Linear Algebra

MA 242 (Spring 2513)

Instructor: M. Chirilus-Bruckner

Linear Algebra and the importance of things

– PageRank –

***A baby internet for illustration** .

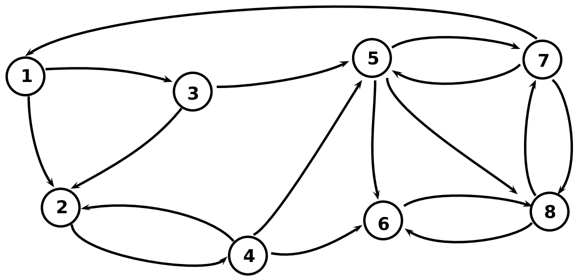


Figure 1: Nodes represent website, directed arrows represent links

***Hyperlink matrix** .

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdot & 0 \\ \cdot & 0 & \cdot & \cdot & 0 & 0 & 0 & 0 \\ \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \cdot & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & \cdot & 0 & 0 & \cdot & 0 \\ 0 & 0 & 0 & \cdot & \cdot & 0 & 0 & \cdot \\ 0 & 0 & 0 & 0 & \cdot & 0 & 0 & \cdot \\ 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 1 & 1/3 & 0 \end{bmatrix}$$

- Begin row-wise:
  1 gets a link from 7, so make a dot at h_{17} and fill the rest of the row with 0's, and so on
- Column-wise:
  1 links to 2 pages, so it gives 1/2 of its importance to each

(Recall: When used for the real-world internet $H \in \mathbb{R}^{n \times n}$ where $n \approx 25$ billion.)

***A way to define importance (implicitly)** .

"The importance of a page is defined by how much importance it gets from other pages."

One can destill this idea into an equation by defining an "importance vector" I (PageRank) whose i-th entry gives the importance of the i-th page. It has to satisfy

$$\boxed{HI = I,}$$

which can be interpreted as an eigenvalue problem: I is the eigenvector of I for the eigenvalue $\lambda = 1$. (How do we now that 1 is an eigenvalue?)

*** Properties of the hyperlink matrix that faciliate the computation of I .**

The matrix H has all non-negative entries and its column entries sum up to 1. It is a so-called stochastic matrix. Assume that from the theory provided for our matrix H you know that its eigenvalues fulfill

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_8|,$$

and that H is diagonalizable. Then

$$H^k x = H^k(c_1 I + c_2 v_2 + \ldots + c_8 v_8) = c_1 I + c_2(\lambda_2)^k v_2 + \ldots + c_8(\lambda_8)^k v_8$$

approaches a multiple of I for large k. (Recall: When used for the real-world internet there are around 25 billion eigenvalues.)

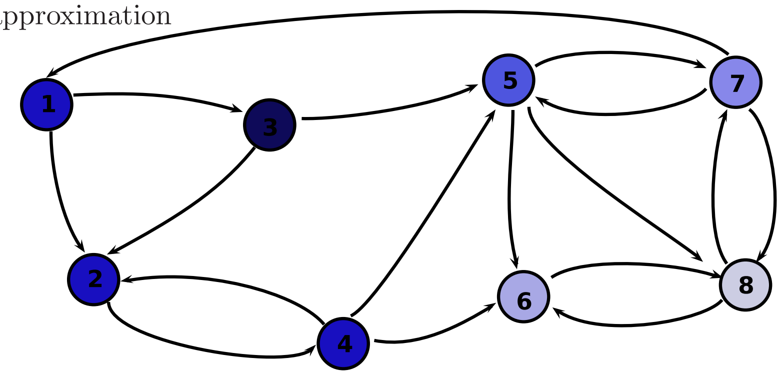
*** Computation of I by the power method .**

Using this theoretical knowledge we can, hence, compute an approximation of I by executing the following loop:

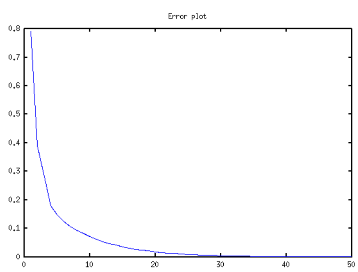
```
Z = [1,0,0,0,0,0,0,0]';
 for kk=1:number_of_iterations
     W = H*Z;
     m = 1/abs(sum(W));
     W = m*W; % the sum of all entries in the importance vector should be 1
     error(kk) = norm(H*W-W); % tracking the error
     Z = W;
 end
```

After 50 iterations we get the approximation

$$I = \begin{bmatrix} 0.060026 \\ 0.067489 \\ 0.029985 \\ 0.067513 \\ 0.097530 \\ 0.202427 \\ 0.0179932 \\ 0.0295098 \end{bmatrix}.$$



Color illustration of importance (according to I) of each website (the lighter, the more important)



Error at each iteration step