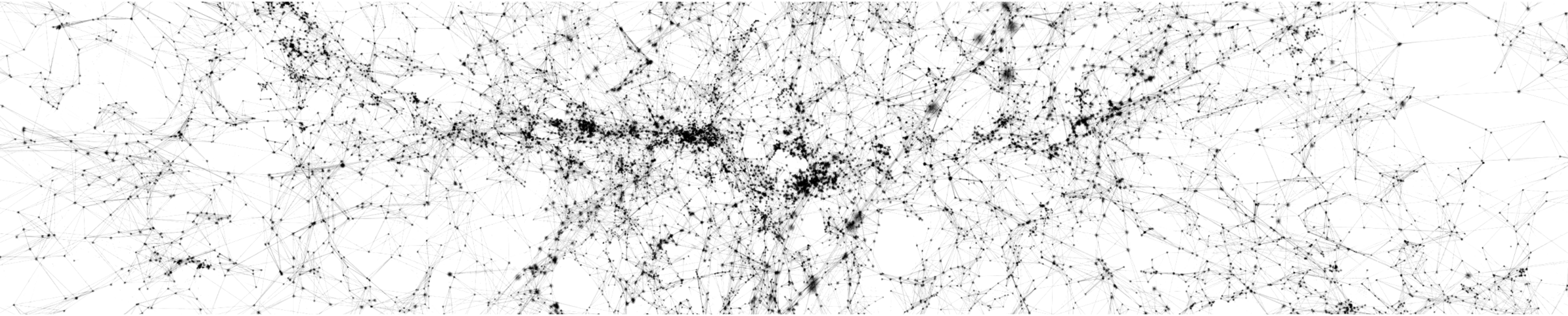
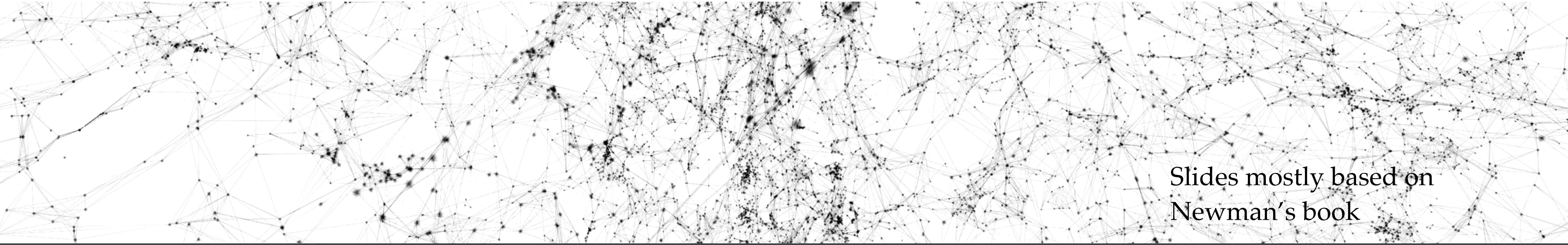


Measures

Analysis of complex interconnected data



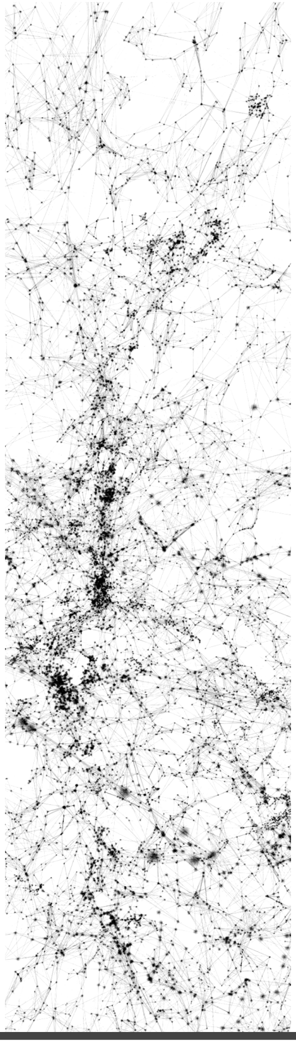
Slides mostly based on
Newman's book

Outline

- **Centrality**
  - Degree Centrality
  - Eigenvalue Centrality
  - Katz Centrality
  - PageRank
  - HITS
  - Closeness centrality
  - Betweenness centrality
- Similarity
  - Common neighbour
  - Cosine similarity
  - Jaccard similarity

$$R : v \mapsto \mathbb{R}$$

$$S : (u, v) \mapsto \mathbb{R}$$

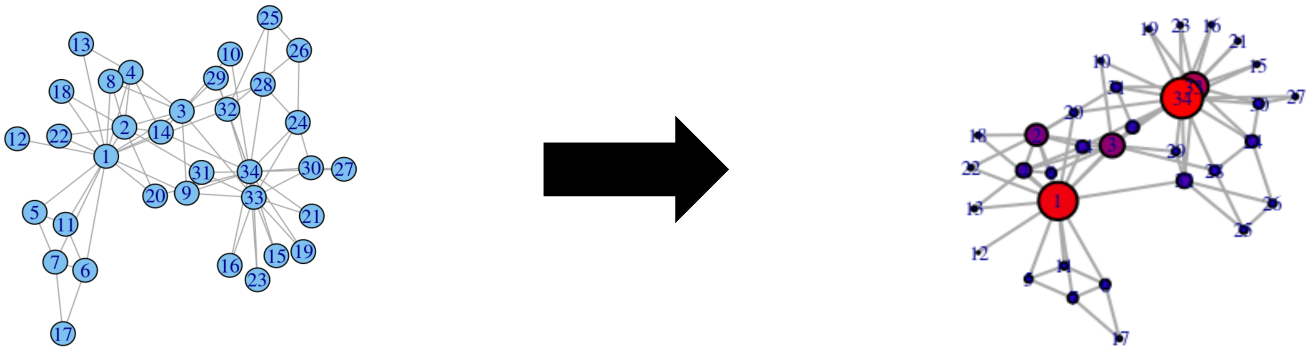


Centrality

Measure the importance of nodes:

$$R : v \mapsto \mathbb{R}$$

maps each node to a value such that ranking by these values ranks the nodes by their importance



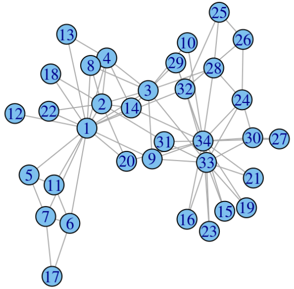
http://www.rpubs.com/shestakoff/sna_lab4

Centrality

Different ways to define importance ⇒

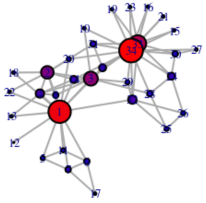
Different centrality measures ⇒

Different ranking of the nodes on the same graph

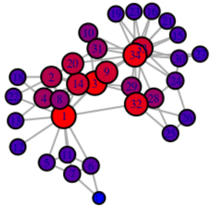


http://www.rpubs.com/shestakoff/sna_lab4

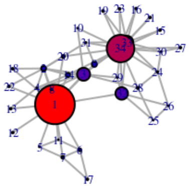
Degree centrality



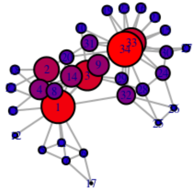
Closeness centrality



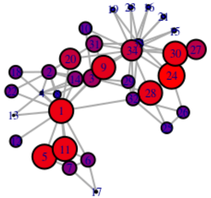
Betwenness centrality



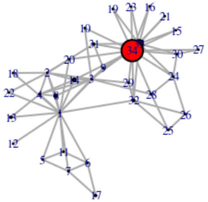
Eigenvector centrality



Bonachich power centrality

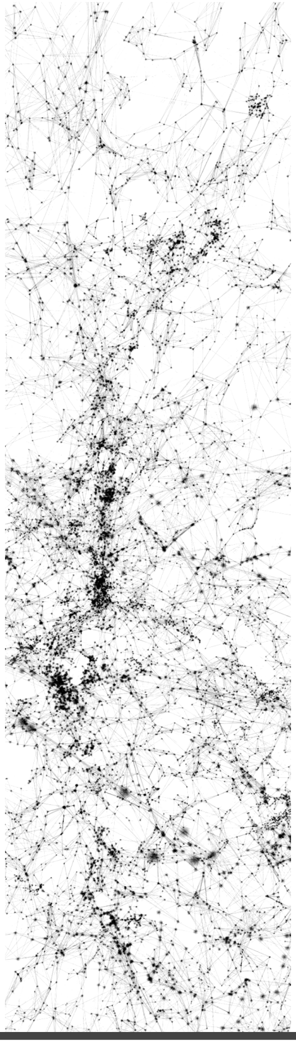


Alpha centrality



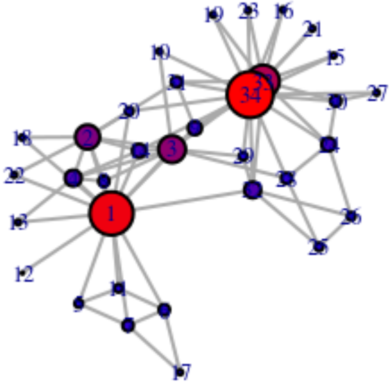
Outline

- Centrality
    - **Degree Centrality**
    - Eigenvalue Centrality
    - Katz Centrality
    - PageRank
    - HITS
    - Closeness centrality
    - Betweenness centrality
- Similarity
    - Common neighbour
    - Cosine similarity
    - Jaccard similarity



Degree centrality

Degree is the simplest centrality measure

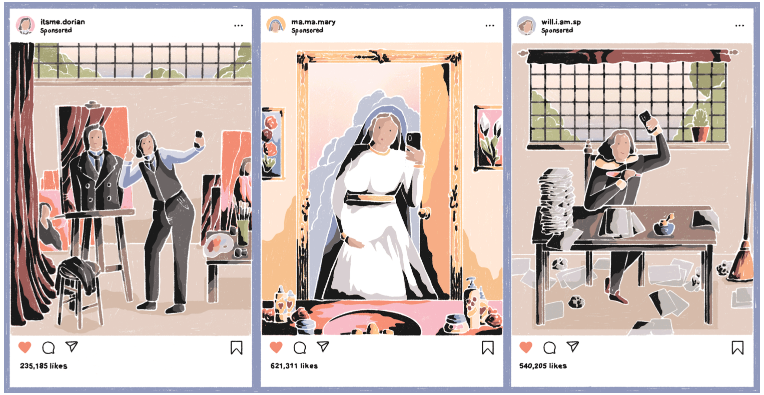


more connections you have (number of edges),
more people you know (number of neighbours),
more important you are

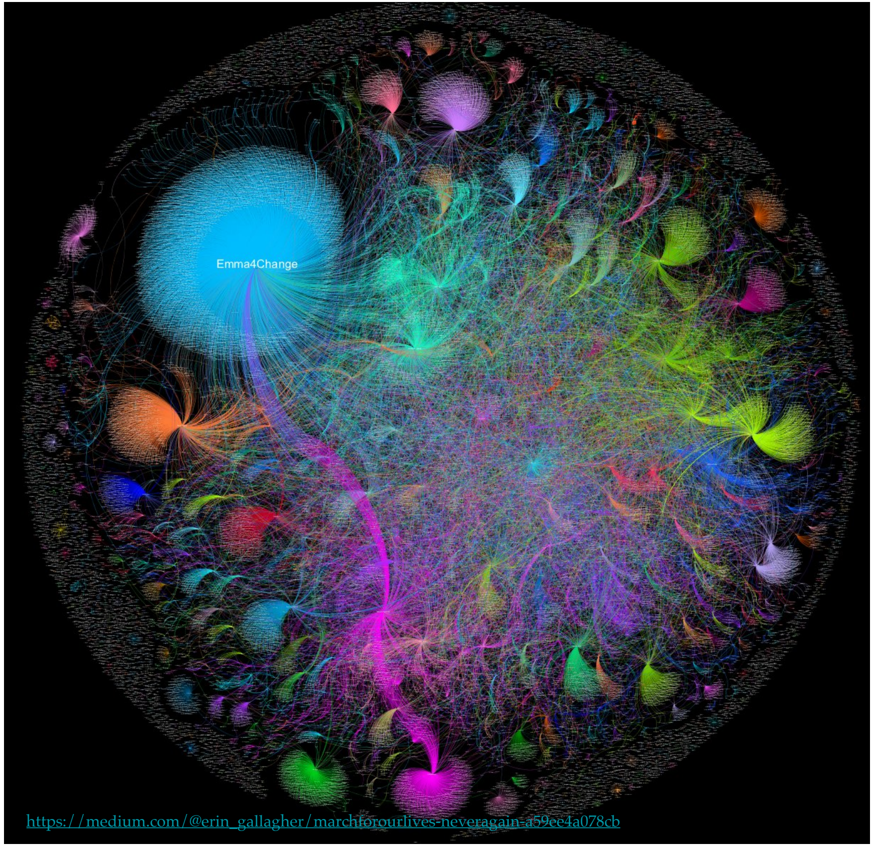
Can you think of a widely used example where people are ranked by degree centrality?

Degree centrality, example

Influencers in social media: number of followers, number of retweets

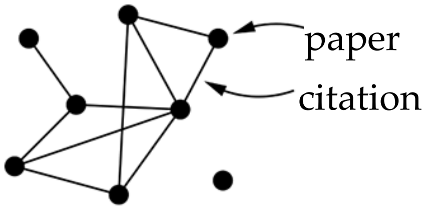


https://www.newyorker.com/culture/annals-of-inquiry/a-history-of-the-influencer-from-shakespeare-to-instagram

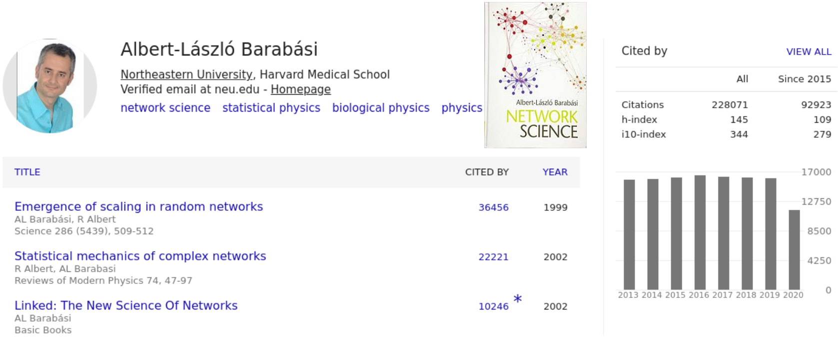


https://medium.com/@erin_gallagher/marchforourlives-neveragain-a59ee4a078cb

Degree centrality, example



Important papers: number of citations, number of time a paper is cited



Albert-László Barabási

Northeastern University, Harvard Medical School
Verified email at neu.edu - Homepage
network science   statistical physics   biological physics   physics

TITLE	CITED BY	YEAR
Emergence of scaling in random networks AL Barabási, R Albert Science 286 (5439), 509-512	36456	1999
Statistical mechanics of complex networks R Albert, AL Barabasi Reviews of Modern Physics 74, 47-97	22221	2002
Linked: The New Science Of Networks AL Barabási Basic Books	10246 *	2002

Cited by   VIEW ALL

	All	Since 2015
Citations	228071	92923
h-index	145	109
i10-index	344	279

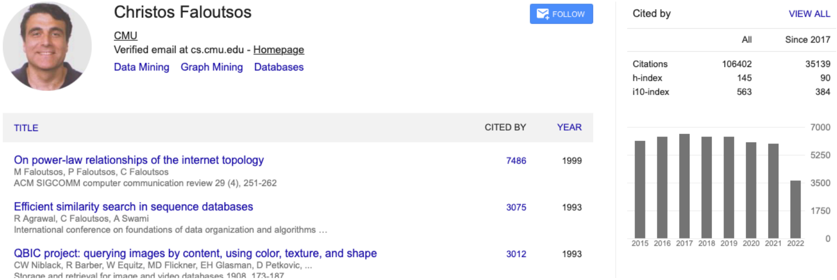
Mark Newman

Professor of Physics, University of Michigan
Verified email at umich.edu - Homepage
Statistical Physics   Networks

TITLE	CITED BY	YEAR
The structure and function of complex networks MEJ Newman SIAM review 45 (2), 167-256	20389	2003
Community structure in social and biological networks M Girvan, MEJ Newman Proceedings of the national academy of sciences 99 (12), 7821-7826	14555	2002
Finding and evaluating community structure in networks MEJ Newman, M Girvan Physical review E 69 (2), 026113	13191	2004

Cited by   VIEW ALL

	All	Since 2015
Citations	189890	90313
h-index	105	81
i10-index	203	174



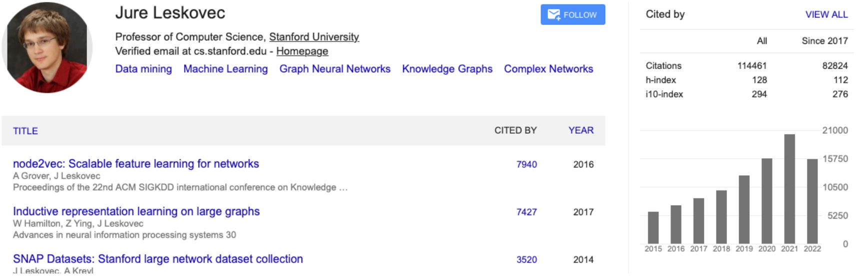
Christos Faloutsos

CMU
Verified email at cs.cmu.edu - Homepage
Data Mining   Graph Mining   Databases

TITLE	CITED BY	YEAR
On power-law relationships of the internet topology M Faloutsos, P Faloutsos, C Faloutsos ACM SIGCOMM computer communication review 29 (4), 251-262	7486	1999
Efficient similarity search in sequence databases R Agrawal, C Faloutsos, A Swami International conference on foundations of data organization and algorithms …	3075	1993
QBIC project: querying images by content, using color, texture, and shape CW Niblack, R Barber, W Equitz, MD Flickner, EH Glasman, D Petkovic, ... Storage and retrieval for image and video databases 1908, 173-187	3012	1993

Cited by   VIEW ALL

	All	Since 2017
Citations	106402	35139
h-index	145	90
i10-index	563	384



Jure Leskovec

Professor of Computer Science, Stanford University
Verified email at cs.stanford.edu - Homepage
Data mining   Machine Learning   Graph Neural Networks   Knowledge Graphs   Complex Networks

TITLE	CITED BY	YEAR
node2vec: Scalable feature learning for networks A Grover, J Leskovec Proceedings of the 22nd ACM SIGKDD international conference on Knowledge …	7940	2016
Inductive representation learning on large graphs W Hamilton, Z Ying, J Leskovec Advances in neural information processing systems 30	7427	2017
SNAP Datasets: Stanford large network dataset collection J Leskovec, A Krevl	3520	2014

Cited by   VIEW ALL

	All	Since 2017
Citations	114461	82824
h-index	128	112
i10-index	294	276

Eigenvector centrality

How to measure having important connections?

You might only have one connection but it can be the president, or the king

Eigenvector centrality

Eigenvector centrality of a node is proportional to the centrality scores of its neighbours

Assume x_i gives the importance of node i, and $N(i)$ gives set of neighbours of i

$$N(i) = \{j \mid A_{ij} = 1\}$$

$$x_i = \frac{1}{\kappa} \sum_{j \in N(i)} x_j$$

You are important if you have **many connections** (of some importance), or a few but **very important connections**

Eigenvector centrality

Eigenvector centrality of a node is proportional to the centrality scores of its neighbours

Assume x_i gives the importance of node i, and $N(i)$ gives set of neighbours of i

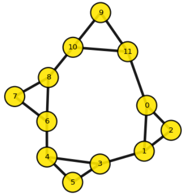
$$N(i) = \{j \,|\, A_{ij} = 1\}$$

$$x_i = \frac{1}{\kappa} \sum_{j \in N(i)} x_j$$

How can we write this in matrix notation?

Note that we have $\sum_{j \in N(i)} x_i = A_{i:}\, x$ here x is a vector of all centrality scores

Example



$$\begin{array}{c|cccccccccccc|c}
 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & \\
\hline
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 3 \\
1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \\
2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\
3 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \\
4 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 3 \\
5 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\
6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 3 \\
7 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 2 \\
8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 3 \\
9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 2 \\
10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 3 \\
11 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 3 \\
\hline
 & 3 & 3 & 2 & 3 & 3 & 2 & 3 & 2 & 3 & 2 & 3 & 3 &
\end{array} \quad A_{4:}$$

$$x^\top = \begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \hline 3 & 3 & 2 & 3 & 3 & 2 & 3 & 2 & 3 & 2 & 3 & 3 \end{array}$$

$$i = 4 \quad \Rightarrow \quad \sum_{j \in N(4)} x_j = A_{4:}\, x = 8$$

Eigenvector centrality

Eigenvector centrality of a node is proportional to the centrality scores of its neighbours

Assume x_i gives the importance of node i, and $N(i)$ gives set of neighbours of i

$$N(i) = \{j \mid A_{ij} = 1\}$$

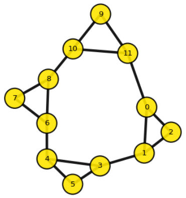
$$x_i = \frac{1}{\kappa} \sum_{j \in N(i)} x_j$$

How can we write this in matrix notation?

here x is a vector of all centrality scores

$$x = \frac{1}{\kappa} A x \quad \Rightarrow \quad \kappa x = A x$$

Example



$$\begin{array}{c|cccccccccccc|c} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & \\ \hline 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 3 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \\ 2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 3 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \\ 4 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 3 \\ 5 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 3 \\ 7 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 2 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 3 \\ 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 2 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 3 \\ 11 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 3 \\ \hline & 3 & 3 & 2 & 3 & 3 & 2 & 3 & 2 & 3 & 2 & 3 & 3 & \end{array}$$

$A_{4:}$

$$x^\top = \begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \hline 3 & 3 & 2 & 3 & 3 & 2 & 3 & 2 & 3 & 2 & 3 & 3 \end{array}$$

$$Ax = \begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \hline 8 & 8 & 6 & 8 & 8 & 6 & 8 & 6 & 8 & 6 & 8 & 8 \end{array}^\top$$

Eigenvector centrality

Eigenvector centrality of a node is proportional to the centrality scores of its neighbours

Assume x_i gives the importance of node i, and $N(i)$ gives set of neighbours of i

$$x_i = \frac{1}{\kappa} \sum_{j \in N(i)} x_j$$

Which in matrix notation is: $\kappa x = A x$

What is x ? an eigenvector of the adjacency matrix

Which eigenvector should we use?
we want **x** to be non-negative then the only choice is the **leading eigenvector**

What is κ ? largest eigenvalue

[Perron–Frobenius theorem]
Any matrix with all non-negative values, such as A, any eigenvector but the leading eigenvector has at least one negative element.

Eigenvector centrality

Eigenvector centrality of a node is proportional to the centrality scores of its neighbours

Assume x_i gives the importance of node i, and $N(i)$ gives set of neighbours of i

$$x_i = \frac{1}{\kappa} \sum_{j \in N(i)} x_j$$

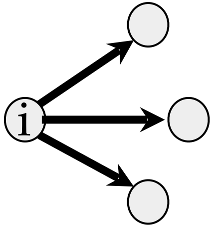
Which in matrix notation is: $\kappa x = Ax$

x is the **eigenvector** corresponding to the largest eigenvalue

Eigenvector centrality ranks the likelihood that a node is visited on a random walk of infinite length on the graph why?

Leading eigenvector is computed with power iteration, $x^{(i+1)} = Ax^{(i)}$, whereas A^k gives number of walks of length k {more on this later}

Eigenvector centrality in directed networks



$$x_i = \kappa^{-1} \sum_j A_{ji} x_j$$

$$\kappa x = xA$$

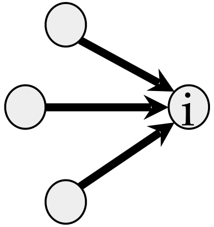
[left]

Can be defined in two ways
⇒ right and left eigenvectors,

and two leading eigenvalues

Which one to use?

Consider the citation network and the www, which one indicates importance?



$$x_i = \kappa^{-1} \sum_j A_{ij} x_j$$

$$\kappa x = Ax$$

[right] ✓

$A_{ij} = 1$if there is an edge from j to i

Eigenvector centrality in directed networks

$$x_i = \kappa^{-1} \sum_j A_{ij} x_j$$

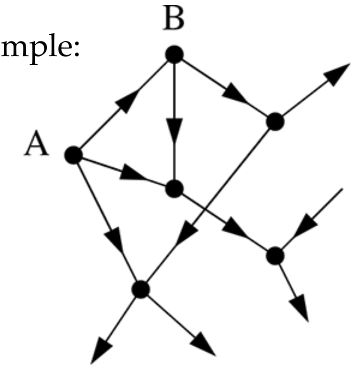
Nodes have non-zero score only if in a strongly connected component of two or more nodes, or the out-component of such a strongly connected component

When will this be a problem? Can you think of an example?

In an **acyclic networks**, such as **citation networks**, where there is no strongly connected components (of more than one node) and all nodes get zero score

How can we fix it? Katz and PageRank variants

Example:



What is the score of A?

a node with no incoming edge

⇒ zero score

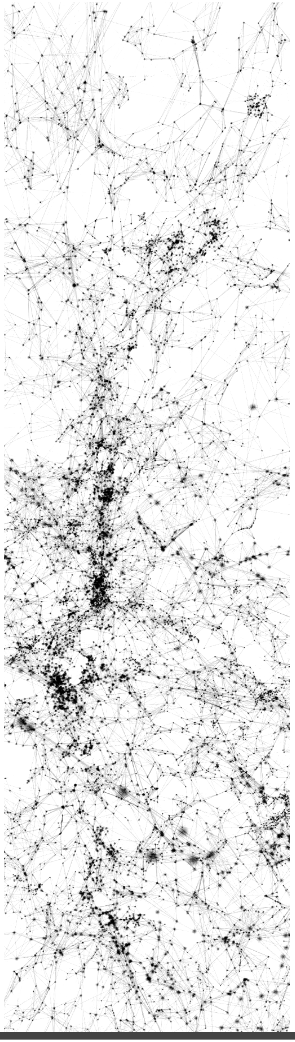
What is the score of B?

only ingoing edge is from A

⇒ zero score

Outline

- Centrality
    - Degree Centrality
    - Eigenvalue Centrality
    - **Katz Centrality**
    - PageRank
    - HITS
    - Closeness centrality
    - Betweenness centrality
- Similarity
    - Common neighbour
    - Cosine similarity
    - Jaccard similarity



Katz centrality

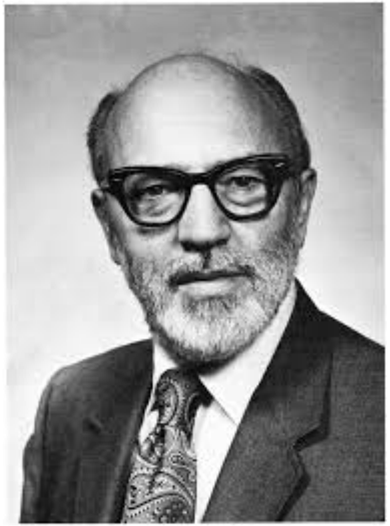
$$x_i = \alpha \sum_j A_{ij} x_j + \beta$$

α and β are positive constants

β **:** *every node gets a basic importance*

"everybody is somebody"

Nodes with zero in-degree gets β and can pass it on ⇒ nodes with high in-degree get high score regardless of being in SCC or pointed by it



Leo Katz (1914-1976)
1953 - Katz centrality

Katz centrality

$$x_i = \alpha \sum_j A_{ij} x_j + \beta$$

$$x = \alpha A x + \beta \, 1$$

$$x = \beta (I - \alpha A)^{-1} 1$$

$$x = (I - \alpha A)^{-1} 1$$ {with β =1}

absolute magnitude of centrality scores are not important, we care about the relative values, so β multiplier is not important

1 is the uniform vector of all ones

I is the identity matrix with diagonal of 1

Example:

$$1 = \begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}$$

$$I = \begin{array}{c|cccccccccccc} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 11 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

What do we get if we set $\alpha = 0$?

Katz centrality

$$x = (I - \alpha A)^{-1} 1$$

What do we get if we set $\alpha = 0$? All nodes have the same importance as β

As we increase α, scores increase and might start to diverge, which happens when

$$det(I - \alpha A) = 0 \Rightarrow det(\alpha^{-1} I - A) = 0 \quad \Rightarrow \quad \alpha^{-1} = \lambda_i \Rightarrow \alpha = 1/\lambda_i$$

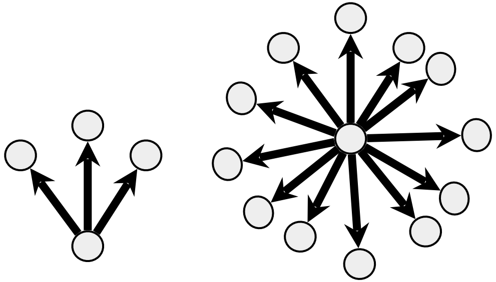
At what α this first happens? largest (most positive) eigenvalue, a.k.a. principal/dominant eigenvalues

To converge then we need: $\alpha < 1/\lambda_1$ In practice α is often set close to this limit

This places the maximum amount of weight on the eigenvector term and the smallest amount on the constant term

The determinant of a matrix is equal to the product of its eigenvalues, and matrix $cI - A$ has eigenvalues $c - \lambda_i$ where λ_i are the eigenvalues of $A \Rightarrow det(cI - A) = (c - \lambda_1)(c - \lambda_2)\ldots(c - \lambda_n)$, with zeros at $c = \lambda_1, \lambda_2, \ldots$, therefore the solutions of $det(cI - A) = 0$ is the eigenvalues of A

Katz centrality



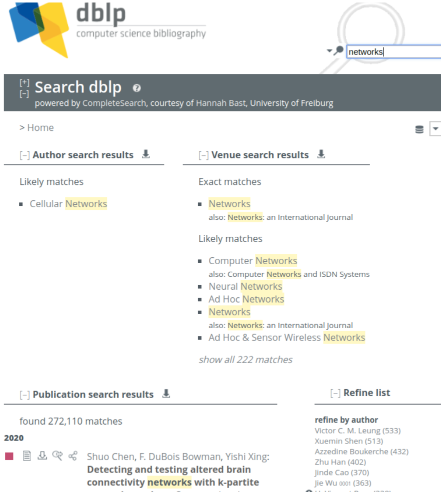
$$x = (I - \alpha A)^{-1}\mathbf{1}$$

$$\alpha < 1/\lambda_1$$

Could this be a good measure to rank pages in the www?

If there is an important directory page, linking to many pages, it passes its importance to all the cited web pages, one can think that the importance should be diluted if shared with many others

amazon.ca prime



dblp
computer science bibliography

networks

[+] [–] Search dblp
powered by CompleteSearch, courtesy of Hannah Bast, University of Freiburg

> Home

[–] **Author search results**

Likely matches

- Cellular Networks

[–] **Venue search results**

Exact matches

- Networks
  also: Networks: an International Journal

Likely matches

- Computer Networks
  also: Computer Networks and ISDN Systems
- Neural Networks
- Ad Hoc Networks
- Networks
  also: Networks: an International Journal
- Ad Hoc & Sensor Wireless Networks

show all 222 matches

[–] **Publication search results**

found 272,110 matches

2020

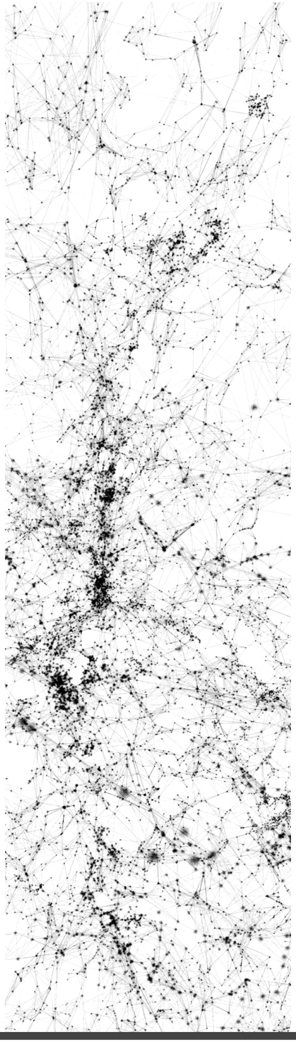
Shuo Chen, F. DuBois Bowman, Yishi Xing: **Detecting and testing altered brain connectivity networks with k-partite**

[–] **Refine list**

refine by author
Victor C. M. Leung (533)
Xuemin Shen (513)
Azzedine Boukerche (432)
Zhu Han (402)
Jinde Cao (370)
Jie Wu 0001 (363)

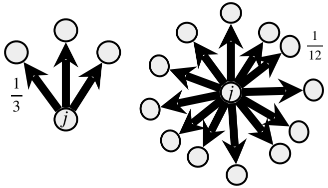
Outline

- Centrality
  - Degree Centrality
  - Eigenvalue Centrality
  - Katz Centrality
  - **PageRank**
  - HITS
  - Closeness centrality
  - Betweenness centrality
- Similarity
  - Common neighbour
  - Cosine similarity
  - Jaccard similarity



PageRank

divide your centrality to your neighbours,
instead of passing to all



$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{d_j^{out}} + \beta \qquad d_j^{out} = \sum_k A_{kj}$$

$$x = \alpha A D^{-1} x + \beta 1$$

$$x = (I - \alpha A D^{-1})^{-1} 1$$

What should α be?

$\alpha <$ *the leading eigenvalue of* AD^{-1}

which is 1 for undirected network but changes for directed ones

The Google search engine uses a value of α=0.85

to avoid 0/0 when $d_j^{out} = 0$ => $D_{ii} = max(d_j^{out},1)$

$d_j^{out} = 0$ when A_{ij}=0 for all *i* then A_{ij} / d_j^{out} = 0/0 , which we want to be 0

Brin, S. and Page, L., The anatomy of a large-scale hypertextual Web search engine, Comput. Netw. 30, 107–117 (1998).

PageRank: iterative algorithm

Relates to power iteration method which computes eigenvalues & eigenvectors of any matrix

Start with equal rank for all

$$x^{(0)} = [\frac{1}{n}, \frac{1}{n} \dots \frac{1}{n}]$$

Update the scores

$$x^{(t+1)} = Mx^{(t)}$$

$$x_i^{(t+1)} = \sum_j A_{ij} \frac{x_j^{(t)}}{d_j^{out}} = \sum_{j \to i} \frac{x_j^{(t)}}{d_j^{out}}$$

Repeat until convergence

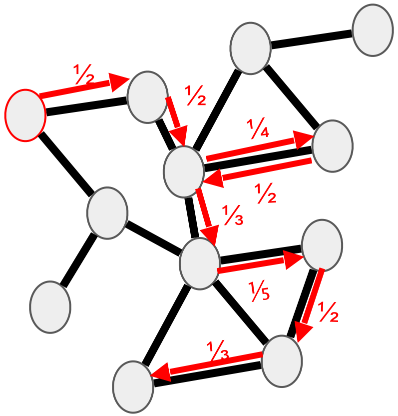
$$\|x^{(t+1)} - x^{(t)}\| < \epsilon$$

What is M? $AD^{-1}, M_{ij} = \frac{A_{ij}}{d_j^{out}}$

M is a column stochastic matrix and columns sum to 1
Therefore the largest eigenvalue of M is 1.

This converges to the leading eigenvalue, for the detailed PageRank version with α see this

PageRank: connection to random walk



A surfer walks on a graph:

At time t, he is at node i and moves out of it through one of the outlinks of i chosen uniformly at random, and ends up at neighbour j, repeats from j at time $t + 1$

$$x_i^{(t+1)} = \sum_{j \to i} \frac{x_j^{(t)}}{d_j^{out}}$$

$p^{(t)}$: a vector of length n (number of nodes) which gives the probabilities of the random walker being at each node

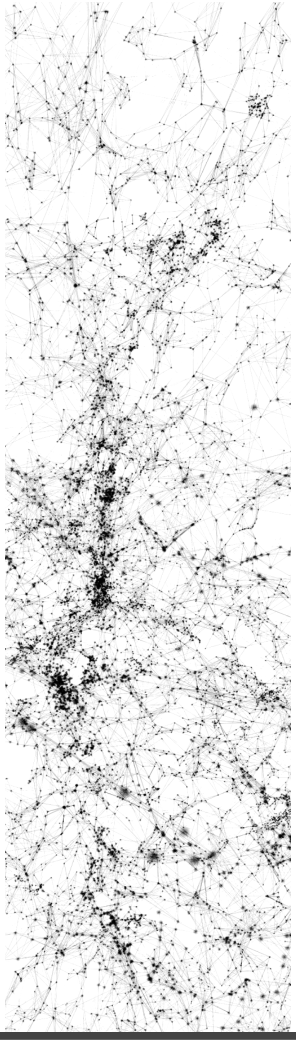
Where is the surfer at time t+1? $p^{(t+1)} = Mp^{(t)}$ Since it follows links uniformly at random

Page ranks are when random walker reaches a stationary state, $p^{(t+1)} = p^{(t)}$

Details of PageRank Formulation by Google explained here & here & here

Outline

- Centrality
    - Degree Centrality
    - Eigenvalue Centrality
    - Katz Centrality
    - PageRank
    - **HITS**
    - Closeness centrality
    - Betweenness centrality
- Similarity
    - Common neighbour
    - Cosine similarity
    - Jaccard similarity



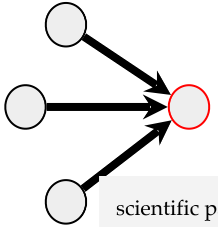
HITS: hyperlink-induced topic search

two different centrality scores

- Highly cited paper [authorities]

  nodes that contain important information

  authority centrality x_i

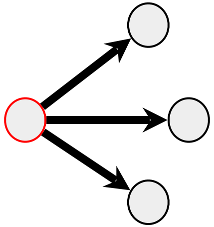


scientific paper is more important if cited by many important reviews

- Survey paper linking to main references [hubs]

  nodes that point us to the best authorities

  hub centrality y_i

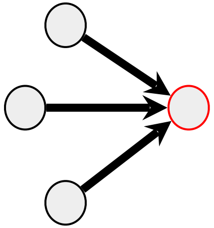


review paper is more important if it cities many important scientific papers

Kleinberg, J. M., Authoritative sources in a hyperlinked environment, *J. ACM* **46**, 604–632 (1999)

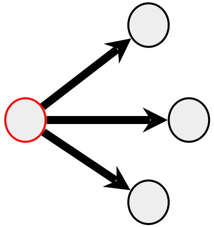
HITS: hyperlink-induced topic search

two different centrality scores



- authority centrality $x_i = \alpha \sum_j A_{ij} y_j$ $\qquad x = \alpha A y$

- hub centrality $y_i = \beta \sum_j A_{ji} x_j$ $\qquad y = \beta A^\top x$



- combining the two we get $AA^\top x = \lambda x$ and $A^\top A y = \lambda y$

What is λ?$(\alpha\beta)^{-1}$

- how to calculate the authority and hub centralities?

  eigenvectors of $AA^\top$ and $A^\top A$ for the largest eigenvalue to not have negative values

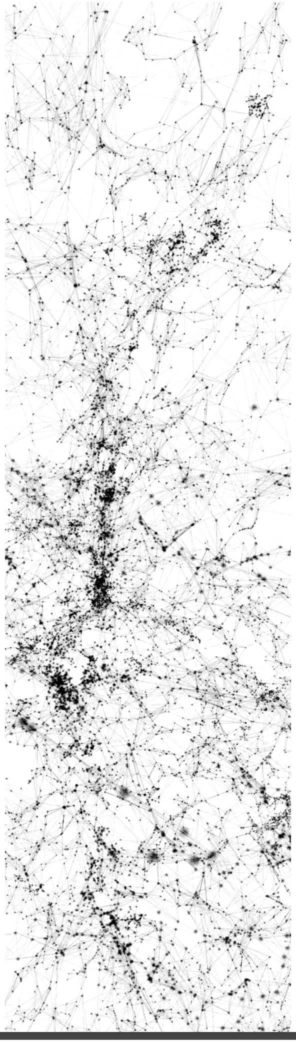
do we have same issue with zero value cascades as katz?

no, since scores flow both ways

$\mathbf{A}^\mathrm{T}\mathbf{A}$ & $\mathbf{A}\mathbf{A}^\mathrm{T}$ have the same eigenvalues, since they are transpose of each other

Outline

- Centrality
    - Degree Centrality
    - Eigenvalue Centrality
    - Katz Centrality
    - PageRank
    - HITS
    - **Closeness centrality**
    - Betweenness centrality
- Similarity
    - Common neighbour
    - Cosine similarity
    - Jaccard similarity



Closeness centrality

the mean distance from a node to other nodes, based on shortest paths

$$s_i = \frac{1}{n}\sum_j s_{ij}$$

{distance}

$$x_i = \frac{n}{\sum_j s_{ij}}$$

{centrality}

What if we have a disconnected graph?

x_i is zero since shortest path to some nodes is infinite

Should we average inside components?

Nodes in smaller components get higher centrality since distances are smaller

guess who has the highest closeness centrality in IMDB?

Christopher Lee, 200 movies, long career

Closeness centrality: reformulation

the mean distance from a node to other nodes, based on shortest paths

$$s_i = \frac{1}{n} \sum_j s_{ij}$$

{distance}

$$x_i = \frac{n}{\sum_j s_{ij}}$$

{centrality}

$$x_i = \frac{1}{n-1} \sum_j \frac{1}{s_{ij}}$$

{reformulation}

Use the harmonic mean distance between nodes instead

What if we have a disconnected graph?

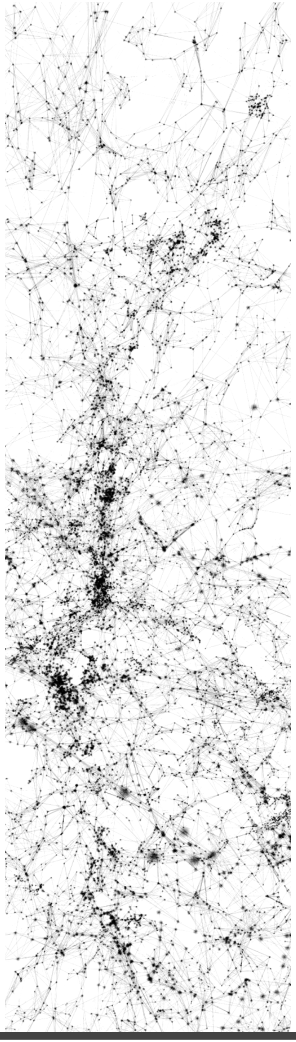
Now it naturally deals with $s_{ij} = \infty$

Is it otherwise same as the original measure?

No, gives more weight to nodes that are close

Outline

- Centrality
    - Degree Centrality
    - Eigenvalue Centrality
    - Katz Centrality
    - PageRank
    - HITS
    - Closeness centrality
    - **Betweenness centrality**
- Similarity
    - Common neighbour
    - Cosine similarity
    - Jaccard similarity



Betweenness centrality

the extent to which a node lies on shortest paths between other nodes

average rate at which traffic passes through node *i*

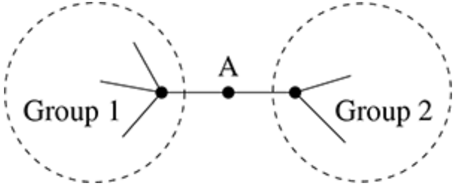
$$x_i = \frac{1}{n^2} \sum_{st} \frac{n_{st}^i}{t_{st}}$$

n_{st}^i : number of shortest paths from s to t that pass through i

t_{st} : total number of shortest paths from s to t

Flow bottlenecks

- control over information passing
- removal from the network will most disrupt communications
  - brokers: low-degree node with high betweenness, lies on a bridge

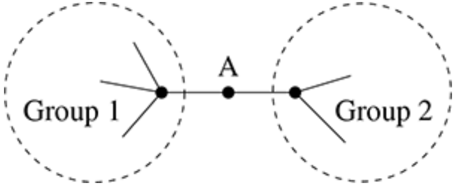


Betweenness centrality

the extent to which a node lies on shortest paths between other nodes

Flow bottlenecks

- control over information passing
- removal from the network will most disrupt communications



Could you guess who has the highest centrality in IMDB?

Fernando Rey

worked extensively in both film and television, in both European and American films, several different languages [in between groups]

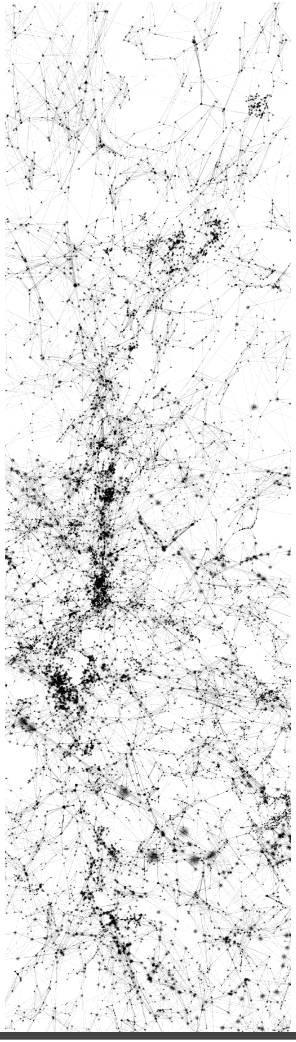
Betweenness centrality has many variants and approximations given its computational complexity and usefulness

Outline

- Centrality
    - Degree Centrality
    - Eigenvalue Centrality
    - Katz Centrality
    - PageRank
    - HITS
    - Closeness centrality
    - Betweenness centrality
- **Similarity**
    - **Common neighbour**
    - Cosine similarity
    - Jaccard similarity

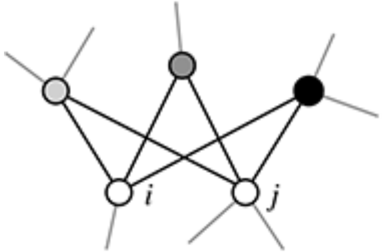
$$R : v \mapsto \mathbb{R}$$

$$S : (u, v) \mapsto \mathbb{R}$$



Similarity Measures

Common Neighbours $$n_{ij} = \sum_k A_{ik} A_{kj}$$



Is 3 a lot or too little? We need to normalize it

Cosine similarity

$$\sigma_{ij} = \sum_k \frac{A_{ik} A_{kj}}{\sqrt{d_i}\sqrt{d_j}} = \frac{n_{ij}}{\sqrt{d_i d_j}}$$

what is σ_{ij} in example?

3 / (√4×√5)

$$cos(A_{i:}, A_{j:}) = \frac{A_{i:} \, . \, A_{j:}}{\|A_{i:}\| \|A_{j:}\|} = \frac{\sum_k A_{ik} A_{jk}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}} = \frac{\sum_k A_{ik} A_{jk}}{\sqrt{\sum_k A_{ik}}\sqrt{\sum_k A_{jk}}} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{d_i}\sqrt{d_j}}$$

Similarity Measures

- Common Neighbours

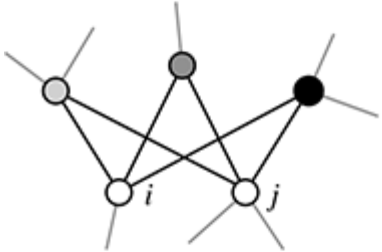
$$n_{ij} = \sum_k A_{ik} A_{kj}$$

- Cosine similarity

$$\sigma_{ij} = \sum_k \frac{A_{ik} A_{kj}}{\sqrt{d_i}\sqrt{d_j}} = \frac{n_{ij}}{\sqrt{d_i d_j}}$$

- Jaccard similarity

$$J_{ij} = \frac{\sum_k A_{ik} A_{kj}}{d_i + d_j - \sum_k A_{ik} A_{kj}} = \frac{n_{ij}}{d_i + d_j - n_{ij}}$$



$n_{ij} = 3$

$\sigma_{ij} = 3/(\surd 4 \times \surd 5)$

what is J_{ij} in example?

$J_{ij} = 3/6$

The (combinatorial) graph Laplacian is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, which is a Symmetric Positive Semi-Definite (SPSD) matrix [28]. Its eigendecomposition gives $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}$, where the columns of $\mathbf{U} \in \mathbb{R}^{N \times N}$ are orthonormal eigenvectors, namely the *graph Fourier basis*, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ with $\lambda_1 \leq \cdots \leq \lambda_N$ and these eigenvalues are also called *frequencies*. Since $\mathbf{L}$ is SPSD, it has N non-negative eigenvalues. Note that it is easy to see $\mathbf{L}\mathbf{1} = \mathbf{0}$. Thus the smallest eigenvalue of $\mathbf{L}$ is 0 and $\mathbf{1}$ is a corresponding eigenvector. A graph signal is a vector $\boldsymbol{x} \in \mathbb{R}^N$ defined on $\mathcal{V}$, where $\mathbf{x}_i$ is defined on the node i. The graph Fourier transform of a graph signal $\boldsymbol{x}$ is defined as $\boldsymbol{x}_{\mathcal{F}} = \mathbf{U}^{-1}\boldsymbol{x} = \mathbf{U}^{\mathsf{T}}\boldsymbol{x} = [\mathbf{u}_1^{\mathsf{T}}\mathbf{x}, \ldots, \mathbf{u}_N^{\mathsf{T}}\mathbf{x}]^{\mathsf{T}}$, where $\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{x}$ is the component of $\boldsymbol{x}$ in the direction of $\boldsymbol{u}_i$. Here $\boldsymbol{x}_{\mathcal{F}}$ is the transformed signal in the spectral domain.

Since $\boldsymbol{u}_i$ is a unit-norm eigenvector, we have

$$\lambda_i = \boldsymbol{u}_i^{\mathsf{T}}\mathbf{L}\boldsymbol{u}_i = \sum_{k,j \in \mathcal{V}, e_{kj} \in \mathcal{E}} (\boldsymbol{u}_{i,k} - \boldsymbol{u}_{i,j})^2$$

A smaller λ_i indicates a smoother basis function $\boldsymbol{u}_i$ defined on $\mathcal{G}$ [34], which means any two elements $\boldsymbol{u}_{i,k}, \boldsymbol{u}_{i,j}$ of $\boldsymbol{u}_i$ corresponding to two connected nodes k, j tend to have more similar values.

In additional to $\mathbf{L}$, some variants of graph Laplacians are commonly used in practice, *e.g.*, the symmetric normalized Laplacian $\mathbf{L}_{\text{sym}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, and the random walk normalized Laplacian $\mathbf{L}_{\text{rw}} = \mathbf{D}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$. $\mathbf{L}_{\text{rw}}$ and $\mathbf{L}_{\text{sym}}$ share the same eigenvalues, which are inside $[0, 2)$, and their corresponding eigenvectors satisfy $\boldsymbol{u}_{\text{rw}}^i = \mathbf{D}^{-1/2}\boldsymbol{u}_{\text{sym}}^i$ [28].

The normalized affinity (transition) matrix derived from $\mathbf{L}_{\text{rw}}$ is defined as $\mathbf{A}_{\text{rw}} = \mathbf{I} - \mathbf{L}_{\text{rw}} = \mathbf{D}^{-1}\mathbf{A}$ and its eigenvalues $\lambda_i(\mathbf{A}_{\text{rw}}) = 1 - \lambda_i(\mathbf{L}_{\text{rw}}) \in (-1, 1]$. Similarly, $\mathbf{A}_{\text{sym}} = \mathbf{I} - \mathbf{L}_{\text{sym}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ is an affinity matrix as well.