Applying theory of Markov Chains to the problem of sports ranking.

A. Govan   C. Meyer

Department of Mathematics
North Carolina State University

AMS Southeastern Section Meeting, March 2007

Outline

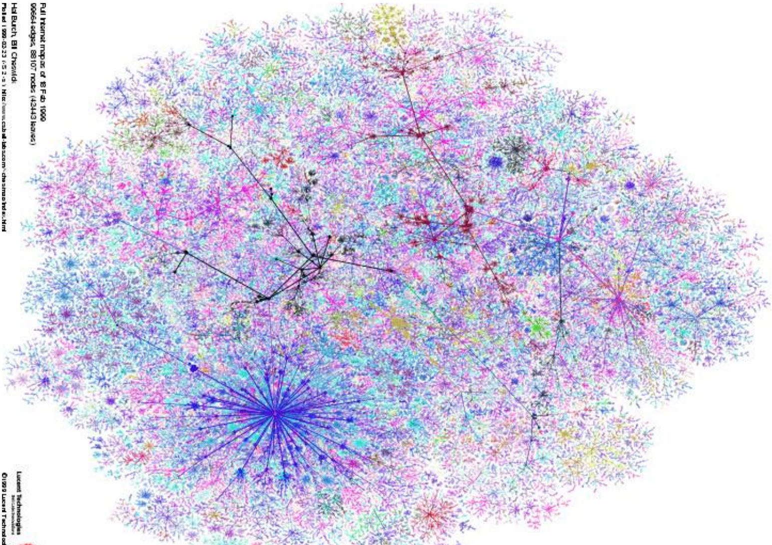
- Google's ranking algorithm.
- Ranking NFL.
- Results and current work.

Basics of PageRank.

- Basic Idea: $r(P) = \sum_{Q \in B_P} \frac{r(Q)}{deg^{-}(Q)}$

  where $r(P)$ is the rank of a webpage P, B_P is the set of web pages pointing to P, and $deg^{-}(Q)$ is the outdegree of a webpage Q.

Web digraph.



Web digraph adjacency matrix.

WWW digraph is represented by an adjacency matrix $\mathbf{A}$.

$$\mathbf{A} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ \vdots \\ P_n \end{array} \begin{array}{c} \begin{array}{ccccc} P_1 & P_2 & P_3 & \cdots & P_n \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 1 \end{pmatrix} \end{array}$$

Web digraph hyperlink matrix.

$$
\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ \vdots \\ P_n \end{array}
\begin{array}{c}
\begin{array}{ccccc} P_1 & P_2 & P_3 & \cdots & P_n \end{array} \\
\begin{pmatrix}
0 & \frac{1}{deg^{-}(P_1)} & 0 & \cdots & \frac{1}{deg^{-}(P_1)} \\
0 & 0 & 0 & \cdots & 0 \\
\frac{1}{deg^{-}(P_3)} & \frac{1}{deg^{-}(P_3)} & \frac{1}{deg^{-}(P_3)} & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\frac{1}{deg^{-}(P_n)} & 0 & \frac{1}{deg^{-}(P_n)} & \cdots & \frac{1}{deg^{-}(P_n)}
\end{pmatrix}
\end{array}
$$

PageRank problem statement.

- Basic Idea: $r(P) = \sum_{Q \in B_P} \frac{r(Q)}{deg^{-}(Q)}$
- Problem restated:
  - π - vector containing the rank scores.

PageRank problem statement.

- Basic Idea: $r(P) = \sum_{Q \in B_P} \frac{r(Q)}{deg^{-}(Q)}$
- Problem restated:
  - π - vector containing the rank scores.
  - $\pi(0)$ - initial rank vector

PageRank problem statement.

- Basic Idea: $r(P) = \sum_{Q \in B_P} \frac{r(Q)}{deg^-(Q)}$
- Problem restated:
  - $\boldsymbol{\pi}$ - vector containing the rank scores.
  - $\boldsymbol{\pi}(0)$ - initial rank vector
  - $\boldsymbol{\pi}^T(k) = \boldsymbol{\pi}^T(k-1)\mathbf{H}$

PageRank problem statement.

- Basic Idea: $r(P) = \sum_{Q \in B_P} \frac{r(Q)}{deg^{-}(Q)}$
- Problem restated:
  - $\boldsymbol{\pi}$ - vector containing the rank scores.
  - $\boldsymbol{\pi}(0)$ - initial rank vector
  - $\boldsymbol{\pi}^T(k) = \boldsymbol{\pi}^T(k-1)\mathbf{H}$
  - $\boldsymbol{\pi}^T(k) = \boldsymbol{\pi}^T(0)\mathbf{H}^k$

PageRank problem statement.

- Basic Idea: $r(P) = \sum_{Q \in B_P} \frac{r(Q)}{deg^{-}(Q)}$
- Problem restated:
  - $\boldsymbol{\pi}$ - vector containing the rank scores.
  - $\boldsymbol{\pi}(0)$ - initial rank vector
  - $\boldsymbol{\pi}^T(k) = \boldsymbol{\pi}^T(k-1)\mathbf{H}$
  - $\boldsymbol{\pi}^T(k) = \boldsymbol{\pi}^T(0)\mathbf{H}^k$
  - $\boldsymbol{\pi}^T(0)\mathbf{H}^k \longrightarrow \boldsymbol{\pi}$?

Google matrix.

- Adjacency Matrix $\mathbf{A}$.

Google matrix.

- Adjacency Matrix $\mathbf{A}$.
- Hyperlink Matrix $\mathbf{H}$.

Google matrix.

- Adjacency Matrix $\mathbf{A}$.
- Hyperlink Matrix $\mathbf{H}$.
- Stochastic matrix $\mathbf{S}$.
  - Replace the zero rows of $\mathbf{H}$ with $(1/n)\mathbf{e}^T$, where $\mathbf{e}$ is a column vector of ones.

Google matrix.

- Adjacency Matrix $\mathbf{A}$.
- Hyperlink Matrix $\mathbf{H}$.
- Stochastic matrix $\mathbf{S}$.
  - Replace the zero rows of $\mathbf{H}$ with $(1/n)\mathbf{e}^T$, where $\mathbf{e}$ is a column vector of ones.
- Google Matrix $\mathbf{G}$.
  - Convex combination: $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T$, $\alpha \in (0,1)$, $\mathbf{v}^T > 0$ and $\mathbf{v}^T\mathbf{e} = 1$.

PageRank vector π.

- $\mathbf{G}$ is the transition probability matrix.

PageRank vector π.

- $\mathbf{G}$ is the transition probability matrix.
- $\mathbf{G}$ is irreducible (and aperiodic).

PageRank vector π.

- $\mathbf{G}$ is the transition probability matrix.
- $\mathbf{G}$ is irreducible (and aperiodic).
- Markov Chains theory implies:
  $\boldsymbol{\pi}^T(0)\mathbf{G}^k \rightarrow \boldsymbol{\pi}^T$
  such that $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T\mathbf{G}$

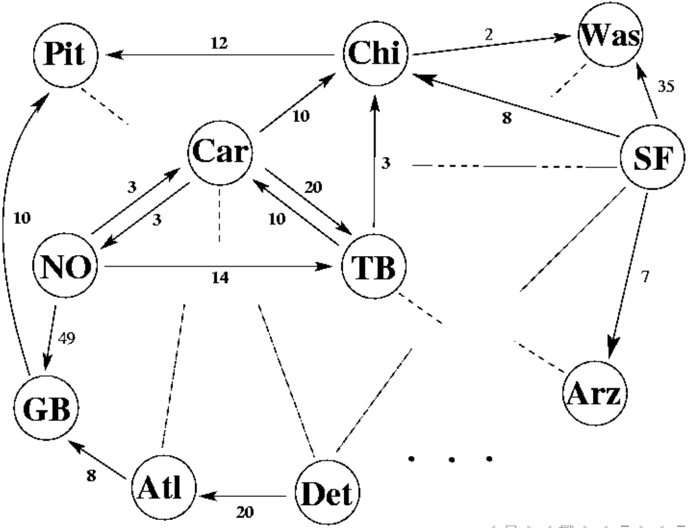
PageRank vector π.

- $\mathbf{G}$ is the transition probability matrix.
- $\mathbf{G}$ is irreducible (and aperiodic).
- Markov Chains theory implies:
  $\boldsymbol{\pi}^T(0)\mathbf{G}^k \rightarrow \boldsymbol{\pi}^T$
  such that $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T\mathbf{G}$
- π is a unique probability distribution vector.

PageRank vector $\boldsymbol{\pi}$.

- $\mathbf{G}$ is the transition probability matrix.
- $\mathbf{G}$ is irreducible (and aperiodic).
- Markov Chains theory implies:
  $\boldsymbol{\pi}^T(0)\mathbf{G}^k \to \boldsymbol{\pi}^T$
  such that $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T\mathbf{G}$
- $\boldsymbol{\pi}$ is a unique probability distribution vector.
- $\boldsymbol{\pi}_i$ is the PageRank score of the web page i.

NFL weighted digraph.



NFL adjacency matrix.

$$\mathbf{A} = \begin{array}{c} \\ \text{Arz} \\ \cdots \\ \text{Car} \\ \text{Chi} \\ \cdots \\ \text{NO} \\ \cdots \\ \text{Pit} \\ \cdots \\ \text{TB} \\ \cdots \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{Arz} & \cdots & \text{Car} & \text{Chi} & \text{NO} & \text{Pit} & \text{TB} & \cdots \end{array} \\ \left(\begin{array}{cccccccc} 0 & \cdots & 4 & 0 & 0 & 0 & 0 & \cdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 10 & 3 & 0 & 20 & \cdots \\ 0 & \cdots & 0 & 0 & 0 & 12 & 0 & \cdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 3 & 0 & 0 & 0 & 14 & \cdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 10 & 3 & 0 & 0 & 0 & \cdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{array} \right) \end{array}$$

GeM (Generalized Markov Method).

- Adjacency matrix $\mathbf{A}$.
- Hyperlink matrix $\mathbf{H}(i,j) = \sum_t w_{ij}^t / (\sum_j (\sum_t w_{ij}^t))$
  where w_{ij}^t is the weight on the edge from team i to team j during week t.

- Stochastic matrix $\mathbf{S}$, dealing with undefeated teams.
- GeM matrix $\mathbf{G} = \alpha_0 \mathbf{S} + \alpha_1 \mathbf{e}\mathbf{v}_1^T + ... + \alpha_k \mathbf{e}\mathbf{v}_k^T$
  where $k \geq 1$.

Feature vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

- Based on the statistical data of the given season.

Feature vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

- Based on the statistical data of the given season.
- Must be nonnegative.

Feature vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

- Based on the statistical data of the given season.
- Must be nonnegative.
- Problem: What statistical data corresponds the most to the performance?

Feature vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

- Based on the statistical data of the given season.
- Must be nonnegative.
- Problem: What statistical data corresponds the most to the performance?
- Start with a matrix containing statistical data for a given season.

Feature vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

- Based on the statistical data of the given season.
- Must be nonnegative.
- Problem: What statistical data corresponds the most to the performance?
- Start with a matrix containing statistical data for a given season.
- SVD $\rightarrow$ no guaranty on nonnegativity.

Feature vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

- Based on the statistical data of the given season.
- Must be nonnegative.
- Problem: What statistical data corresponds the most to the performance?
- Start with a matrix containing statistical data for a given season.
- SVD $\rightarrow$ no guaranty on nonnegativity.
- NMF (nonnegative matrix factorization)

Feature vectors via NMF

Nonnegative matrix factorization: Given $\mathbf{M}_{m\times n} \geq 0$,

$$\mathbf{M} = \mathbf{W}_{m\times k}\mathbf{H}_{k\times n}$$

such that $\mathbf{W} \geq 0$, and $\mathbf{H} \geq 0$

$$\mathbf{M}_j = \sum \mathbf{W}_i h_{ij}$$

Possible uses of NMF:

- Given appropriate $\mathbf{M}$ matrix (e.g. teams by stats) feature vectors could be the nonnegative "basis" of columns of $\mathbf{M}$, i.e. columns of $\mathbf{W}$.

Results.

GeM ranking method:

Sorted Totals '06	Regular Season		Playoffs	
Participant	Games	Spread	Games	Spread
Colley Ranking	141	2035	11	70
Keener Ranking	130	2058	7	89
GeM Ranking	130	2246	6	128
Govan, Vincent	112	2275	6	47
Meyer, Carl	111	2305	5	105
Meyer, Bud	110	2325	6	112
Kelley, Tim	109	2613	3	149
Koh,Gil	106	2039	9	78.5
Glantz-Culver Line	105	2010.4	9	79.5
Rose, Nick	101	2070	3	117
Albright, Russ	90	1996	7	106
Meyer, Becky	88	1991	8	88
Stitzinger, Ernie	83	1886	7	106
Massey Ranking	82	1761	7	112
Kenney, Holly	69	1410	5	100
Kenney, Sean	63	1068	6	98
Meyer, Marty	16	316	0	0
Laake, Kevin	12	217	0	0
Fauntleroy,Amassa	0	0	2	31

Results.

GeM ranking method:

- (without first two weeks) Basic GeM predicts 70% of the games played correctly during 2004 NFL regular season.

Results.

GeM ranking method:

- (without first two weeks) Basic GeM predicts 70% of the games played correctly during 2004 NFL regular season.
- (without first two weeks) Basic GeM predicts 75.9% of the games played correctly during 2005 NFL regular season.

Results.

GeM ranking method:

- (without first two weeks) Basic GeM predicts 70% of the games played correctly during 2004 NFL regular season.
- (without first two weeks) Basic GeM predicts 75.9% of the games played correctly during 2005 NFL regular season.
- (without first two weeks) Basic GeM predicts 62% of the games played correctly during 2006 NFL regular season.

Summary

- Expanding to bigger data set - NCAA men's basketball.
- Experimenting with NMF to obtain feature vectors.
- Moving beyond sports (recommendation systems).