Measures

Analysis of complex interconnected data

Slides mostly based on newman's book

McGill
School of Computer Science

Quick Notes

- Second assignment, questions?
  - Submit single entry as a Group in Mycourses
- Office hours:
  - Me: Thu 12pm-1pm
  - Andy: Wed 1pm-2pm
- Use slack for any questions
  - We should have everyone there now

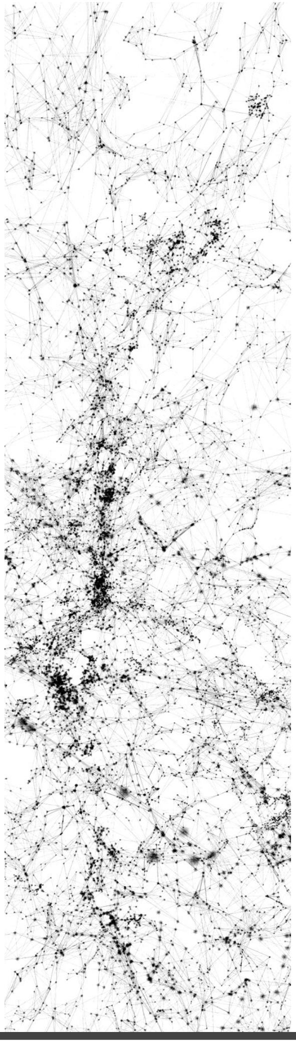
Outline

- **Centrality**
  - Degree Centrality
  - Eigenvalue Centrality
  - Katz Centrality
  - PageRank
  - HITS
  - Closeness centrality
  - Betweenness centrality

R(i) for i in [1..n]

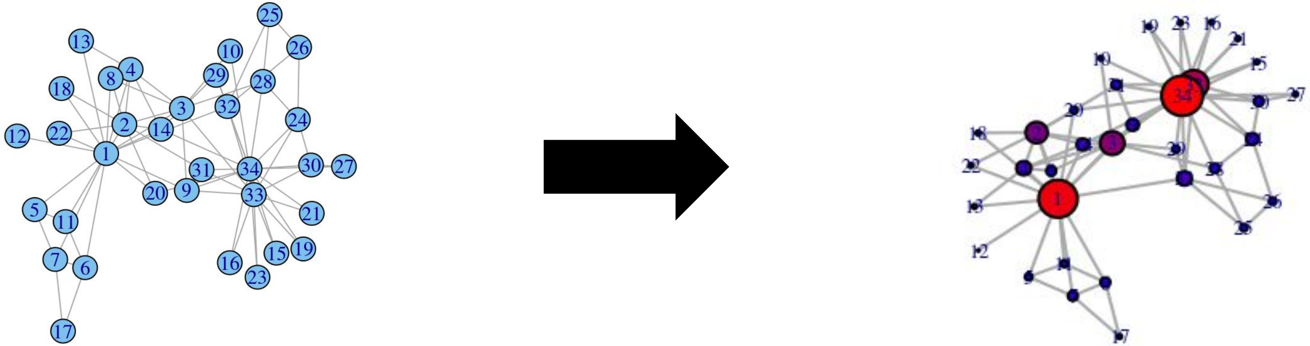
- Similarity
  - Common neighbour
  - Cosine similarity
  - Jaccard similarity

S(i,j) for i,j in [1..n]



Centrality

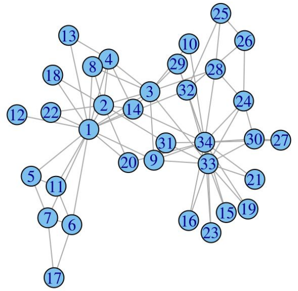
Measure the importance of nodes

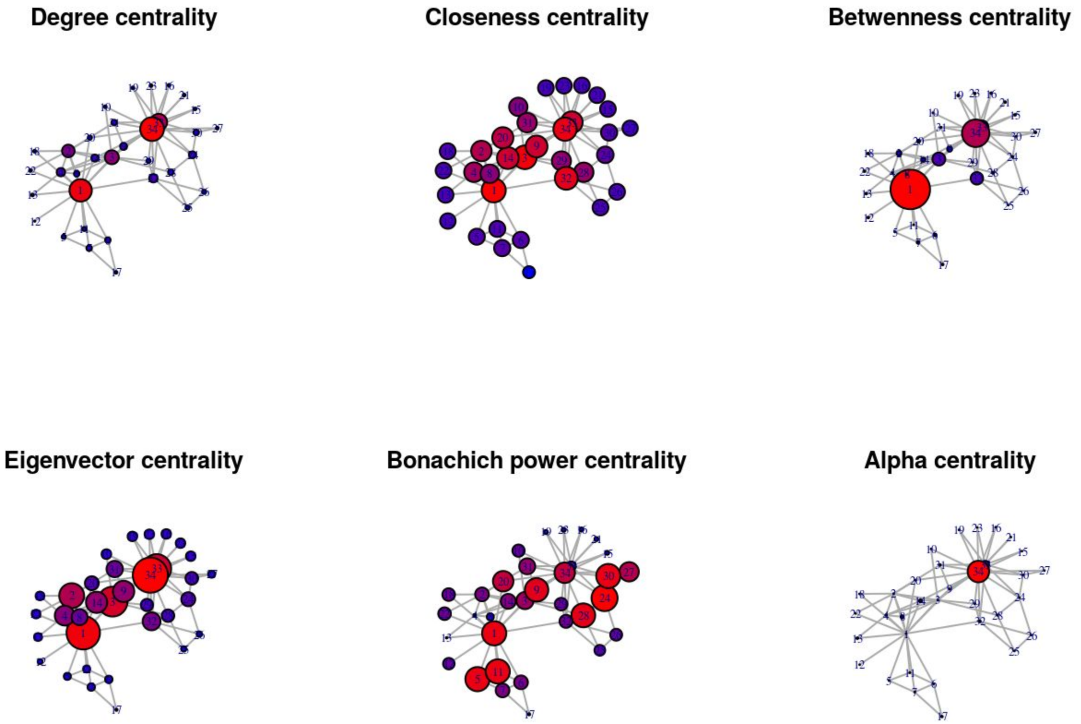


http://www.rpubs.com/shestakoff/sna_lab4

Centrality

Different ways to define importance ⇒ Different centrality measures ⇒ Different ranking of the nodes on the same graph





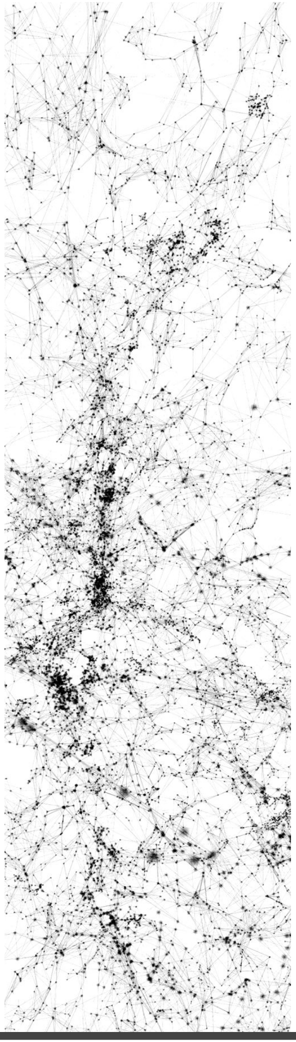
http://www.rpubs.com/shestakoff/sna_lab4

Outline

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

R(i) for i in [1..n]

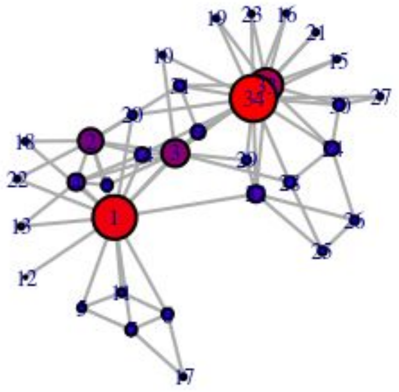
S(i,j) for i,j in [1..n]



Degree centrality

Degree is the simplest centrality measure

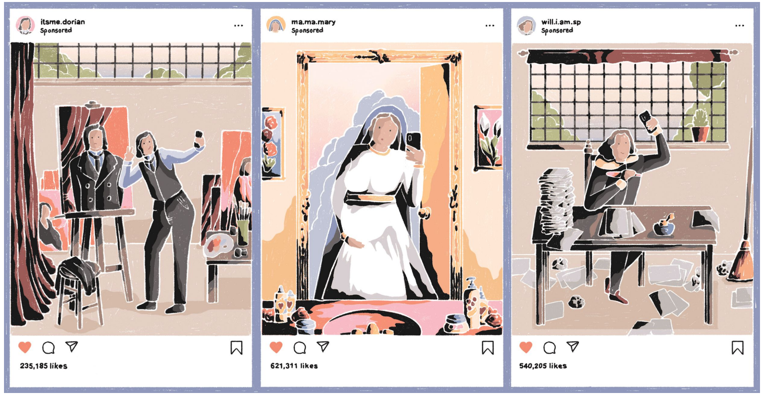
more connections you have (number of edges),
more people you know (number of neighbours),
more important you are



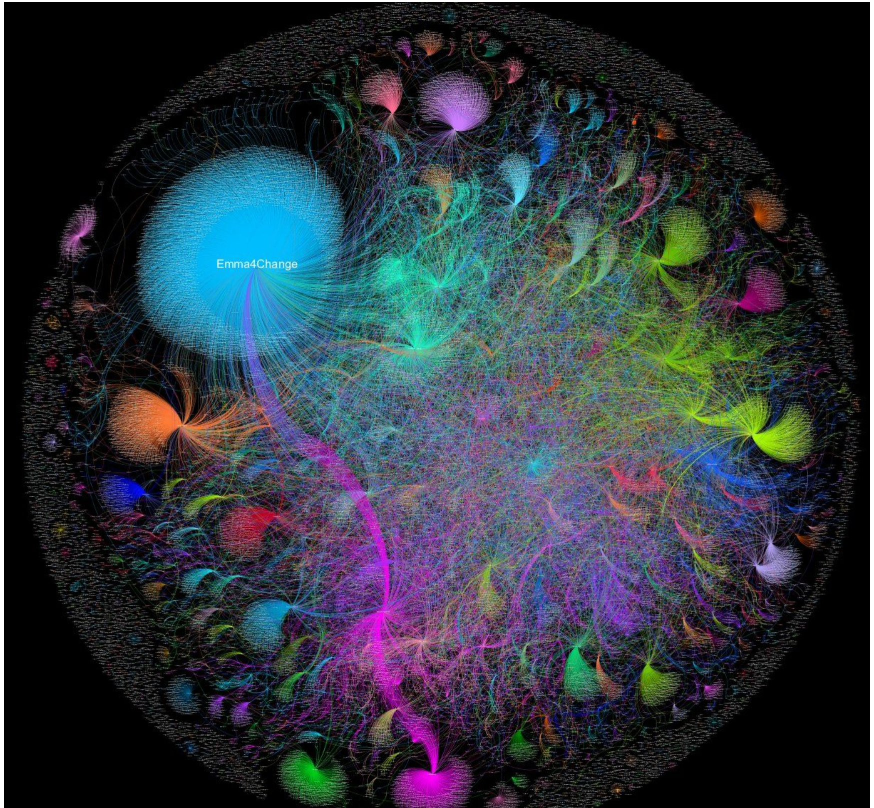
Can you think of a widely used example where people are ranked by degree centrality?

Degree centrality, example

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

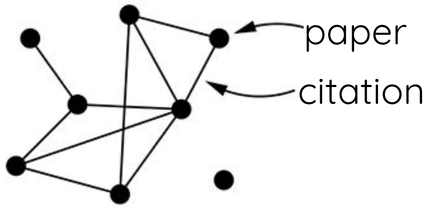


https://www.newyorker.com/culture/annals-of-inquiry/a-history-of-the-influencer-from-shakespeare-to-instagram

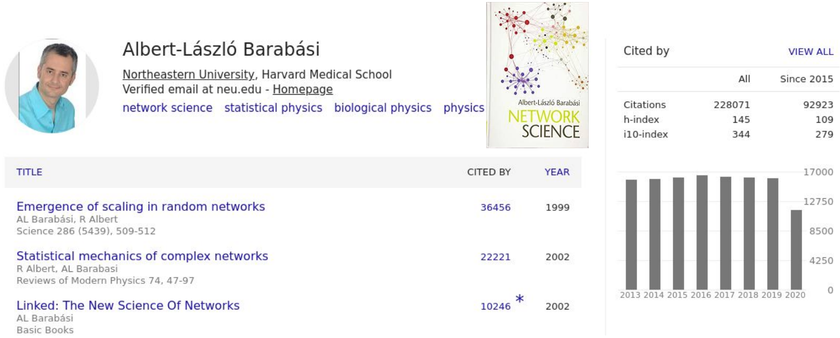


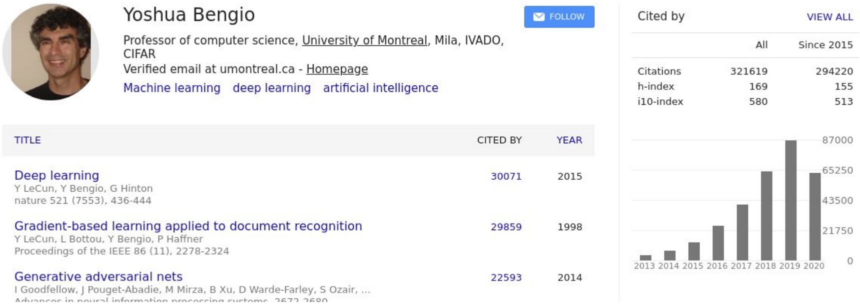
https://medium.com/@erin_gallagher/marchforourlives-neveragain-a59ee4a078cb

Degree centrality, example



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





Eigenvector centrality

How to measure having important connections?

You might only have one connection but it can be the US president

Eigenvector centrality

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

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

$$X_i = \mathbf{K}^{-1} \sum_{j \in N(i)} X_j$$

$$N(i) = \{j \mid \mathbf{A}_{ij} = 1\}$$

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 neighbors

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

$$x_i = \mathbf{K}^{-1} \sum_{j \in N(i)} x_j$$

$$N(i) = \{j \mid \mathbf{A}_{ij} = 1\}$$

How can we write this in matrix notation?

Note that we have $\sum_{j \in N(i)} x_j = \mathbf{A}_{i:} \mathbf{x}$ where $\mathbf{x}$ is a vector of all centrality scores

Eigenvector centrality

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

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

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

$$\mathbf{x} = \kappa^{-1} \mathbf{A} \, \mathbf{x}$$ {Vector notation}

$$\mathbf{A} \, \mathbf{x} = \kappa \, \mathbf{x}$$

What is x?

Eigenvector centrality

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

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

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

$$\mathbf{x} = \kappa^{-1} \mathbf{A} \, \mathbf{x}$$

$$\mathbf{A} \, \mathbf{x} = \kappa \, \mathbf{x}$$

⇒ **x** is an eigenvector of the adjacency matrix

Eigenvector centrality

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

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

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

$$\mathbf{x} = \kappa^{-1} \mathbf{A} \mathbf{x}$$

Which eigenvector should we use?

$$\mathbf{A} \mathbf{x} = \kappa \mathbf{x}$$

x is an eigenvector of the adjacency matrix

Eigenvector centrality

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

x is an eigenvector of the adjacency matrix and $\mathbf{x}_i$ gives the importance of node i

$$\mathbf{A}\,\mathbf{x} = \kappa\,\mathbf{x}$$

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

[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 neighbors

$$\mathbf{A}\,\mathbf{x} = \kappa\,\mathbf{x}$$

$\mathbf{x}$ is the leading eigenvector

what is κ?

Eigenvector centrality

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

$$\mathbf{A}\,\mathbf{x} = \kappa\,\mathbf{x}$$

$\mathbf{x}$ is the leading eigenvector

what is κ? largest eigenvalue

Eigenvector centrality & random walks

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

Why?

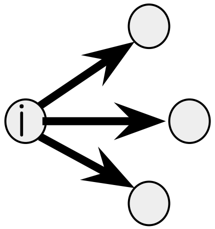
Eigenvector centrality & random walks

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

Why?

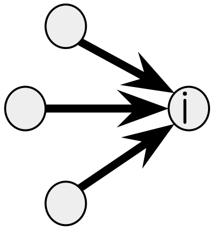
Leading eigenvector is computed with power iteration, $\mathbf{x}^{(i+1)}=\mathbf{A}\mathbf{x}^{(i)}$
A^k gives number of walks of length k

Eigenvector centrality in directed networks



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

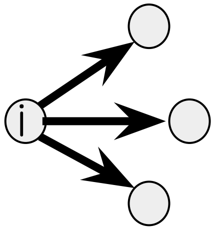
Can be defined in two ways



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

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

Eigenvector centrality in directed networks



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

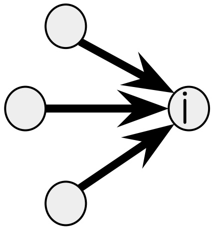
$$\mathbf{x} \mathbf{A} = \kappa \, \mathbf{x}$$

[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?

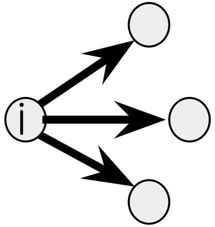


$$x_i = \kappa^{-1} \sum_j A_{ij} x_j$$

$$\mathbf{A} \mathbf{x} = \kappa \, \mathbf{x}$$

[right]

Eigenvector centrality in directed networks



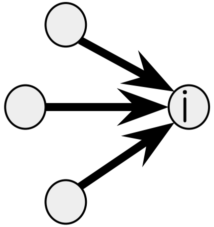
$x_i = \kappa^{-1} \sum_j A_{ji} x_j$

$\mathbf{x} \mathbf{A} = \kappa \mathbf{x}$

[left]

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

Which one to use? Right



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

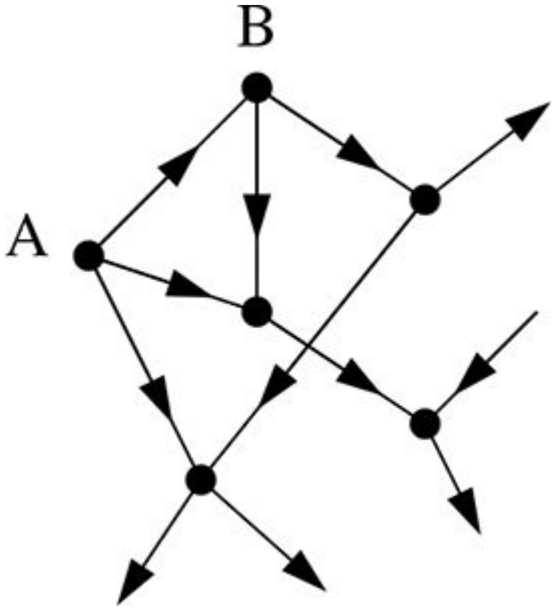
$\mathbf{A} \mathbf{x} = \kappa \mathbf{x}$

[right]

Eigenvector centrality in directed networks

Example:

What is the score of A?

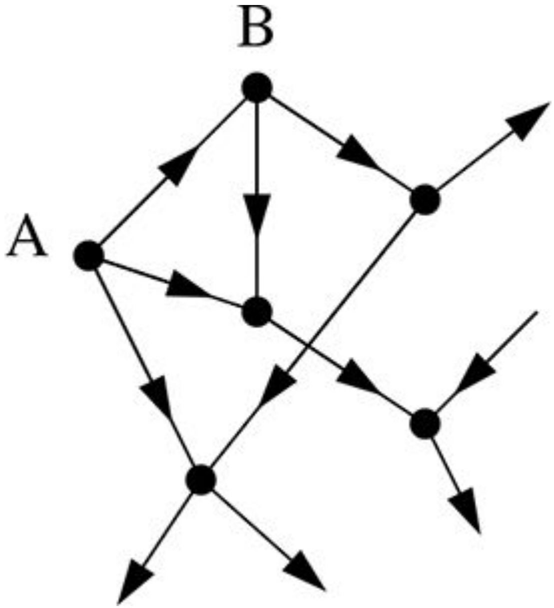


$$x_i = \kappa^{-1} \sum_j A_{ij} x_j$$

Eigenvector centrality in directed networks

Example:

What is the score of A? a node with no incoming edge ⇒ zero score



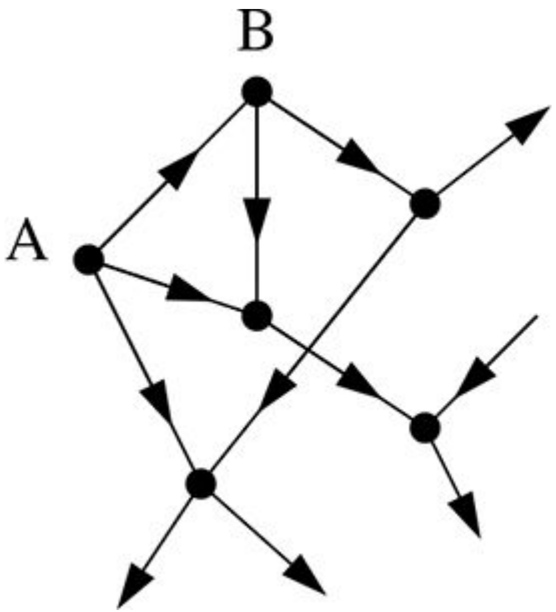
$$x_i = \kappa^{-1} \sum_j A_{ij} x_j$$

Eigenvector centrality in directed networks

Example:

What is the score of A? a node with no incoming edge ⇒ zero score

What is the score of B?



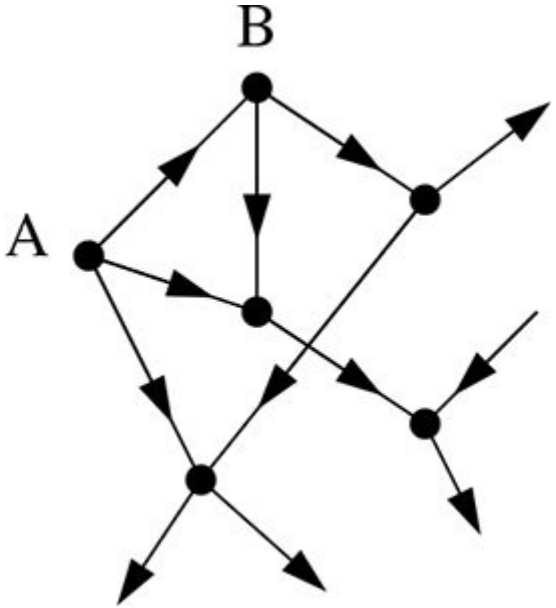
$$x_i = \kappa^{-1} \sum_j A_{ij} x_j$$

Eigenvector centrality in directed networks

Example:

What is the score of A? a node with no incoming edge ⇒ zero score

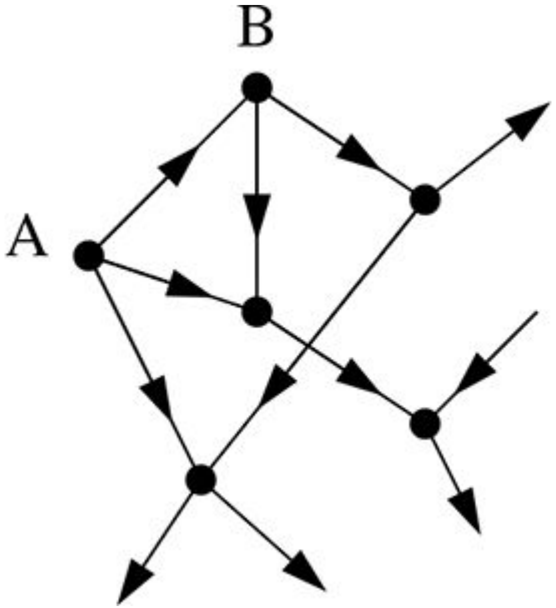
What is the score of B? also zero, only ingoing edge is from A



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

Eigenvector centrality in directed networks

Only non-zero 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?

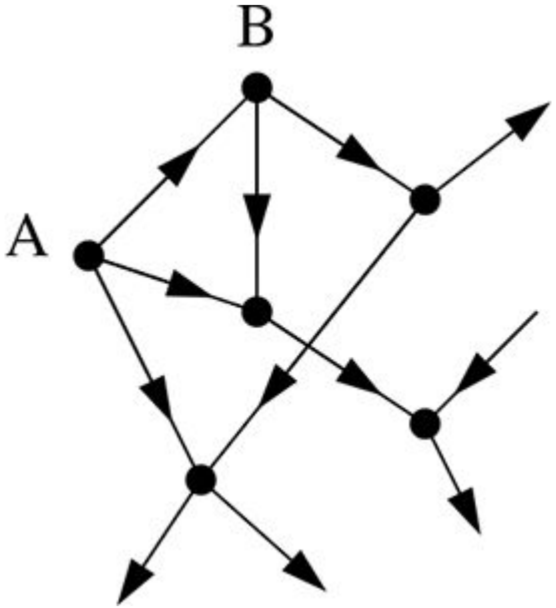
$$x_i = \kappa^{-1} \sum_j A_{ij} x_j$$

Eigenvector centrality in directed networks

Only non-zero 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?

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



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

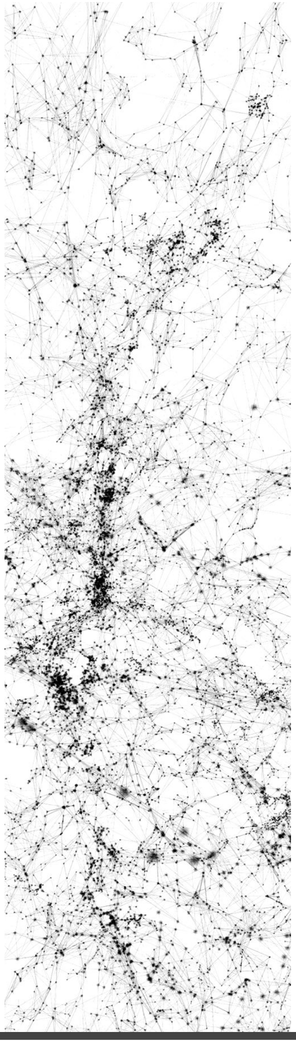
How can we fix it? Katz and PageRank variants

Outline

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

R(i) for i in [1..n]

S(i,j) for i,j in [1..n]



Katz centrality

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

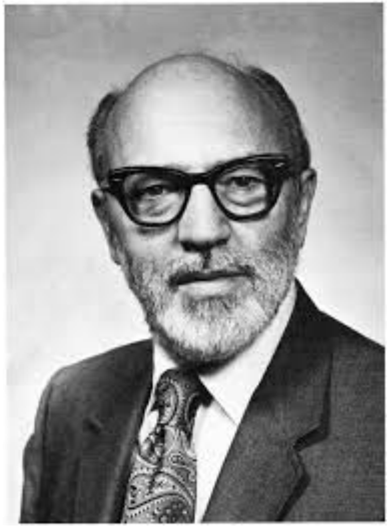
α and β are positive constants

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

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

$$\mathbf{x} = \alpha\, \mathbf{A}\, \mathbf{x} + \beta\, \mathbf{1}$$

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

$$\mathbf{x} = (\mathbf{I} - \alpha\mathbf{A})^{-1}\, \mathbf{1} \qquad \{\text{with } \beta = 1\}$$

1 is the uniform vector of all ones: (1, 1, 1, ...)

I is the identity matrix: diag(1, 1, 1, ...)

absolute magnitude of centrality scores are not important, we care about the relative values

Katz centrality

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

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

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{x} + \beta \mathbf{1}$$

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

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

Katz centrality

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

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{x} + \beta \mathbf{1}$$

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

$$\mathbf{x} = (\mathbf{I} - \alpha\mathbf{A})^{-1} \mathbf{1} \qquad \{\text{with } \beta = 1\}$$

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

All nodes have the same importance as β

Katz centrality

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

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{x} + \beta \mathbf{1}$$

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

$$\mathbf{x} = (\mathbf{I} - \alpha \mathbf{A})^{-1} \mathbf{1} \qquad \{\text{with } \beta = 1\}$$

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 when $(\mathbf{I} - \alpha \mathbf{A})^{-1}$ diverges

Katz centrality

$$x_i = \alpha \sum_j A_{ij} x_j + 1$$

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{x} + \mathbf{1}$$

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

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

All nodes have the same importance of 1

As we increase α, scores increase and might start to diverge when $(\mathbf{I} - \alpha\mathbf{A})^{-1}$ diverges happens when

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

At what **α** this happens?

The determinant of a matrix is equal to the product of its eigenvalues, and matrix $x\mathbf{I}-\mathbf{A}$ has eigenvalues $x-\kappa_i$ where κ_i are the eigenvalues of $\mathbf{A}$ ⇒ $\det(x\mathbf{I}-\mathbf{A})=(x-\kappa_1)(x-\kappa_2)\ldots(x-\kappa_n)$, with zeros at $x=\kappa_1,\kappa_2,\ldots$ ⇒ the solutions of $\det(x\mathbf{I}-\mathbf{A})=0$ give the eigenvalues of $\mathbf{A}$

Katz centrality

$$x_i = \alpha \sum_j A_{ij} x_j + 1$$

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{x} + \mathbf{1}$$

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

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

All nodes have the same importance of 1

As we increase α, scores increase and might start to diverge when $(\mathbf{I} - \alpha\mathbf{A})^{-1}$ diverges happens when

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

At what **α** this happens? $\alpha^{-1} = \kappa_i \Rightarrow \alpha = 1/\kappa_i$

The determinant of a matrix is equal to the product of its eigenvalues, and matrix $x\mathbf{I}-\mathbf{A}$ has eigenvalues $x-\kappa_i$ where κ_i are the eigenvalues of $\mathbf{A}$ ⇒ $\det(x\mathbf{I}-\mathbf{A})=(x-\kappa_1)(x-\kappa_2)\ldots(x-\kappa_n)$, with zeros at $x=\kappa_1,\kappa_2,\ldots$ ⇒ the solutions of $\det(x\mathbf{I}-\mathbf{A})=0$ give the eigenvalues of $\mathbf{A}$

Katz centrality

$$x_i = \boldsymbol{\alpha} \sum_j A_{ij} x_j + 1$$

$$\mathbf{x} = \alpha\, \mathbf{A}\, \mathbf{x} + \mathbf{1}$$

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

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

All nodes have the same importance of 1

As we increase $\boldsymbol{\alpha}$, scores increase and might start to diverge when $(\mathbf{I} - \alpha\mathbf{A})^{-1}$ diverges happens when

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

At what **α** this happens? $\boldsymbol{\alpha}^{-1} = \kappa_i \Rightarrow \boldsymbol{\alpha} = 1/\kappa_i$

At what **α** this first happens?

Katz centrality

$$x_i = \alpha \sum_j A_{ij} x_j + 1$$

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{x} + \mathbf{1}$$

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

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

All nodes have the same importance of 1

As we increase α, scores increase and might start to diverge when $(\mathbf{I} - \alpha\mathbf{A})^{-1}$ diverges happens when

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

At what α this happens? $\alpha^{-1} = \kappa_i \Rightarrow \alpha = 1/\kappa_i$

At what α this first happens?

largest (most positive) eigenvalue

Katz centrality

$$x_i = \alpha \sum_j A_{ij} x_j + 1$$

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{x} + \mathbf{1}$$

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

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

All nodes have the same importance of 1

As we increase α, scores increase
and might start to diverge
when $(\mathbf{I} - \alpha\mathbf{A})^{-1}$ diverges happens when

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

At what α this first happens?
largest (most positive) eigenvalue

$$\alpha < 1/\kappa_1$$

Katz centrality

$$x_i = \alpha \sum_j A_{ij} x_j + 1$$

$$\alpha < 1/\kappa_1$$

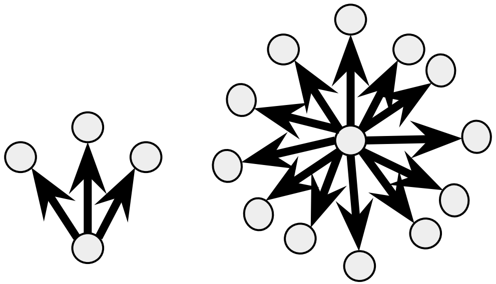
κ_1 is the largest (most positive) eigenvalue

In practice **α** is often set close to this limit

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

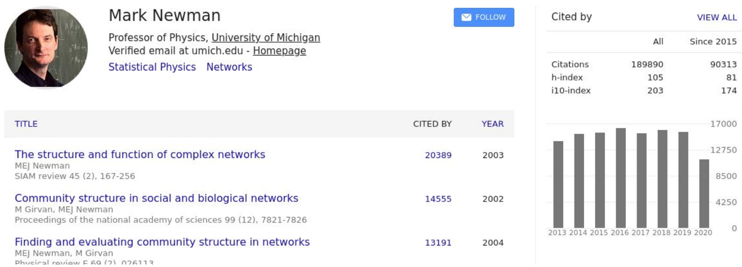
Katz centrality

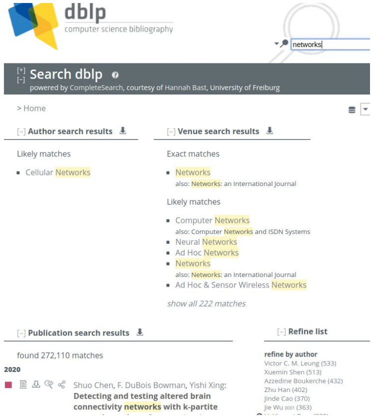
$$x_i = \alpha \sum_j A_{ij} x_j + 1$$



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



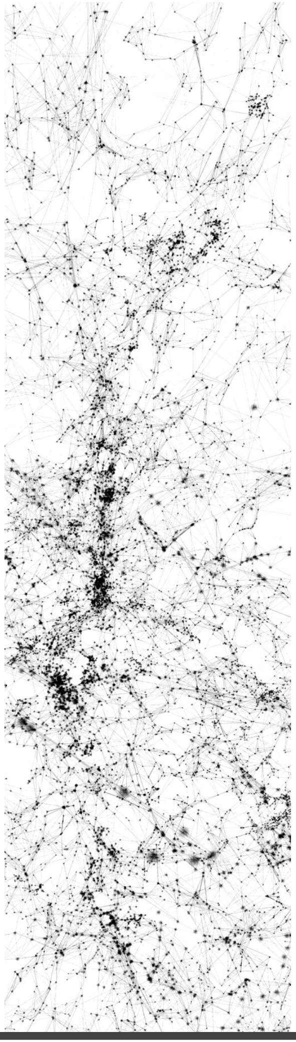


Outline

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

R(i) for i in [1..n]

S(i,j) for i,j in [1..n]



PageRank

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

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

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

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

A_{ij}=1 if there is an edge from j to i ⇒ $d_j^{out} = 0$ A_{ij}=0 for all i, then $A_{ij} / d_j^{out} = 0/0$ which we want to be 0

PageRank

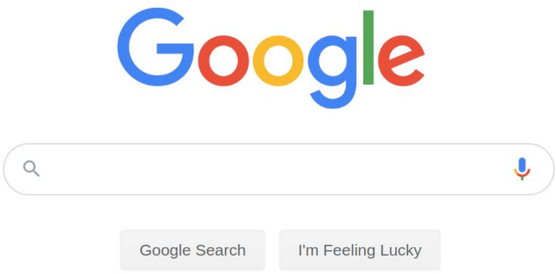
divide your centrality to your neighbours, instead of passing to all

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

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

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

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



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

PageRank

$x_i = \boldsymbol{\alpha} \sum_j A_{ij} / d_j^{out} \; x_j + \beta;$ $\quad\quad \mathbf{x} = (\mathbf{I} - \alpha\, \mathbf{AD}^{-1})^{-1}\, \mathbf{1}$

*What should **α** be?*

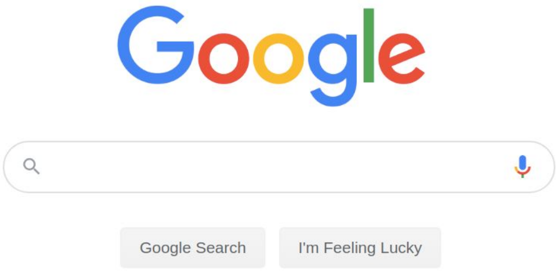
PageRank

$$x_i = \alpha \sum_j A_{ij} / d_j^{out}\ x_j + \beta; \qquad \mathbf{x} = (\mathbf{I} - \alpha\, \mathbf{A}\mathbf{D}^{-1})^{-1}\, \mathbf{1}$$

α < the leading eigenvalue of $\mathbf{AD}^{-1}$

This is 1 for undirected network but changes for directed ones.

The Google search engine uses a value of α=0.85



PageRank

$$x_i = \boldsymbol{\alpha} \sum_j A_{ij} / d_j^{out} \; x_j + \beta; \qquad \mathbf{x} = (\mathbf{I} - \alpha\,\mathbf{AD}^{-1})^{-1}\,\mathbf{1}$$

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

What if undirected and we set $\boldsymbol{\beta} = 0$ and $\boldsymbol{\alpha} = 1$?

PageRank

$$x_i = \boldsymbol{\alpha} \sum_j A_{ij} / d_j^{out}\ x_j + \boldsymbol{\beta}; \qquad \mathbf{x} = (\mathbf{I} - \alpha\, \mathbf{AD}^{-1})^{-1}\, \mathbf{1}$$

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

What if undirected and we set $\boldsymbol{\beta} = 0$ and $\boldsymbol{\alpha} = 1$?

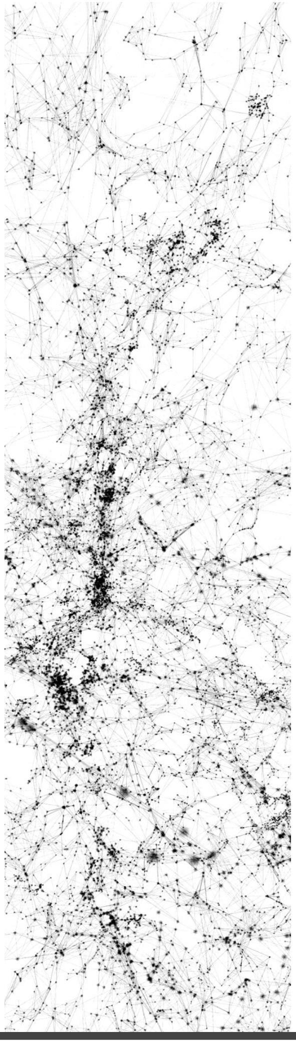
reduces to degree centrality

Outline

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

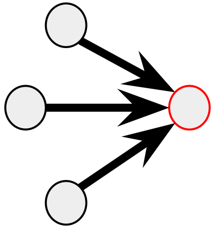
R(i) for i in [1..n]

S(i,j) for i,j in [1..n]

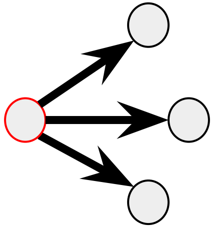


HITS: hyperlink-induced topic search

- Highly cited paper

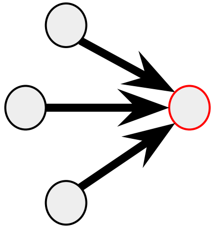


- Survey paper linking to main references



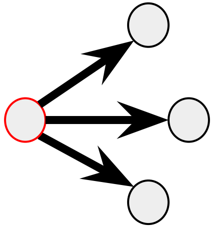
HITS

- Highly cited paper [authorities]



nodes that contain important information

- Survey paper linking to main references [hubs]



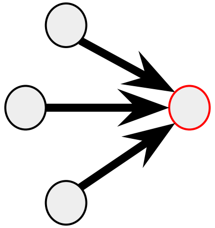
nodes that point us to the best authorities

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

HITS

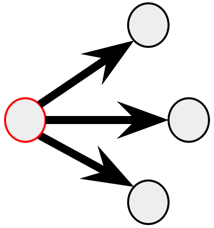
- Highly cited paper [authorities]

  authority centrality x_i



- Survey paper linking to main references [hubs]

  hub centrality y_i



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

important scientific paper (in the authority sense) would be one cited in many important reviews (in the hub sense)

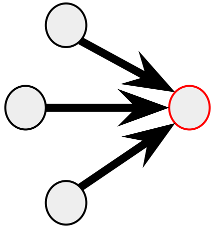
HITS

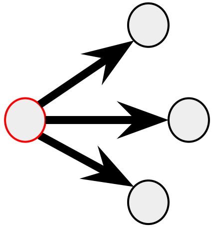
- authority centrality x_i

$$x_i = \alpha \sum_j A_{ij}\, y_j$$

- hub centrality y_i

$$y_i = \beta \sum_j A_{ji}\, x_j$$

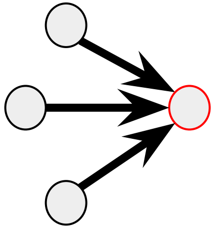




HITS

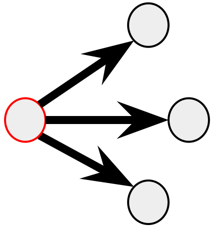
- authority centrality x_i

$$x_i = \alpha \sum_j A_{ij} y_j \qquad \mathbf{x} = \alpha \mathbf{A} \mathbf{y}$$



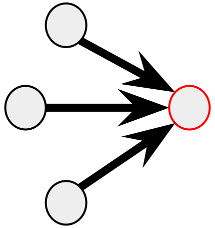
- hub centrality y_i

$$y_i = \beta \sum_j A_{ji} x_j \qquad \mathbf{y} = \beta \mathbf{A}^T \mathbf{x}$$



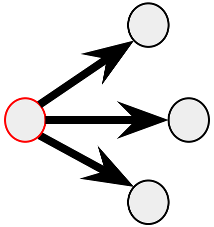
HITS

- authority centrality x_i



$$\mathbf{x} = \alpha \mathbf{A} \mathbf{y} \qquad \mathbf{A}\mathbf{A}^T\mathbf{x} = \lambda \mathbf{x}$$

- hub centrality y_i



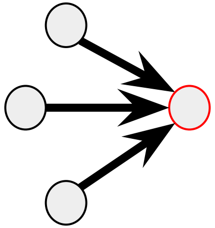
$$\mathbf{y} = \beta \mathbf{A}^T\mathbf{x} \qquad \mathbf{A}^T\mathbf{A}\,\mathbf{y} = \lambda \mathbf{y}$$

What is λ?

HITS

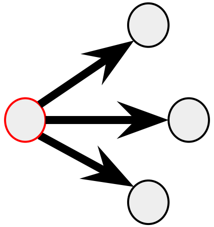
- authority centrality x_i

$$\mathbf{x} = \alpha\, \mathbf{A}\, \mathbf{y} \qquad \mathbf{A}\mathbf{A}^T\mathbf{x} = \lambda\, \mathbf{x}$$



- hub centrality y_i

$$\mathbf{y} = \beta\, \mathbf{A}^T\mathbf{x} \qquad \mathbf{A}^T\mathbf{A}\, \mathbf{y} = \lambda\, \mathbf{y}$$



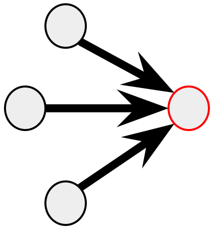
$$\lambda = (\alpha\beta)^{-1}$$

What does it imply?

HITS

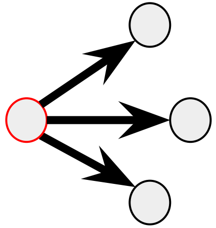
- authority centrality x_i

$$\mathbf{x} = \alpha\, \mathbf{A}\, \mathbf{y} \qquad\qquad \mathbf{A}\mathbf{A}^T\mathbf{x} = \lambda\, \mathbf{x}$$



- hub centrality y_i

$$\mathbf{y} = \beta\, \mathbf{A}^T\mathbf{x} \qquad\qquad \mathbf{A}^T\mathbf{A}\, \mathbf{y} = \lambda\, \mathbf{y}$$



$$\lambda = (\alpha\beta)^{-1}$$

What does it imply?

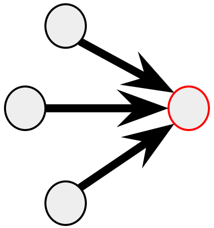
eigenvectors of AA^T and A^TA with the same eigenvalue

HITS

- authority centrality x_i

$$\mathbf{x} = \alpha\, \mathbf{A}\, \mathbf{y}$$

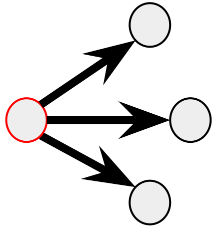
$$\mathbf{A}\mathbf{A}^T\mathbf{x} = \lambda\, \mathbf{x}$$



- hub centrality y_i

$$\mathbf{y} = \beta\, \mathbf{A}^T\mathbf{x}$$

$$\mathbf{A}^T\mathbf{A}\, \mathbf{y} = \lambda\, \mathbf{y}$$



$$\lambda = (\alpha\beta)^{-1}$$

What does it imply?

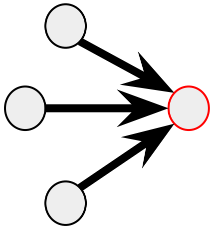
eigenvectors of AA^T and A^TA with the same eigenvalue

largest eigenvalue to get non-negative scores

HITS

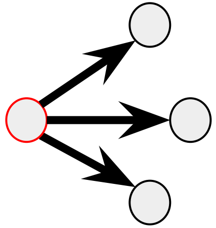
- authority centrality x_i

$$\mathbf{x} = \alpha\, \mathbf{A}\, \mathbf{y} \qquad\qquad \mathbf{A}\mathbf{A}^T\mathbf{x} = \lambda\, \mathbf{x}$$



- hub centrality y_i

$$\mathbf{y} = \beta\, \mathbf{A}^T\mathbf{x} \qquad\qquad \mathbf{A}^T\mathbf{A}\, \mathbf{y} = \lambda\, \mathbf{y}$$



$$\lambda = (\alpha\beta)^{-1}$$

What does it imply?

eigenvectors of AA^T and A^TA with the same eigenvalue

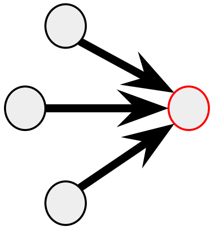
largest eigenvalue to get non-negative scores

$\mathbf{A}^T\mathbf{A}$ & $\mathbf{A}\mathbf{A}^T$ always have the same eigenvalues

HITS

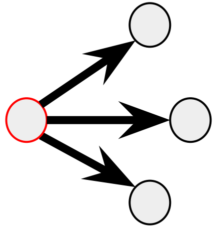
- authority centrality x_i

$$\mathbf{x} = \alpha\, \mathbf{A}\, \mathbf{y} \qquad\qquad \mathbf{A}\mathbf{A}^T\mathbf{x} = \lambda\, \mathbf{x}$$



- hub centrality y_i

$$\mathbf{y} = \beta\, \mathbf{A}^T\mathbf{x} \qquad\qquad \mathbf{A}^T\mathbf{A}\, \mathbf{y} = \lambda\, \mathbf{y}$$

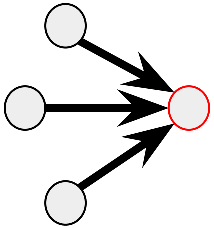


$\lambda = (\alpha\beta)^{-1}$ **largest eigenvalue** to get non-negative scores

Could we have zero scores?

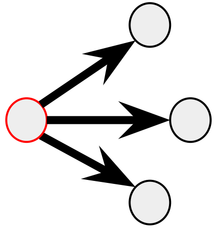
HITS

- authority centrality x_i



$$\mathbf{x} = \alpha\, \mathbf{A}\, \mathbf{y} \qquad\qquad \mathbf{A}\mathbf{A}^T\mathbf{x} = \lambda\, \mathbf{x}$$

- hub centrality y_i



$$\mathbf{y} = \beta\, \mathbf{A}^T\mathbf{x} \qquad\qquad \mathbf{A}^T\mathbf{A}\, \mathbf{y} = \lambda\, \mathbf{y}$$

$\lambda = (\alpha\beta)^{-1}$ **largest eigenvalue** to get non-zero scores

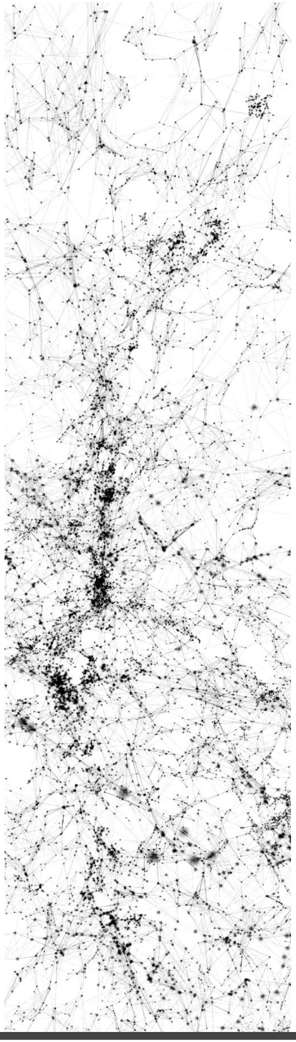
Could we have zero scores? Yes, but no issue since hub could be zero but authority not

Outline

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

R(i) for i in [1..n]

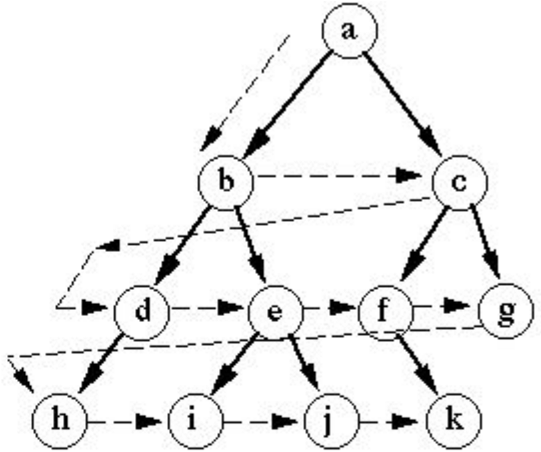
S(i,j) for i,j in [1..n]



Closeness centrality

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

$$s_i = 1/n\ (\Sigma_j\ s_{ij})$$



Breadth-first search

Closeness centrality

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

$s_i = 1/n\ (\Sigma_j\ s_{ij})$ {distance}

$x_i = n/\Sigma_j\ s_{ij}$ {centrality}

Could you guess who has the highest cenerality in IMDB?

Closeness centrality

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

$$s_i = 1/n \left(\sum_j s_{ij}\right)$$

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

Could you guess who has the highest cenerality in IMDB?

Christopher Lee

Closeness centrality

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

$$s_i = 1/n\ (\Sigma_j\ s_{ij})$$

$$x_i = n/\Sigma_j\ s_{ij}$$

What happens if we have many connected components? i.e. disconnected graph?

Closeness centrality

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

$$s_i = 1/n\ (\Sigma_j\ s_{ij})$$

$$x_i = n/\Sigma_j\ s_{ij}$$

What happens if we have many connected components? i.e. disconnected graph?
Infinite

Should we average inside components?

Closeness centrality

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

$$s_i = 1/n\ (\Sigma_j\ s_{ij})$$

$$x_i = n/\Sigma_j\ s_{ij}$$

What happens if we have many connected components? i.e. disconnected graph?

Infinite

Should we average inside components?

Nodes in smaller components get higher centrality

Closeness centrality, reformulation

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

$$s_i = 1/n\ (\Sigma_j\, s_{ij})$$

$$x_i = n/\Sigma_j\, s_{ij} \qquad \Rightarrow \qquad x_i = 1/(n\text{-}1)\ \Sigma_j\, 1/s_{ij}$$

Use the harmonic mean distance between nodes instead

Naturally deals with $s_{ij}=\infty$

Other property?

Closeness centrality, reformulation

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

$$s_i = 1/n \left(\Sigma_j\, s_{ij}\right)$$

$$x_i = n/\Sigma_j\, s_{ij} \quad \Rightarrow \quad x_i = 1/(n\text{-}1)\; \Sigma_j\, 1/s_{ij}$$

Use the harmonic mean distance between nodes instead

Naturally deals with $s_{ij}=\infty$

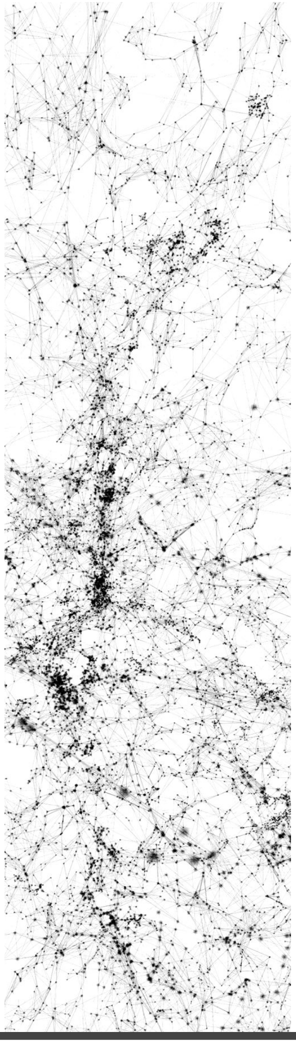
Other property? gives more weight to nodes that are close

Outline

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

R(i) for i in [1..n]

S(i,j) for i,j in [1..n]



Betweenness centrality

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

Flow bottlenecks

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

Betweenness centrality

$$x_i = 1/n^2 \sum_{st} n^i_{st} / t_{st}$$

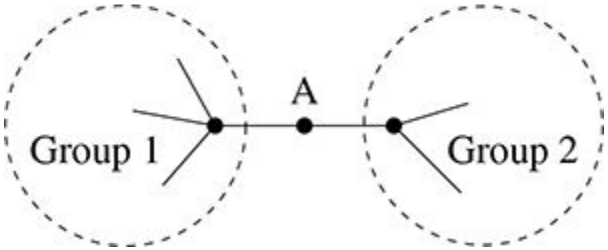
n^i_{st} = the number of shortest paths from s to t that pass through i

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

average rate at which traffic passes through node *i*

Betweenness centrality

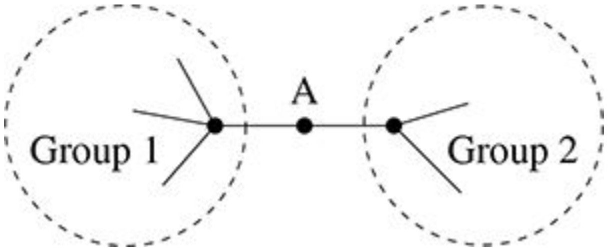
Brokers: low-degree node with high betweenness, lies on a bridge



Could you guess who has the highest betweenness centrality in IMDB?

Betweenness centrality

Brokers: low-degree node with high betweenness, lies on a bridge



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]

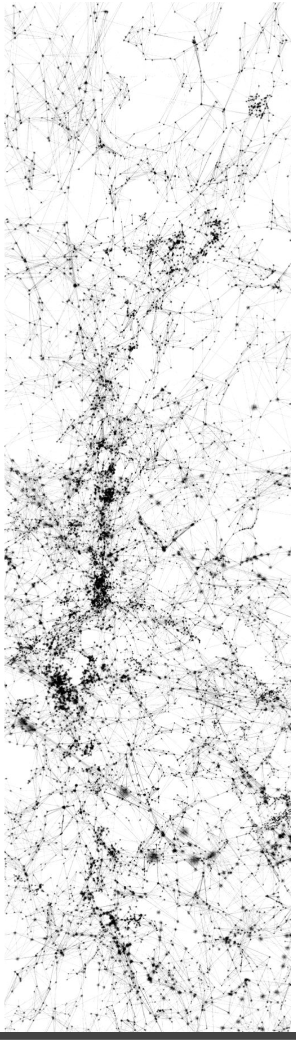
Betweenness centrality has many variants and approximations given its computational complexity and usefulness

Outline

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

$R(i)$ for i in $[1..n]$

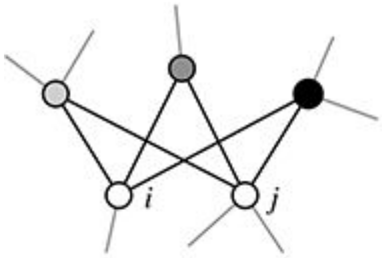
$S(i,j)$ for i,j in $[1..n]$



Number of Common Neighbors

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

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



Cosine similarity

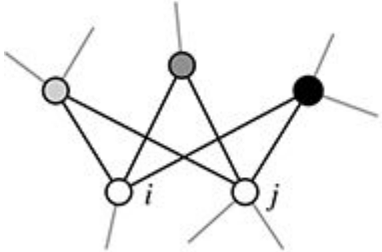
$\sigma_{ij} = \sum_k A_{ik} A_{kj} / (\sqrt{d_i} \sqrt{d_j})$

$$\cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|}$$

$$\sigma_{ij} = \cos\theta = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum_k A_{ik}^2} \sqrt{\sum_k A_{jk}^2}}.$$

$$\sigma_{ij} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{k_i}\sqrt{k_j}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

what is σ_{ij} in example?



Cosine similarity

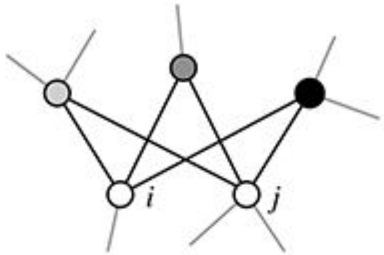
$\sigma_{ij} = \Sigma_k A_{ik} A_{kj} / (\surd d_i \surd d_j)$

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|}$$

$$\sigma_{ij} = \cos\theta = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}}.$$

$$\sigma_{ij} = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{k_i}\sqrt{k_j}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

what is σ_{ij} in example?



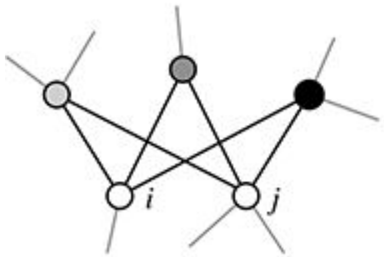
3/(√4×√5)

Other

- Jaccard coefficient

$$J_{ij} = \Sigma_k A_{ik} A_{kj} / (d_i + d_j - \Sigma_k A_{ik} A_{kj})$$ what is J_{ij} in example?

- Pearson correlation coefficient
- Hamming distance
- ....



Other

- Jaccard coefficient

$$J_{ij} = \Sigma_k A_{ik} A_{kj} / (d_i + d_j - \Sigma_k A_{ik} A_{kj})$$ what is J_{ij} in example? 3/6

- Pearson correlation coefficient
- Hamming distance
- ....