Offense-Defense Approach to Ranking, with application to team sports.

Anjela Govan, *Northrop Grumman*
Amy Langville, *College of Charleston*
Carl Meyer, *North Carolina State University*

Ranking and Clustering Workshop,
Charleston

August 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

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

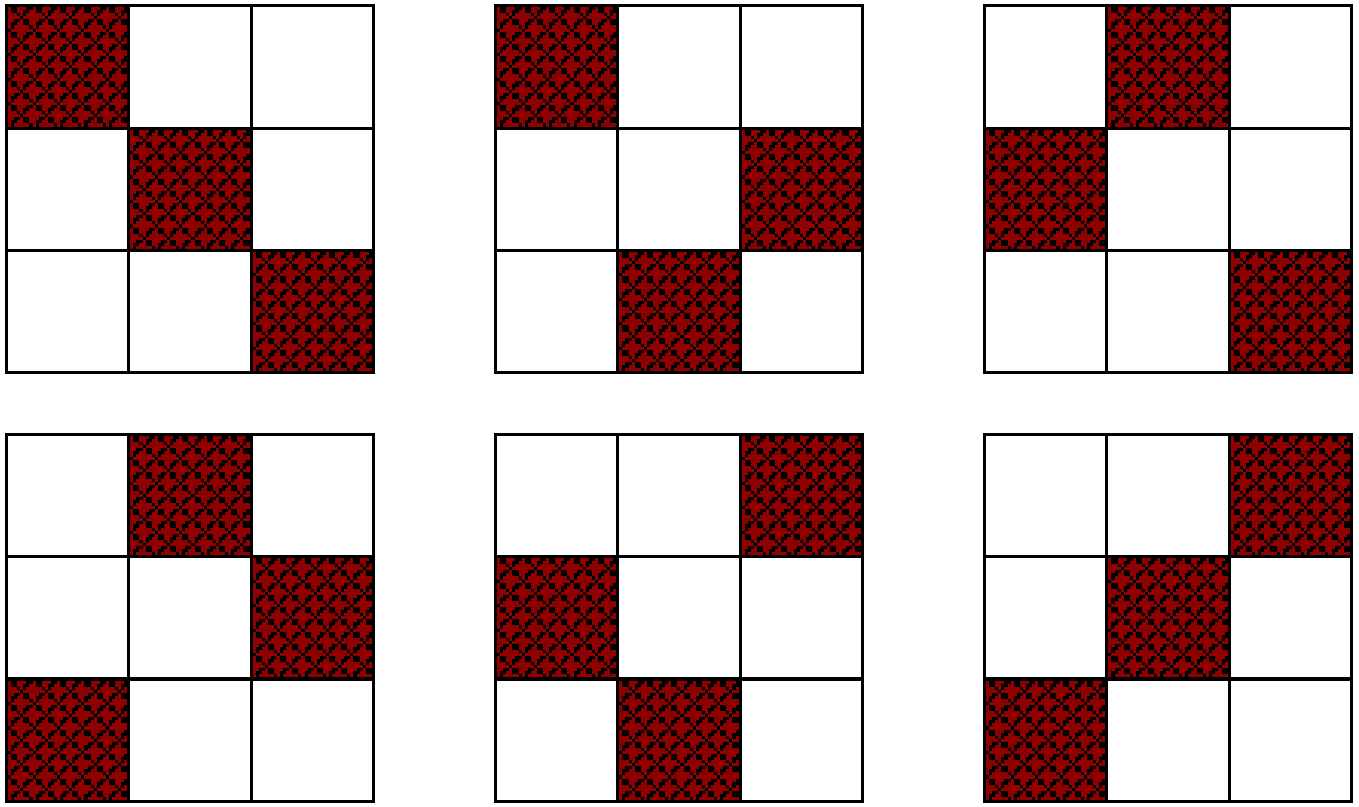
- *defensive rating* of team i

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

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

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

Matrix Structure - Diagonals $a_{1\sigma(1)}, \ldots, a_{n\sigma(n)}$



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

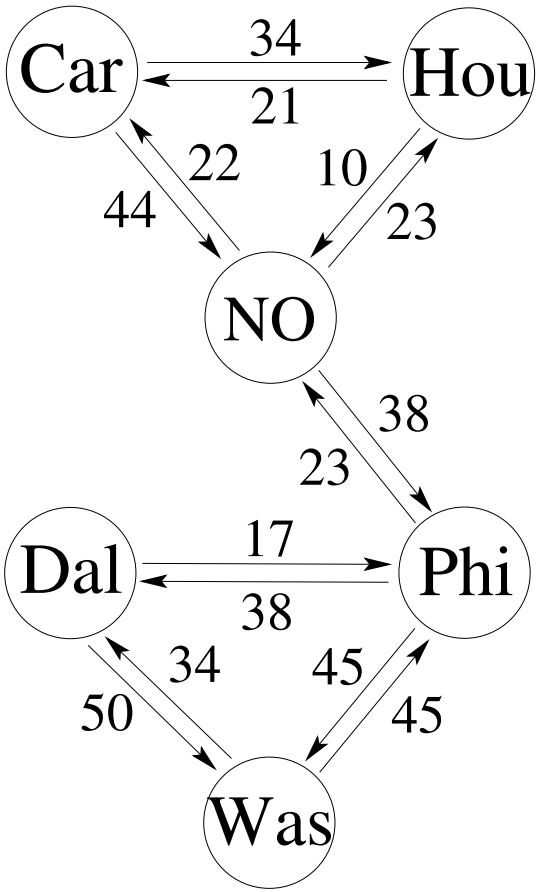
1. Represent the season using a weighted digraph with n nodes. On $i \to j$ the wight w_{ij} = amount of the statistic acquired by team j against team i.
2. Form adjacency matrix $\mathbf{A}$, $\mathbf{P} = \mathbf{A} + \epsilon \mathbf{e}\mathbf{e}^T$.
3. Team i has two rating scores, offensive o_i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

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

  $\mathbf{o} \approx (\ 0.134 \quad 7.043 \quad 0.098 \quad 0.091 \quad 6.396 \quad 12.383\ )^T$

  $\mathbf{d} \approx (\ 827.666 \quad 6.736 \quad 266.663 \quad 403.771 \quad 9.074 \quad 11.912\ )^T$

  $\mathbf{r} \approx (\ 0.00016 \quad 1.0456 \quad 0.00037 \quad 0.00023 \quad 0.705 \quad 1.04\ )^T$

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

  Dal Was Phi Hou NO Car

Matrix Based Ranking Models

- Colley Matrix 2002
  Solves system of linear equations
- Generalized Markov (GeM) 2008 (generalized PageRank 1999)
  Uses dominant eigenvector of an irreducible, primitive, stochastic matrix
- Keener's Method (Keener, 1993)
  Uses dominant eigenvector of a nonnegative irreducible matrix
- Massey Ratings (Massey, 1997)
  Least-squares based method

Colley Method

1. Form Colley matrix $\mathbf{C}$

$$\mathbf{C}_{ij} = \begin{cases} -n_{ij} & \text{if} \quad i \neq j, \\ 2 + n_i & \text{if} \quad i = j, \end{cases}$$

where n_i = total number of games played by team T_i and n_{ij} = number of times T_i played T_j.

2. Form vector $\mathbf{b}$

$$b_i = 1 + (w_i - l_i)/2,$$

where w_i = number of T_i wins and l_i = number of T_i loses.

3. Solve

$$\mathbf{Cr} = \mathbf{b},$$

the vector $\mathbf{r}$ contains rating scores of each team.

Generalized Markov Method (GeM)

1. Form matrix $\mathbf{H}$

$$\mathbf{H}_{ij} = \begin{cases} w_{ij} / \sum_{k=1}^{n} w_{ik} & \text{if } i \text{ played } j \\ 0 & \text{otherwise} \end{cases}$$

2. Form GeM matrix $\mathbf{G}$

$$\mathbf{G} = \alpha[\mathbf{H} + \mathbf{a}\mathbf{u}^T] + (1-\alpha)\mathbf{e}\mathbf{v}^T$$

where $0 < \alpha < 1$, $\mathbf{v} > 0$ and $\mathbf{u}$ are probability distribution vectors and $a_i = 1$ if $\mathbf{H}_i^T = \mathbf{0}$ and 0 otherwise.

3. The vector containing the rating scores is $\boldsymbol{\pi}$ such that

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \mathbf{G}$$

Keener Method

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

   where S_{ij} is the amount of points scored by team T_i against team T_j and

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

2. Rank vector $\mathbf{r}$ is the Perron vector of $\mathbf{A}$.

Massey Ratings

1. Form the Massey matrix $\mathbf{M} = \mathbf{X}^T\mathbf{X}$

$$\mathbf{M}_{ij} = \begin{cases} -n_{j,i} & \text{if} \quad i \neq j, \\ n_i & \text{if} \quad i = j, \end{cases}$$

   n_i - total number of games played by T_i, $n_{j,i}$ - number of times T_i played T_j.
2. Form the vector $\mathbf{d} = \mathbf{X}^T\mathbf{y}$, d_i =total difference in scores for team T_i.
3. Force $\mathbf{M}$ to have full rank, do ONE of the following
   a. Replace $\mathbf{M}$ with $\mathbf{M} + \mathbf{e}^T\mathbf{e}$, $\mathbf{e}$ - vector of all 1's.
   b. Replace one of the rows of $\mathbf{M}$ with $\mathbf{e}$ and the corresponding entry in $\mathbf{d}$ with c.
4. The ratings vector $\mathbf{r}$ is the solution to the resulting system.

Data Gathering and Automation

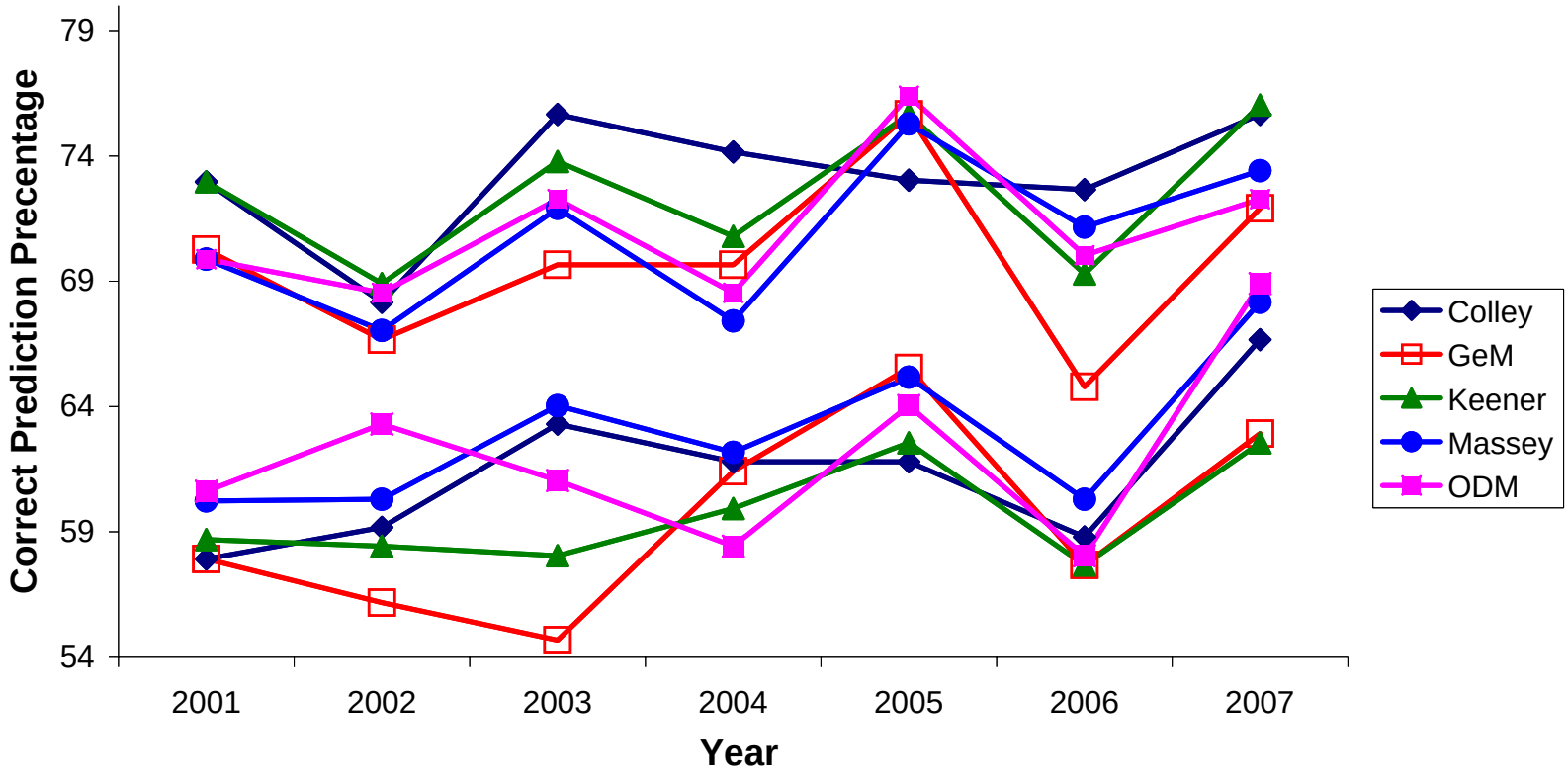
- Reliable data sources
- Data format
- Amount of data

- 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

NFL Game Prediction

- 2001-2007 with preseason padding
- ODM $tol = 0.01$, $\epsilon = 0.00001$
- GeM $\alpha = 0.6$

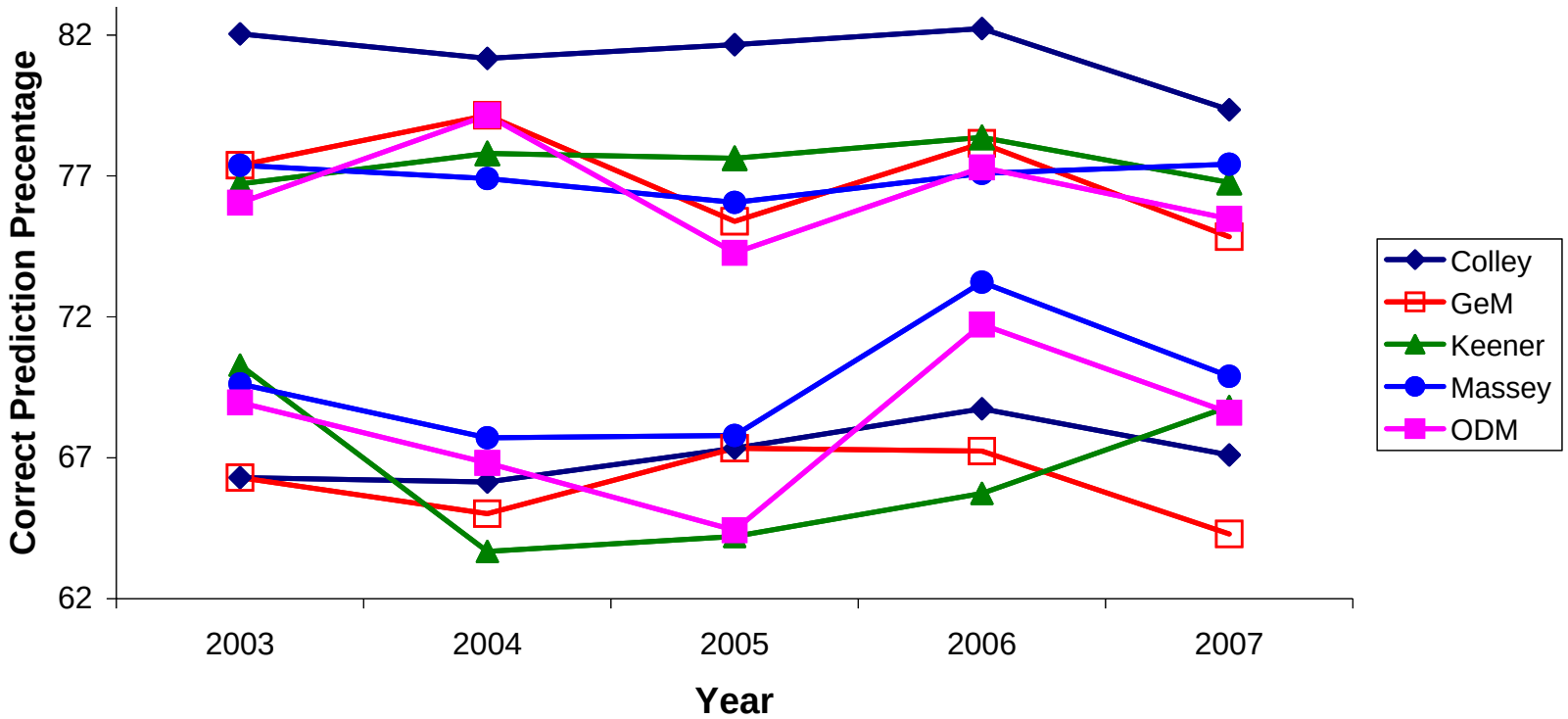
NFL Foresight/Hindsight Prediction Results



NCAA Football Game Prediction

- Div I-A
- 2003-2007 starting week 5
- ODM $tol = 0.01$, $\epsilon = 0.00001$
- GeM $\alpha = 0.6$

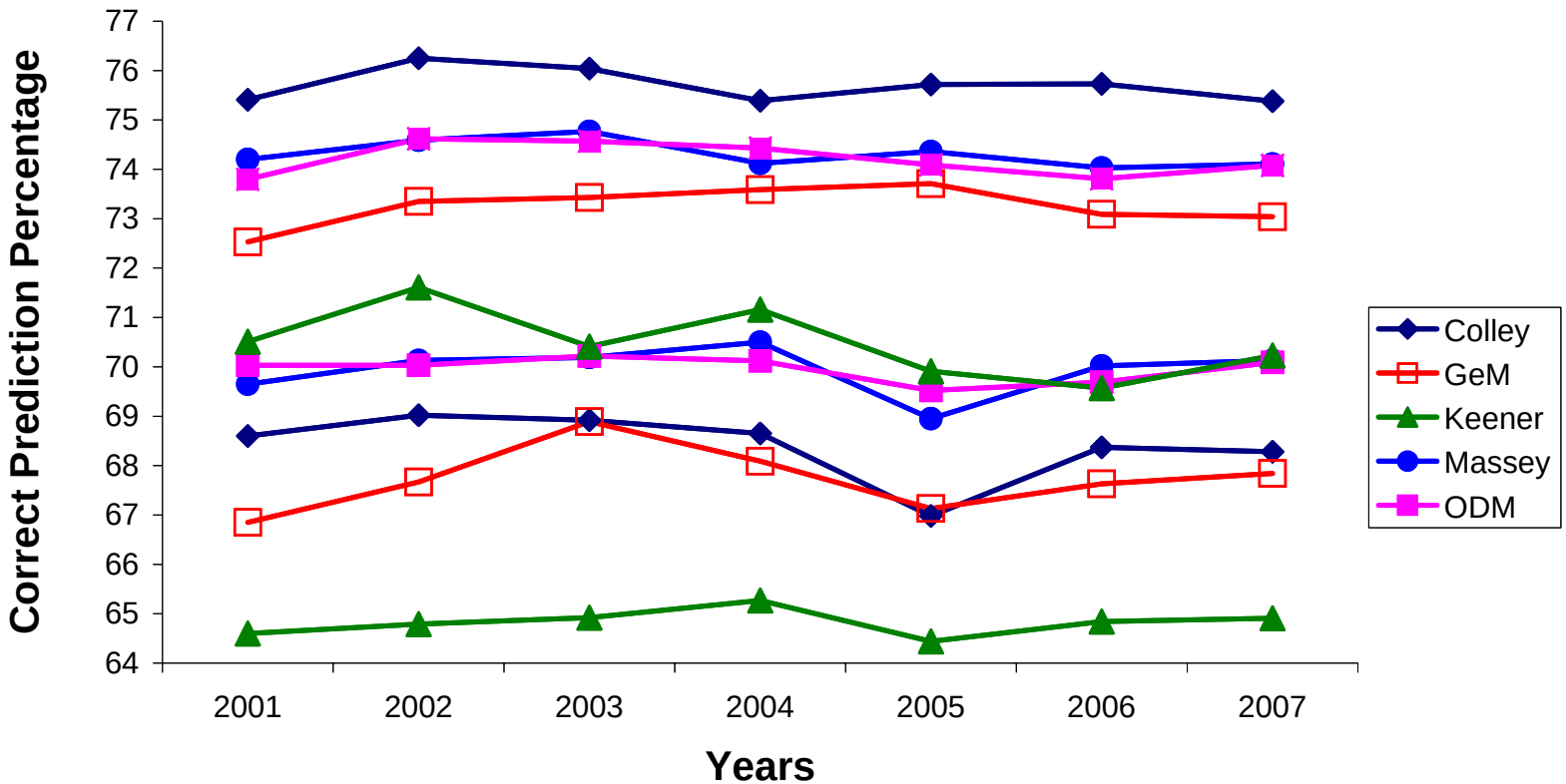
NCAA Football Foresight/Hindsight Prediction Results



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/Hindsight Prediction Results



Ranking Naming Schemes - Problem of Voting

- Have a number of conference rooms to name
- 14 employees voting
- 7 proposed naming schemes
  - College Mascots/Conferences/Teams
  - Dead Giants of Science/Math/Engineering
  - English Premier League Soccer Teams
  - Famous Golf Courses
  - NASCAR Tracks/Races
  - NC Beaches/Beach towns
  - Old-School Arcade Games

Data - round-robin tournament

Pair	LEFT Name Scheme	prefer by	RIGHT Name Schemes	prefer by
1	Famous Golf Courses	3.00	NC Beaches/Beachtowns	0.33
2	Famous Golf Courses	5.00	Old-School Arcade Games	0.20
3	Famous Golf Courses	5.00	NASCAR Tracks/Races	0.20
4	Famous Golf Courses	5.00	College Mascots/Conferences/Teams	0.20
5	Famous Golf Courses	5.00	Dead Giants of Science/Math/Engineeri	0.20
6	Famous Golf Courses	5.00	English Premier League Soccer Teams	0.20
7	NC Beaches/Beachtowns	5.00	Old-School Arcade Games	0.20
8	NC Beaches/Beachtowns	4.00	NASCAR Tracks/Races	0.25
9	NC Beaches/Beachtowns	4.00	College Mascots/Conferences/Teams	0.25
10	NC Beaches/Beachtowns	5.00	Dead Giants of Science/Math/Engineeri	0.20
11	NC Beaches/Beachtowns	5.00	English Premier League Soccer Teams	0.20
12	Old-School Arcade Games	0.25	NASCAR Tracks/Races	4.00
13	Old-School Arcade Games	0.25	College Mascots/Conferences/Teams	4.00
14	Old-School Arcade Games	3.00	Dead Giants of Science/Math/Engineeri	0.33
15	Old-School Arcade Games	0.25	English Premier League Soccer Teams	4.00
16	NASCAR Tracks/Races	2.00	College Mascots/Conferences/Teams	0.50
17	NASCAR Tracks/Races	3.00	Dead Giants of Science/Math/Engineeri	0.33
18	NASCAR Tracks/Races	0.33	English Premier League Soccer Teams	3.00
19	College Mascots/Conferences/Teams	3.00	Dead Giants of Science/Math/Engineeri	0.33
20	College Mascots/Conferences/Teams	0.25	English Premier League Soccer Teams	4.00
21	Dead Giants of Science/Math/Engineerin	0.25	English Premier League Soccer Teams	4.00

Rank according to Colley

1. Famous Golf Courses
2. Dead Giants of Science/Math/Engineering
3. Old-School Arcade Games
4. College Mascots/Conferences/Teams
5. NC Beaches/Beach towns
6. English Premier League Soccer Teams
7. NASCAR Tracks/Races

Rank according to GeM

1. Famous Golf Courses
2. Old-School Arcade Games
3. NC Beaches/Beach towns
4. Dead Giants of Science/Math/Engineering
5. College Mascots/Conferences/Teams
6. English Premier League Soccer Teams
7. NASCAR Tracks/Races

Rank according to ODM

1. Famous Golf Courses
2. NC Beaches/Beach towns
3. Old-School Arcade Games
4. College Mascots/Conferences/Teams
5. Dead Giants of Science/Math/Engineering
6. English Premier League Soccer Teams
7. NASCAR Tracks/Races

The End

Thank You! Questions?