Centrality

Argimiro Arratia & R. Ferrer-i-Cancho

Universitat Politècnica de Catalunya

Version 0.4

Complex and Social Networks (2020-2021)

Master in Innovation and Research in Informatics (MIRI)

Official website: `www.cs.upc.edu/~csn/`

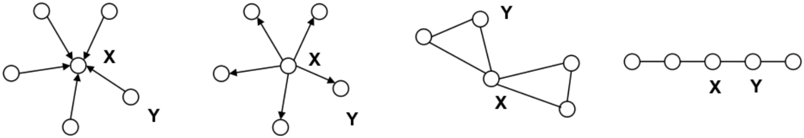
Contact:

- Ramon Ferrer-i-Cancho, rferrericancho@cs.upc.edu, `http://www.cs.upc.edu/~rferrericancho/`
- Argimiro Arratia, argimiro@cs.upc.edu, `http://www.cs.upc.edu/~argimiro/`

What do we mean by centrality?

Centrality is a node's measure w.r.t. others

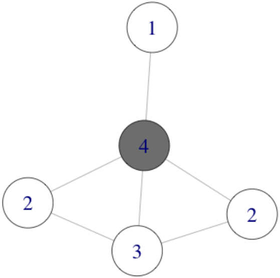
- A central node is *important* and/or *powerful*
- A central node has an *influential position in the network*
- A central node has an *advantageous position in the network*



Degree centrality

Power through connections

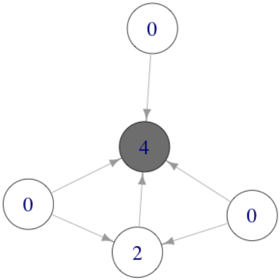
$$degree\_centrality(i) \stackrel{def}{=} k(i)$$



Degree centrality

Power through connections

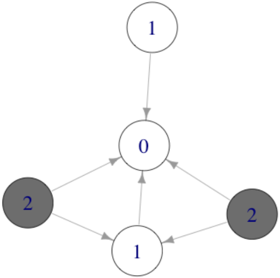
$$in\_degree\_centrality(i) \stackrel{def}{=} k_{in}(i)$$



Degree centrality

Power through connections

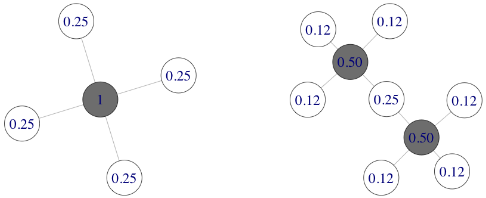
$$out\_degree\_centrality(i) \stackrel{def}{=} k_{out}(i)$$



Degree centrality

Power through connections

By the way, there is a *normalized* version which divides the centrality of each degree by the maximum centrality value possible, i.e. $n-1$ (so values are all between 0 and 1).

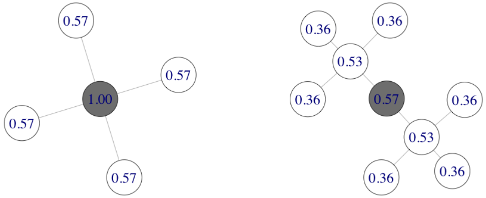


But look at these examples, does degree centrality look OK to you?

Closeness centrality

Power through proximity to others

$$closeness\_centrality(i) \stackrel{def}{=} \left(\frac{\sum_{j \neq i} d(i,j)}{n-1} \right)^{-1} = \frac{n-1}{\sum_{j \neq i} d(i,j)}$$



Here, what matters is to be close to everybody else, i.e., to be easily reachable or have the power to quickly reach others.

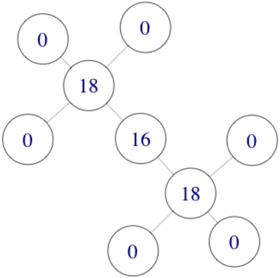
Be aware of ambiguity and failures of this centrality measure!

Betweenness centrality

Power through brokerage

A node is important if it lies in many shortest-paths

- so it is essential in passing information through the network



Betweenness centrality

Power through brokerage

$$betweenness\_centrality(i) \stackrel{def}{=} \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}}$$

Where

- g_{jk} is the number of shortest-paths between j and k, and
- $g_{jk}(i)$ is the number of shortest-paths through i

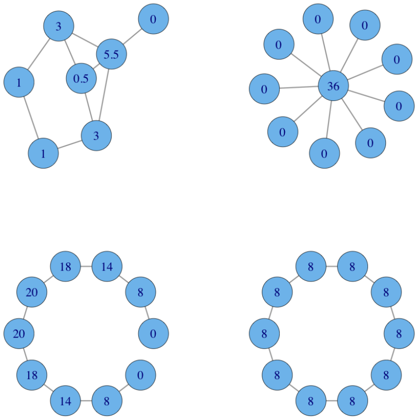
Oftentimes it is normalized:

$$norm\_betweenness\_centrality(i) \stackrel{def}{=} \frac{betweenness\_centrality(i)}{\binom{n-1}{2}}$$

Remarks: i) This measure of centrality offers several advantages ii) [Newman 2010] recommends including extreme points in the count of paths ($j \leq k$): self-paths, etc. But `igraph` implements the fmla. above.

Betweenness centrality

Examples (non-normalized)



Eigenvector centrality

a.k.a. Bonacich centrality, an improvement over degree centrality

Main idea

In degree centrality, each neighbor contributes equally to centrality.
With Bonacich centrality, *important* nodes contribute more.
Namely, a node is central if it is connected to other central nodes.

More precisely, centrality of a node is proportional to the sum of scores of its neighbors.

$$eigenvector\_centrality(i) \propto \sum_j A_{ij} eigenvector\_centrality(j)$$

where A_{ij} is an element of the adjacency matrix, i.e. $A_{ij} = 1$ if i and j share and edge, and $A_{ij} = 0$ otherwise.

Eigenvector centrality I

Computation

To compute, let $x_i = \textit{eigenvector\_centrality}(i)$, for $i = 1, \ldots, n$. Guess an initial value $x_i(0)$ for each $i = 1, \ldots, n$. Then, compute next iteration of values using the formula

$$x_i(t+1) = \sum_{j=1}^{n} A_{ij} x_j(t)$$

Expressed in matrix notation, with $\vec{x} = (x_1, \ldots, x_n)^T$ (as column)

$$\vec{x}(t+1) = \mathbf{A}\vec{x}(t)$$

And so

$$\vec{x}(t) = \mathbf{A}^t \vec{x}(0)$$

Eigenvector centrality II

Computation

Let us express $\vec{x}(0)$ as a linear combination of the eigenvectors $\vec{v}_i$ of $\mathbf{A}$. For the appropriate constants c_i:

$$\vec{x}(0) = \sum_i c_i \vec{v}_i$$

Let λ_i be the eigenvalues of $\mathbf{A}$, and let λ_1 be the largest one. Then

$$\vec{x}(t) = \mathbf{A}^t \vec{x}(0) = \sum_i c_i \lambda_i^t \vec{v}_i = \lambda_1^t \sum_i c_i \left[\frac{\lambda_i}{\lambda_1} \right]^t \vec{v}_i$$

Since $\frac{\lambda_i}{\lambda_1} < 1$ for all $i > 1$, all terms (other than the first) decay exponentially as t grows.

Eigenvector centrality III

Computation

Therefore, in the limit as $t \to \infty$, we have that $\vec{x}(t) \to c_1 \lambda_1 \vec{v}_1$

Eigenvector centrality is *proportional* to the leading eigenvector of **A** (and hence, the name!)
Equivalently, define centrality vector $\vec{x}$ satisfying:

$$\mathbf{A}\vec{x} = \lambda_1 \vec{x}$$

Caveat: Eigenvector centrality does not works in acyclic (directed) networks (asymmetric relations).

Katz or α centrality

An improvement over eigenvector centrality

Main idea: give each vertex a small amount of centrality for free

Define

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

where α and β are positive constants. β is the free contribution for all vertices; hence, no vertex has zero centrality and will contribute at least β to other vertices centrality.
Works in directed acyclic graphs!

Katz or α centrality

In matrix terms:

$$\vec{x} = \alpha \mathbf{A} \vec{x} + \beta \vec{e}$$

where $\vec{e} = (1, 1, \ldots, 1)$. Rearranging for $\vec{x}$ and setting $\beta = 1$:

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

This suggests a good value for α is $0 < \alpha < 1/\lambda_1$, λ_1 the largest eigenvalue of $\mathbf{A}$.[1]
However, instead of computing inverse better to do iterative procedure:

$$\vec{x}(0) = \vec{e}, \quad \vec{x}(t+1) = \alpha \mathbf{A} \vec{x}(t) + \beta \vec{e}$$

[1] We seek α such that $(\mathbf{I} - \alpha \mathbf{A})^{-1}$ does not diverges, i.e. $\det(\mathbf{I} - \alpha \mathbf{A}) \neq 0$, or $\det(\mathbf{A} - \alpha^{-1} \mathbf{I}) \neq 0$. The first value of α that makes this determinant 0 is $\alpha^{-1} = \lambda_1$

Pagerank

An improvement over α centrality

Main idea: the contribution of centrality from each vertex is not the same, it should be diluted in proportion to the amount that is shared with others. **Think:**

- If a very important (central) web page points to my page, as well as to 10 MM other pages, should my web page be equally important (wrto. α centrality), or is my web page just a curiosity as are possibly many of the 10 MM other pages?
- The president of the US connects to all his voters (to keep them informed, etc), is the regular citizen as (political) important as the president of the US?
- The president of the US connects with me (by email or phone) and with no other citizen, am I important?

Pagerank

Definition (Sergey Brin and Larry Page, 1998)

Originally conceived to rank pages in the web (directed graph)

- $V = \{1, .., n\}$ are the nodes (that is, the pages)
- $(i, j) \in E$ if page i points to page j (i.e. $A_{ij} = 1$)
- we associate to each page i, a real value π_i (i's *pagerank*)
- we impose that $\sum_{i=1}^{n} \pi_i = 1$

Define

$$\boxed{\pi_i = \alpha \sum_{j=1}^{n} A_{ji} \frac{\pi_j}{out(j)} + \beta}$$

where $\alpha, \beta > 0$, and $out(j)$ is j's *outdegree*.

Pagerank

Definition (Sergey Brin and Larry Page, 1998)

Brin and Page consider $\beta = \dfrac{(1-\alpha)}{n}$ (and $\alpha = 0.85$). Then

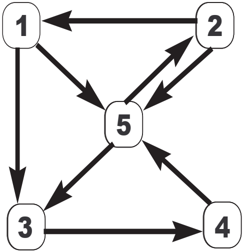
$$\pi_i = \alpha \sum_{j=1}^{n} A_{ji} \frac{\pi_j}{out(j)} + \frac{(1-\alpha)}{n}$$

Note: To avoid indeterminate $(out(j) = 0)$ assume every node has at least $out(j) = 1$ (In graph terms means to allow self-loops)
Then in matrix form

$$(\mathbf{I} - \alpha \mathbf{A}\mathbf{D}^{-1})\pi = \frac{(1-\alpha)}{n}\vec{e}$$

where $\mathbf{D}$ is diagonal matrix with $D_{ii} = \max[out(i), 1]$,
$\pi = (\pi_1, \ldots, \pi_n)^T$ is the Page Rank vector (a probability vector),
and $\vec{e} = (1, 1, \ldots, 1)$.

Pagerank: Example



Want to compute $\pi = (\pi_1, \ldots, \pi_5)$. Solve the system:

$$\pi_1 = \frac{1-\alpha}{5} + \alpha\left(\frac{\pi_2}{2}\right),$$

$$\pi_2 = \frac{1-\alpha}{5} + \alpha\left(\frac{\pi_5}{2}\right),$$

$$\pi_3 = \frac{1-\alpha}{5} + \alpha\left(\frac{\pi_1}{2} + \frac{\pi_5}{2}\right),$$

$$\pi_4 = \frac{1-\alpha}{5} + \alpha\left(\pi_3\right),$$

$$\pi_5 = \frac{1-\alpha}{5} + \alpha\left(\frac{\pi_1}{2} + \frac{\pi_2}{2} + \pi_4\right).$$

For giant network (the WWW) it is unfeasible to do as above.

Pagerank. Example. The power method

Consider in the example the matrix

$$G = \frac{1-\alpha}{5}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} + \alpha \begin{pmatrix} 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 1 & 0 \end{pmatrix}$$

and $\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \end{pmatrix}$, Then previous system of equations is summarize

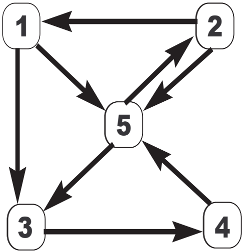
$$\boxed{\pi = G\pi}$$

and in this form we can try solving through the iteration

$$p(k+1) = Gp(k)$$

with initial $p(0) = (p_1, p_2, p_3, p_4, p_5)$, with $0 \leq p_j \leq 1$ and such that $\sum p_j = 1$. (Recall p_j is the probability of being at page j.)

Pagerank. Example. The power method



Approx. solution with $k = 11$ iterations and $p(0) = (0.2, 0.2, 0.2, 0.2, 0.2)$

$$p(11) = \begin{pmatrix} 0.10097776016061 \\ 0.16535594101776 \\ 0.20757694925625 \\ 0.20845457237414 \\ 0.31763477719124 \end{pmatrix}$$

The exact solution by solving the linear system:

$$\pi = \begin{pmatrix} 0.10035700400292 \\ 0.16554589177158 \\ 0.20819761847282 \\ 0.20696797570190 \\ 0.31893151005078 \end{pmatrix}.$$

Pagerank. General matrix form.

In general the Google (or transition) matrix is given by

$$G = \frac{1-\alpha}{n} J + \alpha \mathbf{A}\mathbf{D}^{-1}$$

where J is the $n \times n$ matrix of 1.

And **it is easy** to show that a solution π to
$(\mathbf{I} - \alpha \mathbf{A}\mathbf{D}^{-1})\pi = \dfrac{(1-\alpha)}{n}\vec{e}$, is the same as solving $\pi = G\pi$.
(Hint: note that $J\pi = \vec{e}$ and unravel $(\mathbf{I} - G)\pi = 0$.)

Pagerank. The power method.

So, we seek a solution π for $G\pi = \pi$, and a proposed method is

The Power Method

- Chose initial vector $\vec{p}(0)$ randomly
- Repeat $\vec{p}(t) \leftarrow G\vec{p}(t-1)$
- Until convergence (i.e. $\vec{p}(t) \approx \vec{p}(t-1)$)

Pagerank. The power method.

So, we seek a solution π for $G\pi = \pi$, and a proposed method is

The Power Method

- Chose initial vector $\vec{p}(0)$ randomly
- Repeat $\vec{p}(t) \leftarrow G\vec{p}(t-1)$
- Until convergence (i.e. $\vec{p}(t) \approx \vec{p}(t-1)$)

What guarantees do we have for :

- existence of a solution ?

Pagerank. The power method.

So, we seek a solution π for $G\pi = \pi$, and a proposed method is

The Power Method

- Chose initial vector $\vec{p}(0)$ randomly
- Repeat $\vec{p}(t) \leftarrow G\vec{p}(t-1)$
- Until convergence (i.e. $\vec{p}(t) \approx \vec{p}(t-1)$)

What guarantees do we have for :

- existence of a solution ?
- the power method converges to that solution ?

Pagerank. The power method.

So, we seek a solution π for $G\pi = \pi$, and a proposed method is

The Power Method

- Chose initial vector $\vec{p}(0)$ randomly
- Repeat $\vec{p}(t) \leftarrow G\vec{p}(t-1)$
- Until convergence (i.e. $\vec{p}(t) \approx \vec{p}(t-1)$)

What guarantees do we have for :

- existence of a solution ?
- the power method converges to that solution ?
- The method converges fast to the pagerank solution ?

Pagerank. The power method.

So, we seek a solution π for $G\pi = \pi$, and a proposed method is

The Power Method

- Chose initial vector $\vec{p}(0)$ randomly
- Repeat $\vec{p}(t) \leftarrow G\vec{p}(t-1)$
- Until convergence (i.e. $\vec{p}(t) \approx \vec{p}(t-1)$)

What guarantees do we have for :

- existence of a solution ?
- the power method converges to that solution ?
- The method converges fast to the pagerank solution ?
- The method converges fast to the pagerank solution regardless of the initial vector ?

Pagerank. Guarantee of a solution.

That a solution exists is guaranteed by

Theorem (Perron-Frobenius)

If M is stochastic, then it has at least one stationary vector, i.e., one non-zero vector p such that $M^T p = p$.

(M is stochastic if all entries are in the range $[0, 1]$ and each row adds up to 1)

The transpose of Google matrix is row-stochastic. (check)

Pagerank. Guarantee for convergence of power method

A useful theorem from Markov chain theory

Theorem

If a matrix M is strongly connected and aperiodic, then:

- $M^T\vec{p} = \vec{p}$ *has exactly one non-zero solution such that* $\sum_i p_i = 1$
- *1 is the largest eigenvalue of* M^T
- *the Power method converges to the* $\vec{p}$ *satisfying* $M^T\vec{p} = \vec{p}$, *from any initial non-zero* $\vec{p}(0)$
- *Furthermore, we have exponential fast convergence*

Pagerank. The Google matrix works!

The Google Matrix,

$$G = \frac{1-\alpha}{n} J + \alpha \mathbf{A}\mathbf{D}^{-1}$$

where J is a $n \times n$ matrix containing 1 in each entry.

Pagerank. The Google matrix works!

The Google Matrix,

$$G = \frac{1-\alpha}{n} J + \alpha \mathbf{A}\mathbf{D}^{-1}$$

where J is a $n \times n$ matrix containing 1 in each entry.

- G is stochastic

Pagerank. The Google matrix works!

The Google Matrix,

$$G = \frac{1-\alpha}{n} J + \alpha \mathbf{AD}^{-1}$$

where J is a $n \times n$ matrix containing 1 in each entry.

- G is stochastic
  - ...because G is a weighted average of $\mathbf{AD}^{-1}$ and $\frac{1}{n}J$, which are also stochastic

Pagerank. The Google matrix works!

The Google Matrix,

$$G = \frac{1-\alpha}{n} J + \alpha \mathbf{AD}^{-1}$$

where J is a $n \times n$ matrix containing 1 in each entry.

- G is stochastic
  - …because G is a weighted average of $\mathbf{AD}^{-1}$ and $\frac{1}{n}J$, which are also stochastic
- for each integer $k > 0$, there is a non-zero probability path of length k from every state to any other state of G

Pagerank. The Google matrix works!

The Google Matrix,

$$G = \frac{1-\alpha}{n} J + \alpha \mathbf{AD}^{-1}$$

where J is a $n \times n$ matrix containing 1 in each entry.

- G is stochastic
  - ...because G is a weighted average of $\mathbf{AD}^{-1}$ and $\frac{1}{n}J$, which are also stochastic
- for each integer $k > 0$, there is a non-zero probability path of length k from every state to any other state of G
  - ...implying that G is strongly connected and aperiodic

Pagerank. The Google matrix works!

The Google Matrix,

$$G = \frac{1-\alpha}{n} J + \alpha \mathbf{AD}^{-1}$$

where J is a $n \times n$ matrix containing 1 in each entry.

- G is stochastic
  - ... because G is a weighted average of $\mathbf{AD}^{-1}$ and $\frac{1}{n}J$, which are also stochastic
- for each integer $k > 0$, there is a non-zero probability path of length k from every state to any other state of G
  - ... implying that G is strongly connected and aperiodic
- and so the Power method will converge on G, and fast!

Pagerank

Teleportation in the random surfer view

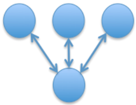
The meaning of α (the damping factor)

- With probability α, the random surfer follows a link in current page
- With probability $1 - \alpha$, the random surfer jumps to a random page in the graph (teleportation)

Pagerank

Excercise.

Compute the pagerank value of each node of the following graph assuming a damping factor $\alpha = 2/3$:



Hint: solve the following system, using $p_2 = p_3 = p_4$

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \left[\frac{2}{3} \begin{pmatrix} 0 & 1 & 1 & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 \end{pmatrix} + \frac{1}{3} \cdot \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \right] \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$$

Eigenvector-based centrality as power series

α-centrality

If α is smaller than the inverse of the spectral radius of $\mathbf{A}$, i.e. $\alpha < 1/\lambda_1$, we have convergence of the series

$$(\sum_{k=0}^{\infty} \alpha^k \mathbf{A}^k)\vec{e} = (\mathbf{I} - \alpha \mathbf{A})^{-1} \cdot \vec{e} = \vec{x}$$

This series is in fact the original form of centrality conceived by Katz (1953): it considers for each vertex i the influence of all the vertices connected by a walk to i.

This suggests other way of computing $\vec{x}$ by taking successive partial sums.

Eigenvector-based centrality as power series

Pagerank on rooted trees

Theorem (Arratia-Marijuán (LAA16))

If a rooted tree has N vertices and height h, then the PageRank of its root r is given by

$$PageRank(r) = \frac{1-\alpha}{N} \sum_{k=0}^{h} \alpha^k n_k \tag{1}$$

where n_k is the number of vertices of the kth–level of the tree. □

This shows that we can do any rearrangements of links between two consecutive levels of a web set up as a rooted tree, and the PageRank of the root will be the same.

Eigenvector-based centrality as power series

Pagerank as power series [Brinkmeier, 2006]

For a given walk $\rho = v_1 v_2 \dots v_n$ in the graph define the **branching factor** of ρ by the formula

$$D(\rho) = \frac{1}{od(v_1)od(v_2)\cdots od(v_{n-1})} \tag{2}$$

Then, for any vertex $a \in V$, we have

$$PageRank(a) = \frac{1-\alpha}{N} \sum_{l \geq 0} \sum_{\rho : w \xrightarrow{l} a} \alpha^l D(\rho) \tag{3}$$

where $\rho : w \xrightarrow{l} a$ denotes a walk ρ from any w to a of length l.
Note: For $D(\rho) = 1$ for all walks ρ, we recover the power series for α-centrality

Centrality measures in `igraph`

- `degree()`
- `betweenness()` , (vertex and edge)
- `alpha.centrality()`
- `page.rank()`