Avaliação de times de basquete baseada em redes

Gustavo H. A. Santos

[1]Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brasil

gustavo@land.ufrj.br

Abstract. *The use of statistical analysis is essential for many sports. In particular, basketball was revolutionized using data analytics. In this work we consider the modelling of basketball teams using graphs. We collect 6 years of data comprising games during the regular season and the playoffs and we extract different metrics from the networks created using this dataset. We propose a new network structure and we show that its distance is correlated with a team's performance. Finally, we show that the network behaviour changes between the regular season and the playoffs.*

Resumo. *A utilização de análise estatística é cada vez mais importante em diferentes esportes. Em especial, o jogo de basquete foi revolucionado de maneira visível a partir da análise de dados. Neste trabalho, consideramos a utilização de redes para modelar times de basquete. Coletamos dados da temporada regular e dos playoffs de 6 anos e extraímos diferentes métricas das redes criadas a partir deste dataset. Propomos uma nova estrutura de rede e mostramos que a distância obtida possui correlação com o desempenho de um time. Por fim, mostramos que o comportamento da rede muda entre as fases da competição.*

1. Introdução

A utilização de análise estatística em esportes tem causado mudanças profundas no processo de contratação de jogadores, na avaliação de desempenho de times e no gerenciamento das equipes. Em particular, o basquete é um esporte no qual a utilização de *data analytics* gerou uma revolução visível na maneira de se abordar o jogo, denominada "revolução dos 3 pontos" [538 2019, ShotTracker 2019]. A valorização da análise de dados gera uma grande quantidade de informação cada vez mais detalhada sobre os eventos de cada jogo. Recentemente, dados extraídos a partir do rastreamento dos jogadores nas partidas da NBA geram informações precisas sobre a quantidade de passes trocadas entre companheiros de time e a porcentagem de acerto de arremessos em diversas condições.

A disponibilização de informação detalhada sobre passes trocados entre jogadores se iniciou na temporada 2013-14. Como a temporada regular da NBA é composta de 82 jogos por equipe, com até 28 partidas adicionais durante os *playoffs*, uma quantidade relevante de amostras para análise estatística pode ser obtida. A partir dos passes entre jogadores e da eficiência de seus arremessos é natural a modelagem utilizando grafos.

Este trabalho se propõe a realizar uma análise de desempenho de times da NBA baseada em redes complexas. Os dados são coletados a partir da implementação de um *crawler* para a coleta de informação do site oficial da nba [NBA 2019]. A partir destes dados, constrói-se para cada uma das 6 temporadas consideradas (2013-14 a 2018-19) uma rede por time para a temporada regular e outra rede para os playoffs. A partir destas

redes, deseja-se analisar a relação entre a estrutura obtida e o desempenho de um time medido utilizando a diferença média de pontos.

O trabalho é organizado como se segue. Na seção 2 são discutidos os trabalhos relacionados. A seção 3 discute a metodologia utilizada neste trabalho. Os resultados obtidos são mostrados na seção 4. Por fim, o trabalho é concluído na seção 5.

2. Trabalhos Relacionados

Outros trabalhos da literatura utilizam abordagens baseadas em redes para a análise de times da NBA. Em [Piette et al. 2011] considera-se um grafo com pesos em que os nós representam jogadores e a existência de uma aresta representa a utilização simultânea de dois jogadores durante algum jogo. Os autores calculam a centralidade de cada jogador utilizando o método de centralidade de autovetor. Em seguida, utiliza-se um procedimento estatístico para se identificar quando um jogador possui uma performance superior ou inferior a esperada a partir da centralidade obtida.

Em [Skinner 2010] utiliza-se o conceito de ”preço da anarquia”para se estimar a eficiência de um time modelado como uma rede em que os nós são jogadores ou resultados de jogadas. O modelo considera que a eficiência de um jogador descresce de acordo com sua utilização, o que representaria o aumento na previsibilidade de um time. Desta forma, o desempenho ótimo é obtido quando o time varia suas jogadas, incluindo aquelas cuja eficiência instantânea é sub-ótima. Estima-se a função de eficiência por utilização de cada jogador a partir do seu histórico de desempenho e mostra-se como calcular o valor ótimo para a utilização de cada jogada de um time.

Os autores de [Fewell et al. 2012] sugerem uma modelagem de grafo que inspira a rede utilizada nesta proposta. Neste trabalho foram coletadas estatísticas sobre cada jogada executada durante um conjunto de 16 partidas da primeira rodada dos *playoffs* de 2010 da NBA. Nesta modelagem os nós representam os jogadores de uma equipe ou os resultados de jogadas, como acerto de arremesso, *turnover* ou falta. Utilizando este grafo são calculadas métricas como entropia, clusterização e fluxo na rede. A partir destas estatísticas analisa-se a relação entre o estilo de jogo de diferentes equipes e a estrutura observada.

O trabalho de [Fewell et al. 2012] considera poucas amostras, não realiza uma análise quantitativa da relação entre o desempenho do time e as métricas de sua rede e considera apenas jogos de *playoffs*. Neste trabalho são extraídas métricas de rede obtidas a partir de 6 temporadas de jogos de *playoffs* e temporada regular, incluindo métricas observadas em [Fewell et al. 2012]. Em seguida, relaciona-se cada métrica de interesse com o desempenho de um time medido através da diferença média de pontos.

3. Metodologia

A metodologia adotada neste trabalho é descrita nesta seção. O processo de extração de dados é mostrado na seção 3.1. As métricas de rede consideradas são detalhadas na seção 3.2. Por fim, o método utilizado para correlacionar a estrutura da rede com o desempenho de um time é explicado na seção 3.3.

3.1. Extração de dados

Utiliza-se neste trabalho o conjunto de dados relativos a *tracking* de jogadores disponível publicamente no site da NBA [NBA 2019]. Em especial, considera-se o *dataset* de passes entre jogadores e eficiência dos arremessos de um jogador condicionada no jogador que realiza o passe. A extração dos dados é realizada utilizando um *crawler* implementado em python a partir da API nba_api [nba_api 2019]. São considerados os dados de todas as 6 temporadas completas disponíveis no site (2013-14 a 2018-19). Também é extraída a quantidade de minutos jogados por cada atleta durante a temporada regular e os *playoffs* para que seja possível criar redes utilizando apenas os jogadores mais usados.

3.2. Métricas de rede consideradas

Consideramos neste trabalho 4 métricas extraídas a partir da estrutura da rede. Avaliamos três métricas propostas por [Fewell et al. 2012] (centralidade, entropia e *uphill/downhill flux*) e propomos uma nova métrica baseada na distância na rede entre jogadores e o acerto de arremessos.

A métrica de centralidade tem como objetivo identificar se um time possui um jogador cujo grau de saída é consideravelmente maior que os demais. A fórmula para o cálculo da centralidade do time é dada por:

$$C = \sum_{v \in V} \frac{deg(v^*) - deg(v)}{|V| - 1} \tag{1}$$

Altos valores para esta métrica indicam um time que concentra a movimentação da bola em um jogador central. Por outro lado, valores baixos indicam times em que a responsabilidade é divida entre múltiplos jogadores. Esta métrica é calculada a partir de um grafo cujos vértices correspondem a jogadores ou a ação de arremesso. Os pesos das arestas são normalizados de forma que a soma dos pesos do grafo é igual a 1.

Outra métrica considerada é a entropia da rede. O objetivo desta métrica é medir o grau de imprevisibilidade de um time. A entropia é calculada por:

$$S = -\sum_{p \in P} plog(p) \tag{2}$$

Em [Fewell et al. 2012] discute-se diferentes possibilidades para o cálculo de entropia de um time. Neste trabalho consideramos um grafo dado por uma cadeia de Markov cujos nós representam jogadores ou a ação de arremesso e as probabilidades de transição correspondem a fração de vezes que cada jogador toma uma ação. Nesta modelagem o estado associado a arremesso é um estado absorvente, o que inviabiliza a utilização da probabilidade em estado estacionário da cadeia para o cálculo da entropia. Desta forma, consideramos a soma das entropias associadas a cada jogador, ou seja, a entropia das probabilidades de transição associadas a cada jogador.

A métrica de *Uphill/downhill flux* captura a tendência de um time em movimentar a bola para os jogadores cujo arremesso é mais eficiente. Esta medida é calculada por:

$$F = \sum_{i \neq j} p_{ij}(x_j - x_i) \tag{3}$$

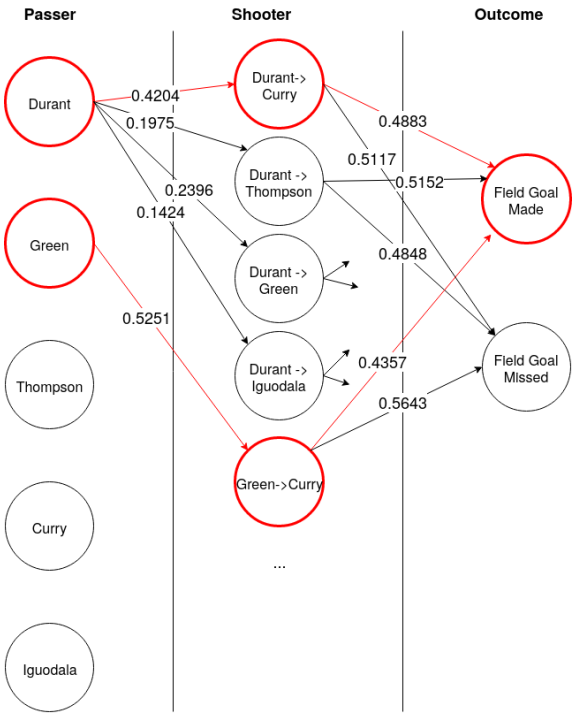


Figura 1. Grafo de distância até a cesta (caminho ótimo em vermelho)

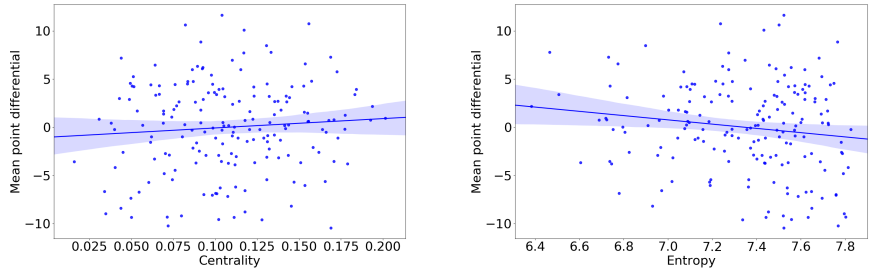
Onde p_{ij} corresponde a probabilidade do jogador i passar a bola para o jogador j e $(x_j - x_i)$ corresponde a diferença de eficiência no arremesso dos jogadores j e i. Para o cálculo desta métrica considera-se um grafo cujos vértices são dados por jogadores ou o resultado de um arremesso (sucesso ou falha) e os pesos das arestas são dados pela fração de vezes em que um jogador realiza cada ação.

A métrica de *uphill/downhill flux* pode penalizar times em que o jogador que realiza o passe também é eficiente [Fewell et al. 2012]. A partir desta observação, propõe-se neste trabalho uma nova métrica baseada na distância na rede entre cada jogador e a cesta (arremesso bem sucedido). Para o cálculo desta métrica considera-se uma nova rede de três camadas exemplificada na Figura 1. Na primeira camada consideramos os jogadores que iniciam as jogadas com passes. A segunda camada considera jogadores que arremessam a bola. Por fim, a terceira camada representa o resultado do arremesso. Os pesos das arestas entre a primeira e a segunda camada são obtidos a partir da fração de passes entre um jogador e um arremessador, enquanto os pesos que conectam a segunda e a terceira camada são dados pela eficiência do arremessador.

A ideia da métrica proposta é extrair da rede quais são os times que são mais eficientes para alcançar o melhor resultado (arremesso bem sucedido). Calcula-se para cada jogador o caminho de menor distância até o arremesso bem sucedido, onde a distância é dada pelo inverso da soma dos pesos das arestas. É importante ressaltar que esta rede considera a eficiência do arremessador condicionada no jogador que passa, uma *feature* importante presente em nosso *dataset* que não foi utilizada por [Fewell et al. 2012].

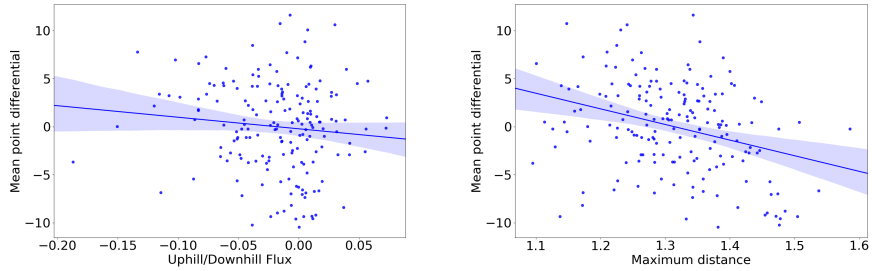
3.3. Relação entre estrutura da rede e desempenho do time

Para comparar a relação entre a estrutura da rede e o desempenho do time realizamos uma regressão linear considerando como *feature* uma das métricas obtida a partir da rede e



(a) Centralidade vs Diferença média de pontos

(b) Entropia vs Diferença média de pontos



(c) *Uphill/downhill flux* vs Diferença média de pontos

(d) Distância máxima vs Diferença média de pontos

Figura 2. *Scatterplot* entre métricas e desempenho do time na temporada regular

como *target* a diferença média de pontos de cada time. Utilizamos a regressão de Ridge para calcular o RMSE (*root mean squared error*) obtido por cada métrica durante a previsão. Os resultados são obtidos utilizando o método de *leave-one-out cross validation* em que cada *fold* corresponde a uma das temporadas, ou seja, aplica-se o método de *cross-validation* considerando 6 *folds*.

4. Resultados

Para a avaliação dos resultados consideramos as três métricas propostas por [Fewell et al. 2012] e a distância máxima utilizando a rede proposta neste trabalho. Outras medidas de distância foram consideradas (como distância mínima e média), mas a distância máxima obteve melhores resultados.

A Figura 2 mostra o *scatterplot* entre cada medida avaliada e a diferença média de pontos. A partir dos gráficos é possível observar que a centralidade apresenta baixa correlação com o desempenho do time, enquanto a entropia e o *uphill/downhill flux* apresentam uma leve correlação. A distância máxima é a métrica que apresenta maior correlação com a diferença média de pontos. Para quantificar a correlação entre cada métrica e o desempenho do time consideramos o erro obtido pela regressão linear, apresentado na Tabela 1. Estes resultados confirmam que a métrica de distância proposta neste trabalho é a que possui o maior poder preditivo.

Para adquirir intuição sobre o funcionamento do modelo consideramos como métrica de acurácia a fração de vezes em que o valor predizido pelo modelo possui o mesmo sinal da diferença média de pontos real. Para isso, ordenamos os times de acordo

Métrica	RMSE
Centralidade	4.695
Entropia	4.648
Uphill/downhill flux	4.609
Distância máxima	4.474

Tabela 1. Erro obtido pela regressão para cada métrica

com o seu diferencial de pontos e medimos a acurácia considerando diferentes subconjuntos de times. As Tabelas 2 e 3 mostram os resultados obtidos. É possível perceber que o modelo possui maior acurácia para os melhores e para os piores times, o que é esperado uma vez que estes times são aqueles que possuem a maior/menor diferença média de pontos. No entanto, quando consideramos times cuja diferença média é próxima de zero a eficácia do modelo diminui.

Conjunto de times	Acurácia	Menor diferença de pontos média
25 melhores	0.64	4.56
50 melhores	0.64	3.07
75 melhores	0.56	0.94
Todos os times	0.567	-10.45

Tabela 2. Acurácia - melhores times

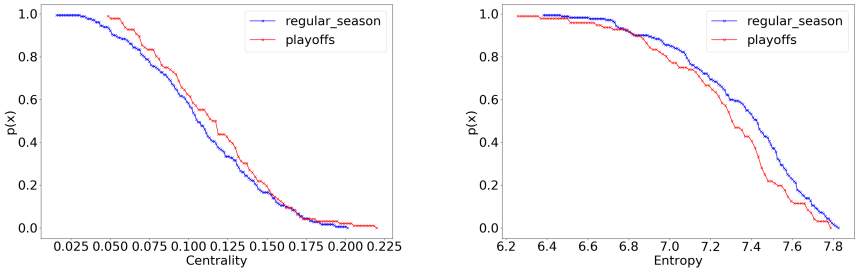
Conjunto de times	Acurácia	Maior diferença de pontos média
25 piores	0.72	-5.72
50 piores	0.64	-2.73
75 piores	0.59	-0.50
Todos os times	0.567	11.63

Tabela 3. Acurácia - piores times

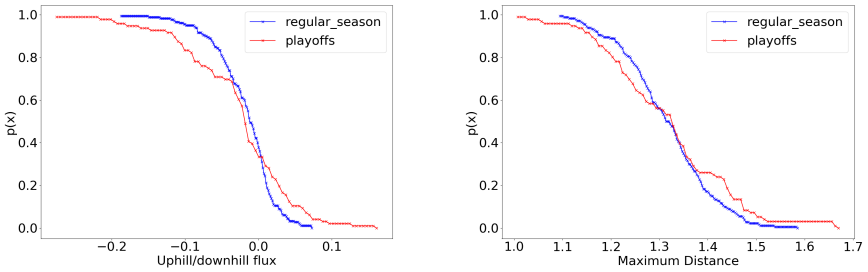
Por fim, consideramos a diferença da rede durante os *playoffs* e a temporada regular. A Figura 3 mostra que a distribuição das diferentes métricas de rede apresenta mudanças entre as fases da competição. Para avaliar de maneira quantitativa, consideramos um modelo treinado utilizando dados de distância máxima da temporada regular e avaliamos o seu desempenho considerando apenas dados dos *playoffs*. O erro do modelo cresce neste cenário, alcançando um RMSE de 4.996. A Figura 4 ajuda a entender o aumento no erro observado. Percebe-se que a correlação entre a distância e a diferença de pontos é mais baixa quando se considera os jogos de *playoffs* e o modelo perde capacidade preditiva, o que indica a mudança de comportamento entre as fases da competição.

5. Conclusão e Trabalhos Futuros

Neste trabalho consideramos a modelagem de times de basquete utilizando redes e avaliamos a relação entre a estrutura desta rede e o desempenho do time. Propomos uma métrica de rede baseada na distância entre jogadores e o sucesso de um arremesso e mostramos que esta nova medida possui maior poder preditivo do que outras métricas propostas na literatura (entropia, centralidade e *uphill/downhill flux*). Mostramos que o modelo preditivo possui acurácia maior para os melhores e para os piores times. Por fim, mostramos que a relação entre a distância máxima e o desempenho do time observada durante a temporada regular não se mantém durante os jogos de *playoffs*.



(a) ECCDF - Centralidade (b) ECCDF - Entropia



(c) ECCDF - *Uphill/downhill flux* (d) ECCDF - Distância máxima

Figura 3. ECCDF das métricas de rede - temporada regular vs playoffs

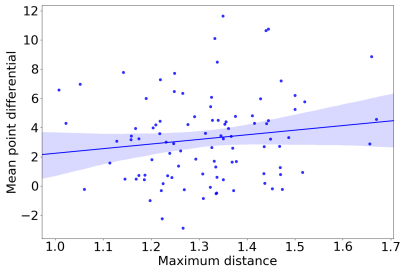


Figura 4. Distância máxima vs Diferença média de pontos - Playoffs

A utilização de redes para a modelagem de partidas de basquete se apresenta muito promissora. Diversas possibilidades podem ser consideradas para trabalhos futuros. A avaliação individual de jogadores utilizando métricas de centralidade como *Page Rank* e *betweeness* parece natural a partir das redes observadas. As redes consideradas neste trabalho tem foco em métricas ofensivas, mas é possível considerar o impacto defensivo de cada time ao se avaliar a mudança da estrutura da rede de acordo com o adversário. Por fim, é de interesse avaliar a evolução temporal da estrutura da rede de diferentes franquias.

Referências

538 (2019). How mapping shots in the nba changed it forever. Acessado em 29/10/2019.

Fewell, J. H., Armbruster, D., Ingraham, J., Petersen, A., and Waters, J. S. (2012). Basketball teams as strategic networks. *PloS one*, 7(11):e47445.

NBA (2019). Nba stats. Acessado em 16/12/2019.

nba_api (2019). Api para coleta de dados da nba. Acessado em 29/10/2019.

Piette, J., Pham, L., and Anand, S. (2011). Evaluating basketball player performance via statistical network modeling. In *The 5th MIT Sloan Sports Analytics Conference*.

ShotTracker (2019). The 3-point revolution. Acessado em 16/12/2019.

Skinner, B. (2010). The price of anarchy in basketball. *Journal of Quantitative Analysis in Sports*, 6(1).