Can Complex Network Metrics Predict the Behavior of NBA Teams?

Pedro O.S. Vaz de Melo
Federal University of Minas Gerais
31270-901, Belo Horizonte
Minas Gerais, Brazil
olmo@dcc.ufmg.br

Virgilio A.F. Almeida
Federal University of Minas Gerais
31270-901, Belo Horizonte
Minas Gerais, Brazil
virgilio@dcc.ufmg.br

Antonio A.F. Loureiro
Federal University of Minas Gerais
31270-901, Belo Horizonte
Minas Gerais, Brazil
loureiro@dcc.ufmg.br

ABSTRACT

The United States National Basketball Association (NBA) is one of the most popular sports league in the world and is well known for moving a millionary betting market that uses the countless statistical data generated after each game to feed the wagers. This leads to the existence of a rich historical database that motivates us to discover implicit knowledge in it. In this paper, we use complex network statistics to analyze the NBA database in order to create models to represent the behavior of teams in the NBA. Results of complex network-based models are compared with box score statistics, such as points, rebounds and assists per game. We show the box score statistics play a significant role for only a small fraction of the players in the league. We then propose new models for predicting a team success based on complex network metrics, such as clustering coefficient and node degree. Complex network-based models present good results when compared to box score statistics, which underscore the importance of capturing network relationships in a community such as the NBA.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: database management—*Database Applications, Data mining*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Statistical computing*

General Terms

Theory

1. INTRODUCTION

The United States National Basketball Association (NBA) was founded in 1946 and since then is well known for its efficient organization and for its high level athletes. After each game played, a large amount of statistical data are generated describing the performance of each player who played in the match. These statistics are used in the United States to move a betting market estimated in tens of billions of dollars. In 2006, the Nevada State Gaming Control Board reported $2.4 billion in legal sports wager [10]. Meanwhile, in 1999, the National Gambling Impact Study Commission reported to Congress that more than $380 billion is illegally wagered on sports in the United States every year [10]. The generated statistics are used, for instance, by many Internet sites to aid gamblers, giving them more reliable predictions on the outcome of upcoming games.

The statistics are also used to characterize the performance of each player over time, dictating their salaries and the duration of their contracts. Kevin Garnett [18] averaged 22.4 Points Per Game (PPG), 12.8 Rebounds Per Game (RPG) and 4.1 Assists Per Game (APG) in the 2006/2007 season, making his salary to be the highest in the 2007/2008 season: $23.75 million. On the other hand, Anderson Varejão [18], who had 6.0 PPG, 6.0 RPG and 0.6 APG in the 2006/2007 season, asked in the following season for a $60 million contract for six years and had his request neglected. Robert Horry [18], who is at the 7th position in the rank of players who won more NBA championships with seven titles for three different teams, has career averages of 7.2 PPG, 4.9 RPG and 2.2 APG and, in the year 2007, of his last title, had a salary of $3.315 million. Two simple questions arise from these observations. First, would Anderson Varejão be overpaid in case his request were accepted? Second, is Robert Horry underpaid, once he wins a title for every team he played?

The first question was answered by Henry Abbot, a Senior Writer of ESPN.com, in his blog True Hoop [1]. He said PPG, RPG and APG only measure the actions of a player within a second or two when someone shoots the ball. The rest of the time, points and rebounds measure nothing. He also said, answering to the first question, that these statistics are against Anderson Varejão, who is one of the best players in the NBA in the adjusted plus/minus statistic. The plus/minus statistic keeps track of the net changes in score when a given player is either on or off the court, and it does not depend on to box scores, such as PPG, APG and RPG [2, 14]. This indicates that Anderson Varejão could have asked for a $60 million contract. Moreover, in the aid of Anderson Varejão, we point out that after he finally reached an agreement with the Cavaliers, the performance of the team went from 9 wins and 11 losses to 15 wins and 7 losses, with Anderson Varejão scoring 7.8 PPG, 8.5 RPG and 1.2 APG before his injury.

For the second question we could not find an answer. Robert Horry has played 14 seasons, averaging a title per

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

two years played and per team played. Is he a lucky guy who always play with the best ones or he really makes a difference? One thing we know for sure is, that simple statistics such as PPG, RPG and APG should not be the only metrics used to predict a player and team success. While the statistics are treated separately and the players are treated individually, little is known whether there is any relationship between them. We have seen in history, players with insignificant box scores statistics playing significant roles on a team success.

A possible way to study the collective behavior of social agents is to apply the theory of complex networks [19, 22]. A network is a set of vertices, sometimes called nodes, with connections between them, called edges. A complex network is a network characterized by a large number of vertices and edges that follow some pattern, like the formation of clusters or highly connected vertices, called hubs [4]. While in a simple network, with at most hundreds of vertices, the human eye is an analytic tool of remarkable power, in a complex network, this approach is useless. Thus, to study complex networks it is necessary to use statistical methods in a way to tell us how the network looks like.

The goal of this work is to model the NBA as a complex network and develop metrics that predict the behavior of NBA teams. The metrics should take into account the social and work relationship among players and teams and should also be able to predict a team success without relying on box score statistics. Before presenting the metrics, we show that the number of players who have made significant impact in the history of the NBA and in their teams is negligible if we draw our conclusions based only on box score statistics. Then, we study the characteristics of the NBA complex network in the direction of understanding how the relationships among players evolve over time. And then, finally, we present and compare the developed metrics that predict the success of NBA teams.

The rest of this paper is organized as follows. Section 2 presents the related work. In Section 3, we show that the box score statistics plays a significant role in only a small fraction of the players in the league. Section 4 describes the NBA as a complex network. The models we develop to predict team behavior are discussed and evaluated in Section 5. Finally, in Section 6, we present the conclusions and future work.

2. RELATED WORK

The growing interest in the study of complex networks has been credited to the availability of a large amount of real data and to the existence of interesting applications in several biological, sociological, technological and communications systems [21]. One of the most popular studies in this area was carried out by Milgram [17] and, from this work, the concepts of "six degrees of separation" and small-world have emerged. In other studies, well known complex networks were investigated, such as the Internet [13], World Wide Web [4], online social networking services [3], scientific collaboration networks [20], food webs [8], electric power grid [22], airline routes [5] and railways [13].

The analysis of the relationship that exists among players of a sports league from the point of view of complex networks is already present in the literature. Onody and de Castro [21] analyzed statistics of the editions from 1971 to 2002 of the Brazilian National Soccer Championship. From the complex network analysis, Onody and de Castro [21] found, among other interesting results, that the connectivity of the players has increased over the years while the clustering coefficient declined. The authors suggest that the possible causes for this phenomenon are the increase in the exodus of players to outside of Brazil, the increasing number of trades of players among national teams and, finally, the increase in the player's career time. Moreover, it was found that the assortativity degree is positive and also increases over time, indicating that exchanges between players are, in most cases, between teams of the same size.

Finally, the work of [7] presents an statistical analysis to quantify the predictability of all sports competitions in five major sports leagues in the United States and England. To characterize the predictability of games, the authors measure the "upset frequency" (i.e., the fraction of times the underdog wins). Basketball has a low upset frequency, which instigated us to look into the basketball league database to find out models to predict the team behavior.

3. MOTIVATION

For every game played in the NBA, a huge amount of statistics is generated. This leads to the existence of a rich historical database that motivates us to discover implicit knowledge in it. The NBA data we used in this work was obtained from the site Database Basketball [11], which was cited by the Magazine Sports Illustrated [9, 11] as the best reference site for basketball. The site makes available all the NBA statistical data in text files, from the year of 1946 to the year of 2006. Among the data, it is available information on 3736 players and 97 teams, season by season or by career. Our main goal is to move beyond the usual box score statistics presented in this database and discover new knowledge in the plain recorded numbers. The first analysis of the database aims to show why we need to look beyond the box score statistics.

In the NBA, players are evaluated according to various box score statistics. The main ones are the points, assists and rebounds a player scores in a game. Figure 1 illustrates the distribution of points, assists and rebounds for the players during their careers in the NBA (with their mid-range and their 99 percentile). We observe that the distributions follow a power law. This means that the majority of the players who played, or are playing, in the NBA contributed with a small fraction of the points, assists and rebounds registered in the history of game. Moreover, the majority of the points, assists and rebounds were scored by a small fraction of the players. Figure 1 shows that more than 99% of the players have scored less points, assists or rebounds than the mid-range value. It is important to point out that there are players that did not score a single point, assist or rebound during their careers. They are likely the players who were signed to a team for an experience period and then were released by the team, without playing a single game in the NBA. Thus, these players correspond to points that fall outside the power law curve showed in Figure 1, for they were not considered as NBA players.

In terms of box score statistics, we also analyze the average performance of the players per game played. The results presented in Figure 1 do not disclose the information about players who had significant performances during a few years and then retired. Figure 2 shows the distribution of efficiency among players during their careers in the NBA. In this work, we define the efficiency of a player as the sum of

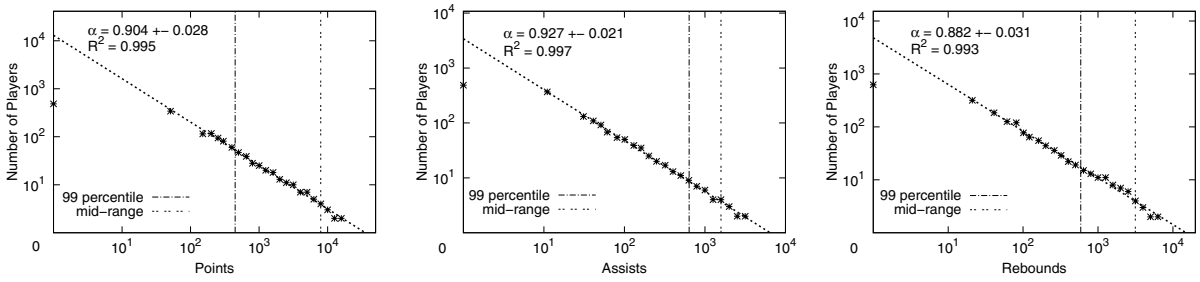


Figure 1: Distribution of points, assists and rebounds of NBA players.

his points, assists and rebounds achieved in a period divided by the total number of games he played in this period. The vertical lines in Figure 2 show the mid-range and the 90 percentile of the distribution. We see again that more than 90% of the players have career efficiencies below the mid-range efficiency value.

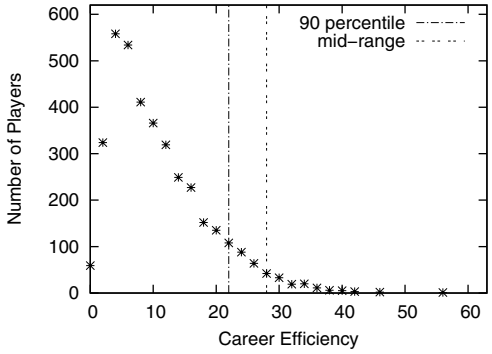


Figure 2: Distribution of the efficiency among players in the NBA.

Figures 1 and 2 lead us to think that only a few players have significantly contributed to a team in terms of the box score statistics analyzed. Moreover, if we consider that the only way to predict a team success is to analyze box score statistics then we are restricted to the analysis of a small fraction of players. Figure 3 illustrates this point. It shows the average rank gain with its standard deviation a player produces when he is transfered from a team to another one. The rank ${r_t}^y$ is a percentage that indicates the amount of teams that had a worse performance than team t in year y. The rank gain g_m indicates how much the team the player left t_{out} lost and how much the team the player joined t_{in} won with the transaction[1] m. The rank gain g_m for a transaction m is defined as $({r_{t_{in}}}^y - {r_{t_{in}}}^{y-1}) + ({r_{t_{out}}}^{y-1} - {r_{t_{out}}}^y)$. The term $({r_{t_{in}}}^y - {r_{t_{in}}}^{y-1})$ indicates how much t_{in} won with the transaction and the term $({r_{t_{out}}}^{y-1} - {r_{t_{out}}}^y)$ shows how much t_{out} lost. High values of the rank gain indicate that the team the player left decay its performance with his departure and the team he joined improves its performance. If the rank gain is zero, no significant change occurred. We observe in Figure 3 that there is no rule for the rank gain based on the player efficiency, that is, the average rank gain is zero for all efficiency values below 40. For the efficiency values greater than 40, we can not state anything. The number of transactions involving players that have efficiency values higher than 40 is not significant, only 20 in the history of the NBA.

[1]in this paper, the term transaction is referred to a exchange of teams by a player

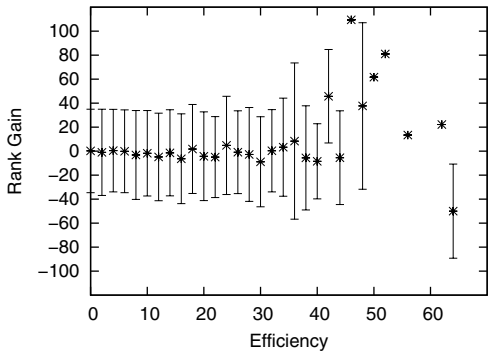


Figure 3: The average rank gain of a transaction.

In summary, we have seen that most of the players do not significantly contribute to box score statistics; they exhibit values that do not differ from the averages. We have also seen that the impact of a transaction on the behavior of the involved teams does not follow any rule when we consider only the efficiency of the involved player. This suggests that a team success does not depend directly and solely on the efficiency of the players it is signing in. Therefore, in the next section, we explore the complex network formed by the teams and the players of the NBA. This complex network allows us to formulate new models to predict team success.

4. THE NBA COMPLEX NETWORK

In order to clarify the understanding of the analysis developed here, we need to briefly explain the history of the NBA. The NBA was founded in 1946 with the name of Basketball Association of America (BAA) and had 11 teams. Prior to that, the American Basketball League and the National Basketball League (NBL) had been earlier attempts to establish professional basketball leagues. The BAA was the first league to attempt to play primarily in large arenas in major cities. The BAA became the National Basketball Association in 1949, when the BAA merged with the NBL, expanding to 17 franchises. From 1950 to 1966, the NBA initiated a process of reducing its teams and, in 1954, it reached its smallest size, with 8 franchises. We will call this time period P_{INI} from now on. From 1966 to 1975, the opposite process was initiated and the NBA grew from 10

franchises, in 1966, to 18, in 1975. During this period the NBA faced the threat of the formation of the American Basketball Association (ABA), which was founded in 1967 with 11 franchises and succeeded in signing major stars, such as Julius Erving. The ABA did not last for too long and, in 1976, both leagues reached a settlement that provided the addition of 4 ABA franchises to the NBA, raising the number of franchises in the NBA to 22. The period the ABA existed, from 1967 to 1975, we will call P_{ABA}. From 1976 to 2006, the number of teams in the NBA kept growing and, in 2006, the NBA reached 30 franchises. We call this period P_{NBA} from now on. Table 1 summarizes the time periods of the NBA. The time evolution graphics present two vertical lines marking the three periods of time.

Period	Description	Label
1950 to 1966	Teams were reduced from 17 to 10.	P_{INI}
1967 to 1975	Teams grew from 10 to 18. ABA founded.	P_{ABA}
1975 to 2006	ABA extinct. Number of teams in the NBA grew from 18 to 30	P_{NBA}

Table 1: Historical periods of the NBA.

In order to confirm the relevance of the periods listed in Table 1 and help us to understand the evolution of the NBA over time, we plotted in Figure 4 the number of active players and transactions per year. We first observe a high correlation (i.e., the correlation coefficient is 0.908) between these factors in a way that the number of transactions grows with the number of active players in the league. We also notice the sharp existence of three different behaviors, representing the peculiar characteristics of three periods described in Table 1.

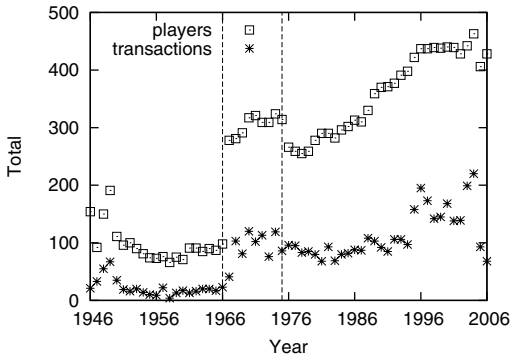
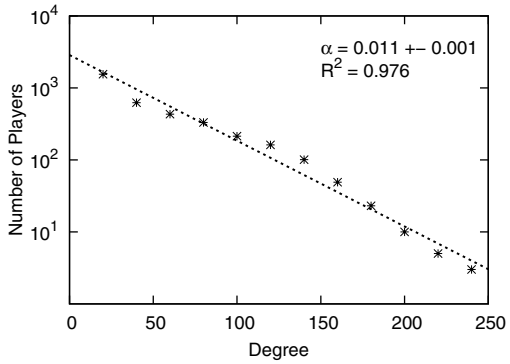


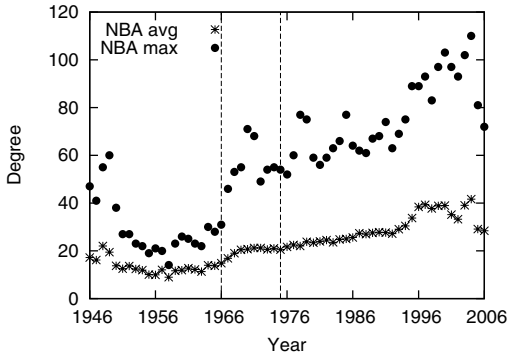
Figure 4: Active players and transactions per year.

The remainder of this work is aimed at studying the NBA as a complex network in evolution, as did in [6]. We construct two networks, the **Yearly NBA Network** and the **Historical NBA Network**, built in a way that the set of players P and the set of teams T are united to form the set of vertices V. Thus, there are two types of vertices: the player vertex and the team vertex. Each network has a different configuration in each year y of the analysis. The vertices of the Yearly NBA Network in the year y are only the players and teams that are active in the league in year y. On the other hand, the vertices of the Historical NBA Network in the year y are every player and team that played in the league before or in the year y. In both networks, there is an edge between two vertices if they had a labor tie. There is an edge between a player vertex p and a team vertex t in the year y if player p played for team t before or in the year y. There is an edge between two player vertices in the year y if they played together for a team before or in the year y. Obviously, there are no edges between two teams.

The first complex network metric we analyze is the degree distribution of each player vertex $p \in P$, which is illustrated in Figure 5-a. This distribution shows an exponential decay. Figure 5-b shows, for the NBA, the evolution of the average (NBA avg) and maximum (NBA max) degrees of the nodes in the Yearly NBA Network. We observe in Figure 5 that there is a significant variability among players degree in the NBA. We also observe that the average and the maximum degree of the nodes in the NBA network follow directly the behavior of the number of transactions showed in Figure 4 (i.e., the correlation coefficient is, respectively, 0.92 and 0.93), having the lowest values during P_{INI}.



(a) Degree distribution.



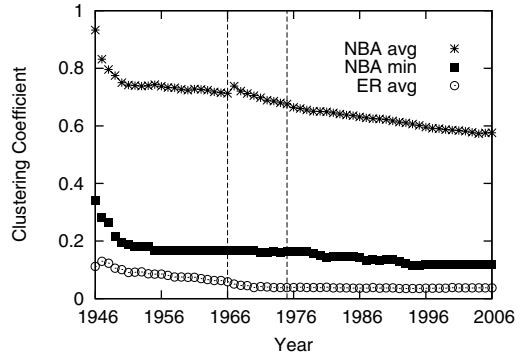
(b) Degree evolution.

Figure 5: Player degree.

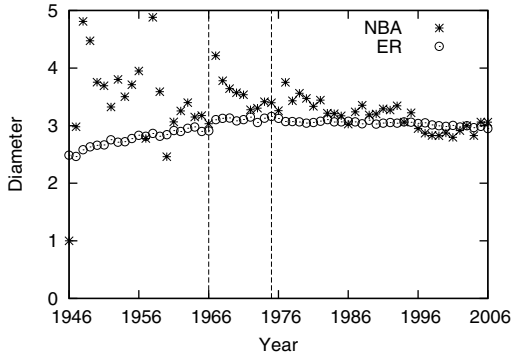
The clustering coefficient cc_i characterizes the density of connections close to vertex i. It measures the probability of two given neighbors of node i to be connected. The clustering coefficient of the network is the average $cc_i, \forall i \in V$. Figure 6-a shows the evolution of Historical NBA Network clustering coefficient (NBA avg), the lowest clustering coefficient $cc_i, \forall i \in V$ (NBA min), and the clustering coefficient of the NBA equivalent ER network (ER avg). The ER network is the one generated by the Erdös-Rényi (ER) model [12], that generates a random graph with the same number of vertices, edges and degree distribution. We observe the NBA clustering coefficient is, on average, one or-

der of magnitude higher than the clustering coefficient of its equivalent ER network. We also observe the average NBA Network clustering coefficient is significantly different from the lowest clustering coefficient, indicating a high variability in the clustering coefficients of the vertices of the network.

An important construction of network science is the small-world network [22]. It is characterized by having a clustering coefficient significantly higher than the one of its equivalent ER network and a diameter as low as the one of the equivalent ER network. The diameter measures the average shortest distance between every pair of nodes. Figure 6-a shows the evolution of the diameter of the Historical NBA Network in comparison to the diameter of the ER network. We observe that the NBA network diameter only stabilizes during P_{NBA}, when the number of active players and transactions kept growing. During this period, the diameters of the NBA and the ER networks are practically the same. The high clustering coefficient, combined with the small diameter, characterizes the NBA network as a small-world network. As a practical consequence, the short distance between NBA players means that new basketball tactics as well as business practices may propagate quickly among NBA players.



(a) Clustering coefficient.



(b) Diameter.

Figure 6: The NBA is a small-world network.

In summary, we have seen that there is a significant variability among the degrees and the clustering coefficients of the vertices of the network. We have also seen that there is a high correlation among the number of active players, the number of transactions and the degree of the nodes. Finally, we have seen that the NBA network is a small-world type of network. In the next section, we show how to apply this knowledge to construct models that predict the NBA team behavior.

5. PREDICTION TECHNIQUES

5.1 Evaluation Metrics

Before analyzing different prediction models, we describe our methodology. Each model calculates a prediction factor Π_t^y for each team t and for year y. After this, we verify if there is any correlation between the prediction factor Π_t^y of team t in year y and its rank in y. The rank ${r_t}^y$ is a number from 0 to 100 that indicates the percentage of teams that had a worse performance than team t in year y. For the calculation of this correlation, we use two correlation coefficients, the *Spearman's* ρ_b and the *Kendall* τ_b rank correlation coefficients [15]. We use the average of the rank correlation coefficients to measure the correlation between all the Π_t^y values and the ${r_t}^y$ for every year y analyzed.

After verifying the correlation, we check whether the given model selects successful teams. Team t_1 that presents the highest (or lowest, depending on the model) value of Π_t^y is selected by the model as a likely successful team in year y. The value of Π varies according to the metrics of each model. After the selection, we look at our database and verify the rank of team t_1 in the year y. The higher the rank of team t_1 the model selects, the better the model is. Another metric used to gauge success of the models is skewness, defined as a measure of the degree of asymmetry of the rank distribution. If the left tail of the distribution is more pronounced than the right tail, the distribution has negative skewness. The more negative the skewnes is the better the model, for it concentrates occurrences on the high rank values.

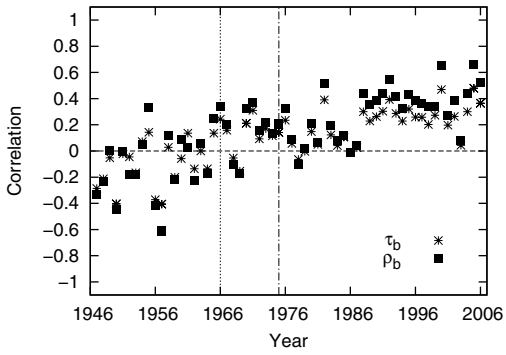
5.2 Efficiency-5 and Efficiency-1 Model

We start with simple prediction models. The **Efficiency-5 Model** is entirely based on box score statistics. The Efficiency-5 Model is very simple one and consists of computing the average of the five highest efficiencies of the players of a given team in the previous year. This average is the Π_i^y value for team t in year y and the team with the highest value of Π_i^y is the one selected by the model. We expect that the higher the efficiency of the players of a given team, the better the team performance is.

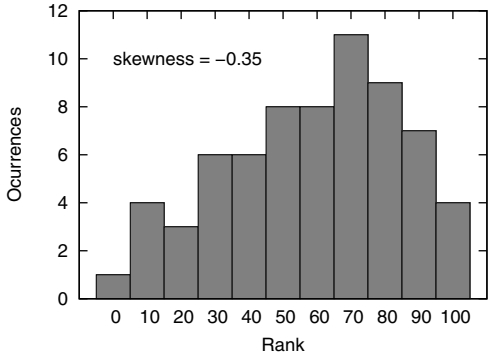
Figure 7 illustrates the validation of this model. Figure 7-a shows the ρ_b and the τ_b correlation coefficients between the prediction factor Π_t^y of team t in year y and its rank ${r_t}^y$. We observe that there is a trend to a positive correlation that begins after the P_{ABA} period. We see the correlation is significantly positive approximately after the half of the P_{NBA} byears. The average ρ_b and τ_b are 0.15 and 0.10, respectively.

Figure 7-b shows the number of times a team of a given rank was selected by the model. The number of teams with rank 70 that were selected is 11, meaning that there were 11 teams with rank greater or equal 70 and lower than 80. The skewness of this distribution is -0.35, indicating that the model selects more teams with a rank higher than 50 than it selects teams with a rank lower than 50. Observing the distribution, we see that it is almost normal, with 34% of the teams selected by the model having a rank lower than 50, 34% of the teams having a rank greater than 80, and 7% of the teams selected as the real best one.

We have seen that only a small fraction of players have box score statistics that can have significant impact on a team success. Therefore, we now show the **Efficiency-1 Model**, that sets the Π_i^y value to the highest efficiency value a player

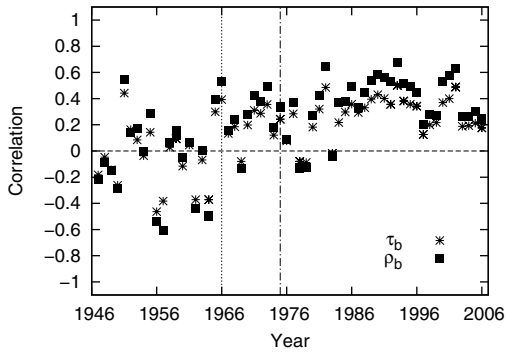


(a) Correlation among Π^y and rank.

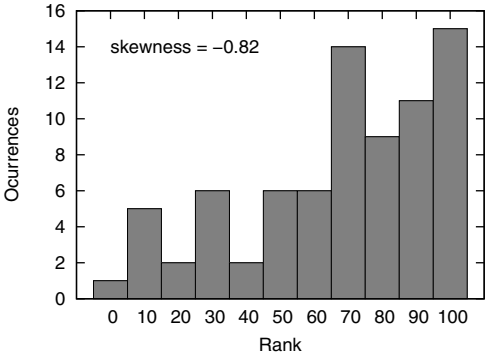


(b) Rank of the teams selected by the model.

Figure 7: Efficiency-5 Model.



(a) Correlation among Π^y and rank.



(b) Rank of the teams selected by the model.

Figure 8: Efficiency-1 Model.

of team t had in the year $y-1$, limiting the prediction to a small fraction of the active players. Figure 8-a shows the correlation coefficients between the prediction factor Π_t^y of team t in year y and its rank in this year. We observe that that this model has more positive coefficients than the Efficiency-5 Model. The average ρ_b and τ_b are 0.23 and 0.16 respectively, indicating that the success of a team is more correlated to the efficiency of its best players than to its best five players. We also observe that the correlation is higher after mid-1980s. This indicates that, during this period, the highest efficiency players of the teams had more impact on success than on the previous periods. This period coincides with the years in which several highly talented players appeared or were active in the league, like Michael Jordan, Magic Johnson, Larry Bird and Isiah Thomas.

Figure 8-b shows the number of times a team of a given rank was selected by the Efficiency-1 Model. The skewness of the distribution is -0.82, significantly higher than the skewness observed in the Efficiency-5 Model, indicating the Efficiency-1 Model selects higher ranked teams than the Efficiency-5 Model, with 27% of the teams selected by the model having a rank lower than 50, 59% of the teams having a rank greater than 80 and 25% of the teams selected as the real best one. In Figure 8-b we show the ranks selected by the Efficiency-1 Model year by year.

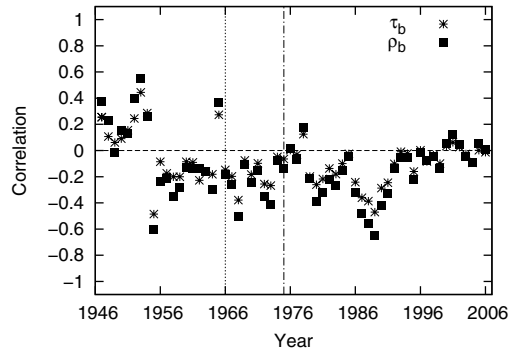
It is interesting to note that a small fraction of players can make significant impact on a team performance. This is may be a consequence of the small-world structure of the NBA network, where new basketball tactics and techniques spread quickly among NBA players, making them more homogeneous. It seems very unlikely that new tactics or techniques cause a significant impact on a team performance, since probably everyone in the league knows how it works and/or how to use them. However, when new techniques and tactics stem from talent, they can not be copied by average players, making the distance among talented players and average players even higher.

5.3 CC Model

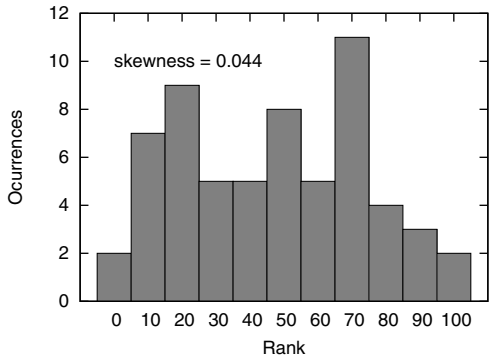
The **CC Model** is based solely on NBA complex network metrics, ignoring the box score statistics. The CC Model is based on the clustering coefficient of the vertices that represent the teams in the Historical NBA Network. A team with a high clustering coefficient is probably a team with new players or a team that does not make transactions frequently. On the other hand, a team with a low clustering coefficient is probably a team that has done a significant amount of transactions, most of the times keeping in its roster only the athletes who improve the team performance. The value Π_i^y is the clustering coefficient of the vertex that represents team t in year y, divided by the highest clustering coefficient of teams in the year y.

Figure 9-a shows the correlation coefficients between the prediction factor Π_t^y of team t in year y and its rank in this year. We observe that there are several years in which there is a negative correlation between the clustering coefficient of a team and its rank. Figure 9-b shows the number of times a team of a given rank was selected by the model. As we observe, this model is the worst so far, with the average ρ_b and τ_b equal to -0.12 and -0.09, respectively. They are less significant than the average ρ_b and τ_b of the previous models, and with 47% of the teams selected by the model

having a rank lower than 50 and with only 15% of the teams having a rank greater than 80 and 3% of the teams selected as the real best one.



(a) Correlation among Π^y and rank.



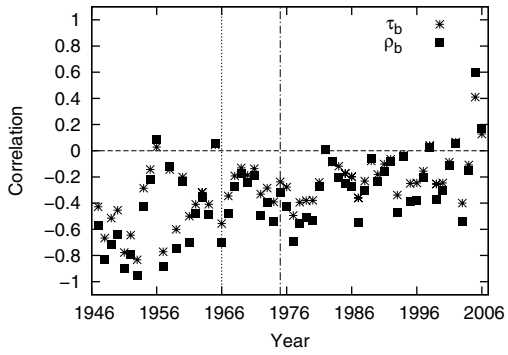
(b) Rank of the teams selected by the model.

Figure 9: CC Model.

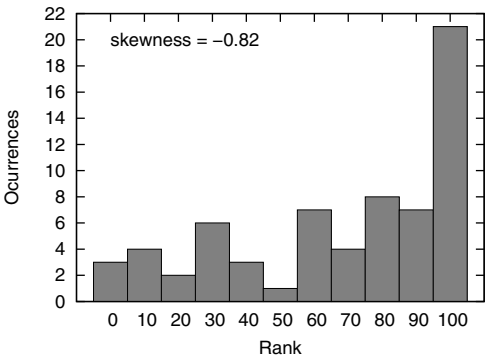
5.4 Degree Model

The second network-based model is the **Degree Model**. It is based on the degree of the players when the network is in the Yearly NBA Network. A player with a high degree is probably a player in the end of his career and/or a player who is traded frequently, that is, a player the teams usually do not want to keep. On the other side, a player with a low degree is a player in the beginning of his career or a player who does not or rarely change the team he plays for, that is, a player the teams want to keep in their roster. The Π_i^y value for team t in year y is the sum of the degrees of each player adjacent to t in year y divided by the highest sum of degrees of the teams in the year y. Thus, we expect the lower Π_i^y, the better the performance of a team.

Figure 10-a shows the correlation coefficients between the prediction factor Π_t^y of team t in year the y and its rank in this year. We observe that there is a clear negative correlation between the sum of the degree of the players of a team and its rank. The average ρ_b and τ_b are -0.35 and -0.27 respectively, indicating that this model is the best so far. In Figure 10-b, we show the number of times a team of a given rank was selected by the Degree Model. We observe that the Degree Model, in this case, is similar to the Efficiency-1 Model, with the same distribution skewness and with 31% of the teams selected by the model having a rank lower than 50, 61% of the teams having a rank greater than 80 and 36% of the teams selected as the real best one, however, this model chooses a third of the times the best team in the year. This is a significant result, once it does not rely on box score statistics. Also, it is important to point out that this model is more efficient before mid-1980s. This observation is consistent with the fact that in the period after mid-1980s there were very talented players that did not follow the logic behind the Degree Model.



(a) Correlation among Π^y and rank.



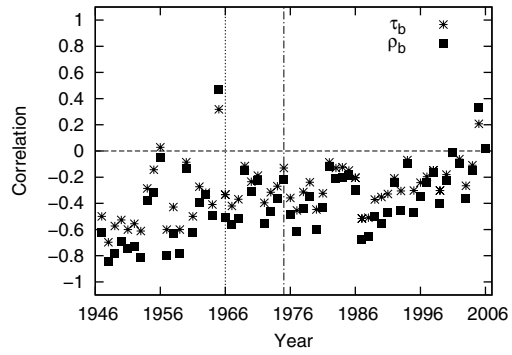
(b) Rank of the teams selected by the model.

Figure 10: Degree Model.

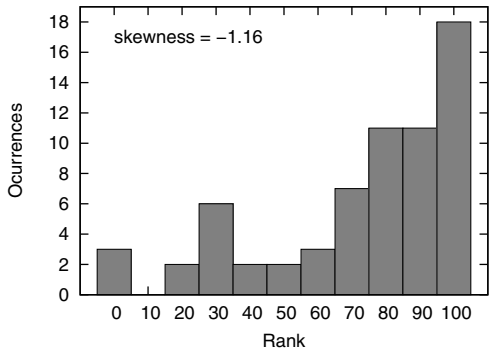
5.5 Mixing Models

In this section, we combine the models of the previous sections in order to obtain better results. The first attempt we do is by combining the CC Model and the Degree Model into the **CC-Degree Model**. The Π_i^y value for this model is the product of the Π_i^y values from the CC Model and the Degree Model. We observe in Figure 11-a that this model shows a slightly more regular behavior on its correlation coefficients than the Degree Model. This is reflected on the average ρ_b and τ_b, which are, respectively, -0.39 and -0.29, the more significant ones so far. This model also shows a better result in the number of successful teams it chooses. In Figure 11-b, we see that the skewness of the distribution is the lowest so far, -1.16, with 22% of the teams selected by the model having a rank lower than 50, 68% of the teams having a rank greater than 80 and 31% of the teams selected as the real best one. This is an impressive result, once this model does not rely on box score statistics, considering only the attributes of the NBA complex network.

Now we combine the CC Model, the Degree Model and the Efficiency-1 Model into the **CC-Degree-Eff Model** to obtain better results. The Π_i^y value for this model is the product of the Π_i^y values from the CC-Degree Model and

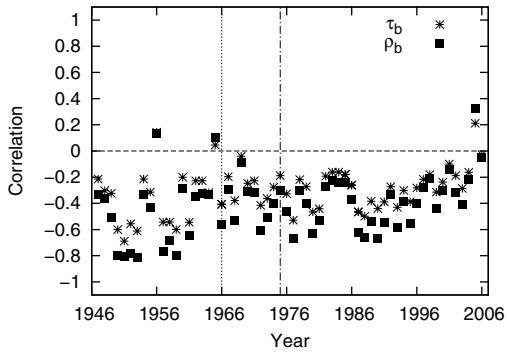


(a) Correlation among Π^y and rank.

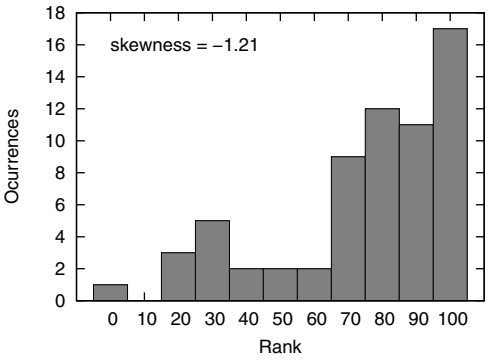


(b) Rank of the teams selected by the model.

Figure 11: CC-Degree Model.



(a) Correlation among Π^y and rank.



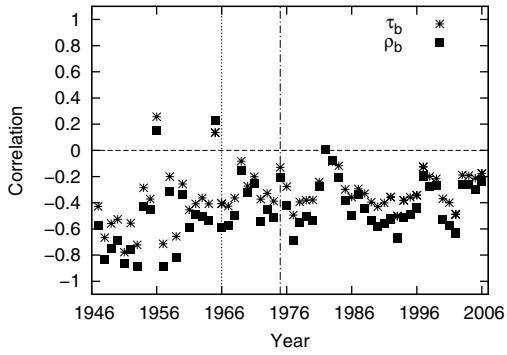
(b) Rank of the teams selected by the model.

Figure 12: CC-Degree-Eff Model.

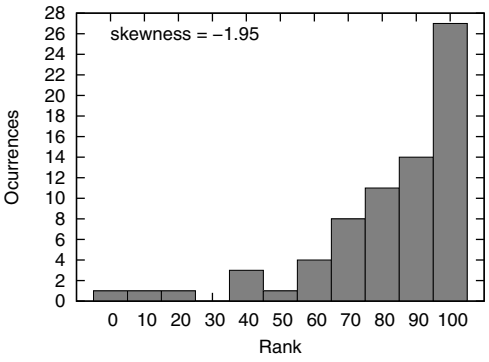
the inverse of the Efficiency-1 Model, since these models correlations have opposite signs. This model is the best so far. We observe in Figure 12-a that this model shows more constant correlation coefficients over the years than the previous models. The average ρ_b and τ_b are, respectively, -0.41 and -0.30, more significants than the previous model ones. Also, in Figure 12-b, we see the lowest skewness of the distribution of the ranks of the teams selected by the model, that is -1.21. It also reduces the teams selected with a rank lower than 50 and raises the number of teams selected with ranks higher than 80, with 19% of the teams selected by the model having a rank lower than 50, 68% of the teams having a rank greater than 80 and 29% of the teams selected as the real best one.

Finally, we modify the CC-Degree-Eff Model to capture the best of each single model. The CC-Degree Model is better in the years before mid-1980s and the Efficiency-1 Model is better after that. In this way, we use the Π_i^y provided by the CC-Degree Model only in the years before the mid-1980s and the Π_i^y provided by the Efficiency-1 Model in the years after. Figure 13 illustrates the rank of the teams selected by this model. We observe that this model presents the best results, as expected. The average ρ_b and τ_b are, respectively, -0.45 and -0.34, with the model selecting the best team 46% of the time. It is obviously biased, but from it we show that the doors that led us to improved prediction models are, as basketball players usually say, wide opened.

The results showed in this section predict the success of a team based on complex network metrics and box score statistics. We also evaluated the models for predicting team failure in the league and the results were practically the



(a) Correlation among Π^y and rank.



(b) Rank of the teams selected by the model.

Figure 13: Modified CC-Degree-Eff Model.

same. This indicates that the models can predict both the team success and the team failure in the league. Also, we observe that the Π value of the models works better to identify teams with performance in the extreme ranks, i.e., the worst and the best ones. When we consider teams on the average, the correlation between Π and the team rank is around 0. The correlation coefficients are near to 1 or -1 only when consider teams with extreme values for Π.

6. CONCLUSIONS AND FUTURE WORK

In this work, the NBA league was analyzed from a complex network standpoint. Initially, we looked into the box score statistics, such as points, assists and rebounds, of the history of the NBA and their teams. We observed this type of statistics captures only the role of a small fraction of players that had significant impact in the NBA. We then analyzed the evolution of the NBA using complex network metrics, calculated from a graph in which the vertices are the players and the teams, and the edges identify labor relationships among them. The principal contributions of our study can be summarized as follows.

- We show that the distribution of the number of points, assists and rebounds a player scores in his career in the NBA follows a power law.
- We show the evolution and growth of the NBA social network, constructed from data obtained in the NBA database.
- We show the NBA network structure can be characterized as a small-world network.
- We use complex network metrics to analyze social relationships among NBA players and discover new knowledge in the NBA database, that was not disclosed by the box score statistics. This new type knowledge, derived from network relationships, improves the understanding of team behavior.
- We construct different models to predict team success in the NBA. We show that the complex network metrics provide good prediction information without using box score statistics. We also show that these metrics may be combined to the box score statistics to improve the prediction efficiency.

As a future work, we plan to develop new prediction models that combine the box score statistics with new kinds of information derived from social networks that naturally emerge from player and team interactions [16]. We also plan to use the complex network metrics to discover knowledge about the player career in the league. Finally, we plan to analyze social networks derived from other sports leagues in the direction of creating theoretic models that explain the different behavior in different modalities of sports.

7. REFERENCES

[1] H. Abbot. Bad use of statistics is killing anderson varejão. *True Hoop*, november 2007.

[2] H. Abbot. Meet adjusted plus/minus. *True Hoop*, october 2007.

[3] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM.

[4] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the World Wide Web. *Nature*, 401:130–131, September 1999.

[5] L. Amaral, A. Scala, M. Barthélémy, and H. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences USA*, 97(21):11149–11152, October 2000.

[6] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, New York, NY, USA, 2006. ACM.

[7] E. Ben-Naim, F. Vazquez, and S. Redner. Parity and predictability of competitions. *Journal of Quantitative Analysis in Sports*, 2(4):1, 2007.

[8] J. Camacho, R. Guimerà, and L. A. Nunes Amaral. Robust patterns in food web structure. *Phys. Rev. Lett.*, 88(22):228102, May 2002.

[9] T. W. Company. Sports illustrated. http://sportsillustrated.cnn.com/, 2007.

[10] C. Cowan. The line on nba betting. *Business week*, July 2006.

[11] databaseSports.com. Database basketball. www.databasebasketball.com, 2007.

[12] P. Erdös and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 7:17, 1960.

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[14] S. Ilardi. Adjusted plus-minus: An idea whose time has come. *82games.com*, october 2007.

[15] M. G. Kendall and J. D. Gibbons. *Rank Correlation Methods*. Oxford University Press, New York, 5th edition, 1990.

[16] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, January 2006.

[17] S. Milgram. The small world problem. *Psychology Today*, 1:60–67, 1967.

[18] nba.com. www.nba.com. 2008.

[19] M. Newman. The structure and function of complex networks, 2003.

[20] M. E. Newman. The structure of scientific collaboration networks. *Proc Natl Acad Sci U S A*, 98(2):404–409, January 2001.

[21] R. N. Onody and P. A. de Castro. Complex network study of brazilian soccer players. *Physical Review E*, 70:037103, 2004.

[22] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.