NBA PLAYER EVALUATION

Submitted by

Kiran Kumar B S	**12IT36**
Suhas H S	**12IT85**
Vinay Rao D	**12IT94**

Under the Guidance of

Mr. Biju R Mohan

Assistant Professor, Information Technology

NITK Surathkal, Mangalore

In partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

In

Information Technology

Department of Information Technology

National Institute of Technology Karnataka Surathkal

2015-16

Department of Information Technology
National Institute of Technology Karnataka Surathkal

DECLARATION

We hereby declare that the project work report entitled **NBA Player Evaluation** which is being submitted to the **National Institute of Technology Karnataka Surathkal** for the award of the degree of Bachelor of Technology in **Information Technology**, is a bonafide report of the work carried out by us. The material content in this project work report has not been submitted in any university or institution for the award of any degree.

Name	**Register No.**	**Signature with Date**
Kiran Kumar B S	12IT36	
Suhas H S	12IT85	
Vinay Rao D	12IT94	

Place: NITK Surathkal, Mangalore
Date: May 2, 2016

Department of Information Technology

National Institute of Technology Karnataka Surathkal

CERTIFICATE

This is to certify that the B.Tech project work report entitled **NBA Player Evaluation** submitted by :

Kiran Kumar B S	12IT36
Suhas H S	12IT85
Vinay Roa D	12IT94

as the record of the work carried out by them, is *accepted as the B.Tech Project Work Report submission* in partial fulfillment of the requirements for the award of degree of **Bachelor of Technology** in **Information Technology**.

Mr. Biju R Mohan
Project Guide
NITK Surathkal, Mangalore

Prof. G. Ram Mohana Reddy
Chairman - DUGC, Dept. of IT
NITK Surathkal, Mangalore

Abstract

The major difficulty in evaluating individual player performance in basketball is adjusting for interaction effects by teammates. With the advent of play-by-play data, the plus-minus statistic was created to address this issue. While variations on this statistic (ex: adjusted plusminus) do correct for some existing confounders, they struggle to gauge two aspects: the importance of a players contribution to his units or squads, and whether that contribution came as unexpected (i.e. over or underperformed) as defined by a statistical model. We try to quantify both in our project by adapting a network-based algorithm to estimate centrality scores and their corresponding statistical significances. We will construct a single network where the nodes are players and an edge exists between two players if they played in the same five-man unit. These edges are assigned weights that correspond to an aggregate sum of the two players performance during the time they played together. We will then determine the statistical contribution of a player in this network by the frequency with which that player is visited in a random walk on the network. Then we will try to predict the results of regular season of NBA using traditional statistic methods such as box-score based models and plus-minus based models and also using the network based model and compare how they perform.

Keywords : Statistical network models, basketball, predict regular season results.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the past, sports organizations transformed historical data into useful knowledge mostly depending on the experience from coaches, scouts and managers. However, relying only on the experts' experience and intuition, organizations could not discover all the value and potential of collected data. A more scientific approach was needed to use the data, so sports analytics has emerged. Currently, sports analytics is being used successfully in many fields, such as Basketball (NBA, NCAA), Baseball(MLB), Soccer, Cricket, Hockey etc.

National Basketball Association (NBA) since its origin has over 68 years. During this organization grow up there are 30 teams formed and divided into Eastern Conference and Western Conference. For the regular season will have 80 games for each team and post season using a best-of-seven series scheme. So a conservative estimate, there will be at least about 12,300 games generated. A mass of data was generated after each NBA game played; those existed data allow us to discover something invisible valuable knowledge. When people pay attention to their favorite team or players, they definitely will concern about the game outcome.

However, predicting the outcomes of competitive sport has always been a challenging and attractive work. The main obstacle analysts in these sports face when evaluating player performance is accounting for interaction effects by fellow teammates, or teamwork. Certain players might find themselves scoring more on a team not because of an increase in scoring ability, but due to the lack of a supporting cast.

So in this project we try to come up for following two questions

1. Given the five-man units of which a player was a member, how important was

that player relative to all other players.

2. How well statistically did that player perform in that role. This project focus on data mining techniques such as network analysis to predict the NBA game outcome.

We choose to build a model inspired by work on social networks. This network is designed to encourage the random walk to visit areas of high gene transcriptional dysregulation. In the same spirit, we implement an algorithm that executes a similar search on a network of basketball players. We will construct a network of individual players, where the nodes are players and two players are connected if they were a member of the same five-man unit at least once. Importantly, the edges are weighted to reflect the interdependency between players with respect to their units performances. Using a random walk, we will be able to determine statistically how central/important a player is relative to all other players in the network, which is referred to as a centrality score [1].

Chapter 2

Literature Survey

This section covers the background work with respect to the related research areas that are useful to address various issues and challenges.

2.1 Related Work

In this section we describe some of the previous work that has been done in the field of basketball analytics, a field which includes both predictive and descriptive analytics. Traditional analysis has focused mostly on evaluative/descriptive statistics, i.e those that focus on describing what has already happened. The goals of those analyses is usually to evaluate players and/or teams and not necessarily to make predictions. However, they are still useful to review, as they are a good starting point in analyzing basketball.

There are lot of data mining models available to make more accurate prediction. The main two categories of models are Descriptive models and Predictive models. The difference between these two models is that Predictive model deals with numerical/continuous target attributes where we predict the possible outcome, whereas Descriptive model deals with discrete/categorical target attributes.

Descriptive models can be generally separated into three categories:

1. Team based analysis: Team based analysis look at factors that drive winning at a team level.

2. Box Score based analysis: Box-Score based analysis generally tries to look at

each box score for players, and assign values to each of the individual box score statistics. The most common box-score based statistics are the following:

(a) Wins Produced: Wins produced was developed by David Berri in 1999. Ultimately, the goal is to estimate the number of wins a player produces for his team, had he played the entire game. A detailed way to calculate it can be found on the wages of wins website [2].

(b) Win Shares: is an extension of Bill James' work on baseball onto basketball. The methods to calculate the offensive portion of win shares is similar to that of wins produced. A detailed method of calculation can be found on the basketball-reference website [3].

(c) PER, or Player Efficiency Rating: PER was developed by John Hollinger, and it assigns coefficients to box score statistics per minute played, and then takes the sum of the coefficients multiplied by the box score statistic, and adjusts for the pace of a team [3].

(d) Wins Above Replacement Player, or WARP: WARP is another system that assigns values to each of the box-score statistics, and the value of a player is the sum of the weights multiplied by the number of times that player performs a box-score statistic.

3. Plus-minus based analysis: The "plus-minus" statistic of a player is simple the number of points his team scored while the player was on the floor minus the number of points the opposing team scored while the player was on the floor.A detailed explanation of the method for adjusting these numbers can be found in [4].

There are some models that attempt to predict outcomes in the NBA. However, most of these models just seem to simply apply machine learning techniques to box score data without doing much else [5] [6]. These models simply take team-level box score data and apply techniques such as logistic regression, naive Bayes', or SVNs. However, they don't seem to present any interesting features or other insights.

Another model by Nate Silver uses a Naive Bayes classifier to predict outcomes in the NCAA tournament [7]. However, as inputs to the model, instead of using team box-score statistics, he uses team strengths from other models, as well as pre-season predictions. Silver also uses other inputs to the model, such as distance travelled by both teams, and discovers that teams that travel more tend to perform worse than they would otherwise. While Silver's model seems to perform well, his methods doesn't quite scale to the problem in hand [8].

2.2 Outcome of Literature Survey

As it can be seen, a lot of work has been done in this field but even then there is a lot of scope for improvement. The individual box-score based models suffer from the following shortcomings:

1. Box-score based models assume that player performance is independent of another player's performance by assigning the value from a box score statistic to the person responsible for that statistic. However, this might not necessarily be the case.

2. Box-score based models might not divide defensive credit (or fault) fairly. Wins produced divides the credit evenly, so if a good defensive player plays with four bad ones, his defense might also look bad.

3. Box-score based models fail to take into account the context of the play; not all statistics of the same type are created equal. For example, a foul is generally a bad play, but a non-shooting foul is generally worse than a shooting foul.

The adjusted plus-minus model also suffers from the first flaw, since it assumes that a team strength is a linear sum of player strengths, ignoring any player-to-player interaction that might be present.

Our project contributes a new approach to a well-researched topic by employing network analysis techniques, rather than traditional regression methods. Our algorithm will provide new and interesting ways of evaluating basketball player performance. We will shed light on statistically significant players who are underhand or over-performing on offense, defense, and in total. We will gain insight on how important certain players are to their units relative to other players. Lastly, by combining these two aspects, we will be able to form a more complete analysis of a players abilities.

2.3 Problem Statement

Evaluate player performance using network model rather than box-score based model (PER Model) which do not take into consideration the interaction between players in the team and predict the regular season results of NBA (win-loss ratio).

2.4 Objectives

1. Evaluate team performance and explain it using box-score based models i.e Player Efficiency Rating(PER) Model and predict regular season results (win-loss ratio).

2. Evaluate player performance via statistical network model and explain its statistical significance.

3. Calculating the actual efficiency of each unit and selecting the best five-man unit for each team.

Chapter 3

Proposed Methodology

Figure 3.1 shows the workflow of the entire project. We use two models to predict the end season results, PER [Player Efficiency rating] Model and Network Model and compare the results. The current industry standard for this analysis is PER Model, that makes use of PER calculated using the Box Score Statistics. PER is a weighted average of the attributes in the box score statistics that is collected from basketball-reference.com. From this PER, Team PER and Win ratio is calculated, which is used to predict the end season results. The next model, Network Model, makes use of Gibbs Sampler to estimate unit efficiency. Next a weighted graph is constructed where nodes are players and units and the player centrality [importance] is calculated. This is used to predict the end season results. The end season results predicted by both the models are then compared.

3.1 PER Model

3.1.1 Data Collection

NBA basketball is highly statistically developed and there is an abundance of sortable basketball statistics, most of which are free and easy to obtain. The data used in this paper are taken from the website basketball-reference.com [3]. This website provides downloadable data for both players and teams from 1946-1947 season till present.

We chose to focus on the most recent 20 regular seasons and base our model on the data from these 20 seasons because as the game evolves over time, the characteristics that distinguish teams winning and losing also change considerably. Therefore, using

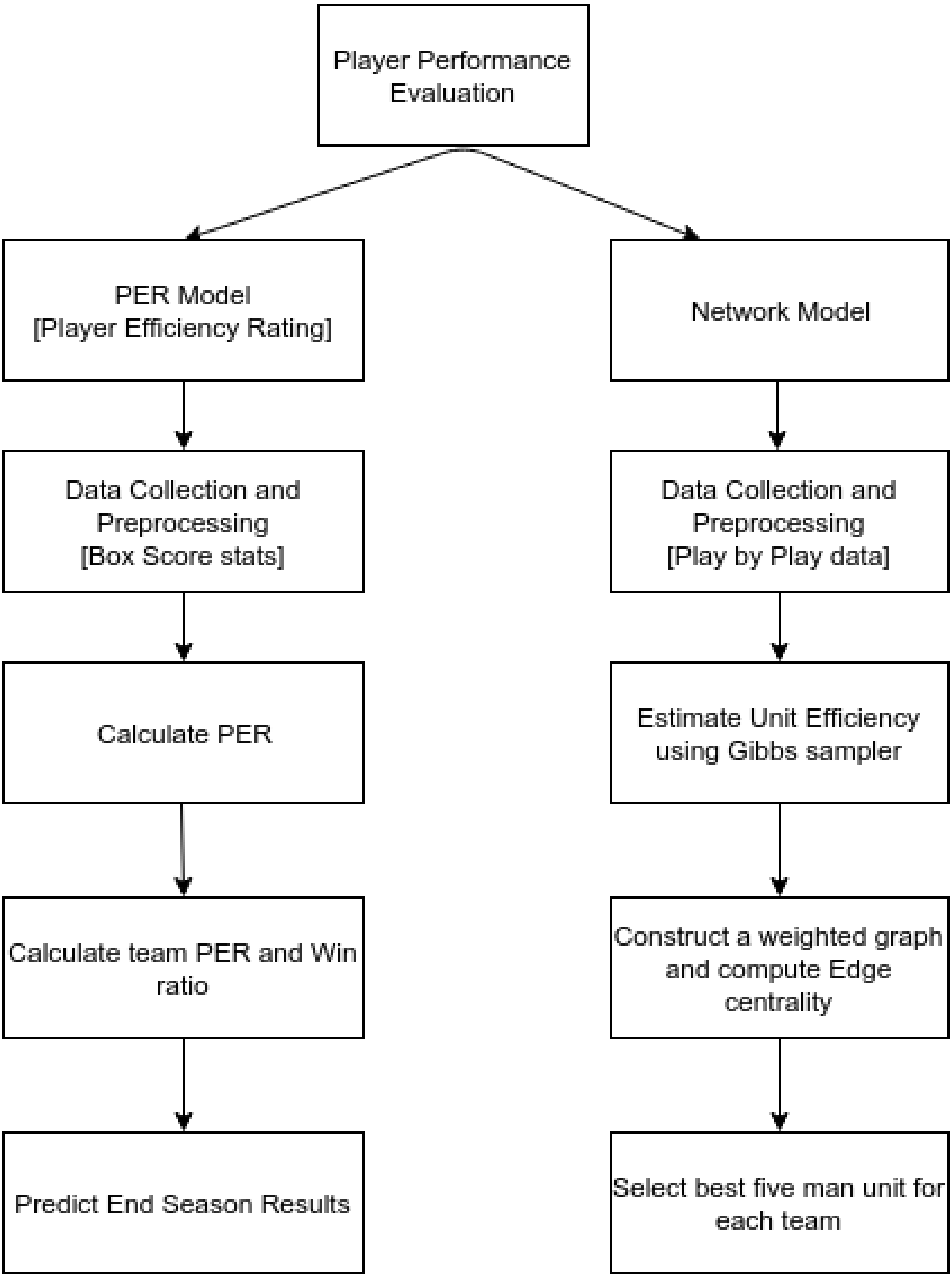
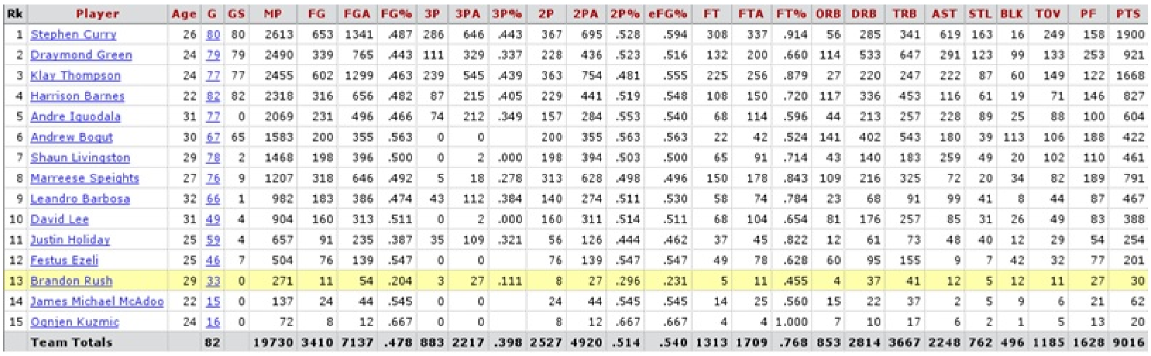

Player Performance Evaluation
PER Model [Player Efficiency Rating]
Network Model
Data Collection and Preprocessing [Box Score stats]
Data Collection and Preprocessing [Play by Play data]
Calculate PER
Estimate Unit Efficiency using Gibbs sampler
Calculate team PER and Win ratio
Construct a weighted graph and compute Edge centrality
Predict End Season Results
Select best five man unit for each team

Figure 3.1: Workflow of the Entire Project

the statistics that were too far back in time might not be as relevant as using more recent data when it comes to developing a model to predict teams performance in the near future. In consequence, we will only use the data from last 20 seasons in this paper, with an emphasis on the data from more recent seasons. The data available at basketball-reference.com is as shown in the figure 3.2:



Rk	Player	Age	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
1	Stephen Curry	26	80	80	2613	653	1341	.487	286	646	.443	367	695	.528	.594	308	337	.914	56	285	341	619	163	16	249	158	1900
2	Draymond Green	24	79	79	2490	339	765	.443	111	329	.337	228	436	.523	.516	132	200	.660	114	533	647	291	123	99	133	253	921
3	Klay Thompson	24	77	77	2455	602	1299	.463	239	545	.439	363	754	.481	.555	225	256	.879	27	220	247	222	87	60	149	122	1668
4	Harrison Barnes	22	82	82	2318	316	656	.482	87	215	.405	229	441	.519	.548	108	150	.720	117	336	453	116	61	19	71	146	827
5	Andre Iguodala	31	77	0	2069	231	496	.466	74	212	.349	157	284	.553	.540	68	114	.596	44	213	257	228	89	25	88	100	604
6	Andrew Bogut	30	67	65	1583	200	355	.563	0	0		200	355	.563	.563	22	42	.524	141	402	543	180	39	113	106	188	422
7	Shaun Livingston	29	78	2	1468	198	396	.500	0	2	.000	198	394	.503	.500	65	91	.714	43	140	183	259	49	20	102	110	461
8	Marreese Speights	27	76	9	1207	318	646	.492	5	18	.278	313	628	.498	.496	150	178	.843	109	216	325	72	20	34	82	189	791
9	Leandro Barbosa	32	66	1	982	183	386	.474	43	112	.384	140	274	.511	.530	58	74	.784	23	68	91	99	41	8	44	87	467
10	David Lee	31	49	4	904	160	313	.511	0	2	.000	160	311	.514	.511	68	104	.654	81	176	257	85	31	26	49	83	388
11	Justin Holiday	25	59	4	657	91	235	.387	35	109	.321	56	126	.444	.462	37	45	.822	12	61	73	48	40	12	29	54	254
12	Festus Ezeli	25	46	7	504	76	139	.547	0	0		76	139	.547	.547	49	78	.628	60	95	155	9	7	42	32	77	201
13	Brandon Rush	29	33	0	271	11	54	.204	3	27	.111	8	27	.296	.231	5	11	.455	4	37	41	12	5	12	11	27	30
14	James Michael McAdoo	22	15	0	137	24	44	.545	0	0		24	44	.545	.545	14	25	.560	15	22	37	2	5	9	6	21	62
15	Ognjen Kuzmic	24	16	0	72	8	12	.667	0	0		8	12	.667	.667	4	4	1.000	7	10	17	6	2	1	5	13	20
	Team Totals		82		19730	3410	7137	.478	883	2217	.398	2527	4920	.514	.540	1313	1709	.768	853	2814	3667	2248	762	496	1185	1628	9016

Figure 3.2: Box Scores of each player of Golden State Warriors.

Each Team and its Players have many attributes. All the attributes are numerical. The attributes are:

1. 2P: 2-Point Field Goals

2. 2P percent: 2-Point Field Goal Percentage; the formula is 2P / 2PA.

3. 2PA: 2-Point Field Goal Attempts

4. 3P: 3-Point Field Goals (available since the 1979-80 season in the NBA)

5. 3P percent: 3-Point Field Goal Percentage (available since the 1979-80 season in the NBA); the formula is 3P / 3PA.

6. 3PA: 3-Point Field Goal Attempts (available since the 1979-80 season in the NBA)

7. AST: Assists

8. BLK: Blocks (available since the 1973-74 season in the NBA)

9. DRB: Defensive Rebounds (available since the 1973-74 season in the NBA)

10. FG: Field Goals (includes both 2-point field goals and 3-point field goals)

11. FGA: Field Goal Attempts (includes both 2-point field goal attempts and 3-point field goal attempts)

12. FT: Free Throws
13. FTA: Free Throw Attempts
14. MP: Minutes Played (available since the 1951-52 season)
15. ORB: Offensive Rebounds (available since the 1973-74 season in the NBA)
16. PF: Personal Fouls
17. PTS: Points
18. STL: Steals (available since the 1973-74 season in the NBA)
19. TOV: Turnovers (available since the 1977-78 season in the NBA)
20. TRB: Total Rebounds (available since the 1950-51 season)

Data for 20 seasons was collected for all the teams and players in the form of comma-separated-value (csv) files.

Once the data collection is ready, we will use Hollingers PER (Player efficiency Rating) along with other basic statistics to value their strengths. PER is one of the most powerful and widely recognized metrics, which is derived from basic player statistics to gauge a players strength and it is both a pace-adjusted and per-minute measure.

3.1.2 Calculation of PER

PER is a metric that aims to measure a players effectiveness with a single number. It takes into account almost all statistics kept by the NBA, and weighs the players production by minutes played per game, and number of team possessions per game. The PER sums up all a player's positive accomplishments, subtracts the negative accomplishments, and returns a per-minute rating of a player's performance.

Unadjusted PER(uPER) : (pace of the team has not yet been taken into account)

$$\begin{split}uPER = (1/MP) * [3P + (2/3) * AST + (2 - factor * (team\_AST/ \\ team\_FG)) * FG + (FT * 0.5 * (1 + (1 - (team\_{A}ST/team\_FG)) + \\ (2/3) * (team\_AST/team\_FG))) - VOP * TOV - VOP * DRB\% * \\ (FGA - FG) - VOP * 0.44 * (0.44 + (0.56 * DRB\%)) * (FTA - FT) + \\ VOP * (1DR\%) * (TRB - ORB) + VOP * DRB\% * ORB + VOP * \\ STL + VOP * DRB\% * BLK - PF * ((lg\_FT/lg\_PF) - 0.44 * (lg\_FTA/lg\_PF) * VOP)]\end{split} \tag{3.1}$$

where

a) $factor = (2/3) - (0.5 * (lg\_AST/lg\_FG))/(2 * (lg\_FG/lg\_FT))$

b) $VOP = lg\_PTS/(lg\_FGA - lg\_ORB + lg\_TOV + 0.44 * lg\_FTA)$

c) $DRB\% = (lg\_TRB - lg\_ORB)/lg\_TRB$

The pace adjustment is:

$$pace\_adjustment = lg\_Pace/team\_Pace \tag{3.2}$$

Adjusted PER(aPER):

$$aPER = (pace\_adjustment) * uPER \tag{3.3}$$

Benefits of PER : It is a huge step up from looking at standard boxscore statistics. It is much more detailed and accurate than anything one can do with raw statistical totals or per-game numbers.

Negatives of PER: One major weakness in the original PER concept is lack of consideration for defense. The formula itself largely measures offensive performance. Though there are blocked shots and steals, the formula doesn't account in any way for players who play great individual or team defense.

3.1.3 Prediction of End Season Results

For each season two major variables are calculated. First is teams win ratio, which is equal to the percentage of games that a particular team won during the season (multiplied by 100 for clarity). Win ratio, from our perspective, is more informative than teams ranking at the end of the season because win ratio is more quantifiable and can show by how much margin one team is better than another for a given season.

$$Teams\_win\_ratio = (Wins/TotalGamesPlayed) * 100 \tag{3.4}$$

The other variable is team PER, which is calculated by :

1. Sorting players on the roster of a specific team by minutes they played for that team over the entire season.
2. Then multiplying individual PER by the minutes they played.
3. Summing the product (of individual PER * minutes played) for the first twelve players (by minutes played) on a given team.

PER is a per-minute statistic, thus multiplying a players PER by his minutes played can serve as an approximation of the total contribution he made towards his team over the entire season. Some teams have made changes in their roster over the season through trade, due to injuries, or at coaching staffs discretion. However, the first 12 players on each teams roster are relatively stable, which indicates that those players are likely to be in the usual rotation of their teams lineup. The huge benefit of calculating team PER the way we did is that the formula automatically puts more weight on players with higher PER because high-PER individuals are, by the way the metric is designed, more capable and are more likely to play a lot more than the players with lower PER.

First, we will establish that there is a correlation between the team's Win Ratio and Team PER. We will use a linear regression model to model this and evaluate the goodness of the fit for each season. Evaluation of the fit is done by using R-sqaured and Adjusted R-Squared values. In statistics, the coefficient of determination R2 is the proportion of variability in a data set that is accounted for by a statistical model. In this definition, the term ”variability” is defined as the sum of squares. Adjusted R-square is a modification of R-square that adjusts for the number of terms in a model. R-square always increases when a new term is added to a model, but adjusted R-square increases only if the new term improves the model more than would be expected by chance.

$R^2 = 1 - \frac{SS_E}{SS_T}$ where, $SS_T = \sum_i (y_i - \overline{y})^2$, $SS_E = \sum_i (y_i - \widehat{y_i})^2$

We will then predict the Win Ratio's for 2015-16 Season using the data from the previous seasons in the same manner.

3.2 Network Model

There are four main steps involved in building a network model. They are:

3.2.0.1 Data Collection and Preprocessing

We use the play-by-play data of each game in this analysis. Play-by-play provides a transcript of the game in a format of individual events. A typical play-by-play data has the following information:

1. The time of the possession
2. The home and away units on the field at that time
3. The player who initiated the possession (in the case of a steal or defensive rebound)
4. The opposing player who initiated the possession (in case of a missed shot or turnover)

We choose to use four seasons of play-by-play data, taken from [9]: 06-07, 07-08, 08-09, and 09-10. We analyze these data to determine for each possession, the two five-man units on court, which unit was home (or away), which unit had possession of the ball, and the number of points scored.

3.2.1 Bayesian Hierarchical Model

Let y_{ij} denote the number of points scored (or allowed, when analyzing defense) by unit i for possession j after adjusting for home court effects. The data likelihood in our model follows

$$y_{ij} \sim Normal(\theta_i, \sigma^2) \tag{3.5}$$

, where σ^2 is the shared variance for each observation and θ_i is the mean efficiency for unit j. We place a prior density on each θ_i of

$$\theta_i \sim Normal(\mu, \tau^2) \tag{3.6}$$

, where μ represents the league-mean efficiency and τ^2 is the corresponding variance. To generate posterior estimates for the parameters of interest (i.e. the θ_i s), we implement a Gibbs sampler for the Bayesian Normal Hierarchical Model.

An example for Bayesian Hierarchical Model - A teacher wants to estimate how well a male student did in his SAT. He uses information on the students high school grades and his current grade point average (GPA) to come up with an estimate. His current GPA, denoted by Y, has a likelihood given by some probability function with parameter θ, i.e.$Y \mid \theta \sim P(Y \mid \theta)$. This parameter θ is the SAT score of the student. The SAT score is viewed as a sample coming from a common population distribution indexed by another parameter ϕ, which is the high school grade of the student. That is, $\theta \mid \phi \sim P(\theta \mid \phi)$. Moreover, the hyperparameter ϕ follows its own distribution given by $P(\phi)$, a hyperprior. To solve for the SAT score given information on the GPA,

$P(\theta, \phi \mid Y) \propto P(Y \mid \theta, \phi)P(\theta, \phi)$

$P(\theta, \phi \mid Y) \propto P(Y \mid \theta)P(\theta \mid \phi)P(\phi)$

All information in the problem will be used to solve for the posterior distribution. Instead of solving only using the prior distribution and the likelihood function, the use of hyperpriors gives more information to make more accurate beliefs in the behavior of a parameter.

3.2.2 Gibbs sampler for the Bayesian Normal Hierarchical Model

We use a Gibbs sampler to estimate the full posterior distributions of all unknown parameters. We follow the implementation in [2]. We assume uniform priors for all remaining parameters, $(\mu, log(\sigma), \tau)$.

We obtain samples for $\theta_i, \mu, \sigma^2, \tau^2$ from the posterior distribution by iteratively sampling from:

1. $p(\theta_i|\mu, \sigma, \tau, y)$, for all units i,

2. $p(\mu|\theta, \tau)$,

3. $p(\sigma^2|\theta, y)$,

4. and $p(\tau^2|\theta, \mu)$.

3.2.3 Constructing a Weighted graph

1. Construction of a bipartite graph with five man unit on one side and players on the other. This bimodal network is represented by incidence matrix W , where the rows are units and the columns are players.

2. A player P_i is adjacent to a unit U_j (i.e. $w_{ij} = 0$) if he has played in that unit. This edge w_{ij} is weighted by the efficiency of unit U_j.

3. Then, we project this bimodal network onto a unimodal network by computing $A = W^T W$.

4. In this final network of just players, the weight of an edge between two player nodes is the sum of the squares of their shared units efficiency scores.

3.2.4 Computing Edge Centrality

We use betweeness centrality with random restart to determine the centrality (or importance) of a edges in a network. Centrality scores of edges correspond to the importance of how two teammates perform together. In this way, we can evaluate the performance on pairs of teammates, which could highlight players who while not successful individually, work great as a pair.

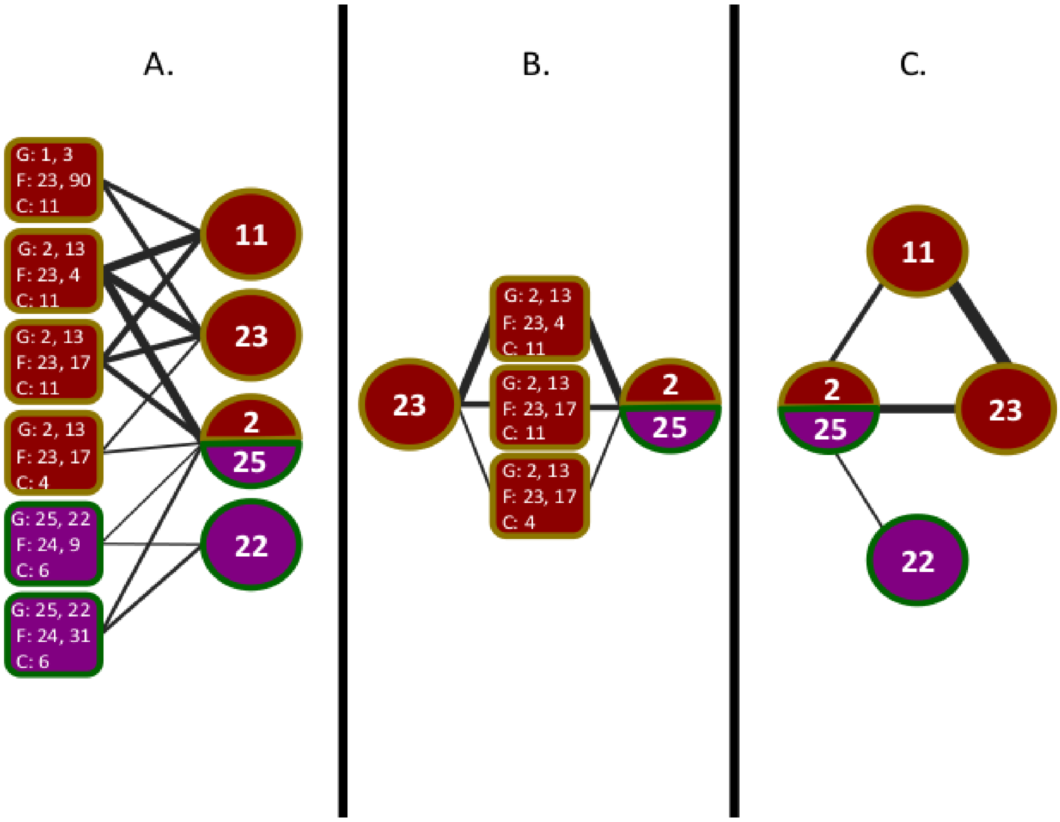


Figure 3.3: An example of construction of 4 node player network

3.2.5 Selecting the best five man unit for each team

For each team, we find the best five man unit among all the possible five man units. This is calculated by selecting the maximum among all the unit efficiencies for that particular team.

Chapter 4

Results and Observations

4.1 PER Model

As given in the methodology, we calculated the PER of each player for each of the 20 seasons. Table 4.1 and 4.2 the top 5 players with a high PER who have played more than 25 games in a season.

Table 4.1: Top Players based on PER for 2014-15 Season.

Player	Team	PER
Anthony Davis	NOP	30.8
Russell Westbrook	OKC	29.1
Stephen Curry	GSW	28.0
Kevin Durant	OKC	27.6
James Harden	HOU	26.7

Even with two players in the top five, OKC - Oklahoma City Thunder, did not win the division league and qualify for the first round. Also, no player from Cleveland Cavaliers is in the top even though they made it to the finals. Even though New Orleans Pelicans had the player with the highest PER in the team, they failed to make to the semifinals.

From these two seasons it can be observed that Kevin Durant and Anthony Davis are in the top five. So the teams can make an effort to buy or trade these players.

Now, we will try to find the correlation between Team Win Ratio and Team PER

Table 4.2: Top Players based on PER for 2013-14 Season.

Player	Team	PER
Kevin Durant	OKC	29.8
LeBron James	MIA	29.3
Kevin Love	MIN	26.9
Anthony Davis	NOP	26.5
DeMarcus Cousins	SAC	26.1

by using scatterplots.

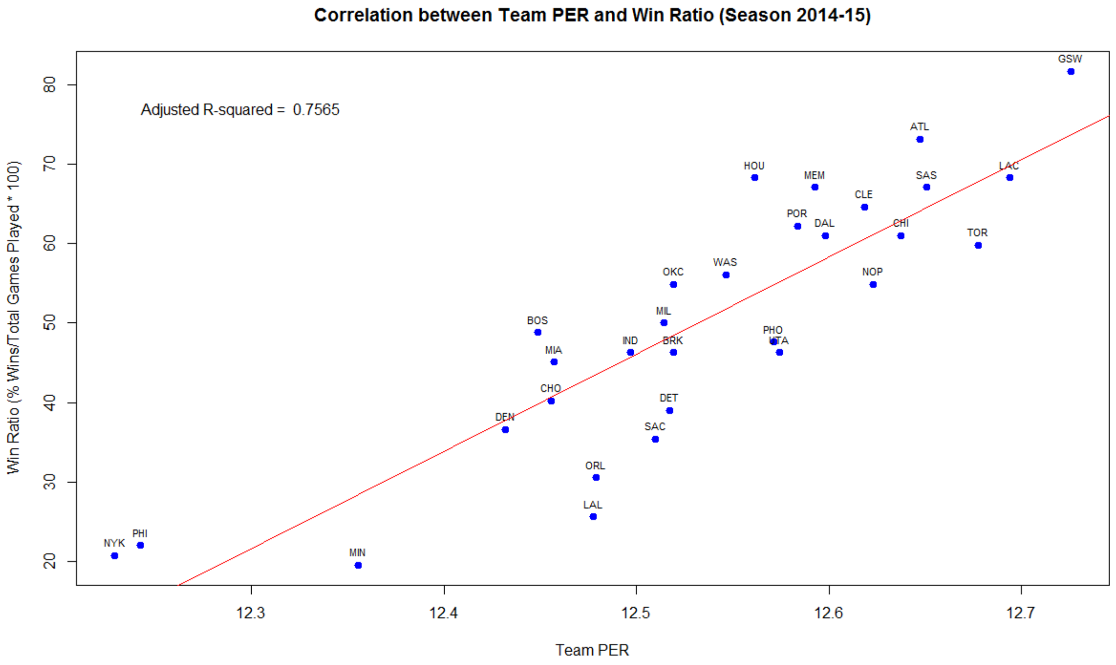


Figure 4.1: Scatterplot of Team Win Ratio versus Team PER for 2014-2015 Season

We can clearly see a linear trend from the scatterplot of figure 4.1, which indicates that the teams with higher PER value are indeed more likely to perform better over the season. We then look at the summary statistics of the linear model for further information:

```
Call:
lm(formula = as.numeric(team_PER) ~ winR)
Residuals:
Min 1Q Median 3Q Max
```

```
-0.119066 -0.033274 -0.004973 0.033922 0.095515
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.232e+01 2.959e-02 416.428 < 2e-16 ***
winR 4.661e-03 5.653e-04 8.244 5.68e-09 ***
---
Signif. codes: 0  ***   0.001   **   0.01   *   0.05   .   0.1       1
Residual standard error: 0.04794 on 28 degrees of freedom
Multiple R-squared: 0.7632, Adjusted R-squared: 0.7565
F-statistic: 67.97 on 1 and 28 DF, p-value: 5.677e-09
```

We can see from the summary statistics that the linear model yields a strong R2 result with an adjusted R2 of 0.6978. GSW(Golden State Warriors) is the team with the highest win ratio and this is substantiated by the fact that they were the Champions.

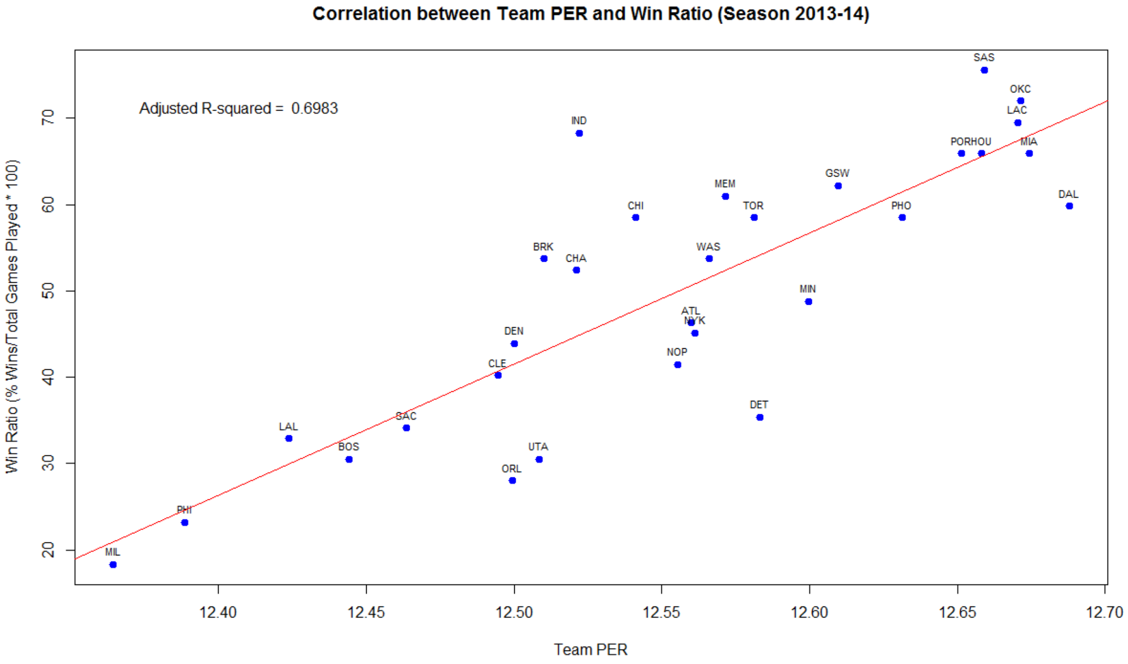


Figure 4.2: Scatterplot of Team Win Ratio versus Team PER for 2013-2014 Season

We can clearly see a linear trend from the scatterplot of figure 4.2, which indicates that the teams with higher PER value are indeed more likely to perform better over the season. We then look at the summary statistics of the linear model for further information:

```
Call:
lm(formula = as.numeric(team_PER) ~ winR)
Residuals:
Min 1Q Median 3Q Max
-0.119066 -0.033274 -0.004973 0.033922 0.095515
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.232e+01 2.959e-02 416.428 < 2e-16 ***
winR 4.661e-03 5.653e-04 8.244 5.68e-09 ***
---
Signif. codes: 0  ***  0.001  **  0.01  *  0.05  .  0.1     1
Residual standard error: 0.04794 on 28 degrees of freedom
Multiple R-squared: 0.7082, Adjusted R-squared: 0.6978
F-statistic: 67.97 on 1 and 28 DF, p-value: 5.677e-09
```

We can see from the summary statistics that the linear model yields a strong R2 result with an adjusted R2 of 0.6978. The significant F-statistics from the Wald test further confirms that the linear model decently fits the data and could potentially be used to make future forecasts. But IND and DET deviate considerably from the line and are anomalies.

Like the scatterplot for 2013-2014 season, the linear trend in the scatterplot of figure 4.3 is also very apparent. In fact, the linearity is more pronounced in this plot than the previous one partially because the anomalies presented in this graph are less extreme than those from last seasons graph. MEM (Memphis Grizzlies) and NOH (New Orleans Hornets), deviating considerably from the regression line, seem to be the ones that do not completely agree with the model this season. Next, we look at the statistical summary of the model:

```
Call:
lm(formula = as.numeric(team_PER) ~ winR)
Residuals:
Min 1Q Median 3Q Max
-0.07810 -0.02433 0.00559 0.02484 0.06892
```

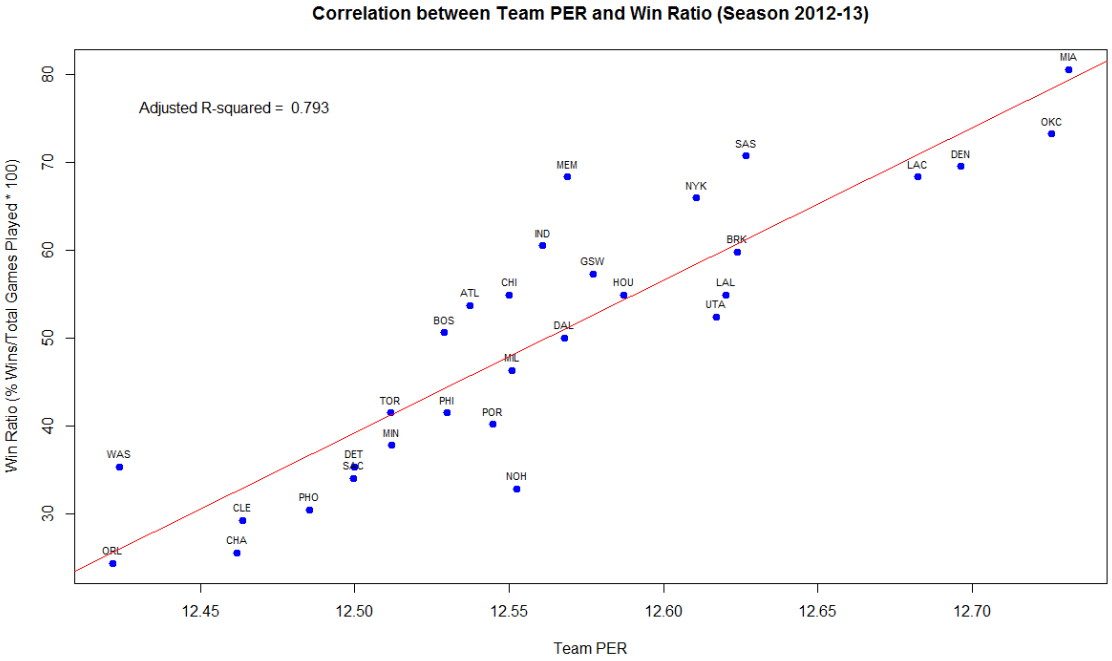


Figure 4.3: Scatterplot of Team Win Ratio versus Team PER for 2012-2013 Season

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.233e+01 2.278e-02 541.2 < 2e-16 ***
winR 4.620e-03 4.358e-04 10.6 2.62e-11 ***
---
Signif. codes: 0  ***   0.001   **   0.01   *   0.05   .   0.1        1
Residual standard error: 0.03642 on 28 degrees of freedom
Multiple R-squared: 0.8005, Adjusted R-squared: 0.7934
F-statistic: 112.3 on 1 and 28 DF, p-value: 2.621e-11
```

The statistical summary of the linear model yields an even stronger R2 value than 2013- 2014 season, with an adjusted R2 of 0.7934. Meanwhile, the F-statistics from the Wald test further confirms that the result is statistically significant.

Similar to the previous two scatterplots, the plot in figure 4.4 also demonstrates clear linearity between team PER and teams win ratio. There are some notable anomalies such as BOS (Boston Celtics) which has a higher win ratio than its team PER would suggest. The R-squared value including BOS is 0.7568, indicating a decent fit of the model. However, if we exclude BOS, the R-squared value will increase

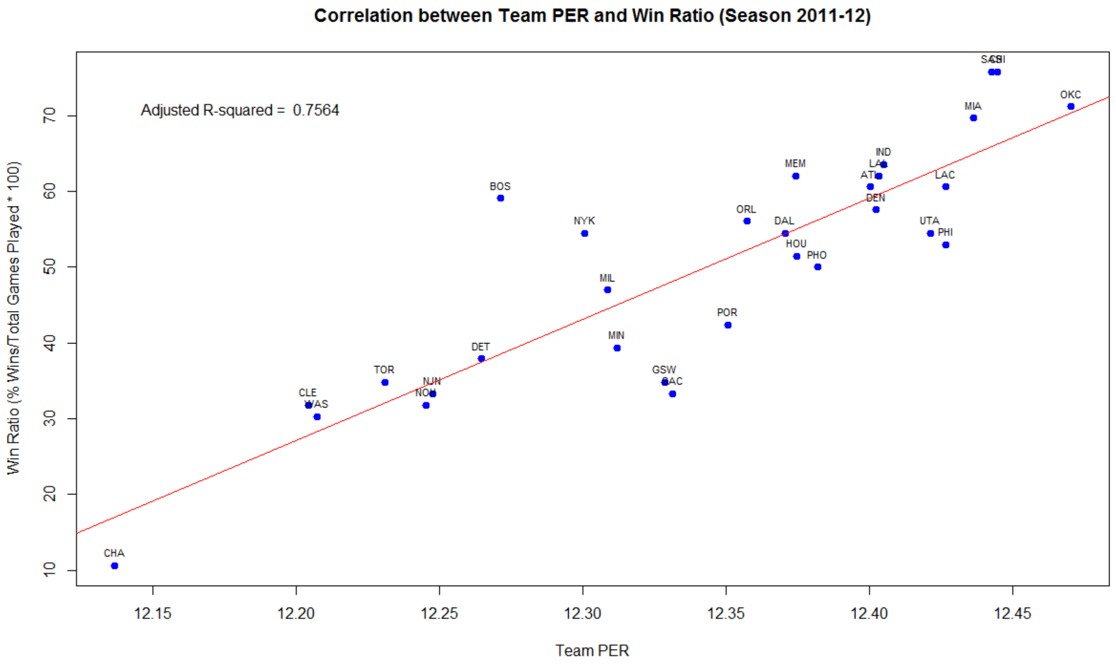


Figure 4.4: Scatterplot of Team Win Ratio versus Team PER for 2011-2012 Season

to 0.8197 (or an improvement of 8.3would be further improved if we take into account other anomalies such as GSW (Golden State Warriors) and SAC (Sacramento Kings) which have lower win ratio than their team PER would indicate. In general, the linear regression line seems to be a good fit graphically. We further confirm the goodness of fit by looking at its statistical summary. Results with BOS included:

```
Call:
lm(formula = as.numeric(team_PER) ~ winR)
Residuals:
Min 1Q Median 3Q Max
-0.114685 -0.020992 -0.001734 0.025915 0.069684
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.210e+01 2.619e-02 462.204 < 2e-16 ***
winR 4.783e-03 5.007e-04 9.552 2.62e-10 ***
---
Signif. codes: 0  ***   0.001  **   0.01  *   0.05  .   0.1       1
Residual standard error: 0.04207 on 28 degrees of freedom
Multiple R-squared: 0.7652, Adjusted R-squared: 0.7568
```

```
F-statistic: 91.24 on 1 and 28 DF, p-value: 2.621e-10
```

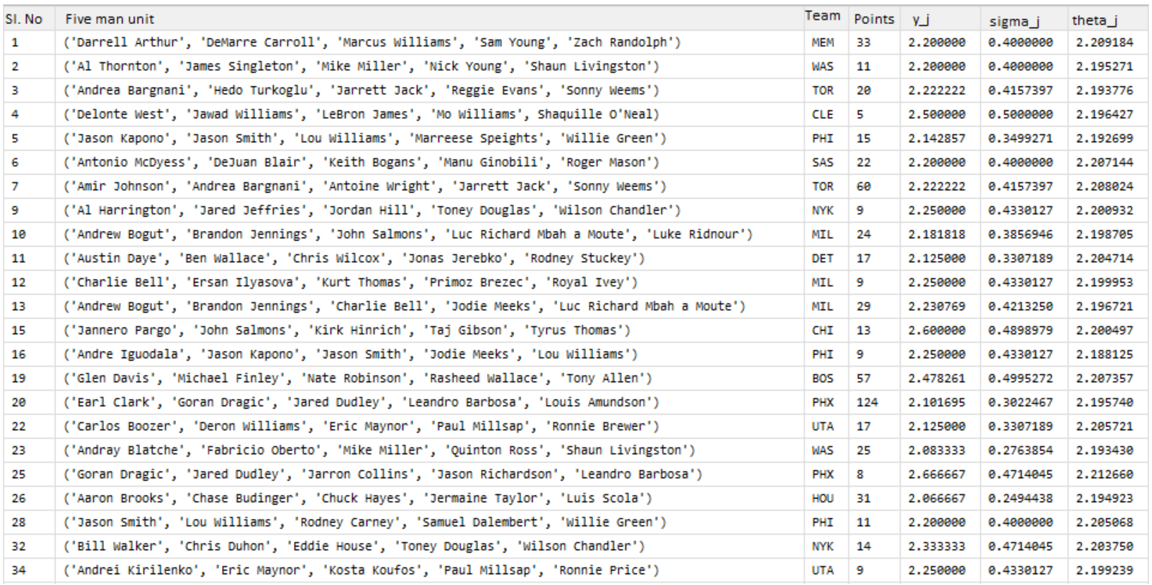
Results with BOS excluded:

```
lm(formula = as.numeric(team_PER[-23]) ~ winR[-23])
Residuals:
Min 1Q Median 3Q Max
-0.068519 -0.023142 -0.007556 0.020585 0.066959
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.210e+01 2.269e-02 533.25 < 2e-16 ***
winR[-23] 4.937e-03 4.359e-04 11.33 9.21e-12 ***
---
Signif. codes: 0  ***  0.001  **  0.01  *  0.05  .  0.1     1
Residual standard error: 0.03641 on 27 degrees of freedom
Multiple R-squared: 0.8261, Adjusted R-squared: 0.8197
F-statistic: 128.3 on 1 and 27 DF, p-value: 9.21e-12
```

As one can see from the statistical summary of the model, it produces significant coefficients as well as F-statistics. The model confirms that, to a large extent, a teams win ratio is linearly correlated to team PER value. The results without BOS indicate that the model sometimes fails to work with teams with high PER that under perform or teams with low PER that overachieve.

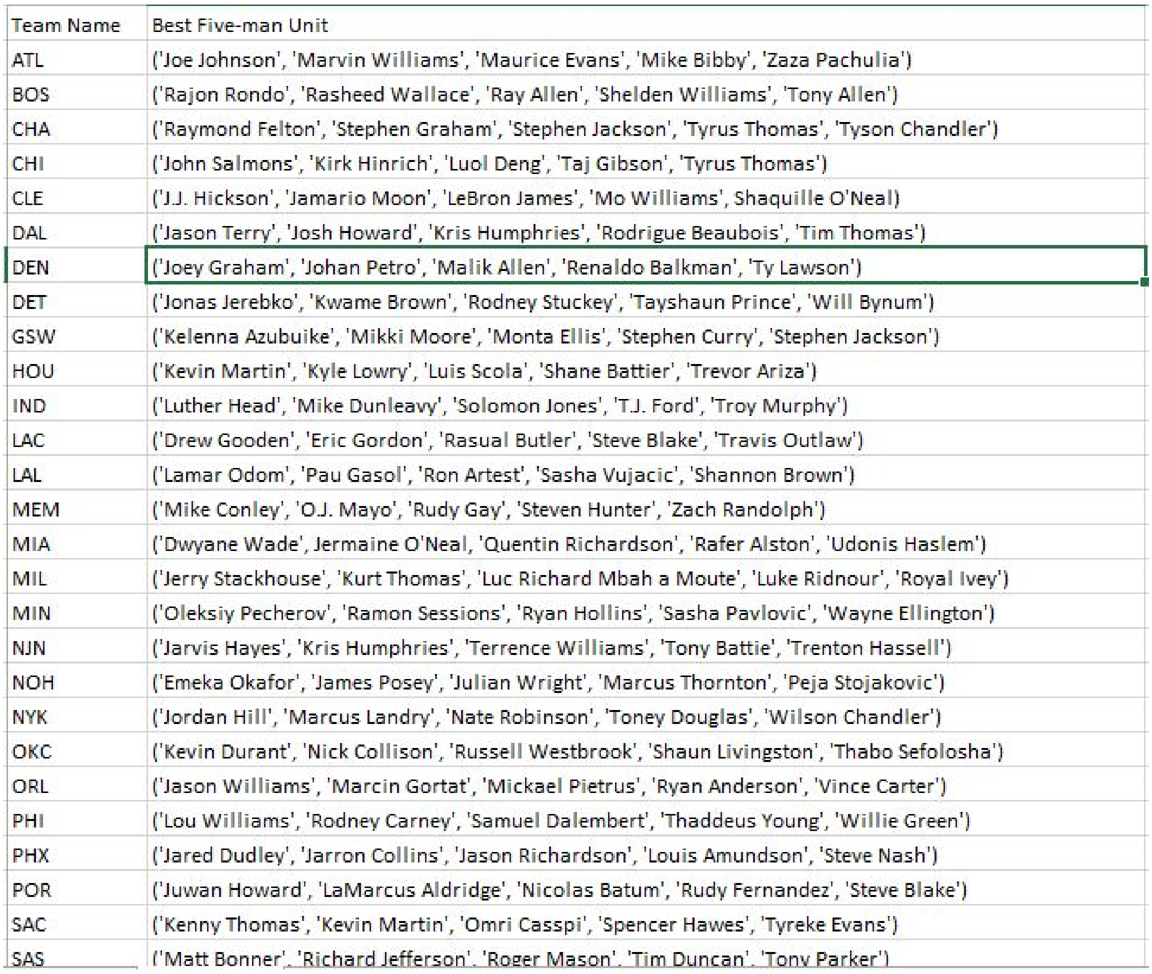
4.2 Network Model

Figure 4.5 shows the output of the Bayesian Normal Hierarchical Model. It shows all the five-man units with their respective team names and actual efficiencies as estimated by the Gibbs Sampler. Figure 4.6 shows the best five-man units for each team.



Sl. No	Five man unit	Team	Points	y_j	sigma_j	theta_j
1	('Darrell Arthur', 'DeMarre Carroll', 'Marcus Williams', 'Sam Young', 'Zach Randolph')	MEM	33	2.200000	0.4000000	2.209184
2	('Al Thornton', 'James Singleton', 'Mike Miller', 'Nick Young', 'Shaun Livingston')	WAS	11	2.200000	0.4000000	2.195271
3	('Andrea Bargnani', 'Hedo Turkoglu', 'Jarrett Jack', 'Reggie Evans', 'Sonny Weems')	TOR	20	2.222222	0.4157397	2.193776
4	('Delonte West', 'Jawad Williams', 'LeBron James', 'Mo Williams', Shaquille O'Neal)	CLE	5	2.500000	0.5000000	2.196427
5	('Jason Kapono', 'Jason Smith', 'Lou Williams', 'Marreese Speights', 'Willie Green')	PHI	15	2.142857	0.3499271	2.192699
6	('Antonio McDyess', 'DeJuan Blair', 'Keith Bogans', 'Manu Ginobili', 'Roger Mason')	SAS	22	2.200000	0.4000000	2.207144
7	('Amir Johnson', 'Andrea Bargnani', 'Antoine Wright', 'Jarrett Jack', 'Sonny Weems')	TOR	60	2.222222	0.4157397	2.208024
9	('Al Harrington', 'Jared Jeffries', 'Jordan Hill', 'Toney Douglas', 'Wilson Chandler')	NYK	9	2.250000	0.4330127	2.200932
10	('Andrew Bogut', 'Brandon Jennings', 'John Salmons', 'Luc Richard Mbah a Moute', 'Luke Ridnour')	MIL	24	2.181818	0.3856946	2.198705
11	('Austin Daye', 'Ben Wallace', 'Chris Wilcox', 'Jonas Jerebko', 'Rodney Stuckey')	DET	17	2.125000	0.3307189	2.204714
12	('Charlie Bell', 'Ersan Ilyasova', 'Kurt Thomas', 'Primoz Brezec', 'Royal Ivey')	MIL	9	2.250000	0.4330127	2.199953
13	('Andrew Bogut', 'Brandon Jennings', 'Charlie Bell', 'Jodie Meeks', 'Luc Richard Mbah a Moute')	MIL	29	2.230769	0.4213250	2.196721
15	('Jannero Pargo', 'John Salmons', 'Kirk Hinrich', 'Taj Gibson', 'Tyrus Thomas')	CHI	13	2.600000	0.4898979	2.200497
16	('Andre Iguodala', 'Jason Kapono', 'Jason Smith', 'Jodie Meeks', 'Lou Williams')	PHI	9	2.250000	0.4330127	2.188125
19	('Glen Davis', 'Michael Finley', 'Nate Robinson', 'Rasheed Wallace', 'Tony Allen')	BOS	57	2.478261	0.4995272	2.207357
20	('Earl Clark', 'Goran Dragic', 'Jared Dudley', 'Leandro Barbosa', 'Louis Amundson')	PHX	124	2.101695	0.3022467	2.195740
22	('Carlos Boozer', 'Deron Williams', 'Eric Maynor', 'Paul Millsap', 'Ronnie Brewer')	UTA	17	2.125000	0.3307189	2.205721
23	('Andray Blatche', 'Fabricio Oberto', 'Mike Miller', 'Quinton Ross', 'Shaun Livingston')	WAS	25	2.083333	0.2763854	2.193430
25	('Goran Dragic', 'Jared Dudley', 'Jarron Collins', 'Jason Richardson', 'Leandro Barbosa')	PHX	8	2.666667	0.4714045	2.212660
26	('Aaron Brooks', 'Chase Budinger', 'Chuck Hayes', 'Jermaine Taylor', 'Luis Scola')	HOU	31	2.066667	0.2494438	2.194923
28	('Jason Smith', 'Lou Williams', 'Rodney Carney', 'Samuel Dalembert', 'Willie Green')	PHI	11	2.200000	0.4000000	2.205068
32	('Bill Walker', 'Chris Duhon', 'Eddie House', 'Toney Douglas', 'Wilson Chandler')	NYK	14	2.333333	0.4714045	2.203750
34	('Andrei Kirilenko', 'Eric Maynor', 'Kosta Koufos', 'Paul Millsap', 'Ronnie Price')	UTA	9	2.250000	0.4330127	2.199239

Figure 4.5: Output of Bayesian Normal Model



Team Name	Best Five-man Unit
ATL	('Joe Johnson', 'Marvin Williams', 'Maurice Evans', 'Mike Bibby', 'Zaza Pachulia')
BOS	('Rajon Rondo', 'Rasheed Wallace', 'Ray Allen', 'Shelden Williams', 'Tony Allen')
CHA	('Raymond Felton', 'Stephen Graham', 'Stephen Jackson', 'Tyrus Thomas', 'Tyson Chandler')
CHI	('John Salmons', 'Kirk Hinrich', 'Luol Deng', 'Taj Gibson', 'Tyrus Thomas')
CLE	('J.J. Hickson', 'Jamario Moon', 'LeBron James', 'Mo Williams', Shaquille O'Neal)
DAL	('Jason Terry', 'Josh Howard', 'Kris Humphries', 'Rodrigue Beaubois', 'Tim Thomas')
DEN	('Joey Graham', 'Johan Petro', 'Malik Allen', 'Renaldo Balkman', 'Ty Lawson')
DET	('Jonas Jerebko', 'Kwame Brown', 'Rodney Stuckey', 'Tayshaun Prince', 'Will Bynum')
GSW	('Kelenna Azubuike', 'Mikki Moore', 'Monta Ellis', 'Stephen Curry', 'Stephen Jackson')
HOU	('Kevin Martin', 'Kyle Lowry', 'Luis Scola', 'Shane Battier', 'Trevor Ariza')
IND	('Luther Head', 'Mike Dunleavy', 'Solomon Jones', 'T.J. Ford', 'Troy Murphy')
LAC	('Drew Gooden', 'Eric Gordon', 'Rasual Butler', 'Steve Blake', 'Travis Outlaw')
LAL	('Lamar Odom', 'Pau Gasol', 'Ron Artest', 'Sasha Vujacic', 'Shannon Brown')
MEM	('Mike Conley', 'O.J. Mayo', 'Rudy Gay', 'Steven Hunter', 'Zach Randolph')
MIA	('Dwyane Wade', Jermaine O'Neal, 'Quentin Richardson', 'Rafer Alston', 'Udonis Haslem')
MIL	('Jerry Stackhouse', 'Kurt Thomas', 'Luc Richard Mbah a Moute', 'Luke Ridnour', 'Royal Ivey')
MIN	('Oleksiy Pecherov', 'Ramon Sessions', 'Ryan Hollins', 'Sasha Pavlovic', 'Wayne Ellington')
NJN	('Jarvis Hayes', 'Kris Humphries', 'Terrence Williams', 'Tony Battie', 'Trenton Hassell')
NOH	('Emeka Okafor', 'James Posey', 'Julian Wright', 'Marcus Thornton', 'Peja Stojakovic')
NYK	('Jordan Hill', 'Marcus Landry', 'Nate Robinson', 'Toney Douglas', 'Wilson Chandler')
OKC	('Kevin Durant', 'Nick Collison', 'Russell Westbrook', 'Shaun Livingston', 'Thabo Sefolosha')
ORL	('Jason Williams', 'Marcin Gortat', 'Mickael Pietrus', 'Ryan Anderson', 'Vince Carter')
PHI	('Lou Williams', 'Rodney Carney', 'Samuel Dalembert', 'Thaddeus Young', 'Willie Green')
PHX	('Jared Dudley', 'Jarron Collins', 'Jason Richardson', 'Louis Amundson', 'Steve Nash')
POR	('Juwan Howard', 'LaMarcus Aldridge', 'Nicolas Batum', 'Rudy Fernandez', 'Steve Blake')
SAC	('Kenny Thomas', 'Kevin Martin', 'Omri Casspi', 'Spencer Hawes', 'Tyreke Evans')
SAS	('Matt Bonner', 'Richard Jefferson', 'Roger Mason', 'Tim Duncan', 'Tony Parker')

Figure 4.6: The best five man units for each team

Chapter 5

Conclusion and Future Work

In this section we will make some concluding remarks about our project along with possible future works.

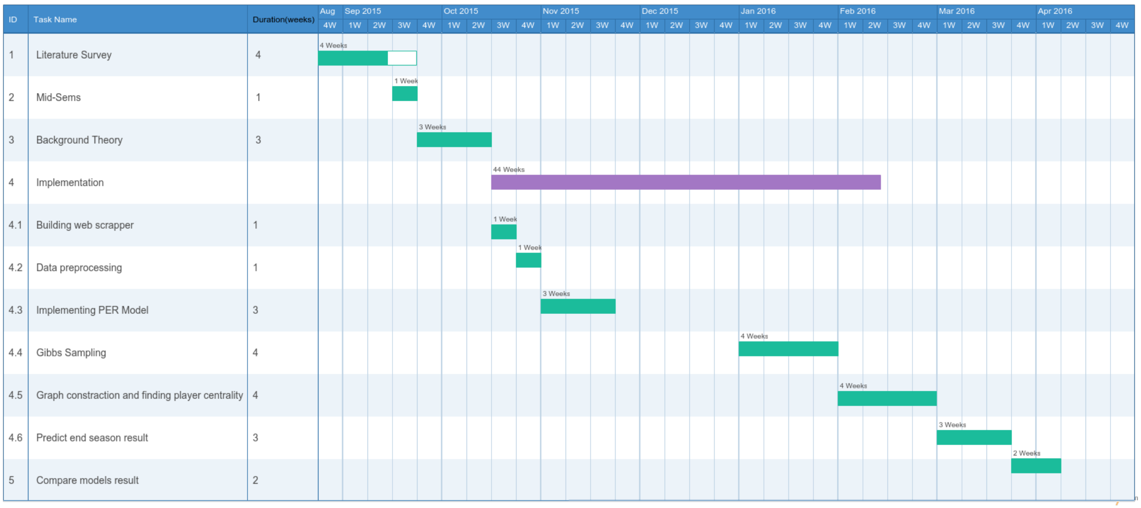
By using network models instead of box-score based models and plus-minus based models,a better rating can be given to a player that is influenced by other players in the team.This rating can be used to select more suitable players for the team and better trades can be made. We calculate centrality scores of edges, these scores correspond to the importance of how two teammates perform together. In this way, we can perform the same type of performance evaluation on pairs of teammates, which could highlight players who while not successful individually, work great as a pair.

One expansion of this algorithm is to use different measures of uint performance(e.g. rebounding and turnover rates) which we can use to gauge other aspects of a basketball player's skill set. Another interesting model extension is to calculate node centrality, instead of edge centrality to determine the centrality (or importance) of a player in a network. With this form of centrality and the design of our weighted network, a player is deemed important if he has important neighbors (i.e. played in a significant number of efficient units).

Appendix I

Project Time Line

A detailed time line of the project with start date and expected duration of each tasks has been shown in Figure I.1



ID	Task Name	Duration(weeks)
1	Literature Survey	4
2	Mid-Sems	1
3	Background Theory	3
4	Implementation	
4.1	Building web scrapper	1
4.2	Data preprocessing	1
4.3	Implementing PER Model	3
4.4	Gibbs Sampling	4
4.5	Graph constraction and finding player centrality	4
4.6	Predict end season result	3
5	Compare models result	2

Figure I.1: Project Timeline.

Appendix II

Plagiarism Report

A plagiarism report was conducted on this report and the result has been attached. An overall similarity score of 19% was observed.

a

ORIGINALITY REPORT

19%	18%	2%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.sloansportsconference.com Internet Source	9%
2	www.stat.berkeley.edu Internet Source	8%
3	www.sloansportsconference.com. Internet Source	1%
4	Dunbrack Jr, Roland. "Bayesian Statistics in Molecular and Structural Biology", Computational Biochemistry and Biophysics, 2001. Publication	1%
5	www.jamesmpiette.com Internet Source	<1%
6	www.scribd.com Internet Source	<1%
7	Kamath, S. Sowmya, and V.S. Ananthanarayana. "A bio-inspired, incremental clustering algorithm for semantics-based web service discovery", International Journal of Reasoning-based Intelligent Systems, 2015.	<1%

Publication

8	Ertug, G., and F. Castellucci. "Who shall get more? How intangible assets and aspiration levels affect the valuation of resource providers", Strategic Organization, 2014. Publication	<1%
9	Lund, Henrik. "Tool", Renewable Energy Systems, 2014. Publication	<1%
10	Schoen, Mary, Mitchell Small, and Jeanne VanBriesen. "Bayesian Load Duration Curves for Bacterial Total Maximum Daily Loads: Urban Case Study", World Environmental and Water Resources Congress 2009, 2009. Publication	<1%
11	LISS 2014, 2015. Publication	<1%

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		

References

[1] J. Piette, L. Pham, and S. Anand, "Evaluating basketball player performance via statistical network modeling," MIT Sloan Sports Analytics Conference, 2011.

[2] "How to calculate wins produced.." `http://wagesofwins.com/how-to-calculate-wins-produced/,`. Accessed: 2015-09-02.

[3] "Basketball-reference." `http://www.basketball-reference.com/`. Accessed: 2015-09-02.

[4] P. Fearnhead and B. M. Taylor, "On estimating the ability of nba players," *Journal of quantitative analysis in sports*, vol. 7, no. 3, 2011.

[5] S. Yang, "Predicting regular season results of nba teams based on regression analysis of common basketball statistics.," Master's thesis, University of California at Berkeley, 2015.

[6] C. Cao, "Sports data mining technology used in basketball outcome prediction," 2012.

[7] D. Miljkovic, L. Gajić, A. Kovacevic, and Z. Konjovic, "The use of data mining for basketball matches outcomes prediction," in *Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on*, pp. 309–312, IEEE, 2010.

[8] K. Puranmalka, "Modelling the nba to make better predictions," Master's thesis, Massachusetts Institute of Technology, 2013.