Lecture 13

Network Science

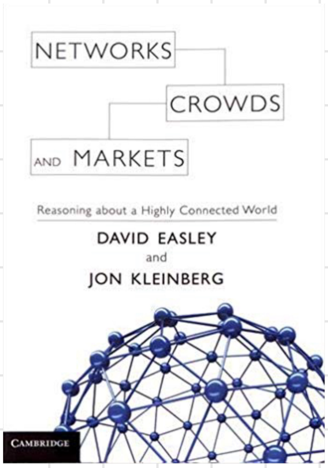
Spectral Analysis, Random Walks and Web Search

Today's topics

- Spectral Analysis of Hubs and Authorities

- Spectral Analysis of PageRank

- Formulation of PageRank using Random Walks



Chapter 14
"Link Analysis and Web Search"

Section 14.6

Introduction

we need to analyze the methods to compute hubs, authorities, and page rank values

Pre-requisites: linear algebra
vector and matrix multiplication

limiting values are coordinates in eigenvectors for given eigen values in matrices derived from our graphs

eigen values / eigenvectors calculation to study the structure of networks → Spectral analysis

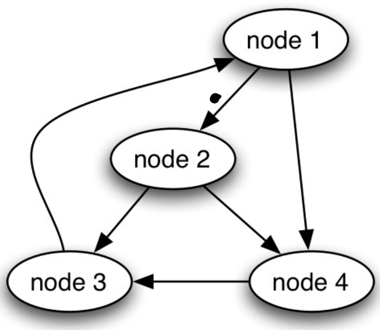
Spectral Analysis of Hubs and Authorities

Def. Adjacency Matrix

nodes: $1, \dots, n$

M : $n \times n$

$$M_{ij} = \begin{cases} 1 & \text{if } (i,j) \text{ is an edge} \\ 0 & \text{otherwise} \end{cases}$$



$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Figure 14.11: The directed hyperlinks among Web pages can be represented using an *adjacency matrix* M: the entry M_{ij} is equal to 1 if there is a link from node i to node j, and $M_{ij} = 0$ otherwise.

not necessarily efficient for computational representations.

h, a : n-dim vectors

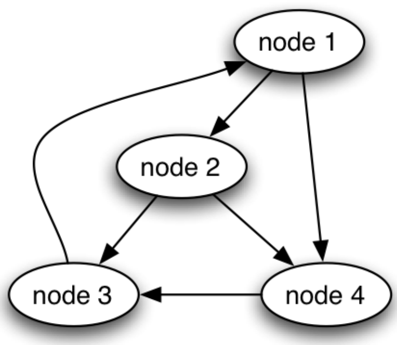
$h \to$ hubs, $a \to$ authority

Hub Update rule

$$h_i \leftarrow \sum_{j=1}^{n} M_{ij} a_j$$

$$= M_{i1} a_1 + M_{i2} a_2 + \dots + M_{in} a_n$$

$$h \leftarrow M a$$



$$\overset{M}{\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}} \overset{a}{\begin{bmatrix} 2 \\ 6 \\ 4 \\ 3 \end{bmatrix}} = \overset{h}{\begin{bmatrix} 9 \\ 7 \\ 2 \\ 4 \end{bmatrix}}$$

Figure 14.12: By representing the link structure using an adjacency matrix, the Hub and Authority Update Rules become matrix-vector multiplication. In this example, we show how multiplication by a vector of authority scores produces a new vector of hub scores.

Authority Update rule

$$a_i \leftarrow \sum_{j=1}^{n} M_{ji} h_i$$

$$* \quad a \leftarrow M^T h$$

Understanding the k-step hub-authority computation

$$a^{<0>},\ h^{<0>} = \underbrace{(1, 1, \dots, 1)}_{n \text{ times}}$$

$a^{<k>},\ h^{<k>}$: after k applications of update rules

1_{st} : $a^{<1>} = M^T h^{<0>}$

$h^{<1>} = M\, a^{<1>} = M M^T h^{<0>}$

2nd : $a^{<2>} = M^T h^{<1>} = (M^T M) M^T h^{<0>}$.

$h^{<2>} = M\, a^{<2>} = (M M^T)(M M^T) h^{<0>} =$

$= (M M^T)^2 h^{<0>}$

...

k_{th} : $a^{<k>} = (M^T M)^{k-1} M^T h^{<0>}$

$h^{<k>} = (M M^T)^k h^{<0>}$

a, h vectors : multiplication of an initial vector $h^{<0>}$ by larger and larger powers of $M^T M$ and $M M^T$

Multiplications and Eigenvectors

normalization

constants c and d

$$\frac{h^{<k>}}{c^k} \quad \text{and} \quad \frac{a^{<k>}}{d^k}$$

$\Rightarrow$ converge to some values for $k \to \infty$

focus on hubs vectors:

if $\dfrac{h^{<k>}}{c^k} = \dfrac{(MM^T)^k \, h^{<0>}}{c^k}$ converges

to a limit $h^{<*>}$, then I can expect that

$$\underset{\uparrow\ \text{eigenvalue}}{\underline{c}}\ \underset{\text{eigenvector}}{\underline{h}^*} = \underline{(MM^T)}\ \underline{h}^{<*>}$$

we need to prove that the sequence of $\dfrac{h^{<k>}}{c^k}$ converges to the eigenvector of MM^T

Square matrix A is symmetric

$A = A^T \qquad (A_{ij} = A_{ji})$

$MM^T = ?$ it is symmetric!

fact 1 "Any symmetric matrix $n \times n$ has a set of n eigenvectors that are orthogonal and all unit vectors - that is they form a basis for the space $\mathbb{R}^n$"

$z_i \cdot z_j = 0 \qquad z_i \cdot z_i = 1$

MM^T is symmetric, we can apply (fact 1) to find

$z_1, \ldots, z_n$ n mutually orthog. eigenvectors

$c_1, \ldots, c_n$ n corresponding eigenvalues

$|c_1| \geq |c_2| \geq \ldots \geq |c_n|$ Assume $|c_1| > |c_2|$

x (n values)

$$x = \rho_1 z_1 + \rho_2 z_2 + \cdots + \rho_n z_n$$

(for $\rho_1, \rho_2, \ldots, \rho_n$ coefficients)

$$(MM^T)x = (MM^T)(\rho_1 z_1 + \cdots + \rho_n z_n)$$

$$= \rho_1 \underline{MM^T z_1} + \cdots + \rho_n \underline{MM^T z_n}$$

$$= \rho_1 c_1 z_1 + \cdots + \rho_n c_n z_n$$

we will use it to analyze multiplication by larger powers of (MM^T)

$$(MM^T)^k x = \rho_1 c_1^k z_1 + \cdots + \rho_n c_n^k z_n$$

Convergence of the hub-Authority Computation

vectors of hub scores:

$$h^{<k>} = (MM^T)^k \, h^{<0>}$$

$$h^{<0>} = q_1 z_1 + q_2 z_2 + \dots + q_n z_n$$

$$h^{<k>} = c_1^k q_1 z_1 + \dots + c_n^k q_n z_n$$

let's divide both sides by c_1^k

$$\frac{h^{<k>}}{c_1^k} = \frac{\cancel{c_1^k} q_1 z_1}{\cancel{c_1^k}} + \dots + \frac{c_n^k q_n z_n}{c_1^k}$$

assumption:

$c_1 > c_2 \Rightarrow$

$\Rightarrow k \to \infty$: $\left(\frac{c_2}{c_1}\right)^k \to 0$

...

$$\lim_{k \to \infty} \frac{h^{<k>}}{c_1^k} = q_1 z_1$$

0

Wrapping up

(i) e limit in the direction of z_1 is reached regardless of initial values of $h^{<*>}$

Let's suppose that $h^{<0>} = x$ and that it is a positive vector:

$$x = p_1 z_1 + \dots + p_n z_n \quad \text{so}$$

$$(\Psi\Psi^T)\, x = c_1^k p_1 z_1 + \dots + c_n^k p_n z_n$$

$$\lim_{k \to \infty} \frac{h^{<k>}}{c_1^k} = p_1 z_1 \qquad \square$$

(ii) coefficient p_1 (or q_1) must be $\neq 0$: assuring that $p_1 z_1$ (or $q_1 z_1$) are non zero vectors in the direction of z_1 $\Rightarrow$ text book!

iii) relax assumption

$|c_1| > |c_2|$

in general we can have $l > 1$ eigenvalues s.t.

$$c_1 = c_2 = \dots = c_l$$

$$c_{l+1} < c_1$$

$$\frac{h^{<k>}}{c_1^k} = \frac{c_1^k q_1 z_1 + \dots + c_n^k q_n z_n}{c_1^k}$$

$k \to \infty$

$$= \boxed{q_1 z_1 + \dots + q_l z_l} + 0$$

this is still a convergence

authority values: the argument is very similar to hubs (multiplication by $M^T M$.)

Spectral Analysis of PageRank

Basic PR update rule

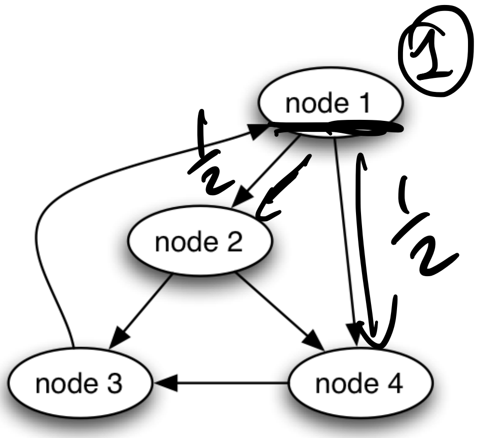
N : $n \times n$ matrix

N_{ij} : the share of i's PR that j should get in one update step

$$N_{ij} = \begin{cases} 0 & \text{if } (i,j) \notin E \\ \dfrac{1}{l_i}, & \text{where } l_i : \#\text{ links out of } i \end{cases}$$

$N_{ii} = 1$, if $l_i = 0$

(to let PR be preserved in nodes with no outgoing links)



$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Figure 14.13: The flow of PageRank under the Basic PageRank Update Rule can be represented using a matrix N derived from the adjacency matrix M: the entry N_{ij} specifies the portion of i's PageRank that should be passed to j in one update step.

r_i : PR of node i

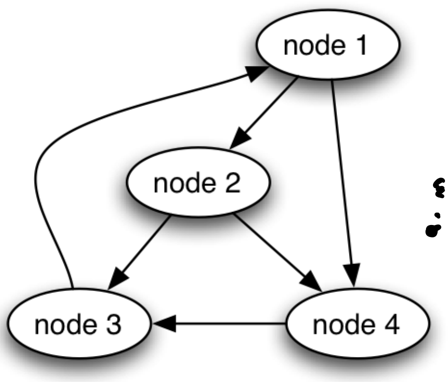
$$r_i \leftarrow N_{1i} r_1 + N_{2i} r_2 + \dots + N_{ni} r_n$$

$$r \leftarrow N^T r$$

Scaled PR update rule

factor s

$$\tilde{N}_{ij} = s N_{ij} + \frac{(1-s)}{n}$$



$$\tilde{N} = \begin{bmatrix} .05 & .45 & .05 & .45 \\ .05 & .05 & .45 & .45 \\ .85 & .05 & .05 & .05 \\ .05 & .05 & .85 & .05 \end{bmatrix}$$

Figure 14.14: The flow of PageRank under the Scaled PageRank Update Rule can also be represented using a matrix derived from the adjacency matrix M (shown here with scaling factor $s = 0.8$). We denote this matrix by $\tilde{N}$; the entry $\tilde{N}_{ij}$ specifies the portion of i's PageRank that should be passed to j in one update step.

Scaled PR update rule application:

$$r_i \leftarrow \tilde{N}_{1i} r_1 + \tilde{N}_{2i} r_2 + \dots + \tilde{N}_{ni} r_n$$

$$\boxed{r \leftarrow \tilde{N}^T r}$$

Repeated Improvement

$r^{<0>}$ initial PR vector

$(1/n, \ldots, \frac{1}{n})$

- $r^{<k>} = (\tilde{N}^T)^k r^{<0>}$

limiting vector $r^{<*>}$

satisfies:

- $\tilde{N}^T r^{<*>} = 1 \cdot r^{<*>}$

$r^{<*>}$ should be an eigenvector of $\tilde{N}^T$ with corresponding eigenvalue of 1

BUT $\tilde{N}^T$ is <u>not</u> symmetric

that means that eigenvalues can be complex numbers and eigenvectors have no clear relationships to one another

Convergence of the Scaled PR Update Rule

$\tilde{N}_{ij} > 0$, $\forall i, j$

Perron's theorem

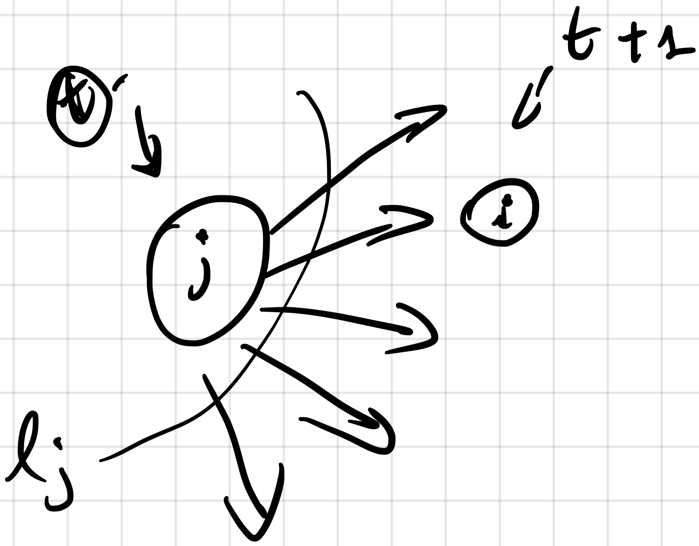
matrix P (entries are >0)

i) P has an eigenvalue $c > 0$ s.t. $c > c'$ ($\forall c'$)
(c' is another eigenvalues)

ii) $\exists$ eigenvector y with real positive values corresponding to c, and y is unique (up to a multiplication by a constant)

iii) if $c = 1$, then for any starting vector $x \neq 0$ with non negative coordinates, the sequence of vectors $P^k x$ converges to a vector in the direction of y ($k \to \infty$) $\square$

Formulation of PageRank using Random Walks



$b_1, b_2, \dots, b_n$: probabilities of being at node i in a given step

$$b_i \leftarrow \sum_{j=1}^{n} \frac{b_j}{l_j} M_{ji}$$

the prob. of being at node i in the "following" step

Let's use matrix N

$$b_i \leftarrow N_{1i} b_1 + \cdots + N_{ni} b_n$$

$$b \leftarrow N^T \cdot b$$

<u>claim</u>

PR of page x is exactly the probability of being at node x after k steps

A scaled version of the Random Walk

For a given probability s:

the walker follows a random edge

with prob. $1-s$:

the walker is "teleported" uniformly at random to another node

$$b_i \leftarrow s \sum_{j=1}^{n} \frac{b_j}{l_j} M_{ji} + \frac{(1-s)}{n}$$

Using matrix $\tilde{N}$:

$$b_i \leftarrow \tilde{N}_{1i}\, b_1 + \dots + \tilde{N}_{ni}\, b_n$$

$$b \leftarrow \tilde{N}^T \cdot b$$

<u>CLAIM</u>

PR is equivalent to the scaled version of random walk.

Take Home Messages

1. We proved that:
   a. hubs, authorities and PageRank scores converge to limiting values after repeated refinements
   b. Hubs and Authorities vectors converge in the direction of vectors no matter of the values of the initial vectors (but they must be positive vectors)

2. Convergence is toward eigenvectors corresponding the greatest eigenvalues of matrices that depend on the structure of the graph (MM^T, M^TM, N, and $\tilde{N}$) (*)

3. PageRank can be formulated in terms of Random Walks.

4. We do not need to iterate our algorithms: we just need to calculate eigenvalues and eigenvector of (*)