Ranking using Matrices.

Anjela Govan,
Amy Langville,
Carl D. Meyer

Northrop Grumman,
College of Charleston,
North Carolina State University

June 2009

Outline

Basics of Ranking

- The *rank* of an object is its relative importance to the other objects in the finite set of size n. The ranks are 1,2,3, etc.
- Ranking models produce ratings.
- Ratings provide the degree of relative importance of each object.
- Applications of ranking include sports and search of web and literature.

ODM Development

$A_{ij} =$ score team j generated against team i
$A_{ij} = 0$ otherwise

- *offensive rating* of team j

$$o_j = A_{1j}(1/d_1) + ... + A_{nj}(1/d_n)$$

- *defensive rating* of team i

$$d_i = A_{i1}(1/o_1) + ... + A_{in}(1/o_n)$$

$$\mathbf{o}^{(k)} = \mathbf{A}^T \frac{1}{\mathbf{d}^{(k-1)}}$$

$$\mathbf{d}^{(k)} = \mathbf{A} \frac{1}{\mathbf{o}^{(k)}}$$

Sinkhorn-Knopp Theorem (1967)

Definition

A square matrix $\mathbf{A} \geq 0$ is said to have total support if $\mathbf{A} \neq 0$ and if every positive element of $\mathbf{A}$ lies on a positive diagonal.

Theorem

For each $\mathbf{A} \geq 0$ with total support there exists a unique doubly stochastic matrix $\mathbf{S}$ of the form $\mathbf{RAC}$ where $\mathbf{R}$ and $\mathbf{C}$ are unique (up to a scalar multiplication) diagonal matrices with positive main diagonal.

A necessary and sufficient condition that the iterative process of alternatively normalizing the rows and columns of $\mathbf{A}$ will converge to a doubly stochastic limit is that $\mathbf{A}$ has support.

ODM convergence

- If $\mathbf{A}$ has total support $\rightarrow \{\mathbf{o}^{(k)}\}$, and $\{\mathbf{d}^{(k)}\}$ converge
- $\mathbf{A}$ may not have total support (but will have support)
- Can force total support

$$\mathbf{P} = \mathbf{A} + \epsilon \mathbf{e}\mathbf{e}^T$$

- As ϵ decreases number of iterations increases

ODM Algorithm

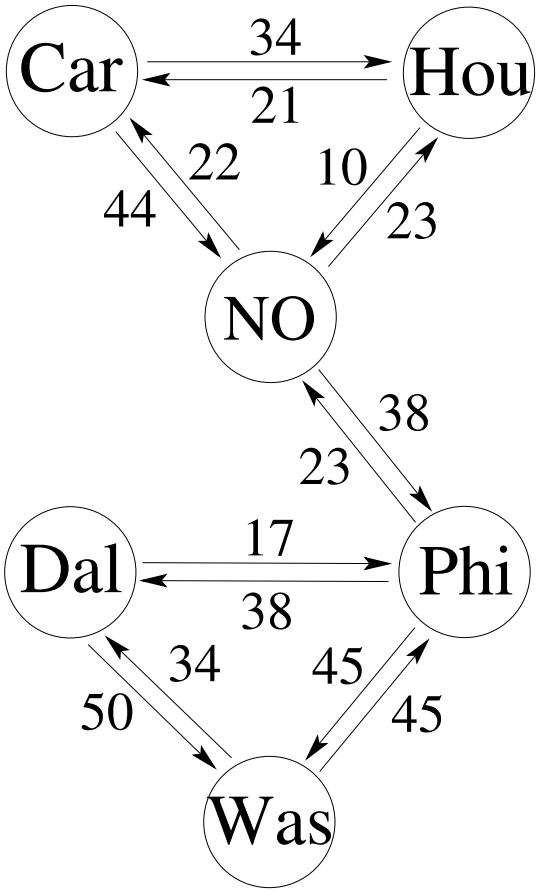
1. Represent the season using a weighted digraph with n nodes. On $i \longrightarrow j$ the wight w_{ij} = amount of the statistic acquired by team j against team i.
2. Form adjacency matrix $\mathbf{A}$, $\mathbf{P} = \mathbf{A} + \epsilon \mathbf{e}\mathbf{e}^T$.
3. Team i has two rating scores, offensive o_i and defensive d_i

$$\mathbf{o}^{(k)} = \mathbf{P}^T \frac{1}{\mathbf{d}^{(k-1)}}$$

$$\mathbf{d}^{(k)} = \mathbf{P} \frac{1}{\mathbf{o}^{(k)}}$$

4. Overall rating score - rank aggregation (e.g. $r_i = o_i/d_i$).

2007 season NFL Example - ODM



Adjacency matrix $\mathbf{A}$:

$$\begin{array}{c} \\ \text{Car} \\ \text{Dal} \\ \text{Hou} \\ \text{NO} \\ \text{Phi} \\ \text{Was} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{Car} & \text{Dal} & \text{Hou} & \text{NO} & \text{Phi} & \text{Was} \end{array} \\ \begin{pmatrix} 0 & 0 & 34 & 44 & 0 & 0 \\ 0 & 0 & 0 & 0 & 17 & 50 \\ 21 & 0 & 0 & 10 & 0 & 0 \\ 22 & 0 & 23 & 0 & 38 & 0 \\ 0 & 38 & 0 & 0 & 0 & 45 \\ 0 & 34 & 0 & 0 & 45 & 0 \end{pmatrix} \end{array}$$

2007 season NFL Example (ODM)-result

- $\mathbf{A} + 0.001\mathbf{e}\mathbf{e}^T$, $tol = 0.01$

$$\mathbf{o} \approx \begin{pmatrix} 0.134 & 7.043 & 0.098 & 0.091 & 6.396 & 12.383 \end{pmatrix}^T$$

$$\mathbf{d} \approx \begin{pmatrix} 827.666 & 6.736 & 266.663 & 403.771 & 9.074 & 11.912 \end{pmatrix}^T$$

$$\mathbf{r} \approx \begin{pmatrix} 0.00016 & 1.0456 & 0.00037 & 0.00023 & 0.705 & 1.04 \end{pmatrix}^T$$

The list of ranked teams (from best to worst) is

Dal Was Phi Hou NO Car

GeM Model

A team is good if it defeated good teams.

- Use game structure of a team sport
- Form directed graph representing the season
  - Each team is a state (node in the digraph)
  - Each game is a directed edge from loser to winner
  - Weight on the directed edge (i, j) is a normalized positive point spread (probability of moving from i to j)

GeM Model

- Choose an edge in the current season digraph to move to the team we'll be rooting for
- The proportion of time spent rooting for the team i is the rating of team i
- Will spend the larger proportion of time on "good" teams, if a team is "good" it has large indegree (many teams lost to it)

GeM Model

- Choose an edge in the current season digraph to move to the team we'll be rooting for
- The proportion of time spent rooting for the team i is the rating of team i
- Will spend the larger proportion of time on "good" teams, if a team is "good" it has large indegree (many teams lost to it)

MARKOV CHAINS

Ranking NFL with GeM.

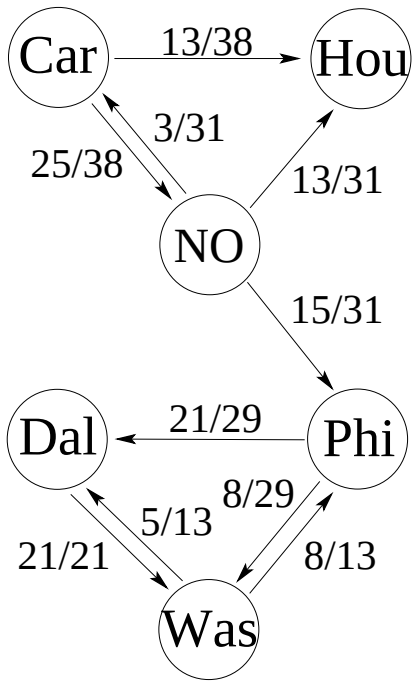
- Each NFL team is a state.

Ranking NFL with GeM.

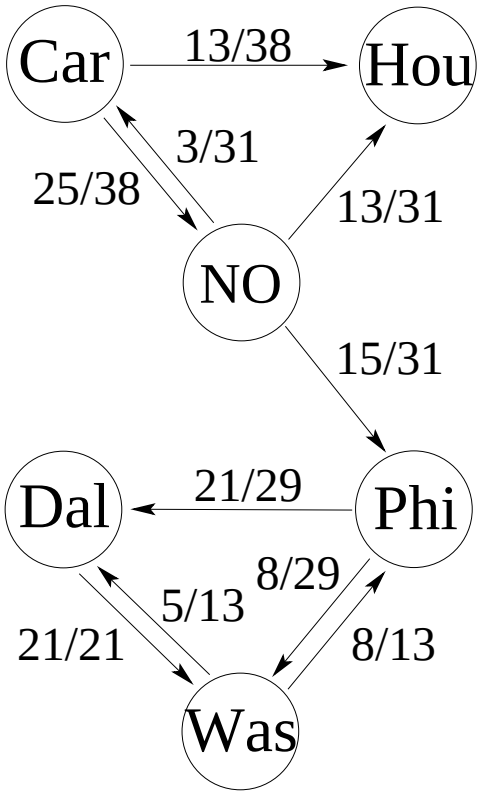
- Each NFL team is a state.
- Score differences determine transition probability

Ranking NFL with GeM.

- Each NFL team is a state.
- Score differences determine transition probability

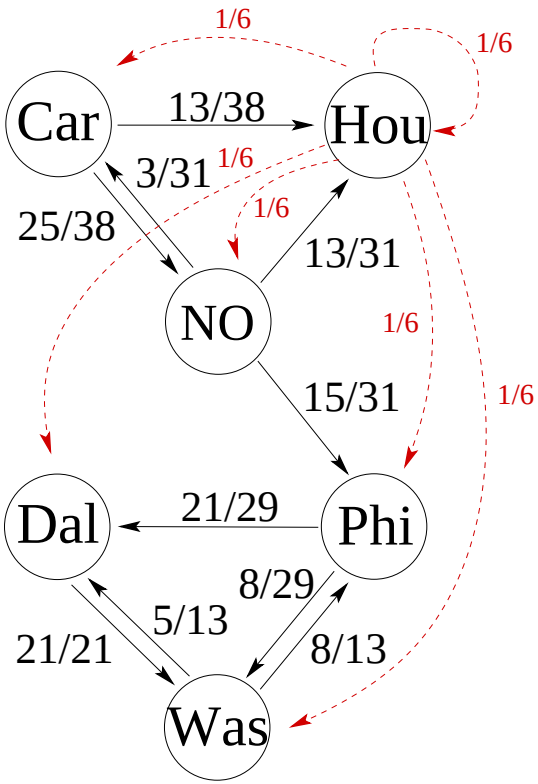


Game Matrix



$$
\begin{array}{c} \\ \text{Car} \\ \text{Dal} \\ \text{Hou} \\ \text{NO} \\ \text{Phi} \\ \text{Was} \end{array}
\begin{array}{c}
\begin{array}{cccccc} \text{Car} & \text{Dal} & \text{Hou} & \text{NO} & \text{Phi} & \text{Was} \end{array} \\
\begin{pmatrix}
0 & 0 & \frac{13}{38} & \frac{25}{38} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\frac{3}{31} & 0 & \frac{13}{31} & 0 & \frac{15}{31} & 0 \\
0 & \frac{21}{29} & 0 & 0 & 0 & \frac{8}{29} \\
0 & \frac{5}{13} & 0 & 0 & \frac{8}{13} & 0
\end{pmatrix}
\end{array}
$$

Game Matrix - Undefeated Teams (stochastic)



$$\mathbf{S} = \begin{array}{c} \\ \text{Car} \\ \text{Dal} \\ \text{Hou} \\ \text{NO} \\ \text{Phi} \\ \text{Was} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{Car} & \text{Dal} & \text{Hou} & \text{NO} & \text{Phi} & \text{Was} \end{array} \\ \begin{pmatrix} 0 & 0 & \frac{13}{38} & \frac{25}{38} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{3}{31} & 0 & \frac{13}{31} & 0 & \frac{15}{31} & 0 \\ 0 & \frac{21}{29} & 0 & 0 & 0 & \frac{8}{29} \\ 0 & \frac{5}{13} & 0 & 0 & \frac{8}{13} & 0 \end{pmatrix} \end{array}$$

Game-Statistic Matrix

- Adding stats (irredicibility and primitivity):

$$\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T$$

where $\mathbf{v} > 0$ is called personalization vector and can be based on teams statistical data, $0 < \alpha < 1$.

2007 season NFL Example - GeM

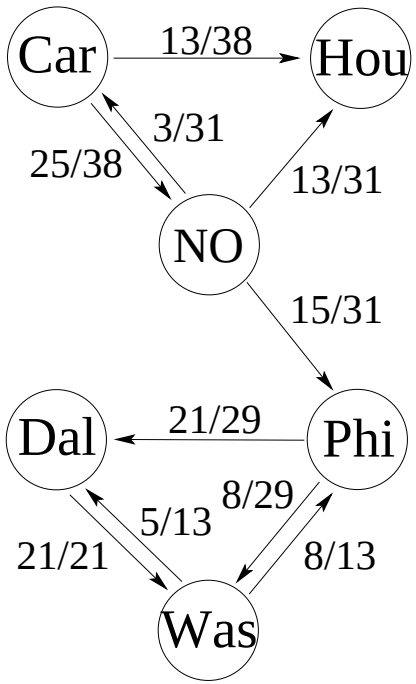
Let $\alpha = 0.85$, $\mathbf{v} = (1/6)\mathbf{e}$, and $\mathbf{S}$ formed using games scores the matrix $\mathbf{G}$ is

$$\mathbf{G} = 0.85\mathbf{S} + 0.15(1/6)\mathbf{e}\mathbf{e}^T =$$

	Car	Dal	Hou	NO	Phi	Was
Car	$1/40$	$1/40$	$6/19$	$111/190$	$1/40$	$1/40$
Dal	$1/40$	$1/40$	$1/40$	$1/40$	$1/40$	$7/8$
Hou	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$
NO	$133/1240$	$1/40$	$473/1240$	$1/40$	$541/1240$	$1/40$
Phi	$1/40$	$743/1160$	$1/40$	$1/40$	$1/40$	$301/1160$
Was	$1/40$	$183/520$	$1/40$	$1/40$	$57/104$	$1/40$

2007 season NFL Example (GeM)-result

$$\boldsymbol{\pi}^T \approx \begin{pmatrix} 0.0389 & 0.2824 & 0.0656 & 0.056 & 0.2289 & 0.3281 \end{pmatrix}$$



The list of the teams in the order of rating scores (from best to worst) is

Was Dal Phi Hou NO Car

Other Matrix Based Ranking Models

- Analytic Hierarchy Process (AHP) (Saaty, 1980)
- Colley Matrix 2002
- Keener 1993
- Massey 1997
- ...

Game Predictions

Point Spread

- Assume that
  point spread for game between team$_i$ and team$_j$ $=$
  $M|$rating team$_i$ $-$ rating team$_j|$
- Use previous results to estimate M (Least Squares)

Data Gathering Challenges

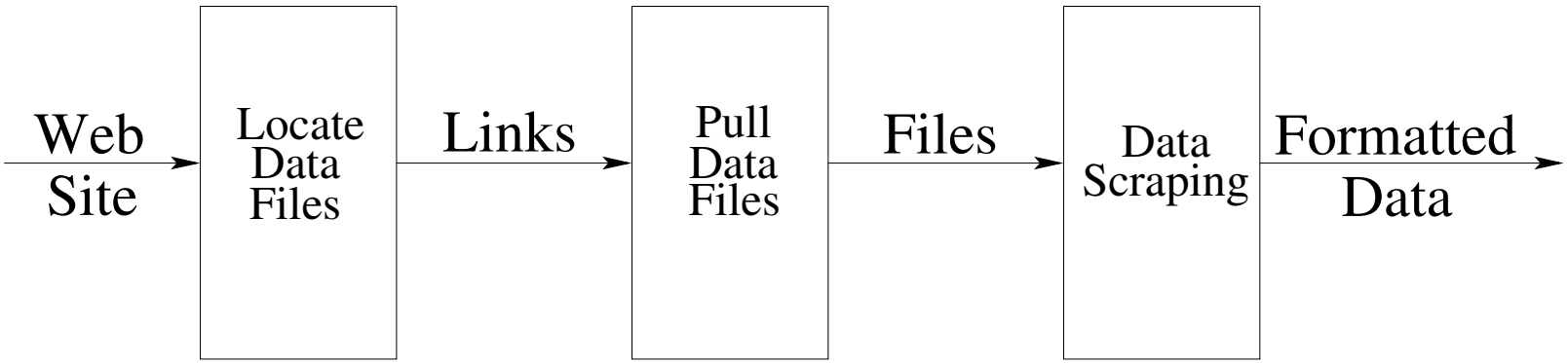
- Reliable data sources
- Data format
- Amount of data
- Team names and league expansions

Data Gathering

- Sources - http://www.jt-sw.com/football/boxes/index.nsf (John M. Troan); http://scores.espn.go.com/ncf/scoreboard (ESPN);

- Data collection and parsing - automated with Perl scripts

WEB SCRAPING



NFL Game Prediction

- 2001-2007 with preseason padding
- ODM $tol = 0.01$, $\epsilon = 0.00001$
- GeM (personalization vector), $\alpha = 0.6$, $\mathbf{v} = (1/n)\mathbf{e}$

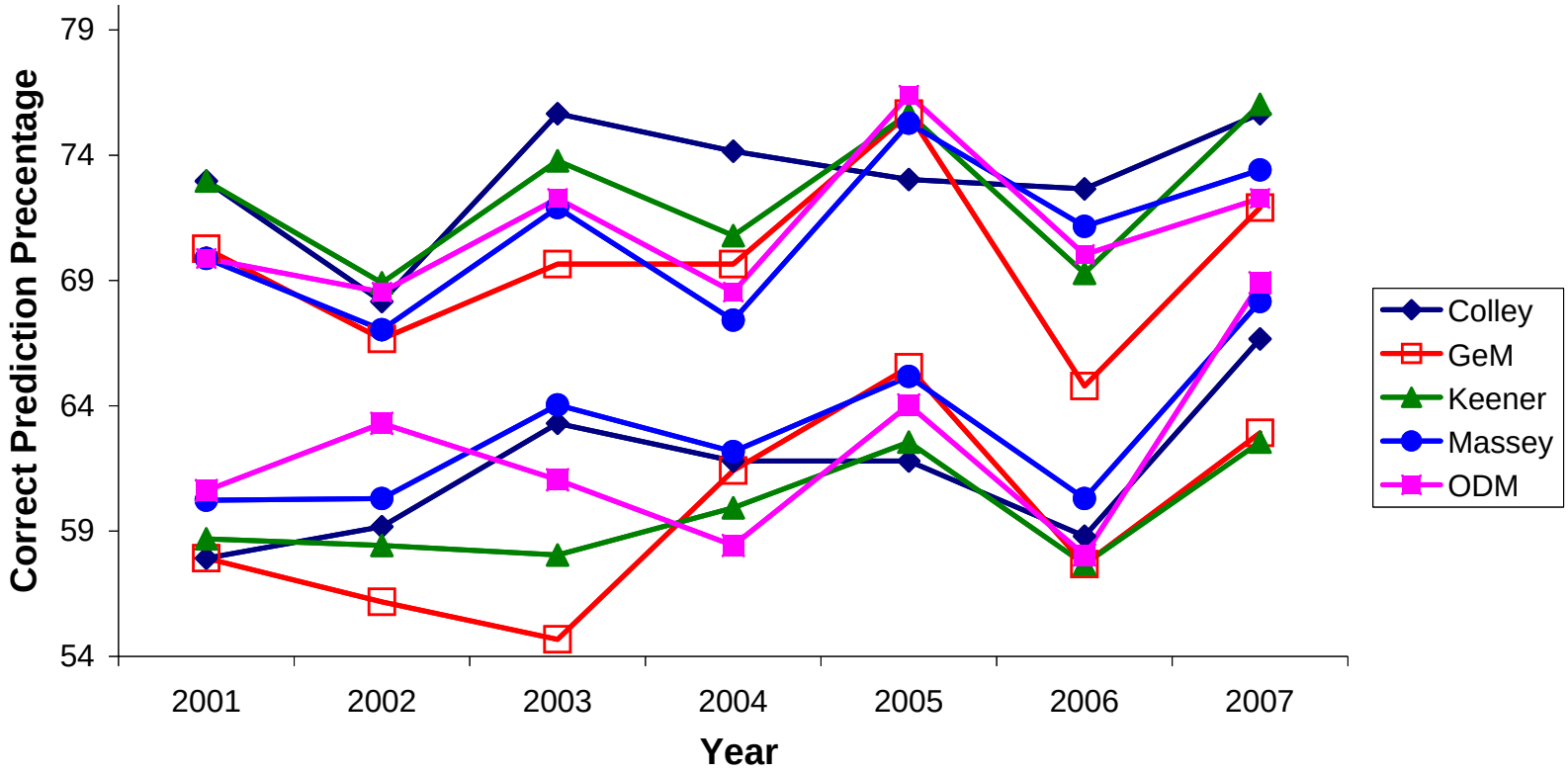
NFL Foresight Prediction Results

	Colley	GeM	Keener	Massey	ODM
2001	57.92	57.92	58.69	60.23	60.62
2002	59.18	56.18	58.43	60.30	63.30
2003	63.30	54.68	58.05	64.04	61.05
2004	61.80	61.42	59.93	62.17	58.43
2005	61.80	65.54	62.55	65.17	64.04
2006	58.80	57.68	57.68	60.30	58.05
2007	66.67	62.92	62.55	68.16	68.91

NFL Hindsight Prediction Results

	Colley	GeM	Keener	Massey	ODM
2001	72.97	70.27	72.97	69.88	69.88
2002	68.16	66.67	68.91	67.04	68.54
2003	75.66	69.66	73.78	71.91	72.28
2004	74.16	69.66	70.79	67.42	68.54
2005	73.03	75.66	75.66	75.28	76.40
2006	72.66	64.79	69.29	71.16	70.04
2007	75.66	71.91	76.03	73.41	72.28

NFL Foresight/Hindsight Prediction Results



NCAA Football Game Prediction

- Division I-A
- 2003-2007 starting week 5
- ODM $tol = 0.01$, $\epsilon = 0.00001$
- GeM (personalization vector), $\alpha = 0.6$, $\mathbf{v} = (1/n)\mathbf{e}$

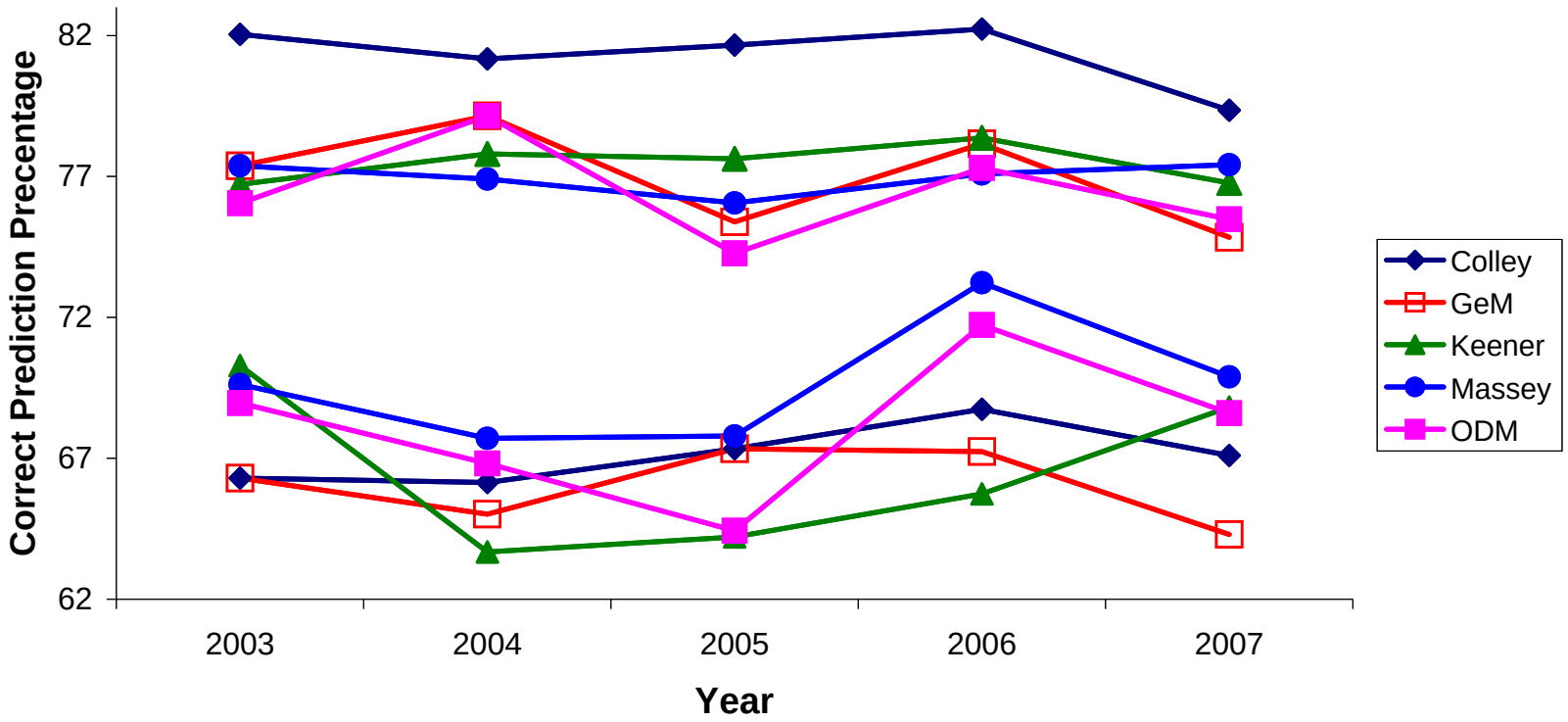
NCAA Football Foresight Prediction Results

	Colley	GeM	Keener	Massey	ODM
2003	66.30	66.30	70.29	69.62	68.96
2004	66.14	65.02	63.68	67.71	66.82
2005	67.34	67.34	64.21	67.79	64.43
2006	68.74	67.24	65.74	73.23	71.73
2007	67.10	64.30	68.82	69.89	68.60

NCAA Football Hindsight Prediction Results

	Colley	GeM	Keener	Massey	ODM
2003	82.04	77.38	76.72	77.38	76.05
2004	81.17	79.15	77.80	76.91	79.15
2005	81.66	75.39	77.63	76.06	74.27
2006	82.23	78.16	78.37	77.09	77.30
2007	79.35	74.84	76.77	77.42	75.48

NCAA Football Foresight/Hindsight Prediction Results



NCAA Basketball Game Prediction

- Division I
- 2001-2007 starting game day 26
- ODM $tol = 0.01$, $\epsilon = 0.00001$
- GeM (personalization vector), $\alpha = 0.6$, $\mathbf{v} = (1/n)\mathbf{e}$

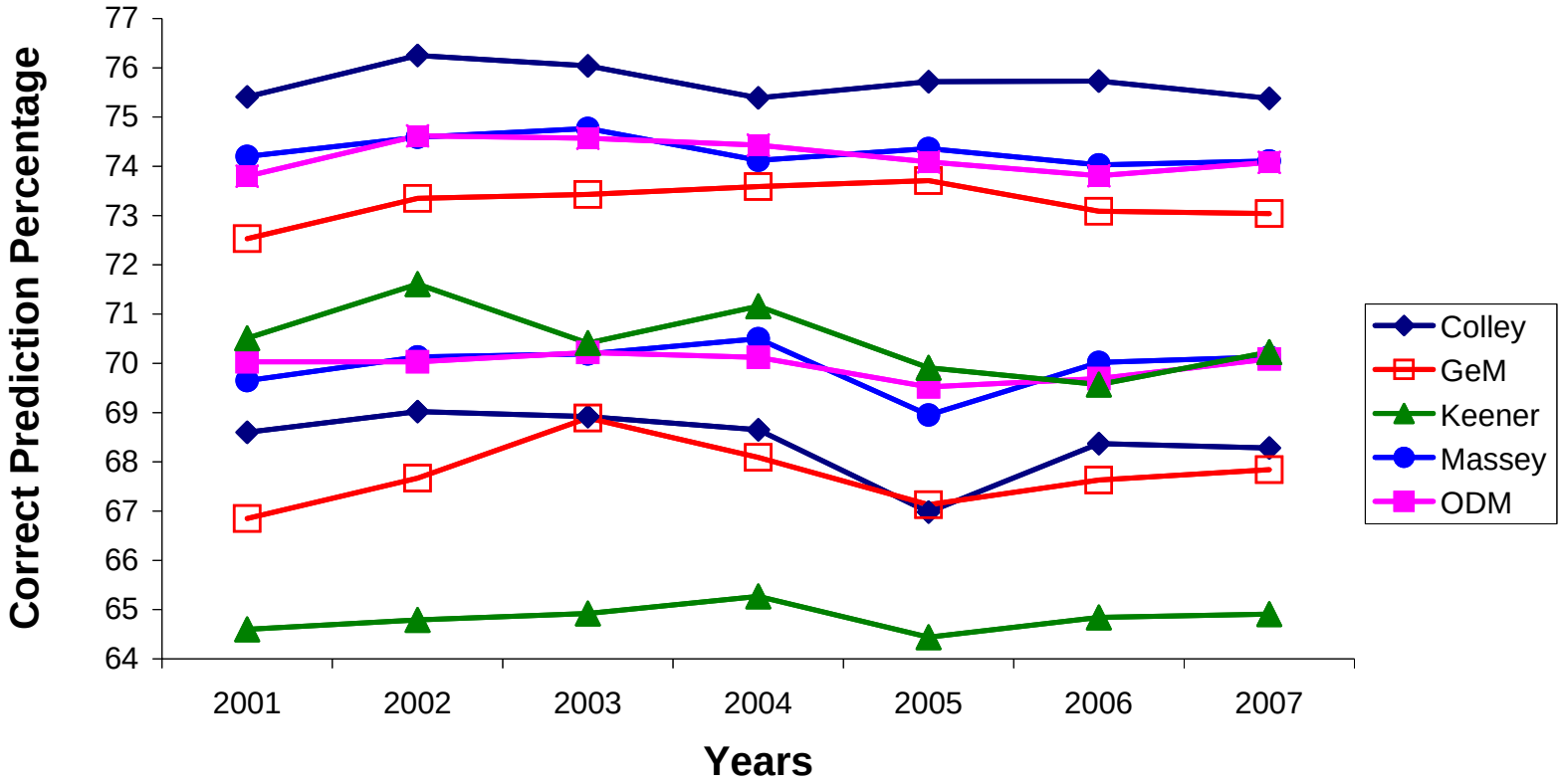
NCAA Basketball Foresight Prediction Results

	Colley	GeM	Keener	Massey	ODM
2001	68.6	66.85	64.60	69.65	70.03
2002	69.02	67.67	64.79	70.13	70.03
2003	68.92	68.89	64.92	70.19	70.22
2004	68.65	68.09	65.27	70.50	70.12
2005	66.98	67.13	64.44	68.95	69.52
2006	68.37	67.63	64.84	70.02	69.69
2007	68.28	67.84	64.91	70.13	70.09

NCAA Basketball Hindsight Prediction Results

	Colley	GeM	Keener	Massey	ODM
2001	75.41	72.53	70.51	74.2	73.8
2002	76.25	73.35	71.61	74.59	74.62
2003	76.04	73.43	70.41	74.77	74.57
2004	75.39	73.59	71.16	74.12	74.43
2005	75.72	73.71	69.91	74.36	74.09
2006	75.73	73.09	69.57	74.03	73.81
2007	75.38	73.04	70.23	74.11	74.08

NCAA Basketball Foresight/Hindsight Prediction Results



The End

Thank You! Questions?