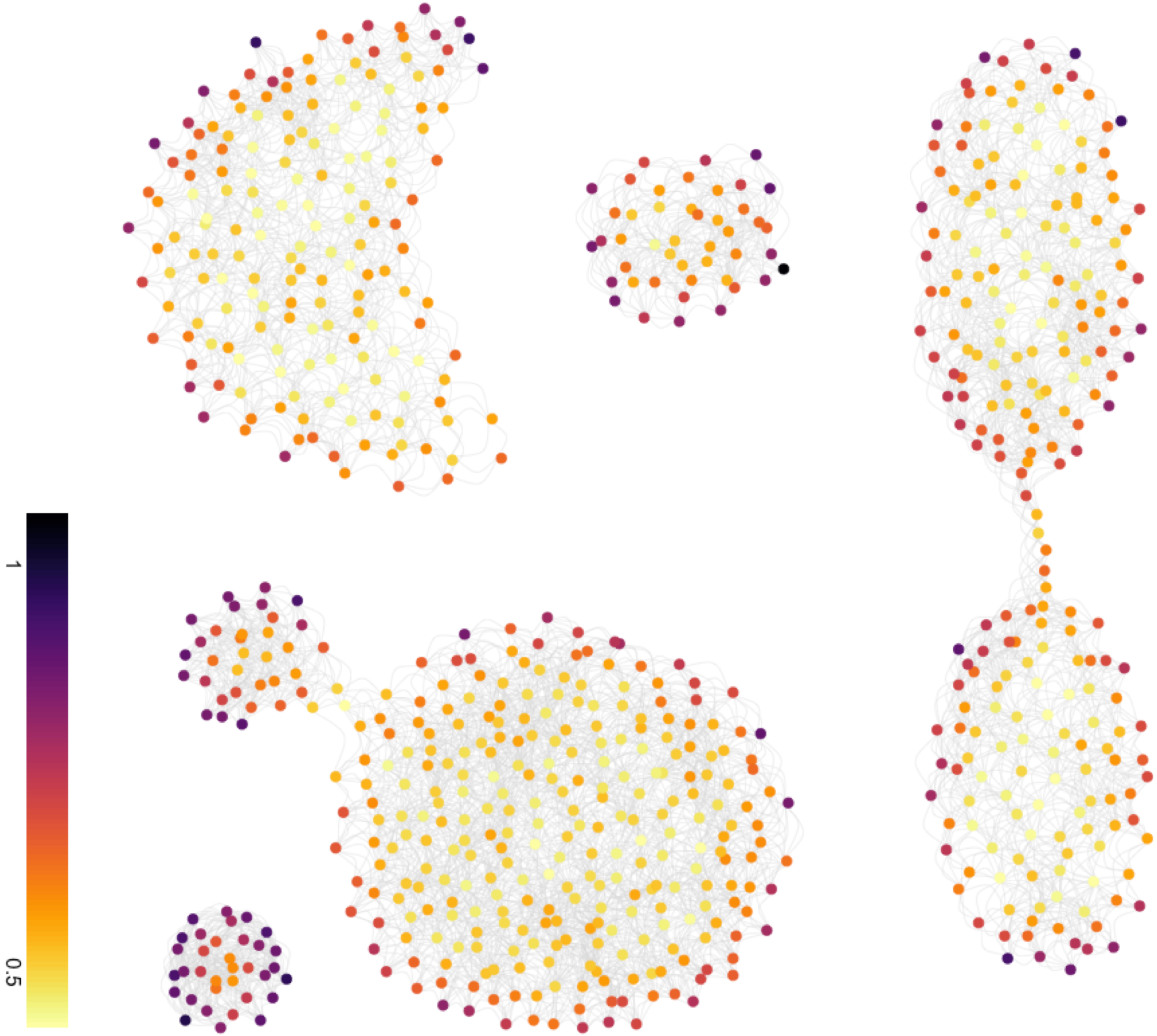


when person has many friends, these friends have less edges among them, which is to be expected since a person with many friends is likely to have friends from more diverse communities, and a paper getting cited many times is likely to be cited by papers from more diverse areas

Warning

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

Visual example



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

Wrap-up

on centrality measures

Takeaways
on centrality measures

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

More on the meaning

https://reticular.hypotheses.org/1745