Ranking Methods.

A. Govan C. Meyer

Department of Mathematics
North Carolina State University

Cha-Cha Days, September 2006

Outline

- Perron Frobenius Theorem
  - Keener
  - Redmond
- Markov Chains
  - Google's ranking algorithm.
- Other Ranking Algorithms
  - Colley
  - HITS
- Future

Basics of Perron Frobenius Theorem

Given a nonnegative irreducible square matrix

- Largest eigenvalue, called Perron root, is positive, real and simple.
- Only one real positive eigenvector corresponding to the largest eigenvalue, called Perron vector.
- If A is primitive, then Perron vector is easy to compute.

Ranking NFL with Keener (SIAM Review,1993)

- *Laplace's rule of succession -*

  $$\frac{S+1}{S+F+2}$$

  probability of a success on the try $n+1$*, and*
  S = $\#$ *of successes,* F = $\#$ *of failures,* $S+F=n$.

- $h(x) = \frac{1}{2} + \frac{1}{2}\textit{sgn}(x - \frac{1}{2})\sqrt{|2x-1|}$

Ranking NFL with Keener

- Keener nonnegative matrix $\mathbf{A}$
  - $$\mathbf{A}(i,j) = \begin{cases} h\left(\dfrac{S_{ij}+1}{S_{ij}+S_{ji}+2}\right) & \text{team } i \text{ played team } j \\ 0 & \text{otherwise} \end{cases},$$

    where S_{ij} is the amount of points scored by team i against team j.
- $\mathbf{A}$ is nonnegative and irreducible
- Rank vector $\mathbf{r}$ is the Perron vector of $\mathbf{A}$.

Ranking NFL with Redmond (Mathematics Magazine,2003)

- Redmond nonnegative matrix $\mathbf{M}$
  - $\mathbf{M}(i,i) = 1/g$, where g is the number of games played
  - $\mathbf{M}(i,j) = \mathbf{M}(j,i) = \begin{cases} 1/g & \text{if team } i \text{ played team } j \\ 0 & \text{team } i \text{ did not played team } j \end{cases}$
- $\mathbf{M}$ is nonnegative, symmetric, and irreducible
- Rank vector is a particular linear combination of normalized (2-norm) eigenvectors of $\mathbf{M}$, excluding the dominant eigenvector.

Basic Markov Chains.

- Markov Chain - stochastic *memoryless process.*
- *Markov Chain* $\equiv$ *Stochastic matrix (nonnegative, rows sum to 1), called transition matrix.*
- *Left Perron vector is called stationary distribution vector*

$$\boldsymbol{\pi}^T = \boldsymbol{\pi} P$$

Google's ranking.

- Think of the internet as a graph.
  - Webpages are nodes of the graph, n nodes.
  - Hyperlinks are directed edges.
- Basic Idea: "a webpage is important if it is pointed to by other important webpages," i.e. rank of a webpage depends on the ranks of the webpages pointing to it.

Google Matrix.

- Hyperlink Matrix $\mathbf{H}$
  - $\mathbf{H}(i,j) = \begin{cases} 1/(\#\text{ links from } i) & \text{there is a link from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$
- Stochastic matrix $\mathbf{S}$
  - Obtained by modifying matrix $\mathbf{H}$.
  - 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: $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T$, $\alpha \in (0,1)$ and $\mathbf{v}^T > 0$
  - Personalization vector $\mathbf{v}$.
- Rank vector is $\boldsymbol{\pi}$, the stationary distribution vector of $\mathbf{G}$.

NFL web.

- Each NFL team is a node in a graph.

NFL web.

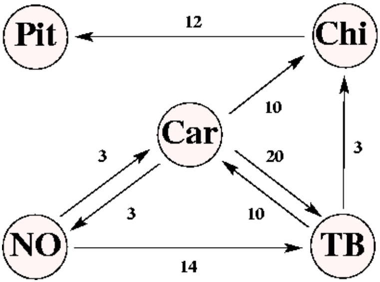
- Each NFL team is a node in a graph.
- A regular season game results in a directed edge from the loser to the winner.

NFL web.

- Each NFL team is a node in a graph.
- A regular season game results in a directed edge from the loser to the winner.
- The edges are weighted, the weight is the score difference of the corresponding game.

NFL web.

- Each NFL team is a node in a graph.
- A regular season game results in a directed edge from the loser to the winner.
- The edges are weighted, the weight is the score difference of the corresponding game.



Variations on PageRank.

- $\mathbf{H}(i,j) = \begin{cases} \dfrac{\sum \text{score differnce for the game where } j \text{ beat } i}{\sum \text{score difference for the game } i \text{ lost}} \\ 0 \end{cases}$
- Dealing with the ith zero row (undefeated team i)

Variations on PageRank.

- $\mathbf{H}(i,j) = \begin{cases} \frac{\sum \text{score differnce for the game where } j \text{ beat } i}{\sum \text{score difference for the game } i \text{ lost}} \\ 0 \end{cases}$
- Dealing with the ith zero row (undefeated team i)
  - $(1/32)\mathbf{e}^T$, equally likely to lose to any other team.

Variations on PageRank.

- $\mathbf{H}(i,j) = \begin{cases} \frac{\sum \text{score differnce for the game where } j \text{ beat } i}{\sum \text{score difference for the game } i \text{ lost}} \\ 0 \end{cases}$

- Dealing with the ith zero row (undefeated team i)
  - $(1/32)\mathbf{e}^T$, equally likely to lose to any other team.
  - $\mathbf{e}_i^T$, team i could "*lose*" *only to team* i.

Variations on PageRank.

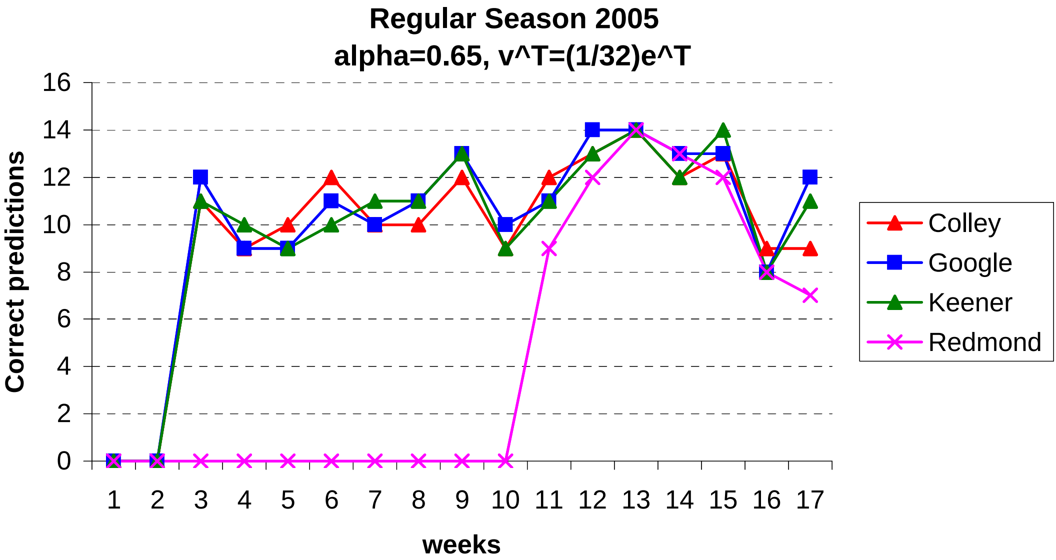
- $\mathbf{H}(i,j) = \begin{cases} \dfrac{\sum \text{score differnce for the game where } j \text{ beat } i}{\sum \text{score difference for the game } i \text{ lost}} \\ 0 \end{cases}$
- Dealing with the ith zero row (undefeated team i)
  - $(1/32)\mathbf{e}^T$, equally likely to lose to any other team.
  - $\mathbf{e}_i^T$, team i could "*lose*" *only to team* i.
  - $\boldsymbol{\pi}_{t-1}^T$, using the ranks from previous week.

Ranking NFL with Colley (Colley's Bias Free Matrix Rankings)

- Colley matrix $\mathbf{C}$
  - Start with *Laplace's rule of succession*, $r_i = \dfrac{n_{w,i}+1}{n_{tot,i}+2}$, rewrite by including strength of schedule.
  - End up with linear system

$$\mathbf{C}\mathbf{r} = \mathbf{b}$$

    where
  - $\mathbf{C}(i,j) = \begin{cases} 2 + n_{tot,i} & i = j \\ -n_{j,i} & i \neq j \end{cases}$
    - $n_{tot,i}$ - total number of games played by team i,
    - $n_{w,i}$ - number of games won by team i,
    - $n_{j,i}$ - number of times team i played team j.
- Ranking vector $\mathbf{r}$ is the solution to the linear system $\mathbf{C}\mathbf{r} = \mathbf{b}$



Regular Season 2005
alpha=0.65, v^T=(1/32)e^T
Correct predictions
0
2
4
6
8
10
12
14
16
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
weeks
Colley
Google
Keener
Redmond

Regular season 2005, α=0.65, $\mathbf{v}^T = (1/32)\mathbf{e}^T$									
	Colley		Google		Keener		Redmond		
	Correct	Spread	Correct	Spread	Correct	Spread	Correct	Spread	games
week 1	0	0	0	0	0	0	0	0	
week 2	0	0	0	0	0	0	0	0	
week 3	11	132	12	109	11	135	0	0	14
week 4	9	163	9	143	10	135	0	0	14
week 5	10	162	9	202	9	167	0	0	14
week 6	12	111	11	126	10	125	0	0	14
week 7	10	124	10	150	11	106	0	0	14
week 8	10	177	11	143	11	148	0	0	14
week 9	12	111	13	140	13	141	0	0	14
week 10	9	109	10	121	9	120	0	0	14
week 11	12	171	11	160	11	159	9	163	16
week 12	13	98	14	111	13	103	12	113	16
week 13	14	134	14	133	14	118	14	116	16
week 14	12	150	13	187	12	172	13	166	16
week 15	13	219	13	217	14	208	12	216	16
week 16	9	149	8	148	8	149	8	149	16
week 17	9	201	12	188	11	202	7	228	16
Total	165	2211	170	2278	167	2188	75	1151	224
	73.7%		75.9 %		74.6 %		67 %		

HITS (Hypertext Induced Topic Search)

- Each webpage gets two scores - authority (depends on inlinks) and hub (depends on outlinks)
- Basic idea: “Good authorities are pointed to by good hubs and good hubs point to good authorities.”

HITS ranking

- x_i = authority score for webpage i
- y_i = hub score for webpage i

$$x_i = \sum_{\text{pages that point to } i} y_i, \qquad y_i = \sum_{\text{pages that } i \text{ points to}} x_i$$

$$\mathbf{x}^{(k)} = \mathbf{L}^T \mathbf{y}^{(k-1)}, \qquad \mathbf{y}^{(k)} = \mathbf{L}\mathbf{x}^{(k)}$$

where

- $\mathbf{L}(i,j) = \begin{cases} 1 & \text{if there is a link from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$

- Two ranking vectors $\mathbf{x}$, $\mathbf{y}$ (authority and hub) are dominant eigenvectors of $L^T L$, and LL^T.

Measuring offence and defence of NFL with HITS

- Based on total yards each team generates
- Authorities = Offence (vector $\mathbf{o}$) and Hubs = Defence (vector $\mathbf{d}$)
- defence score $d_i = \sum_j y_{ij}(1/o_j)$
- offence score $o_j = \sum_i (1/d_i) y_{ij}$

$$\mathbf{d}^{(k)} = \mathbf{Y}[1/\mathbf{o}^{(k-1)}] \qquad \mathbf{o}^{(k)} = (1/\mathbf{d}^{(k)})^T \mathbf{Y}$$

- Converges
- Results are independent of the initial value
- How do ranks of offence and defence correspond to the overall rank?

Regular season 2005				
	Defence		Offence	
	Team Name	defence value	Team Name	offence value
1	Washington	1.4572e+005	Kansas City	0.0387
2	Pittsburgh	1.4790e+005	Denver	0.0366
3	Dallas	1.4828e+005	N.Y. Giants	0.0364
4	Tampa Bay	1.4852e+005	Cincinnati	0.0357
5	SanDiego	1.4914e+005	Seattle	0.0357
6	Baltimore	1.4946e+005	New England	0.0353
7	Carolina	1.4951e+005	San Diego	0.0352
8	Chicago	1.5078e+005	Indianapolis	0.0350
9	Jacksonville	1.5198e+005	St. Louis	0.0340
10	Arizona	1.5226e+005	Arizona	0.0336
11	Philadelphia	1.5392e+005	Washington	0.0333
12	Denver	1.5403e+005	Dallas	0.0327
13	N.Y. Jets	1.5570e+005	Atlanta	0.0326
14	Indianapolis	1.5674e+005	Miami	0.0322
15	N.Y. Giants	1.5838e+005	Philadelphia	0.0319
16	Green Bay	1.5898e+005	Green Bay	0.0319
17	Oakland	1.5988e+005	Oakland	0.0314
18	Seattle	1.6129e+005	Pittsburgh	0.0313
19	Kansas City	1.6359e+005	New Orleans	0.0311
20	Tennessee	1.6579e+005	Tennessee	0.0310
21	Cleveland	1.6586e+005	Jacksonville	0.0310
22	Miami	1.6627e+005	Carolina	0.0306
23	New Orleans	1.6688e+005	Baltimore	0.0289
24	New England	1.6840e+005	Tampa Bay	0.0288
25	Minnesota	1.6909e+005	Minnesota	0.0286
26	Buffalo	1.7087e+005	Cleveland	0.0281
27	Detroit	1.7112e+005	Detroit	0.0273
28	Atlanta	1.7241e+005	Buffalo	0.0252
29	St. Louis	1.8074e+005	Chicago	0.0251
30	Cincinnati	1.8137e+005	Houston	0.0249
31	Houston	1.8430e+005	N.Y. Jets	0.0245
32	San Francisco	1.9346e+005	San Francisco	0.0224

Future work

- Incorporate HITS measure of offence, defence into overall ranking score
- Point spreads