# DMML, 28 Mar 2019

Information retrieval on the Internet

Internet documents are not "refereed"

Commercial value attached to search ranking

HTML makes it easy to add invisible misleading content to misdirect search

Self published documents do not contain "obvious" terms

IBM webpage did not mention "computer"

Additional structure available - hyperlinks

Links within documents refer to other docs

<a href="____">Some text</a>

location

anchor text

describes content

Can treat anchor text more than document itself

- Use anchor text to index target doc

Different approach to indexing

World Wide Web as a gigantic graph

- Reason about the graph as a whole

Social Network Analysis

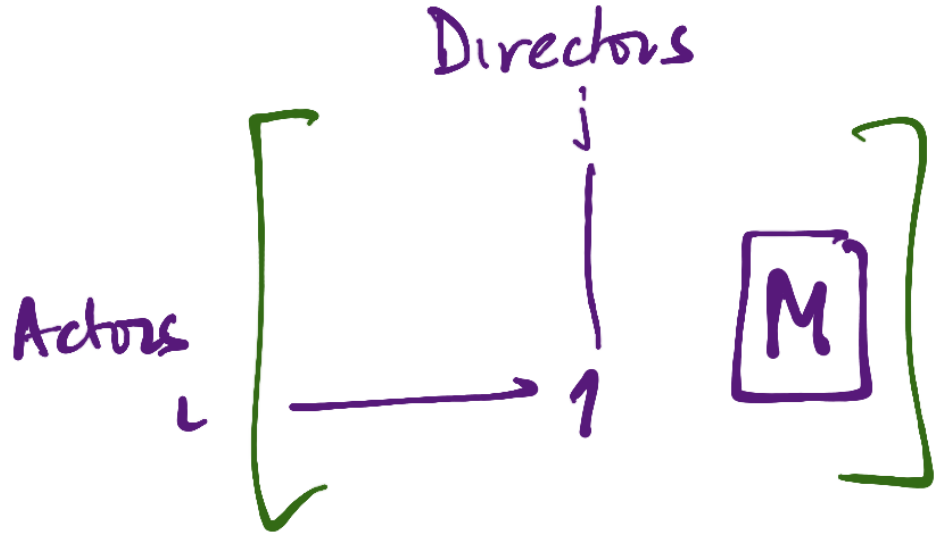
e.g. Film industry

- Actors — Star?
- Directors — Famous?

Good actors want to work with famous directors

Famous directors get stars to work in their films

Graph.




$M[i,j] = 1$ if actor $i$ has worked in a movie of director $j$

Each actor has a star rating $S[i]$

Each director has a "famous" rating $F[j]$

Stars derive their ratings from the directors they work with:

$$S[i] = \sum_j M[i,j] \cdot F[j]$$

Symmetrically

$$F[j] = \sum_i M[i,j] \cdot S[i]$$

$$S[i] = \sum_j M[i,j] \cdot \sum_i M[i,j] \, S[i]$$

$$S = (M \cdot M^T) \cdot S$$

$$F = (M^T \cdot M) \cdot F$$

Solve for $S, F \Longrightarrow$ computes ratings

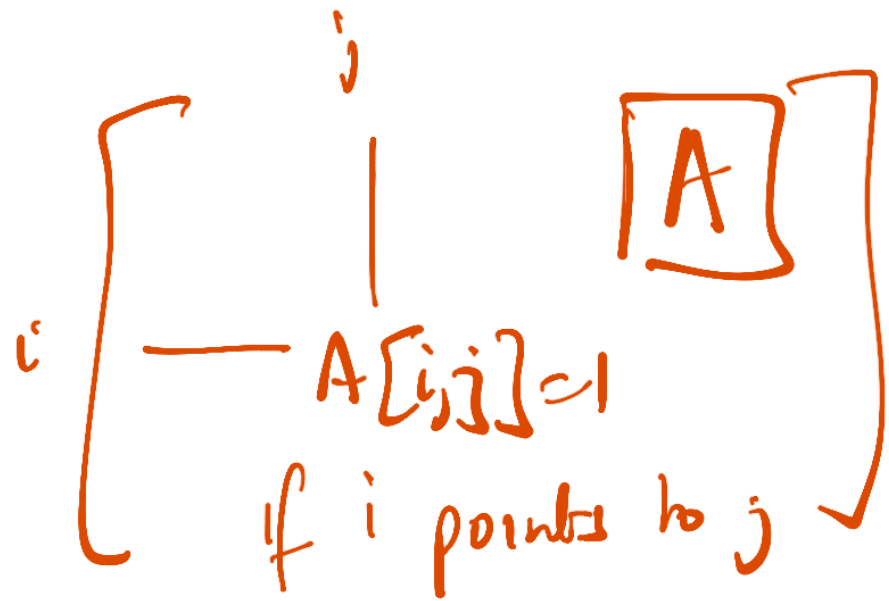
Similarly, use graph structure to derive some conclusions about document ranks

Every document has some "prestige" : P[i]

P[i] is shared (equally) across all outgoing links

Document i derives prestige from incoming links

Adjacency matrix of Internet



If document $i$ has $n$ outgoing links (i.e. $n$ 1's in row $i$ of $A$) - each gets $\frac{1}{n}$ of $P[i]$

3 documents.

$P$ $A^*$

$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix} = \begin{bmatrix} 1.5 & 1 & 0.5 \end{bmatrix}$$

Stable solution.

row entries $\frac{1}{n}$, not 1

$$P^T \cdot A^* = P^T$$

Page Rank — Larry Page

$A^*$ (henceforth A) is a stochastic matrix

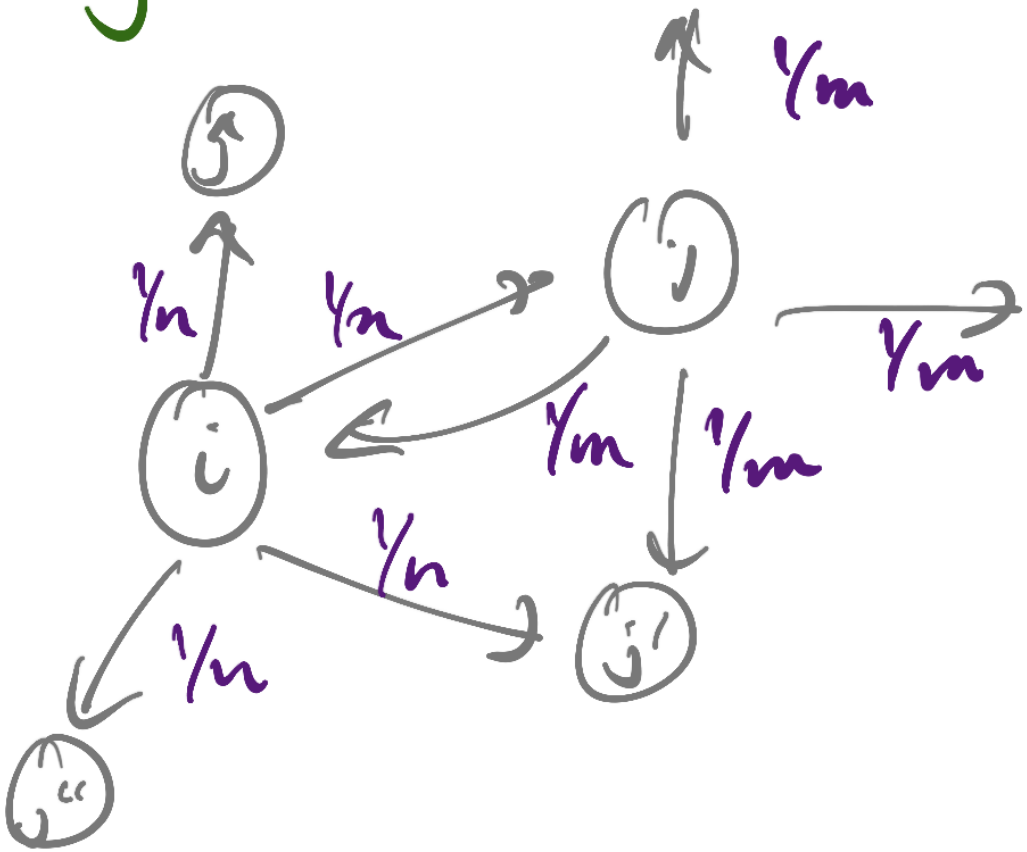
Every row adds up to 1

$$\forall i : \sum_j A[i,j] = 1$$

Interpret the entries as probabilities

"Random surfer" model

Probability $A[i,j]$, move from doc i to doc j



# Markov Chain

Finite collection of "states"

Transition probabilities between states

Start in document 1

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \underbrace{\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}}_{A} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0 \end{bmatrix}$$

$\downarrow$

$P$

After $k$ iterations,

$P[j]$ is probability of being in state $j$ after $k$ steps

In our example

$$[1\ 0\ 0] \to [0\ \tfrac{1}{2}\ \tfrac{1}{2}] \to [\tfrac{3}{4}\ \tfrac{1}{4}\ 0]$$

$$\downarrow$$

$$[\tfrac{7}{16}\ \tfrac{1}{8}\ \tfrac{1}{8}] \leftarrow [\tfrac{1}{4}\ \tfrac{3}{8}\ \tfrac{3}{8}]$$

are all these numbers non-zero after some point?

Markov chain is ergodic if there is some $t_0$ s.t

$$(P A^t)[i] > 0$$

for all $i$, for all $t > t_0$, for all $P$

Not ergodic?

1. Go from $i$ to $j_1, j_2, \ldots, j_k$, no path back
2. Cycle $i \to j \to k \to i \to j \to k$

In an ergodic Markov chain

- there is a stationary distribution $\pi$ s.t.

$$\pi A = \pi$$

- for any starting $P$

$$\lim_{t \to \infty} P \cdot A^t = \pi$$

# Guarantee ergodicity

1. Irreducibility $\Rightarrow$ strongly connected graph

   Any $i, j$ have paths in both directions

2. Aperiodic $\Rightarrow$ "no cycles"

   Any $i, j$, lengths of all paths $i \rightsquigarrow j$

   gcd should be 1

Web graph need not satisfy 1, 2

Also dead ends - docs with no outgoing links

Fix this – "teleportation"

Allow a random jump anywhere!

$T$: $A[i,j] = \frac{1}{N}$ everywhere,

$N$ = # of docs

Transition matrix $M$:

$\alpha$ ⌜ probability of teleportation

$$M = \alpha T + (1-\alpha) A$$

Check that $M$ is stochastic

By construction

M is strongly connected

M is aperiodic – paths of any length

No dead ends

Can solve

$P = MP \rightarrow$ P is Page Rank