Using a combination of the Generalized PageRank Model (GeM) with the Four Factors to Rank NBA Teams

Georgia Twersky [1,2] , George Lyman[1,3], and Annette Pilkington[1,4]

[1] University of Notre Dame, Notre Dame, Indiana, U.S.A.

[2] gtwersky@nd.edu

[3] glyman@nd.edu

[4] Pilkington.4@nd.edu

- Introduction
- Measuring the Strength of a Ranking System
- Generalized Page Rank (GeM)
- Massey’s Method
- The Four Factors
- Performance of Ranking Systems on NBA Tournament

Introduction

- What is the NBA and how do the NBA Playoffs work?
- How does a playoff bracket work?
- Summary of our method used to evaluate brackets
- Summary of Results

The NBA and the Playoffs

- 30 Teams
- 2 Conferences
- 82 Games
- Significant interaction between conferences

- Top eight teams per conference
- Best of seven series
- Single elimination

Playoff Bracket

Step 1

Bracket is filled out before the playoffs begin predicting the winner of each series

Step 2

Brackets are scored with the point value of a correct answer doubling each round

Step 3

The Bracket with the highest score is pronounced the winner

Summary of our Method

Previous Research	Our Focus	Implementation
❖ 2008 Meyer et al. [1] proposed a generalization of Google's PageRank (GeM) ❖ Used the point differential as weights for a weighted graph of games played ❖ Suggested using other game statistics	❖ Use Oliver's Four Factors of basketball with GeM ❖ Compare the results to the regular model	❖ Used data from Bigdataball.com from 2007-2018 ❖ Created brackets using the regular GeM model and the generalized version with the Four Factors

Summary of Results

Contrary to our expectations, no significant improvement resulted when the four factors were incorporated

Introduction

Measuring the Strength of a Ranking System

Generalized Page Rank (GeM)

Massey’s Method

The Four Factors

Performance of Ranking Systems on NBA Tournament

Measuring the Ranking System

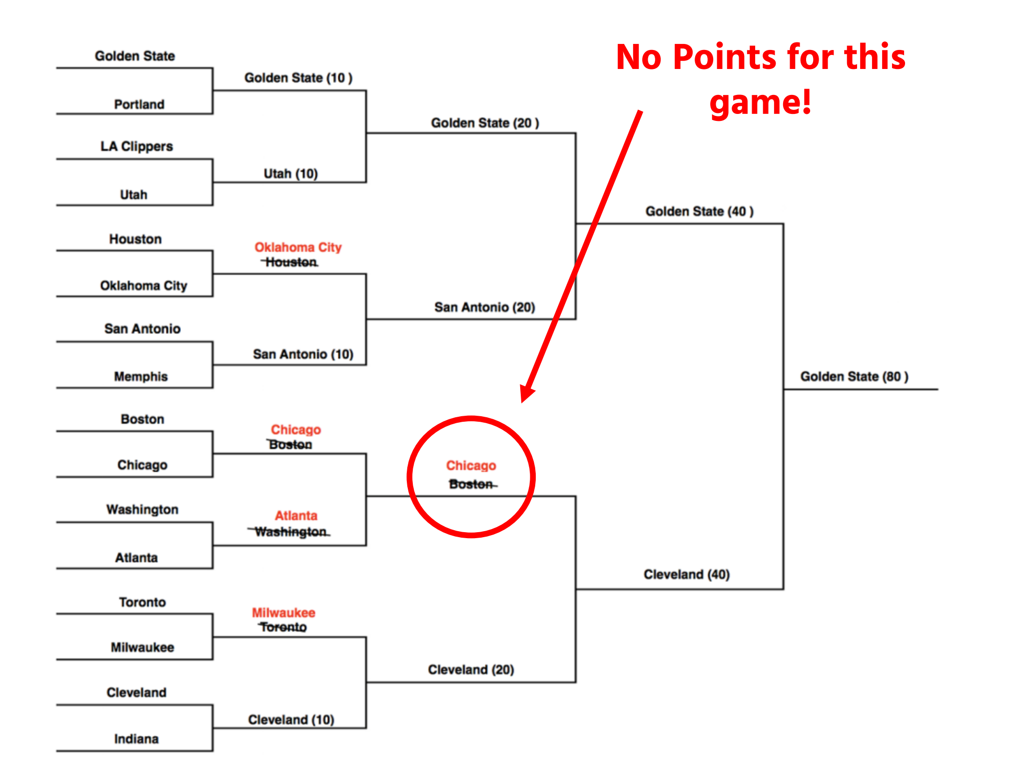
Championship	80 Points →	1 Series
Semifinals	40 Points →	2 Series
2nd Round	20 Points →	4 Series
1st Round	10 Points →	8 Series

320 Total Points

Ranking System Effectiveness is measured by total points

Team	Rank
Golden State	1
Denver	2
San Antonio	3
Utah	4
Oklahoma City	5
Houston	6
LA Clippers	7
Cleveland	8
Chicago	9
Minnesota	10
Atlanta	11
Portland	12
Memphis	13
Boston	14
Brooklyn	15
LA Lakers	16
Sacramento	17
Milwaukee	18
New Orleans	19
New York	20
Miami	21
Washington	22
Indiana	23
Phoenix	24
Philadelphia	25
Detroit	26
Charlotte	27
Toronto	28
Orlando	29
Dallas	30

Ranking from GeM with four factors for 2017 on left
Tournament results below



Generalized Page Rank (GeM)

- Used by Google based on theory of **Markov chains**
- Developed by Google's founders Larry Page and Sergey Brin
- How it works:
  - Web pages are ranked higher through a Google search based on quantity and quality
    - *Quantity- the more links pointing to the page, the higher the ranking*
    - *Quality- if a high quality website refers a page, the higher the ranking*

Sergey Brin and Larry Page

Generalized Page Rank (GeM)

Determines popularity of a web page by modeling internet activity

Certain web pages can link surfers to another page, increasing its visits

Google models everyone as being a random surfer by assuming user randomly chooses links to follow

Reporting the proportion of times each web page is randomly visited, Google is able to develop the PageRank for the web pages in that network

Intuition behind Google Page Rank

Definition If a matrix G is stochastic (rows add to 1), is irreducible and has at least one positive diagonal entry, then the **Perron Frobenius theorem** guarantees that it has a unique eigenvector with eigenvalue 1 and norm 1

The algorithm creates the adjacency matrix of the directed graph of web page links, and normalizes it to get a hyperlink matrix *H*

The matrix *H* is adjusted by adding a rank 1 matrix to deal with dangling nodes (nodes with no arrows pointing outwards, which represents pages with no links)

- Adjustments to ensure irreducibility and positive diagonals.

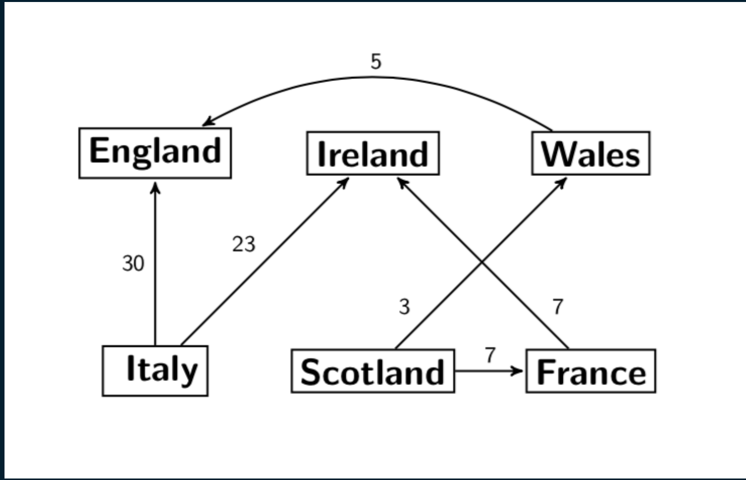
The associated rating vector for the web pages (the Perron vector) is an eigenvector of the transpose matrix, and can be computed by repeatedly applying the matrix to an initial probability distribution vector (using the theory of Markov chains) to get the required vector as the stable vector of the system.

Applying GeM to sports leagues

- We apply an adaptation of Google's PageRank system introduced by Meyer, Albright, and Govan for sports leagues.

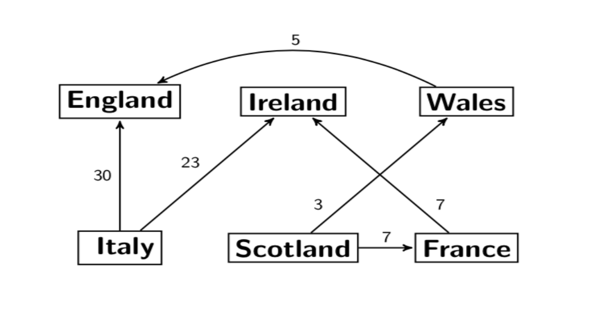
- Sport season
  - A weighted directed graph with n nodes (n= number of sports teams involved)
  - Teams correspond to the nodes
  - Each game is an arrow of loser to winner, with weight w_{ij} , equal to the absolute value of that point differential

Example: Six Nations Rugby, 2015, Rounds 1 and 2.



Steps to constructing stochastic matrix G

- Form the n x n adjacency matrix A of the graph of web pages:
  - $A = \begin{cases} w_{ij} & \text{if team i lost to team j} \\ 0 & \text{otherwise} \end{cases}$



$$
\begin{array}{c|cccccc}
 & \text{Irl.} & \text{Eng.} & \text{Wal.} & \text{Scot.} & \text{Fra.} & \text{Itl.} \\
\hline
\text{Irl.} & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{Eng.} & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{Wal.} & 0 & 5 & 0 & 0 & 0 & 0 \\
\text{Scot.} & 0 & 0 & 3 & 0 & 7 & 0 \\
\text{Fra.} & 7 & 0 & 0 & 0 & 0 & 0 \\
\text{Itl.} & 23 & 30 & 0 & 0 & 0 & 0
\end{array}
$$

Steps to constructing stochastic matrix G

- Form the stochastic "hyperlink" matrix H where
  - $H_{ij} = \{ A_{ij} / \sum_{k=1}^{n} A_{ik}$

  if there is a link between i and j

  0

  otherwise

	Irl.	Eng.	Wal.	Scot.	Fra.	Itl.
Irl.	0	0	0	0	0	0
Eng.	0	0	0	0	0	0
Wal.	0	5	0	0	0	0
Scot.	0	0	3	0	7	0
Fra.	7	0	0	0	0	0
Itl.	23	30	0	0	0	0

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/10 & 0 & 7/10 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 23/53 & 30/53 & 0 & 0 & 0 & 0 \end{pmatrix}$$

H

Steps to constructing stochastic matrix G

- Make an adjustment to H for the dangling nodes (rows of zeros corresponding to unbeaten teams) by adding $1/n\ a\ e^T$ *to get* $H + 1/n\ a\ e^T$. Here a is an n x 1 column matrix with 1's in the j position if j is unbeaten and 0's elsewhere, and e is an n x 1 column matrix of 1's.

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/10 & 0 & 7/10 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 23/53 & 30/53 & 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/10 & 0 & 7/10 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 23/53 & 30/53 & 0 & 0 & 0 & 0 \end{pmatrix}$$

H $\qquad\qquad$ $H + 1/n\ a\ e^T$

Steps to constructing stochastic matrix G

- Finally, the adjustment to ensure irreducibility and primitivity is made to get the basic version of the GeM (Generalized Markov Chain) matrix G, given by:

$$G = \alpha[H + \frac{1}{n}ae^T] + \frac{(1-\alpha)}{n}ee^T, \qquad 0 \le \alpha \le 1$$

Where alpha is a chosen scaling parameter which ensures that the resulting matrix is stochastic. Smaller values of alpha give a larger perturbation of the GeM matrix.

$$\begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/10 & 0 & 7/10 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 23/53 & 30/53 & 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/40 & 7/8 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 7/25 & 1/40 & 31/50 & 1/40 \\ 7/8 & 1/40 & 1/40 & 1/40 & 1/40 & 1/40 \\ 167/424 & 1073/2120 & 1/40 & 1/40 & 1/40 & 1/40 \end{pmatrix}$$

G; alpha = 0.85

Ranking The Teams

- We use the Perron eigenvector (v) of length one with $vG=v$ to give a rating and a ranking for the teams.

Team	Rank	Rating
Ireland	1	0.269
England	2	0.252
Wales	4	0.124
Scotland	5	0.099
France	3	0.158
Italy	5	0.099

The Generalized GeM Model

Note that we can apply the same process with any the differential for any game statistic replacing the point differential.

Meyer et. al. proposed using several game statistics in this way to create a number of stochastic matrices S_{1},S_2, . . . ,S_n , and using weights to create the matrix G for which the Perron vector is used for ranking:

$$G = \alpha_1 S_1 + \alpha_2 S_2 + \cdots + \alpha_P S_P, \quad 0 \leq \alpha_i \leq 1, \quad \sum_{1}^{P} \alpha_i = 1.$$

We decided to test this model using the Four factors for basketball with the weights given by Dean Oliver.

Intuition behind Massey's Method

- Based on the idea that with a perfect set of rating for the teams r_i , the difference in the ratings for the two teams would equal the point differential for each game played between them.
  - Gives us a system of equations $r_i - r_j = p_k$, one for each game.

- Massey algorithm takes the least squares solution to this system to derive a system of equations with infinitely many of solutions.

Generalized Massey

- One can also create a generalized version of Massey's method using a weighted combination of game statistics by simply replacing the point differential in the calculations by the weighted combination of the differentials in the game statistics.

- We also used Dean Oliver's four factors and weights to generalize Massey's method in this way.

Shooting

Turnovers

Rebounds

Free Throws

Intuition behind Massey's Method

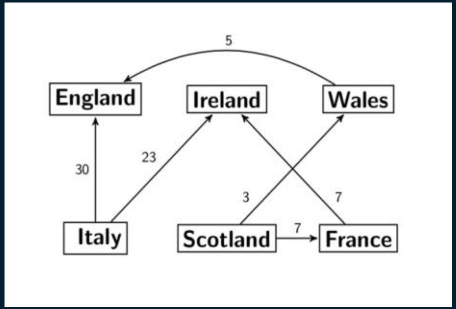
- The result is the system of equations for the *n* teams in the tournament:

$$\left\{ t_i r_i - \sum_{j \neq i} n_{ij} r_j = P_i,\ \sum_j r_j = 0 \right\}_{1 \leq i \leq n-1,\quad 1 \leq j \leq n}$$

t_i = the number of games played by team i,
n_{ij} = minus the number of games played between team i and team j,
P_i denotes the total point differential for team i.

- The last equation in the new system is replaced by the condition that the sum of the associated ratings adds to 0 to get a unique solution, which gives us the Massey ratings.

Massey For Our Running Example



$$\begin{pmatrix} 2 & 0 & 0 & 0 & -1 & -1 \\ 0 & 2 & -1 & 0 & 0 & -1 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ -1 & 0 & 0 & -1 & 2 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \end{pmatrix} = \begin{pmatrix} 30 \\ 35 \\ -2 \\ -10 \\ 0 \\ 0 \end{pmatrix}$$

Team	Rank	Rating
Ireland	2	6.91667
England	1	9.58333
Wales	3	2.41667
Scotland	5	-2.75
France	4	2.08333
Italy	6	-18.25

Introduction

Measuring the Strength of a Ranking System

Generalized Page Rank (GeM)

Massey’s Method

The Four Factors

Performance of Ranking Systems on NBA Tournament

Intuition behind The Four Factors

- The Four Factors are game statistics that correlate very closely with the point differential in basketball games.
  - Shooting
  - Turnovers
  - Rebounding
  - Free throws

- **Effective Field Goal Percentage (eFG%)** ; this statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal. Formula given by:
  - $eFG\% = [(FG + 0.5 * 3P)/FGA] \times 100\%$.

- **Turnover Percentage (TOV%)** estimates turnovers per 100 plays.
  - $TOV\% = TOV/Poss \times 100\%$ $(Poss \approx FGA - OR + TO + 0.44 * \mathrm{FTA})$
    Where *Poss* denotes possessions

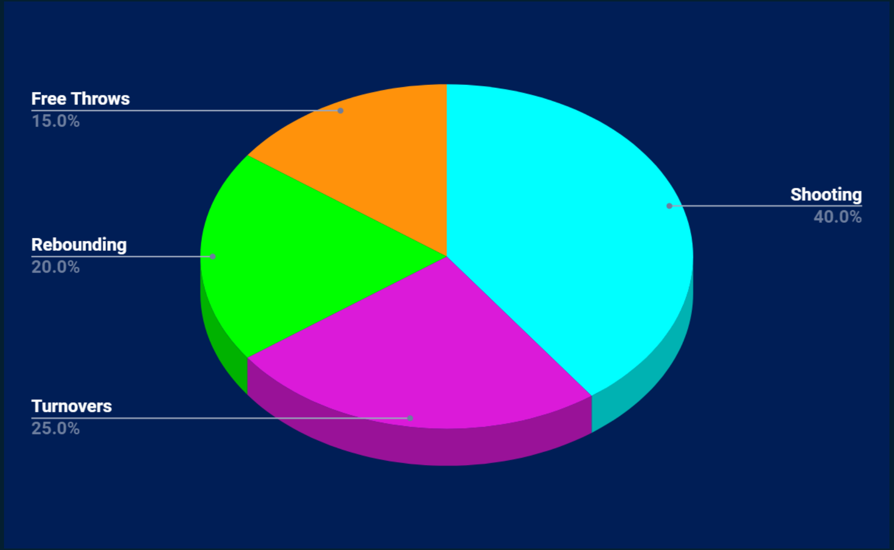
Intuition behind The Four Factors (continued)

- **Offensive Rebound Percentage (ORB%)** is an estimate of the percentage of available offensive rebounds grabbed by a player on the team.
    - *ORB% = ORB.P = ORB/(ORB + DRB)*
      Where *DRB* is the opponents defensive rebounds.

- **Free Throws Per Field Goal Attempt (FTPFGA)** measures the team's ability to get to the foul line by dividing Free Throws Made by Field Goals Attempted
    - *FTR = FTM/FGA*

Intuition behind The Four Factors (continued)

- The factors show a high degree of independence and do not carry equal weight in determining wins and losses
- Oliver’s theory assigns the following linear weights to the factors:
  - Shooting (40%)
  - Turnovers (25%)
  - Rebounding (20%)
  - Free Throws (15%)



Introduction

Measuring the Strength of a Ranking System

Generalized Page Rank (GeM)

Massey’s Method

The Four Factors

Performance of Ranking Systems on NBA Tournament

Performance of Ranking the NBA Tournament

- We ran the regular GeM model on the data from the regular season for each tournament in the study.
- We set the parameter ALPHA to 0.85
  - Ran Massey's algorithm
  - Used linear weights from Oliver's theory to incorporate the Four Factors into the GeM model:

α_1= 0.4, α_2= -0.25, α_3 = .2, α_4=0.15

G = 0.4*S_1,+ -.25*S_2, . . . ,S_n

- Our hope was that by separating these factors in the generalized GeM model, the breakdown of the point differential would lead to better predictions than those obtained from the regular GeM model.

Performance of Ranking on NBA Tournament

The number of points earned on a bracket (maximum 320).

Year	GeM	GeM with Four Factors	Massey	Massey with Four Factors
Year				
2007	210	80	210	210
2008	150	140	290	250
2009	160	180	140	110
2010	80	140	100	110
2011	110	90	130	110
2012	150	140	110	150
2013	130	80	130	60
2014	260	120	270	90
2015	260	300	240	140
2016	210	220	200	180
2017	310	260	250	260
2018	110	220	100	150

Each of the four ranking systems

Bracket scores by year and method (2016 marks end of 2015/2016 season).

Challenges

- No significant difference between the performance of the ranking systems using the four methods
- Some significant differences on the performance of models from year to year.
- Main challenge with using the Generalized GeM model is to decide which parameters are optimal.
    - We used parameters from Oliver's linear method, which may not be optimal since we are dealing with eigenvectors and the eigenvector of a linear combination of matrices
        - Which is not necessarily a linear combination of eigenvectors of the matrices.
- May be beneficial to pursue non-linear methods of combining eigenvectors derived for the individual factors to create a ranking

References

[1] Carl D. Meyer Anjela Y. Govan and Russell Albright. Generalizing google's pagerank to rank national football league teams. SAS Global Forum 2008, Paper 151, 2008.

[2] Sergey Brin and Lawrence Page. The anatomy of a large scale hypertextual web search engine. Computer Networks and ISDN Systems, 33: 107-17, 1998. Available at http://easychair.org/ easychair.zip.

[3] K. Pelton J. Kubatko, D. Oliver and D. Rosenbaum. A starting point for analyzing basketball statistics. Journal of Quantitative Analysis in Sport, Vol 3, Issue 3, 2007, Article 1., 2007.

[4] Konstantinos Kotzias. The four factors of basketball as a measure of success. statathlon.com: https://statathlon.com/four-factors-basketball-success/., 2018.

[5] Kenneth Massey. Statistical models applied to the rating of sports teams. Thesis, Bluefield College, 1997.

[6] R user groups. The comprehensive r archive network. https://cran.r-project.org/

THANKS!