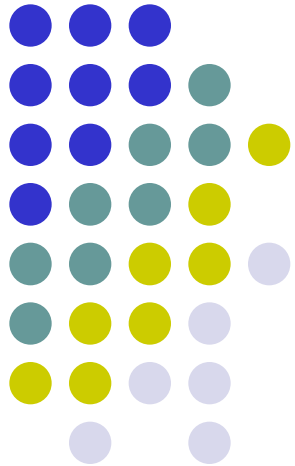
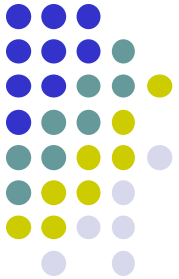
Web Algorithms – Web Search

Part 1: Social Networks and Link Popularity

Eng. Fabio Persia, PhD

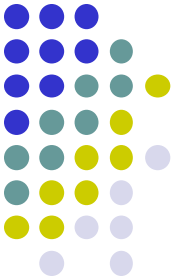


Overview



- Social networks and bibliometry
- Centrality measures
- Spectral analysis and prestige index
- The web graph structure

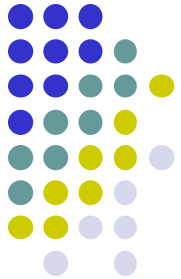
Social science and bibliometry



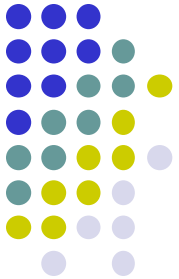
- Initial techniques for web search were borrowed from classical Information Retrieval (occurrence of keywords in text)

- Unfortunately, there are fundamental differences that sometimes make them less effective or partially used in the search engines:
  - Dynamicity of the collection of pages
  - Eterogeneity
  - abundance
  - redundancies and duplications
  - maliciousness of the authors
  - …
  - but above all hyperlinks !

- Web search: classical IR + spectral analysis + ????
  - Spectral analysis: page importance determined as a function of the network structure of Web pages
  - ????: secret + frequent strategy modification for beating competitors and avoid users to adjust for getting relevance



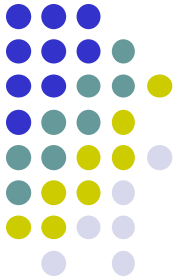
- Hyperlinks provide supplementary information to the normal text that often exceeds the same text for the quality of information that can be inferred
- In particular, they help in establishing measures of authority or popularity of pages that can accompany classical relevance measures
- Such measures originated in so-called social networks, used to model social relations and interactions between individuals and entities



- Networks of social interaction are established between
  - academics through partnerships, participation to councils and committees, …
  - movies staff through directing and acting,
  - musicians, football stars, friends and relatives
  - people who make phone calls or transmit infections ,
  - countries through trade relations, ...

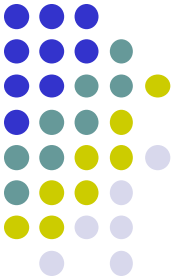
- ... between Web pages via hyperlinks!

- The Web is an example of a social network, but social networks have been the subject of consistent research well before the advent of the Web itself



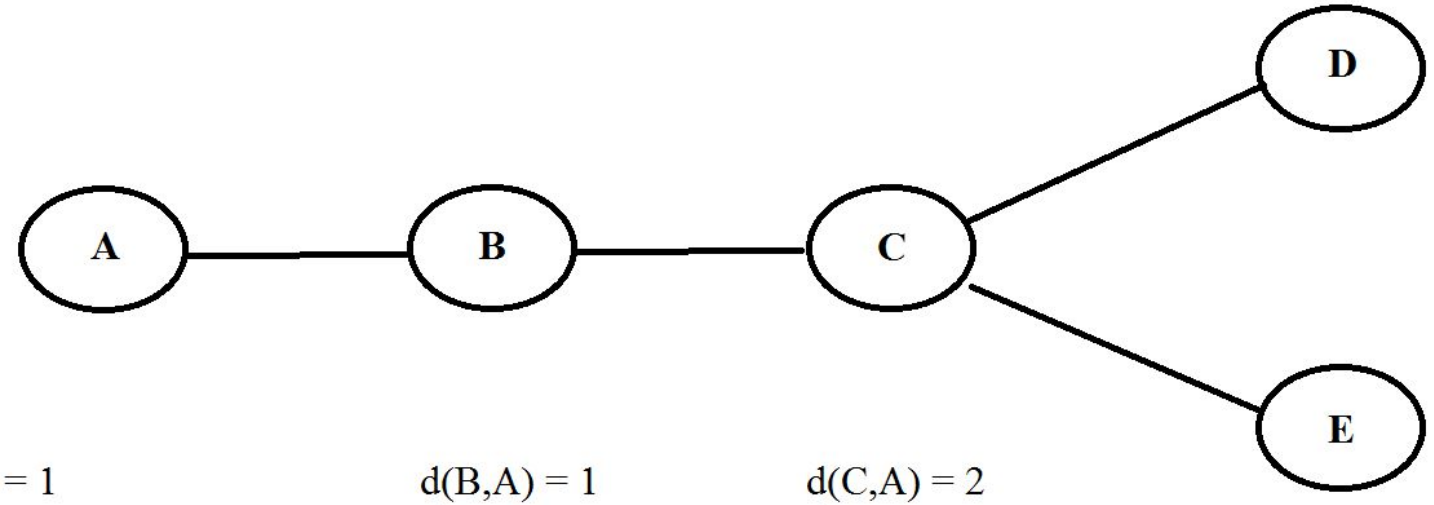
- The applications of the theory of social networks are varied, ranging from epidemiology, to espionage, to citation indices , ...
- In the first two cases we seek for the determination of subsets of nodes whose removal significantly increase the average distance between nodes, while in the third to the identification of articles central and influential in the scientific literature
- In general, the theory of social networks is interested in the determination of properties related to connectivity and distance in graphs
- Among the most relevant we have
  - the centrality
  - co-citation
  - social prestige

Centrality



- Many notions of centrality based on graphs have been proposed in the literature of social networks
- The distance ***d(u,v)*** between two nodes ***u*** and ***v*** in an unweighted graph is the smallest number of links through which you can go from ***u*** to ***v*** (in the case of weighted graphs we add weights to the arcs to derive the length of the path)
- The radius of a node ***u*** is $r(u)=max_v\ d(u,v)$
- The center of the graph is defined as the node with the smallest radius, i.e. with the minimum ***r(u)***
- You might want to search for documents influential in an area of research looking for those documents ***u*** which have smallest ***r(u)***, which means that most of the documents in the research community has a short citation path from path of short quote starting from ***u***

Centrality - Example



d(A,B) = 1
d(A,C) = 2
d(A,D) = 3
d(A,E) = 3
r(A) = 3

d(B,A) = 1
d(B,C) = 1
d(B,D) = 2
d(B,E) = 2
r(B) = 2

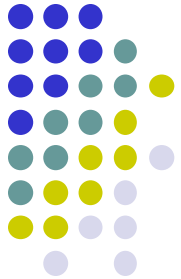
d(C,A) = 2
d(C,B) = 1
d(C,D) = 1
d(C,E) = 1
r(C) = 2

d(D,A) = 3
d(D,B) = 2
d(D,C) = 1
d(D,E) = 2
r(D) = 3

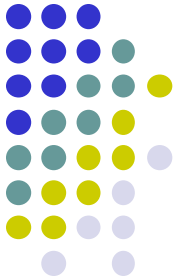
d(E,A) = 3
d(E,B) = 2
d(E,C) = 1
d(E,D) = 2
r(E) = 3

So, the center of the graph is **node C**

Centrality



- Depending on the context other notions of centrality may be more appropriate
- For example, in the study of epidemiology, espionage, or telephone communications suspected of terrorism, it is often useful to identify cuts … whatever it means ☺
- They consist of a small set of arcs which, when removed, disconnect a given pair of vertices
- Or you can look for small groups of vertices that, removed with their incident edges, decompose the graph into two or more connected components
- However, there are a myriad of other formulations and measures although none is well-suited for all possible applications
- The repertoire of the measures is now sufficiently mature and consolidated



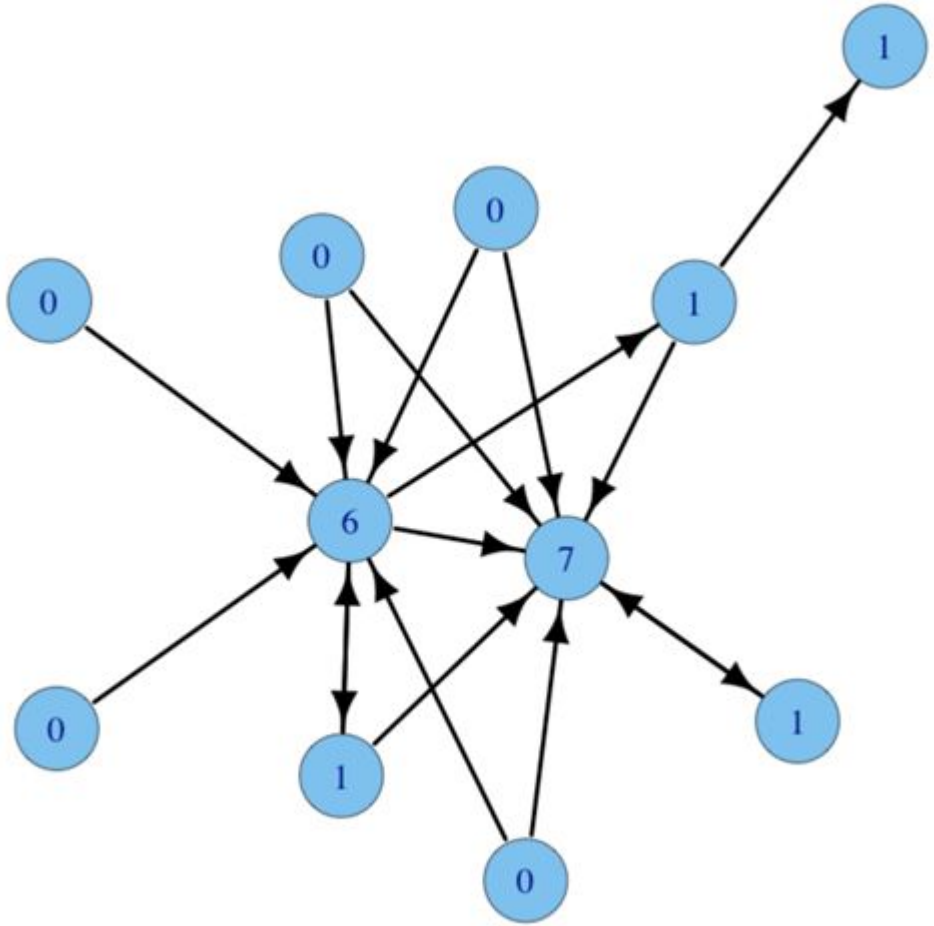
Some classical measures of centrality ***c(v)*** for a node ***v***:

- Degree centrality: ***c[v] = deg(v)*** (degree of *v*)
  - If the graph is directed, it can be **in-degree** and **out-degree**
- Closeness centrality: $c[v] = \Sigma_{t \in V,\, t \neq v} \, 1/d(v,t)$
  - It indicates how close a node is to all other nodes in the network
  - The higher the better
- Betweenness centrality: $c[v] = \Sigma_{s,t \in V,\, s \neq t \neq v} \, \sigma_{st}(v)/\sigma_{st}$

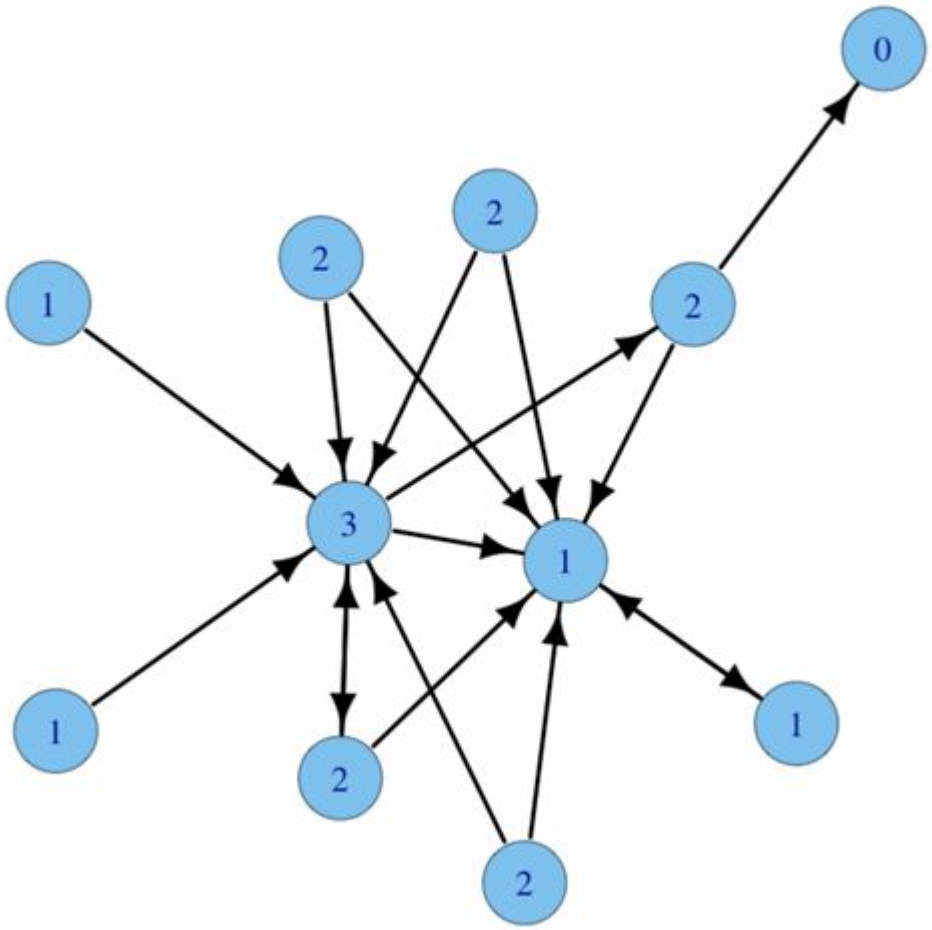
$\sigma_{st}(v)$ = number shortest paths from *s* to *t* containing *v*

σ_{st} = total number shortest paths from *s* to *t*

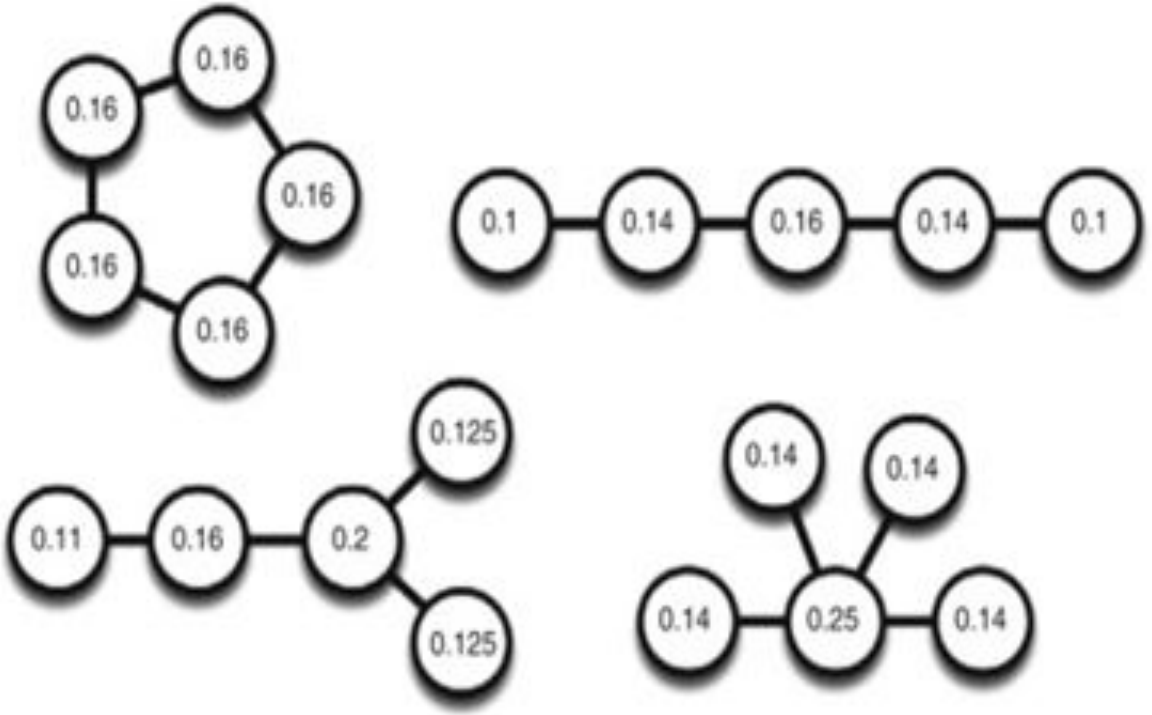
In-degree Centrality



Out-degree Centrality



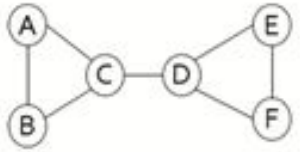
Closeness Centrality



Betweenness Centrality (1/3)

BETWEENNESS CENTRALITY

Calculate the between centrality for node C

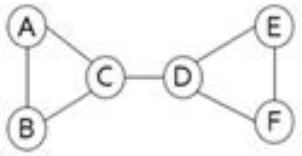


	σ_{uw}	$\sigma_{uw}(v)$	$\sigma_{uw}(v) / \sigma_{uw}$
(A,B)	1	0	0
(A,D)	1	1	1
(A,E)	1	1	1
(A,F)	1	1	1
(B,D)	1	1	1
(B,E)	1	1	1
(B,F)	1	1	1

Betweenness Centrality (2/3)

BETWEENNESS CENTRALITY

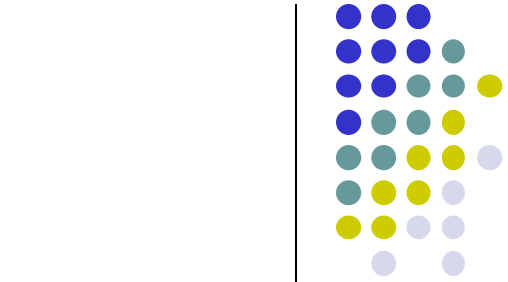
Calculate the between centrality for node C



	σ_{uw}	$\sigma_{uw}(v)$	$\sigma_{uw}(v) / \sigma_{uw}$
(D,E)	1	0	0
(D,F)	1	0	0
(E,F)	1	0	0

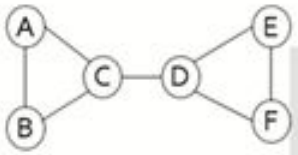
Betweenness Centrality for C =

Betweenness Centrality (3/3)



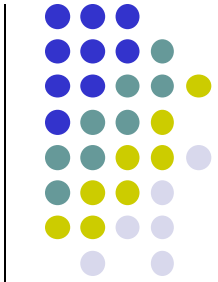
BETWEENNESS CENTRALITY

	σ_{uw}	$\sigma_{uw}(v)$	$\sigma_{uw}(v) / \sigma_{uw}$
(A,B)	1	0	0
(A,D)	1	1	1
(A,E)	1	1	1
(A,F)	1	1	1
(B,D)	1	1	1
(B,E)	1	1	1
(B,F)	1	1	1
(D,E)	1	0	0
(D,F)	1	0	0
(E,F)	1	0	0



Betweenness Centrality for C = 6

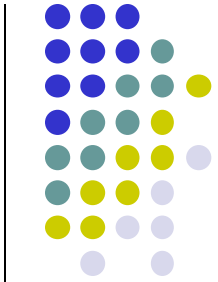
Co-citation



- If document ***u*** cites documents ***v*** and ***w***, then ***v*** and ***w*** are said to be ***co-citated*** by ***u***

- The fact that two documents ***v*** and ***w*** are co-citated by many documents ***u*** gives evidence of the fact that ***v*** and ***w*** are in some way correlated

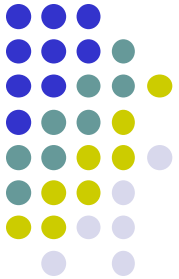
- Consider the **adjacency matrix *E*** of the citation graph of the documents, where ***E[u, v] = 1*** if document ***u*** cites document ***v***, ***0*** otherwise



- Then

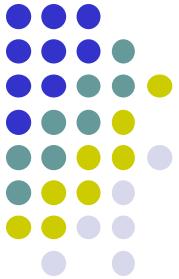
$$(E^T E)[v,w] =$$
$$= \sum_u E^T[v,u]E[u,w] =$$
$$= \sum_u E[u,v]E[u,w] =$$
$$= \left|\{u : (u,v) \in E, (u,w) \in E\}\right|$$

- Entry ***(v, w)*** of matrix $E^T E$ is called ***co-citation index*** of ***v*** and ***w*** and is an indicator of correlation between ***v*** and ***w***

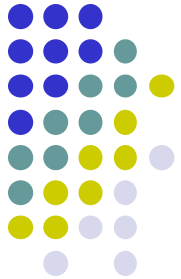


- This index can be used to build clusters of web pages based on their co-citations
- Such clusters reveal important social structures between and within communities

Prestige



- Using weighted directed graphs to model social networks is very common
- In such a model, it is clear that the incoming degree (in-degree) is a good (first order) indicator of the status or prestige
- Since 1949 , Seeley stressed the recursive nature of prestige in a social network : " … The status of an actor is a function of the status of those who choose it; and the status of these is in turn a function of those who had chosen them, and so forth ad infinitum"

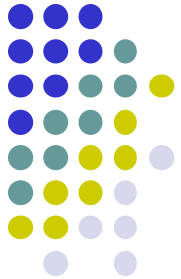


- Consider again the adjacency matrix ***E*** of the citation graph of the documents ***E*** :
  - *E[u, v]=1* if node u has arc toward node *v,*
  - *E[u, v]*= 0 otherwise

- Every node ***v*** has an associated measure of prestige ***p*[*v*]** that is simply a positive real number

- For all the nodes, we represent their prestige scores as a vector ***p***

- Suppose we want to give to each node ***v*** the total sum of the prestige of all the nodes ***u*** having a link to ***v***, thus computing a new prestige vector ***p'***



- This can be easily written in matrix notation as

$$p' = E^T p$$

Since

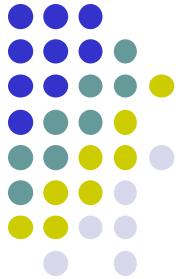
$$p'[v] =$$

$$= \sum_u E[u,v]p[u] =$$

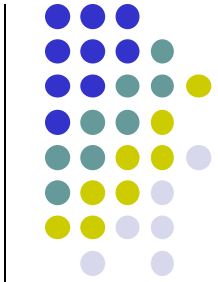
$$= \sum_u E^T[v,u]p[u]$$

- Before continuing with the notion of prestige, let us open a small mathematical (linear algebra) parenthesis

Eigenvectors and eigenvalues



- Given a square matrix ***A*** of ***n*** rows and ***n*** columns and a vector of *n* elements ***p***, ***p*** is an eigenvector of ***A*** if $\boldsymbol{A \cdot p = \lambda p}$
- $\boldsymbol{\lambda}$ is called eigenvalue of *A* with respect to *p*
- Informally, the eigenvalue $\boldsymbol{\lambda}$ espresses the projection or expansion of ***A*** in the direction of ***p***
- Notice that eigenvectors of ***A*** corresponing to a given eigenvalue $\boldsymbol{\lambda}$ are defined up to a scaling factor, in the sense that if ***p*** is an eigenvector, then also all the $\boldsymbol{p' = a\,p}$ obtained by multiplying ***p*** for any scalar ***a*** are also eigenvector of ***A*** with the same eigenvalue
- In the following for the sake of simplicity we will identify an eigenvector both individually and as the class of all the vectors obtained by scaling the same eigenvector ***p***
- The eigenvector of maximum absolute value, if unique (in terms of class), is called the ***principal*** or ***dominant eigenvector*** of ***A***
- The principal eigenvector is in some sense the most significant eigenvector of ***A,*** as it has the maximum “performance”, that is the maximum expansion or amplification of ***A*** in its direction



- QUESTION: how to compute a dominant eigenvector ***p*** of ***A***?

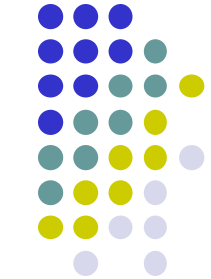
- ***Power iteration*** method from linear algebra: under suitable conditions able to find a solution converging to ***p:***

  - Let initially $\mathbf{p=(1,\ldots,1)^T}$
  - Let iteratively ***p←A p****,* normalizing norm to *1* to avoid overflow, i.e. dividing ***p*** by

$$\|p\|_1 = \sum_u p[u]$$

- Converges effectively if there is only one eigenvalue of maximum absolute value i.e. there exist the principal eigenvector) and if ***A*** is diagonalizable, i.e. there exists a matrix ***T*** such that $\mathbf{A = T^{-1} A' T,}$ with ***A'*** having non null entries only along the main diagonal

Why Power Iteration works?



Let

- $p_0=(1,\ldots,1)^T$
- $p_i = A\ p_{i-1}$ for every $i>0$

Theorem. Sequence p_i converges to the principal eigenvector of A (under particular assumptions)

Proof:

- Assume A has n linearly independent eigenvectors, $x_1, x_2, \ldots, x_n$ with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, where $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$
- Vectors $x_1, x_2, \ldots, x_n$ form a basis and thus we can write:

$$p_0 = c_1\ x_1 + c_2\ x_2 + \cdots + c_n\ x_n$$

- $p_1 = A\ p_0 = A(c_1\ x_1 + c_2\ x_2 + \cdots + c_n\ x_n)$

$$= c_1(Ax_1) + c_2(Ax_2) + \cdots + c_n(Ax_n)$$
$$= c_1(\lambda_1 x_1) + c_2(\lambda_2 x_2) + \cdots + c_n(\lambda_n x_n)$$

- **Repeated multiplication on both sides produces**

$$p_i = c_1(\lambda_1^i x_1) + c_2(\lambda_2^i x_2) + \cdots + c_n(\lambda_n^i x_n)$$

Then

$$p_i = \lambda_1^i \left[c_1 x_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^i x_2 + \cdots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^i x_n \right]$$

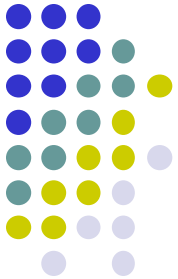
- Since $\lambda_1 > \lambda_2$ then fractions $\frac{\lambda_2}{\lambda_1}, \frac{\lambda_3}{\lambda_1} \ldots < 1$

  and so $\left(\frac{\lambda_k}{\lambda_1}\right)^i = 0$ as $i \to \infty$ (for all $k = 2 \ldots n$).

- Thus $p_i \approx c_1(\lambda_1^i x_1)$

  Note if $c_1 = 0$ then the method won't converge

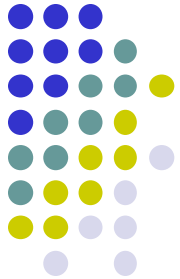
Prestige (continued …)



- Let us now come back to the prestige vector ***p***
- In order to reach a fixed point for the prestige vector, we can use the just introduced power iterations method
  - start initializing $\boldsymbol{p=(1,\ldots,1)^T}$
  - iteratively
    - ***Let*** $\boldsymbol{p \leftarrow E^t p}$***,***
    - normalize ***p*** to avoid overflow (divide *p* by $\|p\|_1 = \sum_u p[u]$)
- Under the already mentioned conditions, the method tends to a converging value of *p*, i.e. the fixed point, that is the **principal eigenvector** of the matrix E^T, that is having the eigenvalue with maximum absolute value
- Notice that if the graph is symmetric (or not directed) E^T is symmetric and therefore diagonalizable, so to obtain the convergence it is sufficient that there exists a unique eigenvalue of maximum absolute value
- There are further improvements of this method obtained using attenuations factors of the type

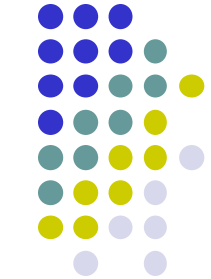
  $\boldsymbol{p' = \alpha E^T p}$
- PROBLEM: in the directed case, like in the web graph, the method hardly converges, since usually we don't have a dominant eigenvector ….

Web and Link popularity



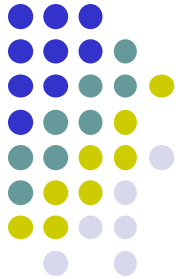
- In the latest generation search engines, resulting pages are sorted not only based on how they approach to the information needs of the user (Information Retrieval), but also according to the **popularity** of the page
- Ex. The official site of Ferrari should appear before the amateur site created by anyone else !
- To do this, there exist modern algorithms, based on the analysis of the links, that allow you to discriminate also according to the "***importance***"of the pages found on the Web
- They are strongly inspired by the notion of prestige defined within the social networks
- In this way they try to cope more effectively with the problem selecting in a set of plenty of pages in the web

Ranking algorithms based on the analysis of the links



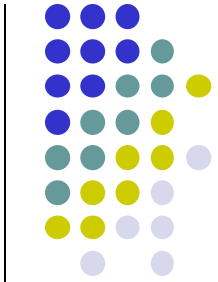
- The basic steps are the following:
  - Start from a collection of web pages
  - Extract from them the graph of the hyperlinks
  - Run the ranking algorithm on the graph
  - Output : a weight of ***popularity*** for each node

- Such algorithms can be distinguished in two broad categories :
  - Query independent: they make the ranking of the entire Web
  - Query dependent: they make the ranking of a relatively small subset of pages related to a specific query



- In the next set of slides we are going to introduce in detail some of the most famous algorithms in this setting:
  - **PageRank** (query independent)
  - **HITS** (query dependent)
  - **SALSA ???** (query dependent)

Ranking algorithms based on link analysis



- We will examine:
  - **PageRank**
  - **Topic-Specific PageRank**
  - **HITS**
  - **SALSA**
  - **TrustRank**