Forecasting in the NBA and Other Team Sports: Network Effects in Action

PEDRO O. S. VAZ DE MELO, VIRGILIO A. F. ALMEIDA, and ANTONIO A. F. LOUREIRO,
Universidade Federal de Minas Gerais
CHRISTOS FALOUTSOS, Carnegie Mellon University

The multi-million sports-betting market is based on the fact that the task of predicting the outcome of a sports event is very hard. Even with the aid of an uncountable number of descriptive statistics and background information, only a few can correctly guess the outcome of a game or a league. In this work, our approach is to move away from the traditional way of predicting sports events, and instead to model sports leagues as networks of players and teams where the only information available is the work relationships among them. We propose two network-based models to predict the behavior of teams in sports leagues. These models are parameter-free, that is, they do not have a single parameter, and moreover are sport-agnostic: they can be applied directly to any team sports league. First, we view a sports league as a network in evolution, and we infer the implicit feedback behind network changes and properties over the years. Then, we use this knowledge to construct the network-based prediction models, which can, with a significantly high probability, indicate how well a team will perform over a season. We compare our proposed models with other prediction models in two of the most popular sports leagues: the National Basketball Association (NBA) and the Major League Baseball (MLB). Our model shows consistently good results in comparison with the other models and, relying upon the network properties of the teams, we achieved a $\approx$ 14% rank prediction accuracy improvement over our best competitor.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Statistical computing*

General Terms: Theory

Additional Key Words and Phrases: Complex networks, social networks, sports analytics

ACM Reference Format:
Vaz de Melo, P. O. S., Almeida, V. A. F., Loureiro, A. A. F., and Faloutsos, C. 2012. Forecasting in the NBA and Other Team Sports: Network effects in action. ACM Trans. Knowl. Discov. Data. 6, 3, Article 13 (October 2012), 27 pages.
DOI = 10.1145/2362383.2362387 http://doi.acm.org/10.1145/2362383.2362387

1. INTRODUCTION

It is well known that playing and watching sports is one of the preferred forms of entertainment for people. However, besides being a diversion, sports also move a multi-million dollar market. In 2006, the Nevada State Gaming Control Board reported $2.4 billion in legal sports wagers [Cowan 2006]. 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 [Cowan 2006]. At the same time, there is a lot of discussion on who should control and, therefore, receive a

Authors' addresses: P. O. S. Vaz de Melo (email: pedro.olmo@gmail.com), V. A. F. Almeida, and A. A. F. Loureiro, Universidade Federal de Minas Gerais, and C. Faloutsos, Carnegie Mellon University.

DOI 10.1145/2362383.2362387 http://doi.acm.org/10.1145/2362383.2362387

share of this amount [Hambach and Schottle 2006; Weinberg 2003]. Moreover, there are also those who say that sports betting and predictions can save the world [Lightman 2010], that is, predict hurricanes, markets, wars, and even climate change.

The sports betting market [Spann and Skiera 2009; Luckner et al. 2008] is mostly fed by the uncountable number of statistics generated after each game, which are used to describe and characterize the performance of teams and players. But is this characterization accurate? Henry Abbot, a senior writer at ESPN.com, in his blog True Hoop [Abbot 2007a], made a comment about the use of these statistics in the United States National Basketball Association (NBA). He wrote that these so-called box-score statistics, such as points per game (PPG), rebounds per game (RPG) and assists per game (APG), only measure the actions of a player within a second or two when someone shoots the ball. The rest of the time, points, rebounds, and assists measure nothing. Moreover, Michael Lewis wrote an article in the New York Times about the player Shane Battier, who has marginal box-score statistics values, but consistently makes his teammates better and his opponents worse when he is on the court [Lewis 2009]. For Shane Battier and other players, such as Anderson Varejão [Abbot 2007a], the box-score statistics are not able to characterize their performances.

In order to improve the characterization of players, other descriptive statistics were proposed without relying on the box-score statistics. In the NBA, one of the most used statistic is the adjusted plus/minus statistic [Rosenbaum 2004; Abbot 2007b; Ilardi 2007], which keeps track of the net changes in a score when a given player is either on or off the court. It does not depend on box scores, but still presents some bias which may jeopardize the characterization of a players performance [Barzilai and Ilardi 2008].

The fact is that collective sports are extremely dynamic, and to accurately capture in numbers the whole performance of a player and a team is impractical. Thus, in this work, we propose the use of network features to describe the performance of players and teams. We view a team sports league as a network of players, coaches, and teams in evolution, and we propose network-based models that use implicit feedback [Kelly and Teevan 2003] of network changes to aid in the prediction of team's behavior in a season. Implicit feedback used in information retrieval research because it unobtrusively obtains information about users by watching their natural interactions with the system, removing the cost from the user of explicitly providing feedback. In the case of a sports league, the use of network implicit feedback removes the cost of registering the uncountable number of box-score statistics that are registered after each game. Moreover, it disregards the explicit feedback given by the teams, players, and their agents when a transaction happens. It is possible that the public comments about the transactions are mostly moved by commercial interests, not picturing the reality behind them.

Initially, we focus our work on the NBA dataset. We view the National Basketball Association as a complex network, and develop metrics that are correlated with the behavior of NBA teams, taking into account only the social and work relationships among players, coaches, and teams. Then, based on these metrics, we propose models to predict how well a team will perform in the following season. We show, as seen in Figure 1, that one of our proposed network-based models, the NetForY, performs surprisingly well on either the NBA and on the Major League Baseball (MLB) datasets. In summary, the main contributions of this article are as follows.

—We show that in the NBA the regular box-score statistic distributions are highly skewed, with only a few players having high values. Moreover, acquiring players who presented high values in box-score statistics in previous seasons do not guarantee an improvement in team performance;

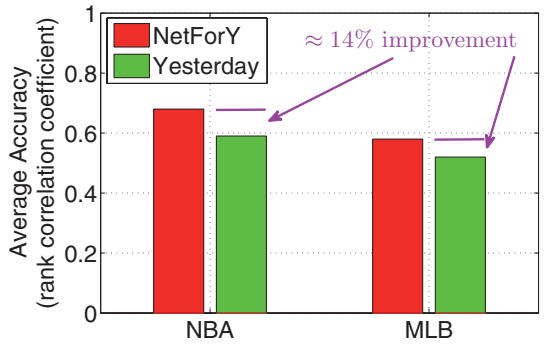


Fig. 1. Average accuracy of the network-based models and the Yesterday model (currently, state-of-the-art). Accuracy is measured using Spearman's rank correlation coefficient ρ. While the NetFor gives a similar performance, NetForY consistently beats Yesterday, with up to 14% improvement.

—We propose general network features that are good indicators of team performance in sports leagues. As an example, we show that high team volatility is not good for team performance;
—We propose network-based models to predict the behavior of teams in sports leagues, which present surprisingly good results when compared to other approaches.

We emphasize that the proposed models are generic and may be applied to any team sports league, that is, they do not rely on any particular box-score statistics.

The rest of this article is organized as follows. Section 2 presents related work. In Section 3, we show our method for modeling a sports league into a network and, in Section 4, we show the network implicit feedback features that give valuable information on the teams' performance. From these features we propose the network-based prediction models in Section 5, which are evaluated in Section 6. In Section 7, we directly apply the proposed models to predict the behavior of the teams in the MLB dataset, showing that they are general models. Finally, in Section 8, we present the conclusions and future work.

2. RELATED WORK

The use of networks to model collective sports is particularly interesting because of their dynamics. Teams constantly change players, players may exchange techniques among them, synergy among players may influence the performance of their teams, and so on. However, little attention has been given in the literature to such dynamic systems in terms of networks.

Girvan and Newman [2002] modeled the United States college football league as a network, where teams are the vertices and edges represent regular-season games between the two teams they connect. They used the college football league network to check whether their algorithm detects the communities of teams, that is, the league conferences. Later, Park and Newman [2005] proposed a ranking system for the United States college football league based on the network properties derived from this league. Moreover, Fast and Jensen [2006] modeled the National Football League (NFL) as a network to analyze the influence that coaching mentors have on their protégés. They identified notable coaches and characterized championship coaches, using this information and the social relationships to predict which teams will make the playoffs in a given year. Finally, Onody and de Castro [2004] analyzed the statistics of the Brazilian National Soccer Championship, and concluded that the players' connectivity has increased over the years while the clustering coefficient declined. They suggested

that the possible semantic reasons for this phenomenon are the exodus of players going abroad, the increasing number of players traded among national teams, and, finally, the increase in the players' career time.

Another interesting study relevant to our proposal is the work of Ben-Naim et al. [2007], which presents a 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 measured the "upset frequency" (i.e., the fraction of time the underdog wins). While basketball and American football leagues have the lowest upset frequencies, soccer and baseball leagues have the highest ones, meaning that the former ones are easier to predict than the later ones. Given that, we start our analysis based on a basketball database and then move to a baseball database, which is supposedly harder to predict and, therefore, a better test for our proposed network-based models.

Finally, in terms of forecasting in sports leagues, many different types of models have been constructed to predict the outcomes of sporting events, but, unfortunately, many of these models have never been used in forecasting beyond the period of fit [Stekler et al. 2010]. We believe that the main reason is the difficulty of competing with one of the simplest but most powerful models, the Yesterday model. In a 2006 article for the feature section of ESPN.com *Page 2* [Easterbrook 2006], Gregg Easterbrook, a NFL specialist, made a prediction for the upcoming season: "I predict that every NFL team will end the 2006 season with the same record as it did in 2005". What he proposed is the exact idea of the Yesterday model. He commented that this "obviously won't be right, but it will be closer than the countless pseudo-scientific forecasts floating around". In advance, we point out that he was partially right. The Yesterday is, to the best of our knowledge, the current state of the art for forecasting in generic team sports leagues and, as we will show in this work, present consistently good results. If there is one model to be beaten, it is the Yesterday model.

3. THE NBA NETWORK

3.1. Problem Definition

The main goal of this work is to show the high potential of network effects on predicting the behavior of complex social systems. Thus, we define the following general problem:

Problem 1. Network-Based Forecasting. Given a network G of players, coaches and teams with edges signifying how long they played together, how can we use G to predict the future behavior of this system?

Despite the fact that in this work we are analyzing team sports leagues, we strongly believe that the relevance of Problem 3.1 goes beyond these systems. While the main sports leagues have uncountable explicit data available, real-world systems frequently have missing or erroneous data. For example, in online auctions websites, for example, eBay, some users may provide erroneous information to improve their credibility. In this case, network effects can be used to spot anomalies and fraudulent users [Pandit et al. 2007]. In the same direction, network effects can be used to detect frauds and other violations among brokers [Neville et al. 2005]. Moreover, as another example, the social connections a person has may tell more about him/her than his/her personal attributes explicitly described in his/her curriculum vitae [Shetty and Adibi 2005]. These examples show the general power of the network effects of social systems, being completely independent of any kind of individual attribute of their entities.

Thus, we begin our analysis by modeling the historical NBA database as a network in evolution. The NBA data we used in this work is publicly available at the site DataBase Basketball [databaseSports.com 2010]. This site provides all the NBA statistical data

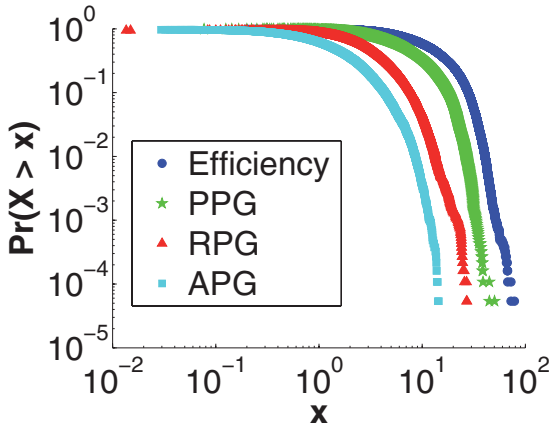


Fig. 2. The majority of players contribute marginally to their teams in terms of box-score statistics in comparison with a few players who make significant contributions. Observe the complementary cumulative distribution of the players' efficiencies and averages in points, assists, and rebounds per season.

in text files, from 1946 to 2008 and, among this data, provides information on 3863 players, 265 coaches, and 71 teams, season by season or by career. Our main goal is to move beyond the ordinary individual box-score statistics presented in that database and use network theory to discover new knowledge in the simple recorded numbers.

3.2. Motivation

The United States National Basketball Association (NBA) was founded in 1946, and since then is well known for its efficient organization and its high-level athletes. After each game played, a large amount of individual statistical data, such as points per game (PPG), assists per game (APG) and rebounds per game (RPG), are generated describing the performance of each player in the match. These statistics, called box-score statistics, are used to characterize the performance of each player over time, dictating their salaries and the duration of their contracts. They are also used by many services to provide more reliable predictions on the outcome of upcoming games. But one question that naturally arises is: are the box-score statistics the only ones capable of aiding in the prediction of team behavior? Again, we point out that PPG, RPG, and APG only measure the actions of a player within a second or two, while a player shoots the ball. The rest of the time, points, rebounds, and assists measure nothing [Abbot 2007a].

In Figure 2, we show, by plotting the complementary cumulative distribution (CCD), the probability of a player having points, assists, and rebound averages in a season greater than a determined value. We also show the CCD for the general efficiency of the players for every season analyzed, which is computed as

$$\text{Efficiency} = \frac{\text{points} + \text{rebounds} + \text{assists}}{\text{games played}}. \tag{1}$$

We observe that the CCDs for these box-score statistics have almost the same shape, characterized by a significant drop after the 90th percentile, that is, $Pr(X > x) = 10^{-1}$. We see that for all seasons analyzed, 90% of the players have marginal box-score statistical averages, lower than 11 PPG, 4 APG, 8 RPG, and an efficiency of 12. This means that the majority of the players contribute to their teams in ordinary ways, if we only look at box-score statistics. As an example, considering that an average team

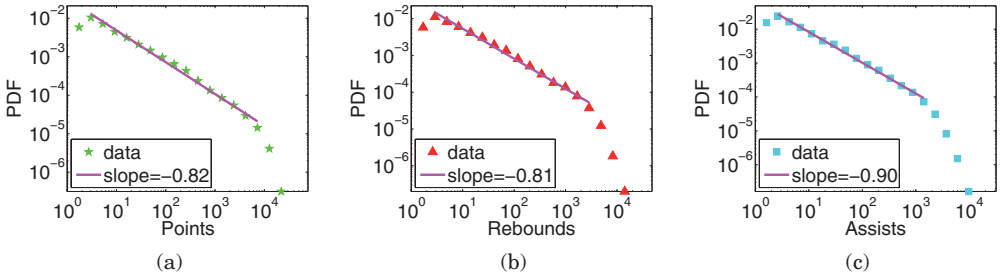


Fig. 3. The majority of players in their careers have scored marginal values in box-score statistics compared with a few players who made significant contributions. Observe the probability density function (PDF) of the number of points, rebounds, and assists the players achieved in their careers. The distributions follow a power law with a cutoff in the tail.

scores 90 points per game, the probability of having a player in a season scoring more than one-third of the team's points per game is less than 0.4%.

Moreover, we show in Figure 3 the probability density function (PDF) of the number of points, rebounds, and assists the players achieved in their careers in the NBA. We observe that the distributions follow a power law with an exponential cutoff in the tail, which is approximately in the 90th percentile for the three distributions. This indicates that, in general, 90% of the players have not scored more points, rebounds or assists than the cutoff value that is, respectively, 22%, 14%, and 11% of the maximum value in points, rebounds, and assists. Once again, we conclude that the majority of players have ordinary careers in terms of box-score statistics in comparison with a few players who make significant contributions.

Figures 2 and 3 lead us to conclude that only a few players contribute significantly to a team in terms of box-score statistics. Thus, if we consider that the only way to predict team success is to analyze box-score statistics, then we are restricted to the analysis of a small fraction of players.

Now we analyze the impact of acquiring or losing players with a determined efficiency value on the performance of the teams. But before that, we define the performance metric $f(t, y)$ of team t in year y as

$$f(t, y) = \text{ number of victories of team } t \text{ in regular season } y. \tag{2}$$

We chose this performance metric because it characterizes the teams' performance with fairness, since it judges the teams based on their performances over the entire season, with every team playing the same number of games.

Thus, given the performance metric $f(t, y)$, we define the performance ${r_\tau}^y$ of team τ in year y as

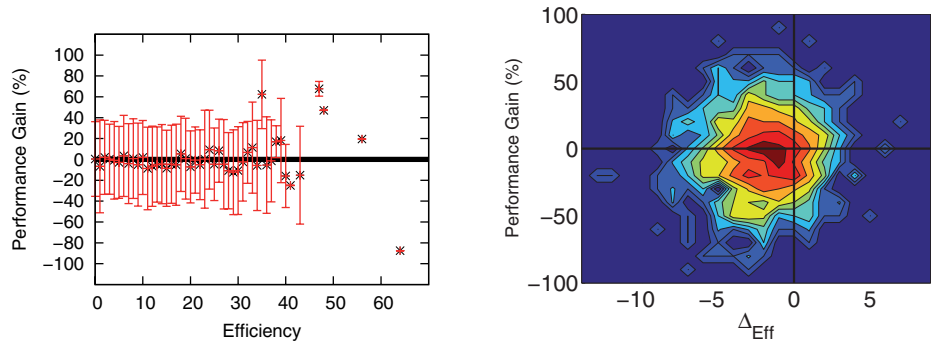
$${r_\tau}^y = \text{the percentage of teams } t, \text{ such that } f(t, y) < f(\tau, y), \tag{3}$$

or simply the percentage that indicates the number of teams that had a lower number of victories than team τ in regular season y. The best performance team has ${r_t}^y = 100\%$ and the worst performance team has ${r_t}^y = 0\%$.

In Figure 4 we show the average performance gain, with its standard deviation, a player produces when he is transferred from a team t_{out} to another team t_{in}. The performance gain g_m indicates how much the team the player who left t_{out} lost and how much the team the player who joined t_{in} won with the transaction m,[1] being defined as

$$g_m = \left({r_{t_{in}}}^y - {r_{t_{in}}}^{y-1}\right) + \left({r_{t_{out}}}^{y-1} - {r_{t_{out}}}^y\right), \tag{4}$$

[1]In this work, the term transaction refers to a exchange of teams by a player.



(a) performance gain of a transaction involving a player with a determined efficiency.

(b) Isocontours of the performance gain according to Δ_{Eff}.

Fig. 4. The efficiency of players acquired by a team has, in general, little effect on its future performance.

where the term $({r_{t_{in}}}^{y} - {r_{t_{in}}}^{y-1})$ refers to how much t_{in} won with the transaction and the term $({r_{t_{out}}}^{y-1} - {r_{t_{out}}}^{y})$ refers to how much t_{out} lost. High values for the performance gain indicate that the team the player left decayed in its performance with his departure and the team he joined improved its performance. If the performance gain is zero, no significant change occurred.

We observe in Figure 4(a) that there is no rule for the performance gain based on player efficiency, that is, the average performance gain is zero for all efficiency values below 40. For the efficiency values greater than 40, we cannot state anything, since the number of transactions involving players who have efficiency values higher than 40 is not significant—only 20 in the history of the NBA. This also corroborates the fact that individual box-score statistics of players are rarely relevant for predicting the performance of a team.

Finally, we also investigate whether a team will improve or decrease its performance when it changes its roster. In order to do this, we look at the efficiency values of the players this team had in the previous season and currently has. Thus, we define the ΔEff_t^y for a team t in year y as

$$\Delta Eff_t^y = avg\left(\sum_{\forall p_t^y} eff(p_*^y)\right) - avg\left(\sum_{\forall p_t^{y-1}} eff(p_*^{y-1})\right), \tag{5}$$

which is the difference between the averages of the efficiency values of the players this team had in year y and in year $y-1$. A high value of ΔEff_t^y means that team t had higher-valued players in year y compared to year $y-1$. In Figure 4(b) we show the isocontours of the performance gain according to ΔEff. Most of the transactions are in the dark red region where ΔEff is between -5 and 0, and the performance gain is approximately 0. Moreover, it is clear that there is no correlation between ΔEff and the performance gain, since the performance gain values are almost symmetric to the $\Delta Eff = 0$ line. This also implies that simply acquiring players with high box-score statistics is not a guarantee for team success. Besides this, it is interesting to note that most of the times the teams have negative ΔEff values from year to year, which suggests that changing the team roster is usually not a good strategy. This will become clearer when we describe the network-based features in Section 4.

3.3. The Network Definition

In the previous section, we showed that the usual box-score statistics are not good predictors of the future performance of NBA teams over the years. Moreover, we showed that box-score statistics alone cannot predict whether a transaction will be good or bad for a team. Thus, in order to explain and better understand the consequences and causes instigating the transactions dynamics, we propose modeling the NBA as a network (or graph) in evolution. We believe that metrics which capture network changes over time may provide implicit feedback signals that might characterize the temporal situation of the teams and players in the NBA. We emphasize that the following model can be extended to any team sports league which involves teams and players and, in fact, it will be used in Section 7 to model the MLB database.

In the network we construct from the NBA, the set of vertices V contains three types of vertices: the set of players vertices P, the set of coaches vertices C, and the set of team vertices T. Since in the NBA, the players, coaches, and teams change over time, we construct one network $G^y(V^y, E^y)$ per year, each one containing a set of player vertices $P^y \subset V^y$, a set of coach vertices $C^y \subset V^y$ and a set of team vertices $T^y \subset V^y$. But before this, we initially construct the Yearly NBA Networks (YNN) $G'^y(V'^y, E'^y)$ for every year y, where V'^y is the set of players P'^y, coaches C'^y, and teams T'^y that are active in season y. The set of edges E^y are defined according to the labor relationships that exist between players, coaches, and teams in V'^y. We link a player p or a coach c to a team t in E'^y if and only if p or c played or coached for t in y. Moreover, we link a player p to other player q in G'^y if and only if p played together with q for a common team in y. Moreover, we link a player p to a coach c in G'^y if and only if p was coached by c in a common team in y. Clearly, the data is temporal and the characteristics of players and teams change each year, and thus we use t^y, c^y, and p^y to denote, respectively, the nodes of team t, coach c, and player p in year y.

After constructing all YNNs, we are now able to construct the NBA networks that contain the information about the historical relationships among players and teams. In this way, we recursively define the NBA network $G^y(V^y, E^y)$ as the graph that contains the information in G'^y and also the information in G^{y-1}. The use of historical information enables us, for instance, to search for players who have played for a great many teams, or to search for teams that frequently change that their rosters significantly over the years. There are several ways to propagate the information contained in G^y to G^{y+1}, and we describe some possible propagation models here.

(1) *Historical*: $G^{y+1} = G'^{y+1} \cup G^y$. This model propagates all the vertices and edges through the years, never removing them from the network. We use this model in this work.
(2) *Yearly*: $G^{y+1} = G'^{y+1} \cup G^y(V'^{y+1}, E^y)$. This model removes from G^{y+1} all vertices that left the NBA in $y + 1$, that is, it considers only players and teams that are active in y.
(3) *Delayed*: $G^{y+1} = G'^{y+1} \cup (G^y(V^y, E'^{y+1} \cup (E^y - E^{y+1-\Delta_y}))$. This model removes from the G^{y+1} the edges that are Δ_y years old. When Δ_y is equal or higher than the total number of years in the dataset, the *Delayed* model propagates in the same way as the Historical model. On the other hand, when $\Delta_y = 1$, it propagates as the *Yearly* model does.

Thus, our goal is to move beyond the usual box-score statistics and discover new knowledge in the network properties of the NBA, which are defined by the work and social relationships among players, coaches, and teams. We are interested in knowing, for instance, whether a team that always had players who had already played with each other in previous seasons is good for the team's performance; or, as another example,

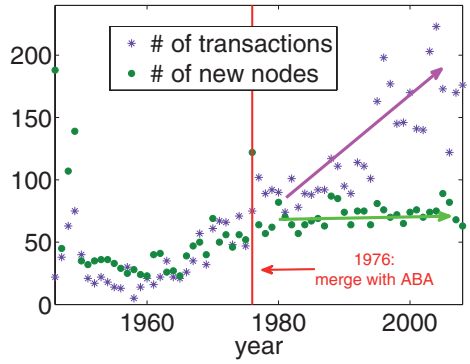


Fig. 5. NBA network changes: the number of new nodes and transactions (new edges) over the years. Note that after 1976 there is much more mobility (transactions), with the number of new nodes being around 70, while the number of transactions keeps on growing.

if a player who has already played with a large number of other players in previous seasons is one who will help his team improve its performance.

The work and social relationships among players, coaches, and teams are defined by network changes, which basically occur in two situations: (i) when a new player or team joins the NBA or (ii) when a transaction occurs, that is, an exchange of teams by a player. Semantically, Dilger [2002] showed that a player may leave a team when he is not performing well or when his salary is high enough to force his team to free some space in its payroll budget due to a salary cap.[2] Besides these reasons, we point out that a player may also leave a team when he becomes a free agent and wishes to join another team which, in his judgment, has more chances to win the championship [ESPN.com 2009].

In this direction, we show in Figure 5 the evolution of the two situations that cause network changes, that is, the number of transactions (new edges) and new players, coaches, and teams (new nodes). First, we observe a high correlation between the number of new nodes and new edges for the years before the late seventies. Then, the number of new nodes staid practically the same while the number of new edges grew practically linearly. This suggests that the semantic reasons and implicit feedback associated with network changes for these two periods might be different due to, for instance, changes in regulations and player demands. The results presented in Section 6 reflect this observation.

4. PROPOSED NETWORK FEATURES

4.1. Preliminaries

Once we define the NBA network and how it evolves over time, we are able to define and tackle the following problem:

Problem 2. Feature Extraction. Given a system in which the only information available is its historical network $G^y(V^y, E^y)$ for year y, which features can we extract from G^y that provide relevant information on its future behavior?

The features that can be extracted from G^y are exclusively network features that obviously do not contain any sort of explicit information about their nodes. Therefore, in order to solve Problem 4.1, considering the NBA dataset, we must rely on the implicit feedback that arises from network effects, which indicate whether a team will

[2]The salary cap is the maximum dollar amount that teams can spend on player contracts.

Table I. Table of Symbols

Description	Symbol
node of team t in year y	t^y
node of player p in year y	p^y
node of coach c in year y	c^y
performance of team t in year y	r_t^y
first year of node v	$y0_v$
age of node v in year y	$a_v^y = y - y0_v$
roster of team t in year y	R_t^y
degree of node v in year y	d_v^y
clustering coefficient of node v in year y	cc_v^y
team volatility	$\Delta d_t^y = d_t^y - d_t^{y-\epsilon}$
roster aggregate volatility	$\Sigma\,\Delta d_t^y = \sum_{\forall v \in R_t^y} \frac{d_v^y}{y - y0_v}$
team inexperience	$ixp_t^y = cc_t^y$
roster aggregate coherence	$\overline{cc_t^y} = avg(cc_v^y \times (y - y0_v)), \forall v \in R_t^y$
roster size	$s_t^y = \sum \forall p \in R_t^y$

have a good or bad performance in a year. Formally, each implicit feedback is a function $f_i(G^y, t^y) \Rightarrow \Re$ that receives as parameters the NBA network G^y for year y and the team node t^y, in which we want to know its future performance in year $y+1$. Moreover, each implicit feedback f_i is associated with a specific observable network effect which can be used as an implicit measure of interest—in our case, verifying if a team is likely to have a good or bad performance in the following season. As an example, a team that significantly increased its degree from G^{y-1} to G^y is one that has significantly changed its roster and, as a consequence, might not perform well in season y. On the other hand, if the degree between seasons is similar, the team is probably satisfied with its roster and might perform well.

In the following sections, we show our proposed network features. These features are based on the degree d_v^y and on the clustering coefficient [Newman 2003][3] cc_v^y of the vertices v of the NBA network $G^y(V^y, E^y)$, being the same metrics reported by Vaz de Melo et al. [2008] as those that correlate with the performance of NBA teams. In Table I, we show a summary of these features and other relevant symbols.

4.2. Team Volatility

Our first feature is based on the well-known quote, "never change a winning team" by Sir Alf Ramsey, manager of the English national football team from 1963 to 1974. As we mentioned in Section 3.3, a player may leave a team when (i) he is not performing well; (ii) his salary is too high for the team; or (iii) he becomes a free agent and wishes to join another team, which, in his judgment, has more chances to win the championship [ESPN.com 2009]. Because reasons (ii) and (iii) are more related to high-performance players and, as we showed in Section 3.2, most players have average box-score performances, we conjecture that the majority of transaction occur due to reason (i).

Thus, when the degree d_t of a team t is constantly increasing and at a high rate over the years, it probably means that t has not yet found a good roster. Therefore, a highly increasing rate for a team node t before season y starts is an implicit feedback that this team will not perform well in y. Then, we define the feature *team volatility* Δd_t for team t in year y as

$$\Delta d_t^y = d_t^y - d_t^{y-\epsilon}, \tag{6}$$

where ϵ is a time window parameter that should be small, in order to capture the most recent behavior.

[3]The clustering coefficient is the probability of two given neighbors of a certain node being connected.

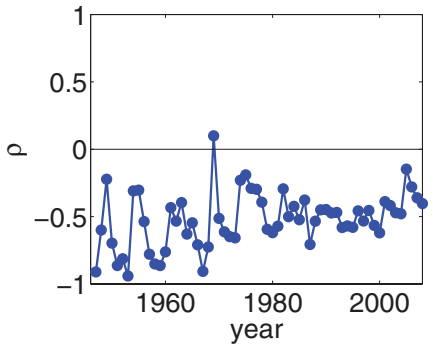


Fig. 6. *Team volatility* hurts the performance of the teams, that is, the more a team hires and fires players and coaches, the worse its performance. Observe the regularly negative Spearman's rank correlation ρ between the *team volatility* Δd and the performance of the team. Note that, unlike box-score statistics, our network-based features can capture such a correlation.

In Figure 6, we show Spearman's rank correlation ρ [Kendall and Gibbons 1990] between the *team volatility* Δd and the performance of the teams, with $\epsilon = 2$.[4] The coefficient ρ measures how well the relationship between two variables can be described using a monotonic function as being appropriate to compare two ranks. The values of ρ vary from -1, when the two ranks are completely opposite, to 1, when the two ranks are the same. When $\rho \approx 0$, there is no relationship between the ranks. As we expected, we can observe that, with the exception for the year 1969, the correlation is always negative, with an average $\overline{\rho} = -0.52$. This indicates that *team volatility* hurts the performance of the teams, that is, the more a team hires and fires players and coaches, the worse for its future performance. Thus, it is reasonable to state that the *team volatility* feature is a good indicator of team performance.

4.3. Roster Aggregate Volatility

At the same time a team that constantly changes its roster is probably one that is not achieving satisfactory results, a player who constantly changes his team is probably one who is not adapting. Thus, this player is probably one who will not provide any significant improvement to his current and future teams; in the same way, this is also valid for coaches. Therefore, a team with a large number of members who have played for several teams is an implicit feedback that this team will not perform well in the season. Thus, in order to spot such teams we define the *roster aggregate volatility* feature as

$$\Sigma\Delta d_t^y = \sum_{\forall v \in R_t^y} \frac{d_v^y}{y - y0_v}, \tag{7}$$

where $y0_v$ is the year of the first season played by the team member v who can be a player or a coach. In summary, $\Sigma\Delta d_t^y$ is the sum of the degree of increasing rates for all the team members in the roster R_t^y of team t in year y, counting from the first season for each member. Again, the use of the *historical* propagation model is justified because we want to know the history of players and coaches who played with every team member, $v^y \in t^y$.

In Figure 7, we show Spearman's rank correlation ρ between the *roster aggregate volatility* $\Sigma\Delta d$ and the performance of the teams. As we observe, the correlation is

[4]Other small values of ϵ were tested, and the results are similar.

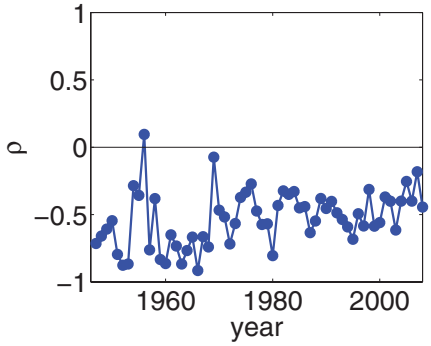


Fig. 7. Again, note that the *roster aggregate volatility* hurts the performance of the teams, that is, the more a team hires players and coaches who are constantly changing their teams, the worse the performance. The Spearman's rank correlation ρ between the *roster aggregate volatility* $\Sigma\Delta d$ and the performance of the teams is consistently negative. This is another observation that cannot be drawn from the usual box-score statistics.

negative for the vast majority of years, with an average $\overline{\rho} = -0.52$. Although this feature is similar to the *team volatility*, it has subtle but significant differences. For instance, a team that is changing its roster at a slow rate, according to the *team volatility* metric, is likely to have a good performance. However, if this team is hiring players that are changing their teams at a high rate, according to the *roster aggregate volatility* metric, this team is likely to have a bad performance.

4.4. Team Inexperience

It is well known that when a new team arrives at a sport league, the general expectation is that the team does not perform well [Reheuser 2010]. One of the reasons is that the team managers may not have the proper number of connections and notoriety to raise a significant number of funds from sponsors to hire well-known good players. Then, we expect that the experienced teams will perform better than the newly arrived inexperienced teams.

The usual and most natural way to describe a team t experience in year y is by its age a_t^y, that is, the number of years a team has played in the NBA. However, the amount of experience a team acquires during the years is not a linear function, that is, the amount of experience a team gets after playing its first season is probably significantly higher than the amount it will get by playing its 20th season. Thus, as a measure of experience we propose the clustering coefficient of a node that has a nonlinear relationship with the age of the team (for more details, see the Appendix). Moreover, the clustering coefficient also indicates whether a team has not experimented with a wide variety of players in its roster, which in most cases may suggest that this team is not a good one team. Thus, we define the *team inexperience* feature as

$$ixp_t^y = cc_t^y. \tag{8}$$

Thus, when a team t joins the NBA in year y, $ixp_t^y = cc_t^y = 1$ and, as t makes transactions and become more experienced, cc_t^y decreases. Despite being a natural intuitive network feature, we can see in Figure 8 that the rank correlation ρ between the *team inexperience* feature and the performance of the teams is not regular, averaging only -0.14. However, a closer look indicates that the correlation fluctuates smoothly, which suggests that the temporal scenario plays an important role in this feature—in some subsequent years the correlation being significantly negative, and in other subsequent years being close to zero. Moreover, the median performance of inexperienced teams,

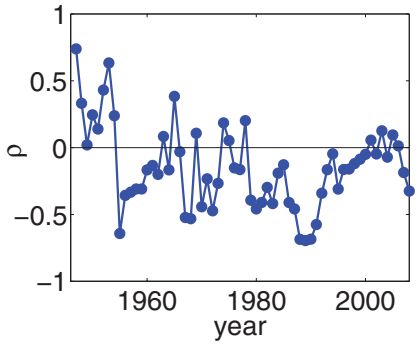


Fig. 8. There is no clear correlation between the *team inexperience* and the performance of the teams. However, a closer look indicates that the correlation fluctuates smoothly, which suggests that the temporal scenario plays an important role in this feature.

that is, teams that played a season with a team experience $ixp_t^y > 0.95$, is 21%. This indicates that, in general, their performance is significantly below average, that is, 50%. However, as we have shown in Figure 19 (see the Appendix) and in Vaz de Melo et al. [2008], the clustering coefficient carries valuable information about the teams and players which can be used jointly with other information to predict the future behavior of teams and players. This is shown in the description of our next feature, which uses the clustering coefficient, and is the one that shows the highest correlation with the performance of the teams.

4.5. Roster Aggregate Coherence

It is commonsense among sports analysts that good chemistry among players is essential for the success of a team. This so-called chemistry involves relationship among the players and also knowledge of the future moves and plays of each teammate. Moreover, it is also important that the coach knows his players well, being able to request from them what he knows as their best techniques. Thus, since the best way to build chemistry among players and the coach is through time, we define the metric *roster aggregate coherence* $\overline{cc_t^y}$ for team t in year y as

$$\overline{cc_t^y} = avg(cc_v^y \times (y - y0_v)), \forall v \in R_t^y, \tag{9}$$

which is the average of the clustering coefficients cc_v^y for every team member v_t^y who is in team t in year y multiplied by the current member's age $a_v^y = y - y0_v$. In summary, a high value of $\overline{cc_t^y}$ means that the team roster played together for a substantial amount of time and that few changes occurred. On the other hand, a low value of $\overline{cc_t^y}$ means that the roster recently faced a large number of changes.

In Figure 9, we show Spearman's rank correlation ρ between the *roster aggregate coherence* $\overline{cc_t^y}$ and the performance of the teams. As we expected, the correlation is positive for every year, with an average $\overline{\rho} = 0.55$, the highest among all the proposed features. This indicates that high coherence and rapport among the players and that the coach is a good indicator that a team will succeed. Again, this is a result that cannot be extracted from the usual box-score statistics, showing the usefulness of network effects.

4.6. Roster Size

Finally, we present the last feature we extracted from the NBA network, which is the *roster size*. The size of the roster of a team t in year y is the number of players playing

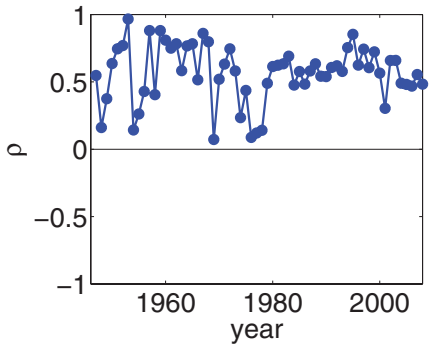


Fig. 9. A high degree of coherence and rapport among players and the coach is a good indicator that a team will succeed. Note that Spearman's rank correlation ρ between the *roster aggregate coherence* $\overline{cc}_t^y$ and the performance of a team is regularly positive.

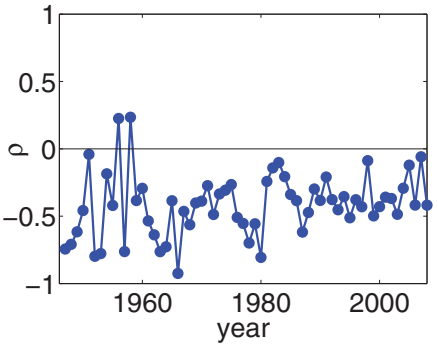


Fig. 10. The larger the roster size, the worse for the team. The Spearman rank correlation ρ between the *roster size* s_t^y and the performance of the teams is consistently negative.

(or have contracts) for t in y. Given that a salary cap[5] is in the NBA regulation since the first season [nba.com 2008], teams with a big number of players in their rosters tend, on average, to pay less to their players than teams with fewer players. Thus, it is expected that teams with fewer players will have more valuable players, that is, stars, since these players usually demand higher salaries. From this, we define the *roster size* feature s_t^y for team t in year y as

$$s_t^y = \sum \forall p^y \in R_t^y. \tag{10}$$

In Figure 10, we show Spearman's rank correlation ρ between the *roster size* s_t^y and the performance of the teams. We observe that the correlation is negative for every year, except the years of 1956 and 1958, with an average $\overline{\rho} = -0.41$. Thus, the larger the roster size, the worse for the team. This also enables us to consider the *roster size* feature as a good indicator of the team's performance.

For an analysis on the independence of the features presented in this section, please see the Appendix.

[5]The salary cap is the total amount of money the teams can spend on player salaries.

5. NETWORK-BASED PREDICTION MODELS

5.1. Problem Definition

Once the features to be extracted from the NBA network are defined, we were able to tackle the main problem of this work:

Problem 3. Rank Prediction. Given a NBA Network $G^y(V^y, E^y)$ for year y, predict the performance $r_t{}^{y+1}$ of every team t in year $y+1$?

In Section 5.2, we present the *Network-based Forecaster (NetFor)*, which is our solution for Problem 5.1 and the main contribution of this article. The NetFor is a prediction model that uses only network features to predict the behavior of the teams. Also, in Section 5.3, we propose the *Network-based Plus Yesterday Forecaster (NetForY)*, which is a prediction model that uses both the network features and the information on the previous performances of the teams.

5.2. NetFor

We begin to solve Problem 5.1 by defining the basic operation of a prediction model M. First, we determine that a prediction model M_A has a function $F_A(t_y)$ to calculate a prediction score $\Pi_t^y = F_A(t_y)$ for each team t in every year y, where Π_t^y measures the likelihood of team t performing well in year $y+1$. Thus, the descending order of the Π_t^y values for every team t in year y gives the predicted rank of the teams for year $y+1$, where the team t^* with the highest prediction score $\Pi_{t^*}^y$ is the most likely team to win the championship in year $y+1$.

Thus, we propose as a solution to Problem 5.1 the *Network-based Forecaster (NetFor)* M_{NetFor}, which is based solely on network features. The function F_{NetFor} that calculates the Π_t^y value for each team t in year y is given by the features shown in the previous section. We showed that the *team volatility* Δd_t^y, the *roster aggregate volatility* $\Sigma\Delta d_t^y$, the *team inexperience* ixp_t^y, the *roster aggregate coherence* $\overline{cc_t^y}$, and the *team volatility* s_t^y are good indicators of how well a team t will perform in a season y. Thus, we simply gather these five features in the function F_{NetFor} to compute the prediction score Π_t^y for each team t in year y, in a way that

$$\begin{aligned}\Pi_t^y &= F_{NetFor}(t^y, S)\\ &= \Delta d_t^{y-w1} \times \Sigma\Delta d_t^{y-w2} \times ixp_t^{y\,w3} \times \overline{cc_t^y}^{w4} \times s_t^{y-w5}\\ &= -w1 \times \log(\Delta d_t^y) - w2 \times \log(\Sigma\Delta d_t^y) - w3\\ &\quad \times \log(ixp_t^y) + w4 \times \log(\overline{cc_t^y}) - w5 \times \log(s_t^y),\end{aligned} \tag{11}$$

where the set of parameters $S = \{w1, w2, w3, w4, w5\}$ are real values that weigh the features to which they are associated. The sign of the parameter is negative if the correlation is negative and positive if the correlation is positive. We do not use a plain linear model because the scale of the feature values may differ in orders of magnitude, which may make the size of the parameter space impractical. A plain multiplicative model removes from the parameters any dependency on the scale of the features they represent. In this case, the parameter values only measure, quantitatively, how relevant a feature is compared to another one. A linear alternative to the multiplicative model is the liner model using the logarithm of the feature values, which can be derived directly from the multiplicative model, as described in Eq. (11).

In order to assign values to the feature parameters $w1, w2, w3, w4$, and $w5$, we propose a simple method. Since we want to predict the performance of the teams for year y, we can use all the information from the previous years $y-1, y-2, \ldots, y-W_Y$ as the training set used to determine a good set of values $S_i^y = \{w1_i^y, w2_i^y, w3_i^y, w4_i^y, w5_i^y\}$ for the feature parameters. Thus, for a prediction for year y, our training set is a

window of the W_Y previous seasons $y-1, y-2, \ldots, y-W_Y$, and we use it to find a set S_*^y that gives the best results for the W_Y years.

Before explaining what a good result is and how we find the best one, we list two classes of people who would benefit directly from a sports prediction model.

(1) *Sports analyst: the whole rank is important*. He/she wants to know how all teams will perform, that is, he/she wants to know all the future team performances.
(2) *Gambler: only the top predicted team t^* is important*. He/she has only one bet for which team will be the champion, that is, he/she only wants that the top team predicted by the model to also be the champion.

Given these classes, we could use as the evaluation metric to find S_*^y, Spearman's rank correlation coefficient to satisfy the *sports analyst* class or the number of times the prediction model selects the best performing team to satisfy the *gambler* class. However, for simplicity, from now on we will use the weighted Spearman's rank correlation ρ_W [da Costa and Soares 2004] as our main evaluation metric for determining how good a parameter set S for a training set is. The ρ_W weighs the distance between two ranks using a linear function of those ranks, giving more importance to higher ranks than lower ones, satisfying the two classes of people described above. Moreover, according to da Costa and Soares [2004] the main application for the proposed rank correlation coefficient is in the evaluation of rank prediction methods, which are mostly applied to recommendation systems, information retrieval, stock trading support, and, in our case, sports outcome predictions.

Thus, we use the weighted Spearman's rank correlation ρ_W to determine how good a set of feature parameter values S is. In order to find the quasi-optimal set of values S_*^y for the parameters, we search the parameter space for a set of values $S_*^y = \{w1_*^y, w2_*^y, w3_*^y, w4_*^y, w5_*^y\}$ that maximizes the value of ρ_W for the years $y-1, y-2, \ldots, y-W_Y$ when we want to predict the team's performance in year y, in a way that

$$S_*^y = \arg\max_{S^y} \left(\sum_{i=1}^{W_Y} \rho_W \left(r_t^{y-i}, f_{NetFor}(t^{y-i}, S^y) \right) \right). \tag{12}$$

In our case, since the parameter values only measure the relative temporal importance from one feature to another, we can search for parameter values in a reduced parameter space. Thus, we execute a linear local maxima search in the parameter space using the coordinate ascent optimization algorithm [Fessler and Hero 1994] to find the quasi-optimal set S_*^y for every year $y-1, y-2, \ldots, y-W_Y$ of our dataset. Each parameter wi starts with the value 0.1 and linearly grows by 0.1 until the local maxima point is found. We emphasize that we tried different parameter spaces and different optimization algorithms and the results were similar.

5.3. NetForY

One advantage of the NetFor is that it is entirely based on implicit measures, which are generally thought to be less accurate than explicit measures [Nichols 1998], but as large quantities of implicit data can be gathered at no extra cost to the user, they are attractive alternatives. Moreover, implicit measures can be combined with explicit data to obtain a more accurate representation of the analyzed system [Kelly and Teevan 2003]. Thus, in this direction, we propose the *Network-based Plus Yesterday Forecaster (NetForY)*.

The NetForY is a simple extension of the NetFor, which together with the network features described in Section 4, also uses the information on the previous performance of the teams. Considering that the network features inform the amount of change a

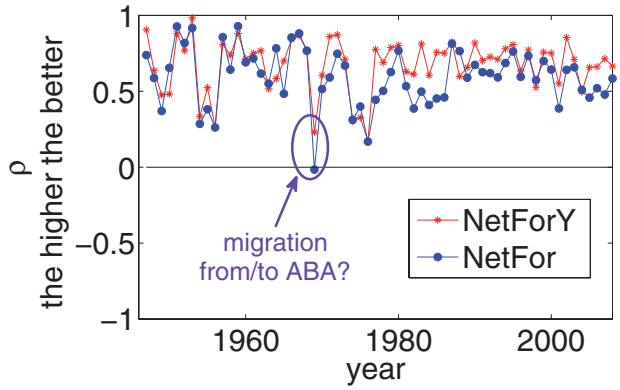


Fig. 11. The proposed network-based models had good results in predicting the entire rank of the teams. Observe Spearman's rank correlation coefficient ρ between Π_t^y and r_t^y. The average correlation $\overline{\rho}$ for the NetForY is 0.68 and for the NetFor it is 0.59, significantly high values.

team made in its roster in a nontrivial way, they can also indicate, using the previous performance information, whether a team will maintain its performance. For instance, if a team had a good performance in year y and the network features indicate that the network changes made in its roster are not significant, then it is very likely that this team will continue to give a good performance in year $y+1$. Thus, we simply gather the five network features used in the NetFor together with the previous performance of the teams r_t^{y-1} in the function $F_{NetForY}$ to compute the prediction score Π_t^y of each team t in year y, in a way that

$$\begin{aligned}\Pi_t^y &= F_{NetForY}(t^y)\\ &= -w1 \times \log(\Delta d_t^y) - w2 \times \log(\Sigma \Delta d_t^y)\\ &\quad + w3 \times \log(ixp_t^y) + w4 \times \log(\overline{cc_t^y})\\ &\quad - w5 \times \log(s_t^y) + w6 \times \log(1 + r_t^{y-1}),\end{aligned} \tag{13}$$

where $F_{NetForY}(t^y)$ is identical to $F_{NetFor}(t^y)$, with the addition of the extra term $w6 \times \log(1 + r_t^{y-1})$. Moreover, the method for assigning values to the parameters $w1$, $w2$, $w3$, $w4$, $w5$, and now $w6$ is the same as the one described in the previous section. To see the parameter values and how they change over time for both network-based prediction models, see the Appendix.

6. RESULTS AND VALIDATION

In this section we describe the results of our proposed network-based models. In Section 6.1, we show the individual results of our models for metrics that are based on the two classes of people who are described in Section 5.2. Moreover, in Section 6.2, we compare the models with other models based on explicit descriptive statistics. For the W_Y parameter, we used a fixed sliding window of 10 years, that is, $W_Y = 10$. We tried other values for W_Y and also the exponential weighted window, which gives more weight for more recent years, but the results are similar. For a W_Y sensitivity analysis, see the Appendix.

6.1. Results

As we mentioned earlier in this section, the basic evaluation metrics are targeted to the two classes of people described in Section 5.2 who can benefit directly from an NBA prediction model. First, we address the *sports analyst* class by showing, in Figure 11,

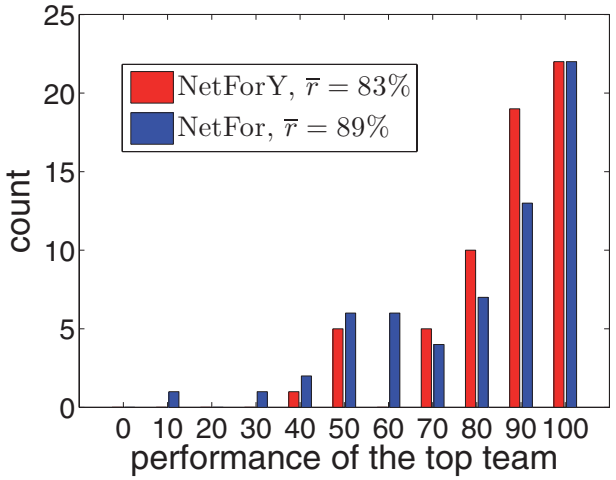


Fig. 12. The network-based models consistently indicate a high performing team as their top team. They correctly predicted the top performing team $\approx 35\%$ of the time. Note that the higher the count values at the right of the distribution, the better.

Spearman's rank correlation ρ^y between the performance r_t^y and the prediction score Π_t^y of every team t in year y. We observe a significative positive rank correlation for either the NetFor and the NetForY, with the NetForY performing better, on average. The average correlation $\overline{\rho}$ for the NetForY is 0.68 and for the NetFor it is 0.59. We also performed, for both models, a linear regression for every pair (y, ρ^y), and we verified that the slope for both regressions is 0. This indicates that there is no particular trend for the performance of the models over time.

Second, we analyze how the network-based models perform according to the interests of the *gambler* class. For this, we verify the performance $r_{t_1}^y$ of the team t_1^y that got the highest prediction score $\Pi_{t_1}^y$ in year y. A perfect prediction model gives, for every year y, the highest prediction score to the team t_*^y that gave the best performance in year y, that is, $\Pi_{t_1}^y = \Pi_{t_*}^y$. Thus, we verify the percentage $hit_{\%}$ of times t_1^y was the actual best performing team t_*^y and also the average performance $\overline{r}$ of the teams t_1^y selected by the model.

In Figure 12, we show the histogram of the performance of the teams t_1^y selected by the model over the 62 years of our dataset $1947 \leq y \leq 2008$. As an example, in the histogram, the number of teams with a performance of 80% selected by the NetFor is 7—meaning that in 7 years the model selected as a t_1^y a team with performance greater or equal to 80% and lower than 90%. Thus, we observe that the vast majority of the teams selected by the models are concentrated on the right side of the graph, indicating that the models generally select good performing teams as the most likely ones to perform well in the following season. Moreover, for the NetFor, $hit_{\%} = 35\%$ of the teams selected by the NetFor model are the actual best performing teams and $\approx 68\%$ had, at least, a performance higher than 80%. On the other hand, the NetForY also selected $hit_{\%} = 35\%$ of the times the actual best performing team, but the model is more stable, only once selecting a team with a performance lower than 50%, with 90% of the teams t_1^y having a performance higher than 80%. The average performance $\overline{r}$ of the teams selected by the NetFor is $\overline{r} \approx 83\%$, and by the NetForY it is $\overline{r} \approx 89\%$.

6.2. Comparison

In this section we compare our network-based models with other prediction models, using the two metrics described in the previous section, the average Spearman's rank

correlation coefficient $\overline{\rho}$ and the average performance $\bar{r}$. Before presenting the results, we describe the models we use in the comparison.

Yesterday. This is currently the state of the art for forecasting team sport leagues in general, as we mentioned in Section 2. Yesterday states that a team's performance in year y will be the same as its performance in year $y-1$. Thus, the prediction score Π_t^y of team t in year y is simply the performance r_t^{y-1} of team t in year $y-1$.

Aggregated Box-Score (ABS) Model. This model uses the aggregated value of the three main box-score statistics of team t in year $y-1$: points, rebounds, and assists. Therefore, the prediction score Π_t^y of team t in year y is $\text{points}(t^{y-1}) + \text{rebounds}(t^{y-1}) + \text{assists}(t^{y-1})$. It is important to point out that when we use more sophisticated formulas, such as the APBRmetrics team efficiency difference rating [APBRmetrics], the prediction factor becomes significantly correlated to the prediction factor of the Yesterday model, giving similar results.

Efficiency-based models. We introduce two efficiency-based models, the Eff-1 and Eff-5 model. These models solely use the box-score efficiency of the players in the year $y-1$ to calculate the prediction score for the teams in the year y. The only difference between these models is related to the number of players each one uses to compute the prediction score Π_t^y of team t in year y. In the Eff-1 model, the prediction score Π_t^y of team t in year y is the highest efficiency of a player $p \in t^y$, and in the Eff-5 model it is the average of the five highest efficiencies of players $p_j \in t^y$. Moreover, when $y < 1973$, we compute the efficiency of a player as the sum of his points, assists, and rebounds achieved in a period divided by the total number of games he played in this period. When $y \geq 1973$, we compute the efficiency by using the NBA efficiency per minute (EPM) formula [Paulsen 2006], which can only be computed using more detailed box-score statistics, which were not recorded before 1973.

The adapted Page model. Page et al. [2007] used the 1996-1997 NBA season data to propose the use of Bayesian hierarchical modeling and box-score statistics to determine how each player's position needs to perform in order for his team to be successful. Page et al. showed the most important box-score statistics for each player position in order to improve team performance. Thus, we adapt his model, which predicts the winner game by game, to a season prediction model. We use the player's data from year $y-1$ in the equation of the model described in Page et al. [2007], using the same parameters, to compute the prediction score Π_t^y of team t in year y. In this way, we expect that a team in year y that has more players with high values of box-score statistics in year $y-1$ favorable to their respective positions, according to Page's model, will be more likely to be successful in year y.

As shown in Figure 5, the network's evolutionary behavior differs before and after the year 1976. Thus, we divided the results into three blocks: (i) before the year 1976; (ii) after the year 1976; and (iii) the overall results, between 1947 and 2008. This division is also useful for another two reasons. First, after 1973, more box-score statistics were recorded after each game, such as blocks, steals, and offensive and defensive rebounds, which may collaborate to measure the efficiency of a player more accurately by using the NBA EPM metric, for instance. Hence, this probably helps the efficiency-based models. More importantly, the Pages model uses these box-score statistics to evaluate which team has the highest probability of winning a match, and because of this, we can only evaluate the Pages model after 1973. Second, the work relationships between players and teams – this is exactly what the edges of our networks measure – were significantly different [Bradley 2009] before 1976 than now. For example, before 1976, player contracts had the option clause [Looney 1976] that practically bound a player to his team, with the player released to play for another team only when his

Table II. Comparing the Network Model and Other Prediction Models

Model	year $\leq$ 1976		year $>$ 1976		Overall		
	$\overline{\rho}$	$\overline{r}$	$\overline{\rho}$	$\overline{r}$	$\overline{\rho}$	$\overline{r}$	$hit_{\%}$
NetForY	**0.66**	**88**	**0.69**	**90**	**0.68**	**89**	**35%**
NetFor	*0.62*	**88**	0.57	79	*0.59*	83	**35%**
"Yesterday"	0.57	85	*62*	*88*	*0.59*	*87*	26%
ABS	0.44	78	0.43	74	0.43	76	18%
Page's	–	–	0.52	80	0.52	80	21%
Eff-1	0.37	74	0.41	79	0.39	77	24%
Eff-5	0.40	71	0.42	74	0.41	72	18%

In **bold**, the best result and in *italic* the runner up. Note that our proposed NetForY always achieved the best result. Moreover, NetFor gave a better performance than Yesterday in the early years and in selecting the best performing team.

team let him do so. This means that the implicit feedbacks of network change before and after 1976 may be significantly different.

In Table II we compare the models described in this section with our proposed network-based models. We observe that the network-based models are the ones with the best results. The NetForY has the best results for all metrics in all periods, sometimes tying with NetFor. In comparison with our best competitor, the Yesterday, the NetFor performs better in the first period and worse in the second. This could be explained because the implicit feedback behind network changes before 1976 were much clearer and simpler. In this period, a player would move from one team to another only if he was not performing well for the team.

In general, the NetForY is always the best model, and both the NetFor and the Yesterday perform similarly, which is a fascinating result, since NetFor considers only network features. More specifically, the network-based models are the ones that can achieve the best results—35% of the time—being the best choices for the *gambler* class. Considering the *sports analyst* class, while NetFor has a similar performance to Yesterday, the NetForY achieved $\approx$ 14% improvement, which is an impressive result in terms of predictive power. This shows the potential of combining implicit network features with explicit features.

Concerning the models that look to box-score statistics to calculate their prediction scores, they all presented worse results than the network-based models. The adapted Page model, as expected, presented better results than the efficiency models and the ABS model, since it is significantly more sophisticated. However, the fact that the efficiency models improved their results after 1973 is quite interesting. There are two possible reasons for this. First, more high box-score statistics players have grown in the NBA over the years, and are more determinant to their team's success. Second, the new box-score statistics being recorded over the years are improving the characterization of the players' performance in a game and, consequently, their importance to their team's success.

7. GENERALITY

In the previous section, we showed that the proposed network-based prediction models give significantly good results when compared to other approaches. Moreover, since the models are parsimonious, that is, they require a single parameter,[6] and do not rely on any particular box-score statistics, so they are ready to be used in any system that shares similar characteristics with the NBA. Thus, in this section, we verify the performance of the network-based models in the network formed from the North

[6]The wi parameters are calculated deterministically using this single parameter W_Y.

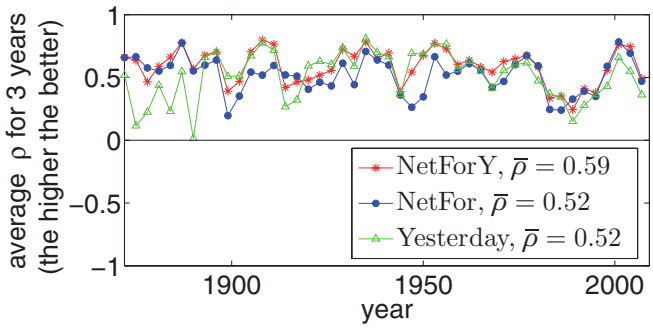
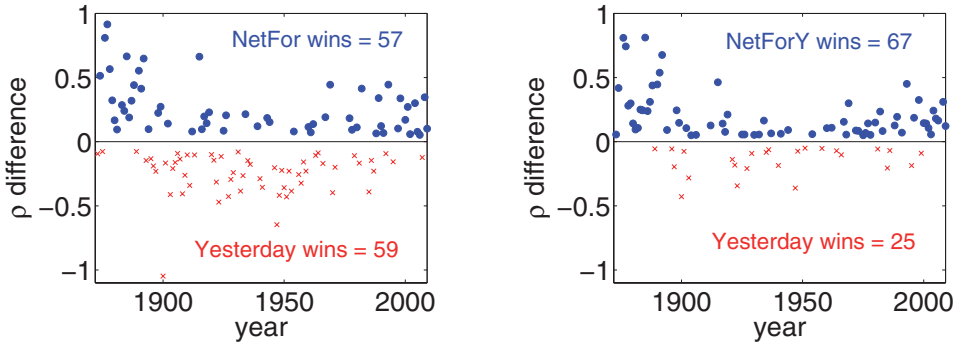


Fig. 13. The average ρ between Π_t^y and r_t^y for three consecutive years. The global average correlation $\overline{\rho}$ for NetForY is 0.59 and for NetFor and Yesterday models is 0.52.



(a) NetFor vs. *"Yesterday"*. (b) NetForY vs. *"Yesterday"*.

Fig. 14. The difference $\Delta\rho = \rho_{M1} - \rho_{M2}$ between the ρ coefficients achieved by the two models $M1$ and $M2$.

American Major League Baseball (MLB) dataset using the same value for the W_Y parameter, that is, $W_Y = 10$.

We use the MLB dataset which is publicly available in Sean Lahman's Baseball Archive site [Lahman 2008]. It contains a huge amount of information about all the teams and players from 1871 to 2009, a total of 139 years of data! In our case, the only information we need is the roster and the ranking of each team per year. With this, we are able to construct the MLB networks and also verify the performance of our proposed network-based prediction models, as we did for the NBA dataset.

In Figure 13, we show the average ρ between Π_t^y and r_t^y for three consecutive years[7] for NetForY, NetFor, and Yesterday models (the latter was our best competitor in the NBA dataset). In a first look, we observe that the behavior of the three models is very similar, with constant positive correlation values for the 138 years of the analysis. However, a closer look reveals that the NetForY model is robust to the oscillations of both the NetFor and the Yesterday models, showing a more regular behavior. This can be verified by computing the global average correlation $\overline{\rho}$, which is ≈ 0.59 for NetForY, whereas it is 0.52 for NetFor and the Yesterday models.

The better performance of NetForY can also be verified from Figure 14, which shows the difference $\Delta\rho = \rho_{M1} - \rho_{M2}$ between the ρ coefficients achieved by two models $M1$ and $M2$. We can observe that, while NetFor and Yesterday performed, similarly NetForY had almost three times higher rank correlation coefficients ρ than the Yesterday.

[7]This was done to ease visualization.

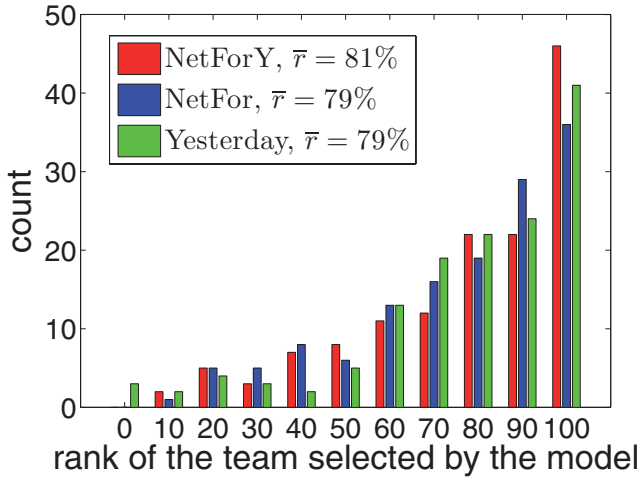


Fig. 15. Performance of the network-based models and of Yesterday model in identifying the best performance team.

Finally, as verified when analyzing the NBA dataset, in the MLB dataset, while the Yesterday model presented an irregular behavior, the NetFor is the best model for the first years of the league, showing constantly high correlation values. This may indicate that for leagues where the rules and dynamics are not yet established, the NetFor may be the preferred choice for predicting the behavior of the teams.

Moreover, in Figure 15, we show the behavior of the network-based models and the Yesterday model in identifying the best performing team; we also show the histogram of the performance for the teams t_1^y selected by the model over the 138 years of our dataset, $1872 \leq y \leq 2009$. Again, we observe that NetForY presents the best results, with an average performance $\overline{r} = 0.81$, while NetFor and Yesterday again had a similar performance, with $\overline{r} = 0.79$. Moreover, NetForY could identify the best performing team, on average, once every three years, with $hit_{\%} = 33.3\%$, while NetFor identified it 26% of the time and Yesterday, 29% of the time.

In summary, we can see that the network-based models gave good results, even when applied to the MLB dataset that comprises 139 years of data and is related to a sport that was one of the most difficult to predict [Ben-Naim et al. 2007]. We believe that this test validates our proposed models, and puts the NetForY model as the state of the art for automated predictive models for general team sports leagues, since it consistently showed better results when compared to the Yesterday model that was, to our knowledge, the best so far. More important is the finding that network effects have an enormous potential to describe the evolution of complex and dynamic social systems.

8. CONCLUSIONS AND FUTURE WORK

In this work, we proposed the use of network implicit feedbacks to aid in the prediction of team behavior in sports leagues. We proposed five temporal network features and, from them, we described two network-based models: the NetFor, which is entirely based on these features, and the NetForY, which also considers the information on the previous performance of the teams. We analyzed the proposed models in two of the most popular professional sports leagues in the United States, the National Basketball Association (NBA) and the Major League Baseball (MLB). In both leagues, the network-based models presented consistently good results, with the NetForY being the best

Table III. Pearson's Correlation Coefficient between Each Pair of Network Features

Features	ixp_t^y	Δd	$\Sigma\Delta d$	$\overline{cc_t^y}$	s_t^y
ixp_t^y	1	0.12	0.01	0.47	−0.19
Δd	0.12	1	0.69	0.54	0.58
$\Sigma\Delta d$	0.01	0.69	1	0.56	0.88
$\overline{cc_t^y}$	0.47	0.54	0.56	1	0.21
s_t^y	−0.19	0.58	0.88	0.21	1

model for all metrics considered for both datasets when compared to other models in the literature.

In summary, the main contributions of this article are as follows.

—We showed that only a small fraction of players have significantly high box-score statistical values and, moreover, their acquisition does not guarantee a team's performance improvement.
—We proposed five general network features that are good indicators of team performance in sports leagues. For instance, we showed that a high team volatility and roster size are not good for the team performance.
—We proposed the NetFor and the NetForY network-based models to predict the behavior of teams in sports leagues. The models presented surprisingly good results when compared to other approaches. The NetForY model presented better results than the current best model, with $\approx$ 14% accuracy improvement in predicting the next years rank.
—The network-based models are generic and may be applied to any team's sports league, that is, they do not rely on any particular box-score statistic.

As future work, we plan to apply network implicit feedback features to analyze other systems besides sports leagues, like recommendation systems or business-oriented social networks, like the *LinkedIn* network. We believe that the concepts presented in this work may be directly applied to other competitive systems and may lead to expressive results in different kinds of applications.

APPENDIX

A.1. Feature Independence

Table III shows the Pearson's correlation coefficient between each pair of network features. Since they are correlated the performance of the teams, it is expected that they are all correlated with each other. However, an exception is made for the *roster aggregate volatility* and the *roster size*, since all the correlations are not high enough for us to say that one metric is fully dependent on another one. In the case of the *roster aggregate volatility* and the *roster size* metrics, we show in Figure 16 Pearson's correlation coefficient over time. We observe that, although in some years the correlation is almost 1, in other years the correlation is small enough for us to say that both metrics are fairly independent and may contribute individually to the network-based prediction models.

A.2. Random Model Comparison

Here we validate the network-based models by comparing them with a null model. The objective of this is to verify whether it is possible that the results of our models came from randomness, that is, overfitting. We define the null model M_0 as a model that sets a uniformly random distributed number to the prediction score Π_t^y of team t in year y,

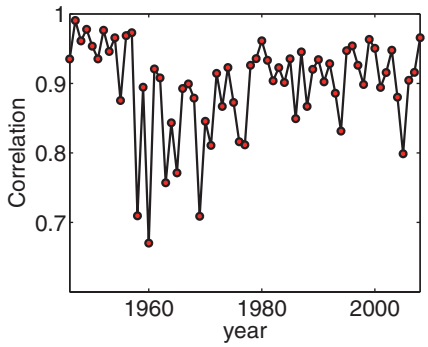
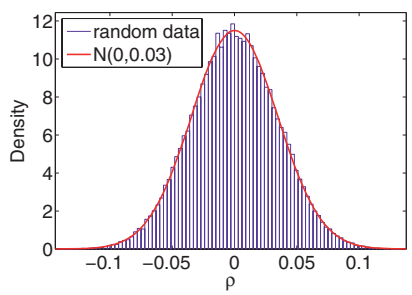
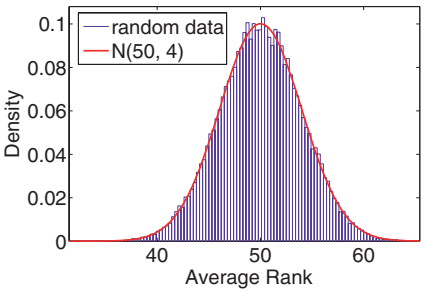


Fig. 16. Pearson's correlation coefficient between the *roster aggregate volatility* and the *roster size* features over time.



(a) Spearman $\overline{\rho}$ distribution. The distribution was fitted to a normal distribution $N_{\overline{\rho}}(0, 0.03)$.



(b) Average performance $\overline{r}$ distribution. The distribution was fitted to a normal distribution $N_{\overline{r}}(50, 4)$.

Fig. 17. Distributions of the null model results over 100,000 simulations.

in a way that

$$\Pi_t^y = f_0(t_y) = U(0, 1).$$

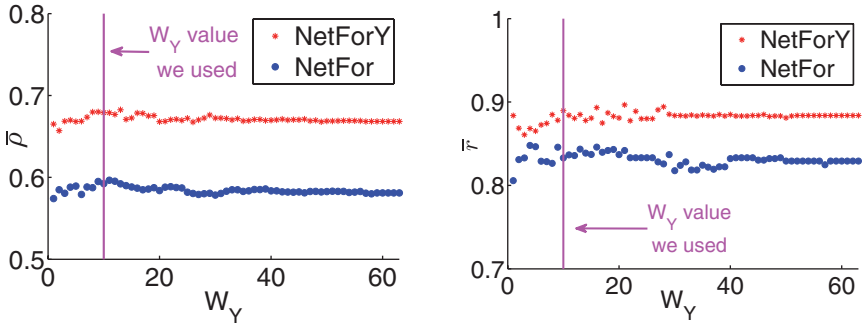
For this model, we ran 100,000 simulations in a way that one simulation is the prediction of every season between 1947 and 2008 using the null model. In order to analyze the behavior of the null model, we define two random variables, X_ρ and $X_{\overline{r}}$, in a way that

- $[X_\rho =]$ the average Spearman rank correlation coefficient $\overline{\rho}$ of a simulation;
- $[X_{\overline{r}} =]$ the average selected performance $\overline{r}$ of a simulation;

Figure 17 shows the distribution of the results of the null model for 100,000 simulations. Figure 17(a) shows the density function of the random variable X_ρ and Figure 17(b) shows the density function of the random variable $X_{\overline{r}}$. In red lines, we show the maximum-likelihood fitting of these density functions, where the random variables X_ρ and $X_{\overline{r}}$ were fitted, respectively, to the normal distributions $N_\rho(0, 0.03)$ and $N_{\overline{r}}(50, 4)$.

Now we validate the network model by verifying whether the results of the network-based models could be generated from the null model. Since the average results of NetForY are always better than the results of NetFor, we will only compare the null model with NetFor. Thus, we define the null hypothesis H_0 which states that the NetFor is an instance of the null model, in a way that

$$H_0 : M_{NetFor} \subset M_0.$$



(a) W_Y vs. $\overline{\rho}$ (b) W_Y vs. $\overline{r}$

Fig. 18. The proposed network-based models are not significantly sensitive to W_Y. Observe that we could even have used values of W_Y that would give slightly higher results than the ones we used in the previous sections.

In order to reject H_0, we verify the probability of making a simulation with the null model that leads to an average Spearman rank correlation coefficient $\overline{\rho} > 0.59$ and an average selected performance $\overline{r} > 0.83$ that were the results obtained by NetFor. Thus, based on the fitted distributions N_ρ and N_γ, we calculate the following probabilities:

$$P(X_\rho > 0.59) = 0$$
$$P(X_{\overline{r}} > 0.83) = 0$$

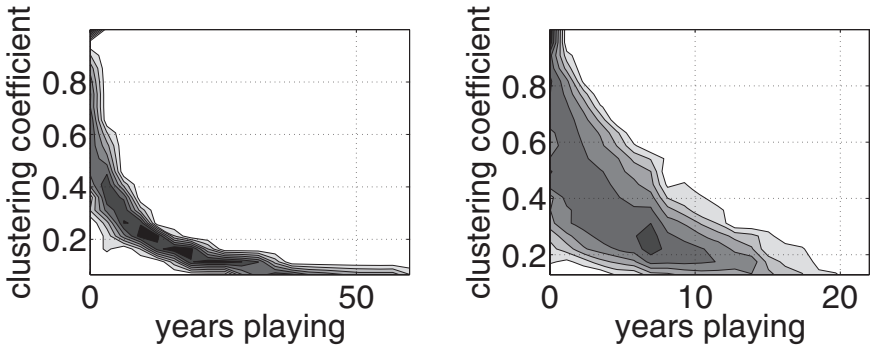
By these probabilities, we can conclude that the network model result cannot be a result that came from randomness or from the null model, and therefore we reject the null hypothesis H_0. We performed same test for the MLB dataset, and the conclusions were the same.

A.3. Parameter Sensitivity Analysis

We also investigate the sensitivity of the models with their only parameter W_Y, since there is the possibility that the results shown so far had came from a lucky W_Y. In Figure 18, we plot the $\overline{\rho}$ of the NetForY and NetFor when the parameter W_Y is varied. We observe that these models are not significantly sensitive to W_Y. Observe that we could even have used values of W_Y that would give slightly higher results than the ones we used in the previous sections. This shows that the task of selecting a good value for W_Y is quite simple. We should only avoid significantly small values that could give high importance to anomalous years, or significantly high values which may fail to capture the temporal effects. However, as we show in Figure 18, even extreme values of W_Y give results that are, in general, better than the results achieved by our competitors. We emphasize that we performed same analysis for the MLB dataset, and the results were similar.

A.4. Network Features to Node Attributes

Another interesting application for the network features is to reverse engineer from network effects to individual node attributes. Such reverse engineering can aid in the analysis of systems that have missing or erroneous explicit data such as online social networks or auction networks. In our case, we found a significative correlation between the age of a node and its clustering coefficient when using the *historical* propagation model, as we can observe in Figure 19. In fact, this nonlinear correlation is one of the main reasons for the *team inexperience* feature to aid in the prediction of team behavior.



(a) Team age vs. clustering coefficient. (b) Player age vs. clustering coefficient.

Fig. 19. Relationship between the clustering coefficient and the age of teams and players. In this case, the age is the number of years a team or player was active in the league. Note that there is a clear nonlinear correlation between the clustering coefficient and the age of the nodes.

Although the usual and most natural way to describe a team t experience in year y is by its age a_t^y, that is, the number of years a team has played in the NBA, the amount of experience a team acquires during the years is not a linear function; that is, the amount of experience a team gets after playing its first season is probably significantly higher than the amount it will get by playing its 20th season. When we consider the clustering coefficient as a measure of experience, the experience difference between team $t1$ that has $a_{t1}^y = 0$ and another team $t2$ that has $a_{t2}^y = 2$ is higher than the experience difference between a team $t3$ that has $a_{t3}^y = 20$ and another team $t4$ that has $a_{t4}^y = 22$.

ACKNOWLEDGMENTS

The authors would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico.

REFERENCES

Abbot, H. 2007a. Bad use of statistics is killing Anderson Varejao. *True Hoop*.

Abbot, H. 2007b. Meet adjusted plus/minus. *True Hoop*.

APBRMETRICS. www.apbrmetrics.com.

Barzilai, A. and Ilardi, S. 2008. Adjusted plus-minus: 2007-2008 midseason results. *82games*.

Ben-Naim, E., Vazquez, F., and Redner, S. 2007. Parity and predictability of competitions. *J. Quant. Anal. Sports 2,* 4, 1.

Bradley, R. 2009. Labor pains nothing new to the NBA. *APBR.org*.

Cowan, C. 2006. The line on NBA betting. *Business Week*.

Da Costa, J. P. and Soares, C. 2004. A weighted rank measure of correlation. *Australian New Zealand J. Stat. 47,* 4, 515–529.

Databasesports.com. 2010. Database basketball. www.databasebasketball.com.

Dilger, A. 2002. Never change a winning team: An analysis of hazard rates in the NBA. *SSRN eLibrary*.

Easterbrook, G. 2006. The five-month NFL forecast. *ESPN.com*. http://sports.espn.go.com/nba/playoffs/2009/news/story?id=4135263.

Fast, A. and Jensen, D. 2006. The NFL coaching network: Analysis of the social network among professional football coaches. In *Proceedings of the AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection*.

Fessler, J. A. and Hero, A. O. 1994. Space-alternating generalized expectation-maximization algorithm. *IEEE Trans. Signal Process. 42,* 10, 2664–2677.

Girvan, M. and Newman, M. E. 2002. Community structure in social and biological networks. *Proc. the Nat. Acad. Sci. 99,* 12, 7821–7826.

HAMBACH, W. AND SCHOTTLE, H. 2006. The German sports-betting market: Uncertainty and chaos for private providers like BWIN. *Gaming Law Rev. 10,* 6.

ILARDI, S. 2007. Adjusted plus-minus: An idea whose time has come. *82games.com*.

KELLY, D. AND TEEVAN, J. 2003. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum 37,* 2, 18–28.

KENDALL, M. G. AND GIBBONS, J. D. 1990. *Rank Correlation Methods* 5th Ed. Oxford University Press, Oxford, UK.

LAHMAN, S. 2008. The Lahman baseball database. baseball1.com.

LEWIS, M. 2009. The no-stats all-star. NYTimes.com.

LIGHTMAN, A. 2010. Open prediction: How sports fans can help save the world. *h+ Mag.*

LOONEY, D. S. 1976. The start of a chain reaction? *Sports Illustrated*.

LUCKNER, S., SCHRDER, J., AND SLAMKA, C. 2008. On the forecast accuracy of sports prediction markets. In *Negotiation, Auctions, and Market Engineering*, LNBIP, vol. 2, Springer, Berlin, 227–234.

NBA.COM. 2008. www.nba.com.

NEVILLE, J., SIMSEK, O., JENSEN, D., KOMOROSKE, J., PALMER, K., AND GOLDBERG, H. 2005. Using relational knowledge discovery to prevent securities fraud. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD'05)*. ACM, New York, 449–458.

NEWMAN, M. 2010. The structure and function of complex networks. *ACM Trans. Embed. Comput. Syst. 9*, 4, Art. 39.

NICHOLS, D. M. 1998. Implicit rating and filtering. In *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*. 31–36.

ONODY, R. N. AND DE CASTRO, P. A. 2004. Complex network study of Brazilian soccer players. *Physical Rev. E 70*, 037103.

PAGE, G., FELLINGHAM, G., AND REESE, C. 2007. Using box-scores to determine a position's contribution to winning basketball games. *J. Quant. Anal. Sports 3,* 4, 1.

PANDIT, S., CHAU, D. H., WANG, S., AND FALOUTSOS, C. 2007. Netprobe: A fast and scalable system for fraud detection in online auction networks. In *Proceedings of the 16th International Conference on the World Wide Web (WWW'07)*. ACM, New York, 201–210.

PARK, J. AND NEWMAN, M. E. J. 2005. A network-based ranking system for us college football. *J. Stat. Mech. Theory Exper. 10*, P10014.

PAULSEN, J. 2006. Efficiency per minute. *The Scores Report*.

REHEUSER, R. 2010. Bucking the trend. *NBA Encyclopedia*.

ROSENBAUM, D. T. 2004. Measuring how NBA players help their teams win. *82games*.

SHETTY, J. AND ADIBI, J. 2005. Discovering important nodes through graph entropy the case of Enron email database. In *Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD'05)*. ACM, New York, 74–81.

SPANN, M. AND SKIERA, B. 2009. Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *J. Forecast. 28,* 1, 55–72.

STEKLER, H., SENDOR, D., AND VERLANDER, R. 2010. Issues in sports forecasting. *Int. J. Forecast. 26,* 3, 606–621.

VAZ DE MELO, P. O., ALMEIDA, V. A., AND LOUREIRO, A. A. 2008. Can complex network metrics predict the behavior of NBA teams? In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*. ACM, New York, 695–703.

WEINBERG, A. 2003. The case for legal sports gambling. Forbes.com.

Received August 2010; revised October 2011; accepted April 2012