Aplicação de Modularidade e *Ball Tree* para Validação e Segmentação de Comunidades e Funções Táticas na NBA

*

1st João Antônio Melo Zacarias
Engenharia De Computação
CEFET-MG, Campus V
Divinópolis, Brasil
joaoantmeloz@gmail.com

2nd Humberto Henrique Lima Cunha
Engenharia De Computação
CEFET-MG, Campus V
Divinópolis, Brasil
humberto17henrique@gmail.com

***Resumo*—A evolução tática da NBA em direção ao "basquete sem posições" tornou obsoletas as classificações tradicionais de jogadores. Este trabalho propõe uma abordagem computacional híbrida para segmentar e validar arquétipos de atletas, combinando Teoria dos Grafos e estruturas de dados espaciais. A metodologia modela a liga como uma rede complexa, onde a detecção de comunidades é realizada através da otimização de Modularidade (algoritmo de Clauset-Newman-Moore), permitindo a emergência orgânica de grupos funcionais. Diferente de abordagens puramente aglomerativas, este estudo introduz uma etapa de validação cruzada utilizando *Ball Trees* para indexação espacial eficiente em alta dimensionalidade. O sistema verifica a consistência dos agrupamentos ao calcular a interseção entre a comunidade atribuída a um jogador e a comunidade de seus k-vizinhos mais próximos (k-NN). Os resultados demonstram que esta arquitetura não apenas identifica clusters latentes de desempenho, mas também quantifica a "pureza" tática dos atletas, distinguindo jogadores representativos de seus arquétipos daqueles com funções híbridas.**

***Palavras-chave*—Detecção de Comunidades; Modularidade; *Ball Tree*; NBA Analytics; k-Nearest Neighbors; Indexação Espacial.**

I. Introdução

No cenário contemporâneo da *National Basketball Association* (NBA), a taxonomia clássica de posições — armador, ala e pivô — tornou-se insuficiente para capturar a complexidade funcional dos atletas. O surgimento de jogadores polivalentes exige métricas que transcendam rótulos nominais e foquem em vetores de desempenho real. No entanto, a simples aplicação de algoritmos de clusterização gera um desafio de validação: como garantir que um grupo matemático represente uma função tática real no jogo?

Este trabalho aborda esse problema através de uma arquitetura dupla. Primeiramente, utilizamos a *Teoria dos Grafos* para mapear a estrutura global da liga. Ao modelar jogadores como nós e similaridades estatísticas como arestas, aplicamos a maximização de **Modularidade** para segmentar a rede em comunidades densamente conectadas. Esta técnica supera limitações de métodos baseados em centroides (como K-Means), pois não exige a definição a priori do número de grupos e respeita a topologia das relações entre os atletas.

Em segundo lugar, abordamos o desafio da recuperação eficiente e validação de similaridade. Em um espaço de alta dimensionalidade (múltiplas estatísticas), a busca linear por jogadores similares é computacionalmente custosa. Implementamos, portanto, a estrutura de dados *Ball Tree* (Árvore de Esferas), que organiza os vetores de características em hiperesferas aninhadas, permitindo consultas de vizinhos mais próximos (*k-NN*) com eficiência logarítmica.

A contribuição central deste artigo reside na interseção desses métodos: utilizamos a *Ball Tree* para validar a robustez das comunidades geradas pela Modularidade. Se um jogador pertence à Comunidade A, mas seus vizinhos mais próximos (identificados pela árvore) pertencem majoritariamente à Comunidade B, o sistema classifica este atleta como um "Híbrido". Essa validação cruzada oferece uma nova ferramenta analítica para *scouting*, permitindo distinguir especialistas de jogadores de transição com precisão quantitativa.

II. Trabalhos Correlatos

A aplicação de computação avançada na análise esportiva apoia-se em três pilares fundamentais: a definição de métricas, a indexação de dados multidimensionais e a análise de redes complexas.

Em 2004, Oliver [5] estabeleceu as bases quantitativas para a avaliação de eficiência, enquanto Alagappan [7] foi pioneiro ao demonstrar matematicamente a insuficiência das cinco posições clássicas do basquete, propondo novas classes através de agrupamento hierárquico. Este trabalho expande essa premissa, substituindo a hierarquia fixa por uma detecção dinâmica de comunidades.

Do ponto de vista algorítmico, a eficiência na recuperação de dados similares é crítica para sistemas de recomendação (no âmbito esportivo, o chamado *scouting*). Embora Bentley [8] tenha introduzido as *k-d trees* para busca associativa,

o desempenho dessa estrutura degrada em altas dimensões. Dolatshah *et al.* [6] discutem a superioridade das *Ball Trees* em espaços métricos complexos, fundamentando nossa escolha por essa estrutura para otimizar as consultas de vizinhança Euclidiana necessárias para a validação dos jogadores.

Por fim, na análise estrutural, Newman [1] formalizou a **Modularidade** como a métrica padrão-ouro para qualidade de partição em redes. A viabilidade computacional dessa métrica em grandes grafos foi possibilitada pelo algoritmo guloso de Clauset, Newman e Moore [2]. Adotamos este método específico para garantir que as comunidades detectadas representem grupos com máxima coesão interna estatística, servindo como a verdade contra a qual as consultas da *Ball Tree* são validadas.

III. Fundamentação Teórica

A metodologia proposta fundamenta-se na combinação de técnicas de pré-processamento estatístico, geometria analítica e teoria dos grafos. Esta seção detalha os modelos matemáticos utilizados para transformar dados brutos de performance em estruturas de conhecimento tático.

A. Pré-processamento de Dados

Os dados da NBA apresentam grandezas heterogêneas. Métricas de volume, como Pontos por Jogo (PPG), possuem magnitude muito superior a métricas de eficiência defensiva, como Roubos de Bola (SPG). Para evitar que variáveis de maior escala enviesem o cálculo de similaridade, foram aplicadas duas técnicas de normalização distintas.

1) Padronização (Z-Score): Utilizada na construção do modelo de grafos e na estrutura de busca. Transforma os dados para que tenham média $\mu = 0$ e desvio padrão $\sigma = 1$. Para um valor x_{ij} (estatística j do jogador i):

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{1}$$

Exemplo Prático: Se a média da liga em tocos é $\mu = 0.5$ e um jogador possui média 2.0, seu Z-Score será alto (ex: $+3.0\sigma$). Isso sinaliza ao algoritmo que este atributo é uma anomalia positiva relevante, pesando tanto quanto um cestinha que faz 30 pontos (também $+3.0\sigma$ em sua respectiva distribuição), equalizando a importância tática de defesa e ataque.

2) Normalização Min-Max: Aplicada exclusivamente para a visualização gráfica de comparação de jogadores (Gráficos de Radar), reescala os dados para o intervalo $[0, 1]$:

$$x'_{ij} = \frac{x_{ij} - \min(X_j)}{\max(X_j) - \min(X_j)} \tag{2}$$

B. Métrica de Similaridade (Distância Euclidiana)

A dissimilaridade tática entre dois atletas é quantificada pela Distância Euclidiana (L_2 *Norm*) no espaço vetorial multidimensional de n estatísticas padronizadas. Dados dois vetores de jogadores p e q:

$$d(p, q) = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2} \tag{3}$$

Exemplo de Aplicação: Considere um cenário simplificado com duas dimensões: Pontos (z_{pts}) e Assistências (z_{ast}).

- **Jogador A:** Alto volume de pontos ($z_{\text{pts}} = 22.0$) e baixo volume de assistências ($z_{\text{ast}} = 3.2$).
- **Jogador B:** Baixo volume de pontos ($z_{\text{pts}} = 11.0$) e alto volume de assistências ($z_{\text{ast}} = 7.8$).

$$d(A, B) = \sqrt{(-11)^2 + (3.3)^2} = \sqrt{142.16} \approx 11.92$$

Quanto menor este valor, maior a similaridade funcional entre os atletas, independentemente de suas posições nominais.

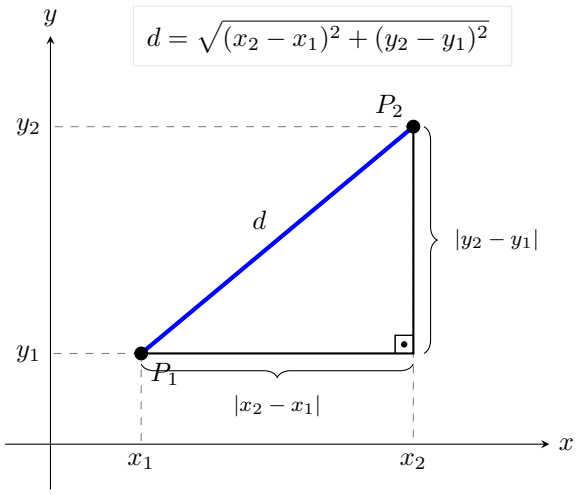

Fig. 1. **Representação Geométrica da Distância Euclidiana.** A distância d entre dois jogadores P_1 e P_2 é calculada como a hipotenusa do triângulo formado pelas diferenças de suas estatísticas normalizadas.

C. Modelagem de Redes e Detecção de Comunidades

A liga (*National Basketball Association*) foi modelada como um grafo não-direcionado $G = (V, E)$, onde cada nó $v \in V$ representa um jogador. A construção das arestas baseou-se no algoritmo k-Nearest Neighbors (k-NN), onde cada jogador se conecta aos seus k pares mais similares.

Para identificar os arquétipos, utilizou-se o método de Maximização de Modularidade Gulosa (Clauset-Newman-Moore). A modularidade Q avalia a qualidade da divisão da rede em comunidades, conforme a formulação definida por Clauset *et al.* [2]:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{4}$$

Onde A_{ij} é a matriz de adjacência, k o grau dos nós, m o número total de arestas e δ a função de Kronecker (1 se i e j estão na mesma comunidade, 0 caso contrário).

Conceito Aplicado: O algoritmo busca maximizar conexões intra-grupo. Se jogadores conhecidos como "Pivôs de Garrafão" (ex: Rudy Gobert) se conectam frequentemente entre si devido a estatísticas similares de rebotes e tocos, a modularidade aumentará ao agrupá-los na mesma comunidade, separando-os naturalmente de "Alas Arremessadores".

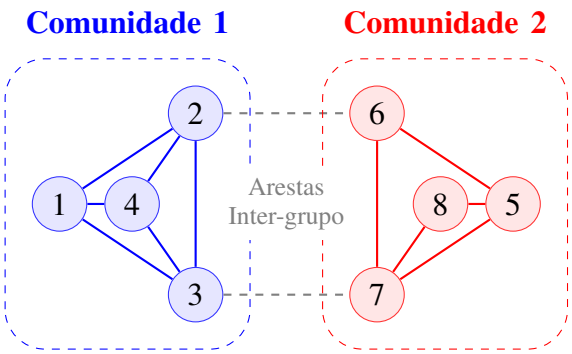

Fig. 2. **Conceito de Modularidade.** A detecção busca maximizar conexões internas (sólidas) e minimizar externas (tracejadas).

D. Indexação Espacial (Ball Tree)

Para viabilizar o sistema de *scouting* em tempo real, os vetores de características foram indexados utilizando uma *Ball Tree* (Árvore de Esferas). Esta estrutura de dados particiona o espaço métrico em hiperesferas aninhadas.

Durante uma busca por vizinhos mais próximos (k-NN), a estrutura utiliza a desigualdade triangular $|d(x, piv\hat{o}) - d(y, piv\hat{o})| \leq d(x, y)$ para realizar a poda (*pruning*) de ramos da árvore.

Vantagem Computacional: Ao descartar hiperesferas inteiras que não podem conter candidatos viáveis (por estarem matematicamente distantes demais do jogador alvo), o algoritmo evita o cálculo de força bruta, reduzindo a complexidade da busca de $O(N)$ para aproximadamente $O(\log N)$.

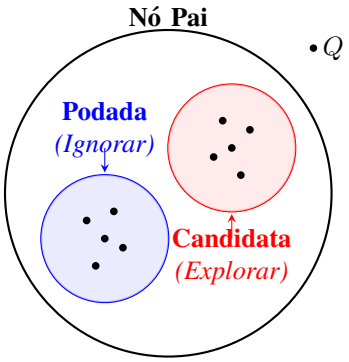

Fig. 3. **Estrutura *Ball Tree*.** O espaço é particionado em esferas. Regiões distantes (esfera em azul) do alvo Q são descartadas integralmente, enquanto regiões próximas (esfera em vermelho) do alvo Q são exploradas, otimizando a busca.

IV. Metodologia

A abordagem proposta segue um pipeline de quatro estágios: (A) Curadoria de dados, (B) Normalização vetorial, (C) Modelagem de redes complexas e (D) Validação cruzada via indexação espacial. A Figura 1 ilustra o fluxo completo do sistema.

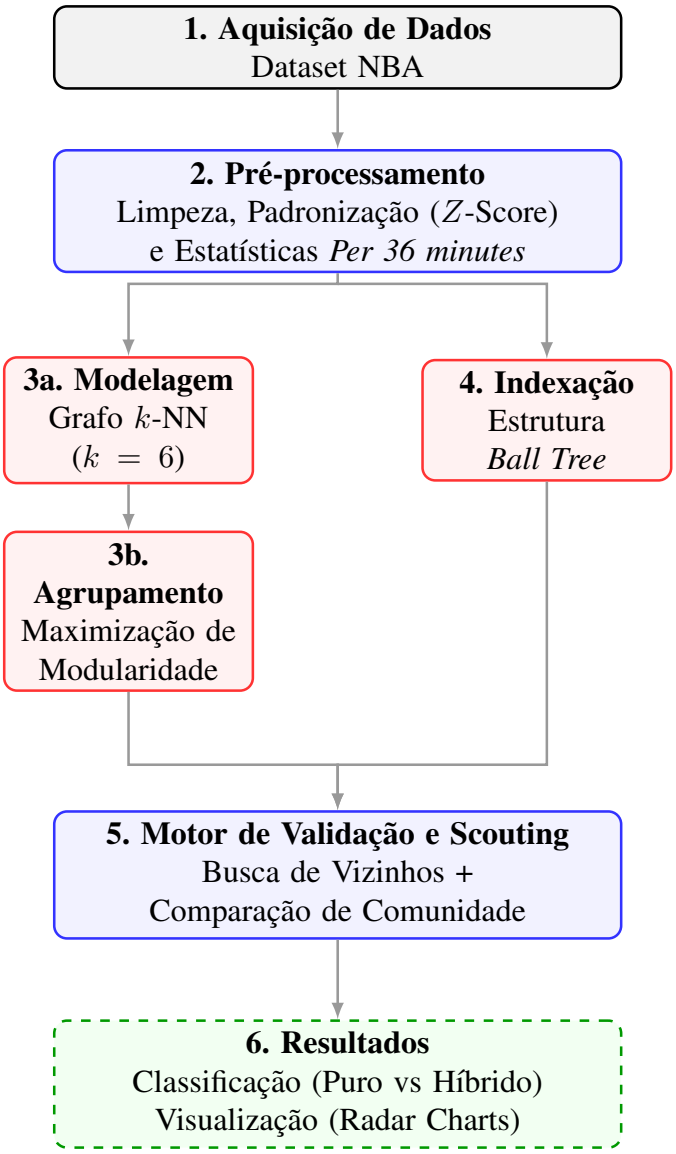

Fig. 4. **Pipeline da Metodologia.** O fluxo inicia com a normalização dos dados. Em seguida, o processo se divide simetricamente: a via da esquerda define os arquétipos (comunidades) e a via da direita indexa o espaço (scouting). Ambas convergem no motor de validação.

A. Aquisição de Dados

O conjunto de dados bruto foi obtido através de coleta automatizada (*web scraping*) do repositório estatístico *NBAStuffer*[1]. Utilizou-se a biblioteca *Pandas* da linguagem Python para a extração direta das tabelas HTML estruturadas disponíveis na página.

[1]Disponível em: https://www.nbastuffer.com/2024-2025-nba-player-stats/

A coleta focou especificamente nos dados da **Temporada Regular 2024-2025**. O *dataset* original contém registros individuais de atletas, abrangendo métricas acumuladas e médias por jogo. Para garantir a consistência da análise tática e evitar distorções causadas por pequenas amostragens (jogadores com tempo de quadra insignificante), aplicou-se uma filtragem inicial nos dados brutos antes da etapa de normalização.

Foram extraídos atributos fundamentais como Pontos (PPG), Rebotes (RPG), Assistências (APG), além de métricas avançadas de eficiência, que serviram de base para o cálculo das estatísticas normalizadas *Per 36 Minutes* detalhadas na etapa de pré-processamento.

B. Pré-processamento e Filtragem

1) Filtragem Inicial: A fim de garantir a relevância estatística e eliminar ruídos causados por atletas com amostragem insuficiente, aplicou-se uma etapa rigorosa de filtragem nos dados brutos. Foram mantidos no *dataset* apenas os jogadores que atenderam simultaneamente aos seguintes critérios durante a temporada regular:

1) Participação em, no mínimo, 41 partidas ($GP \geq 41$), o que corresponde a 50% da temporada regular.
2) Média de tempo em quadra superior ou igual a 20 minutos por jogo ($MPG \geq 20$).

Esta filtragem removeu jogadores que entravam somente ao fim de partidas (*Garbage Time*), jogadores em contratos temporários (contratos de 10 dias) e jogadores que se lesionaram e tiveram baixa amostragem, cuja variância estatística poderia distorcer a formação das comunidades.

2) Normalização Temporal (Per 36 Minutes): Uma limitação crítica das estatísticas médias por jogo é o viés da minutagem: jogadores titulares tendem a acumular mais volume estatístico simplesmente por permanecerem mais tempo em quadra. Para equalizar a comparação entre diferentes funções táticas (ex: um reserva eficiente versus um titular), todas as métricas de volume foram normalizadas para uma base comum de 36 minutos.

A transformação aplicada a cada estatística S de um jogador foi dada pela Equação 5:

$$S_{36} = \left(\frac{S_{jogo}}{M_{jogo}} \right) \times 36 \tag{5}$$

Onde S_{jogo} é a média estatística por partida e M_{jogo} é a média de minutos por partida. A Tabela I apresenta o dicionário das principais variáveis resultantes deste processo, utilizadas posteriormente na construção do grafo.

TABLE I
DICIONÁRIO DE VARIÁVEIS ESTATÍSTICAS (*per 36 min*) UTILIZADAS

Sigla	Descrição
PpG	Pontos por Jogo
RpG	Rebotes por Jogo
ApG	Assistências por Jogo
SpG	Roubos de Bola (*Steals*)
BpG	Tocos (*Blocks*)
$TOpG$	Desperdícios de Bola (*Turnovers*)
$TO\%$	Taxa de Desperdícios (*Turnover Percentage* por posse)
$2PA$	Tentativas de Arremesso de 2 Pontos
$3PA$	Tentativas de Arremesso de 3 Pontos
$3P\%$	Aproveitamento de Arremessos de 3 Pontos
$USG\%$	Taxa de Uso (*Usage Rate*) - Estima a porcentagem de jogadas da equipe usadas por um jogador enquanto ele está em quadra.
$TS\%$	Aproveitamento Real (*True Shooting*) - Mede a eficiência de arremesso levando em conta bolas de 2, de 3 e lances livres.
FTA	Tentativas de Lance Livre
$FT\%$	Aproveitamento de Lances Livres
$eFG\%$	Aproveitamento Efetivo (*Effective Field Goal*) - Ajusta o FG% considerando que a bola de 3 vale mais que a de 2.
$ORtg$	Índice Ofensivo (*Offensive Rating*) - Estimativa de pontos produzidos por 100 posses de bola.
$DRtg$	Índice Defensivo (*Defensive Rating*) - Estimativa de pontos permitidos por 100 posses.

Após este processamento, os dados foram salvos em um arquivo estruturado, servindo de entrada para o algoritmo de normalização escalar (Z-Score) descrito na seção seguinte.

3) Padronização de Escala (Z-Score): Após a normalização temporal, os dados ainda apresentavam heterogeneidade de grandezas. Atributos como Pontos (PpG) possuem amplitude numérica muito superior a atributos defensivos como Roubos de Bola (SpG). Em algoritmos baseados em distância espacial (como k-NN), variáveis com maior magnitude dominam o cálculo do gradiente, enviesando o agrupamento.

Para mitigar este efeito e atribuir peso igualitário a todas as 15 dimensões táticas, aplicou-se a padronização Z-Score em todas as variáveis. O valor transformado z_{ij} para o jogador i na estatística j é dado por:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{6}$$

Onde x_{ij} é o valor original, μ_j é a média da liga para a estatística j, e σ_j é o desvio padrão. O resultado é um espaço vetorial onde todas as métricas possuem média 0 e desvio padrão 1, ideal para o cálculo da Distância Euclidiana subsequente.

Após este tratamento, o fluxo se bifurca em dois processos paralelos: modelagem estrutural e indexação espacial.

C. Etapa 3: Modelagem e Agrupamento (Ramificação A)

Esta etapa foca na análise macroscópica da liga, correspondendo aos blocos 3a e 3b do diagrama.

1) 3a. Modelagem de Grafo k-NN: A matriz de dados normalizados foi convertida em um grafo não-direcionado $G(V, E)$. As arestas foram construídas utilizando o algoritmo *k-Nearest Neighbors* e métrica Euclidiana. Isso significa que cada jogador (nó) foi conectado aos seus k pares estatisticamente mais similares.

2) 3b. Maximização de Modularidade: Sobre a estrutura do grafo, aplicou-se o algoritmo de Clauset-Newman-Moore para detecção de comunidades. O método otimiza a modularidade (Q), particionando a rede em grupos densamente conectados que representam os arquétipos funcionais dos atletas.

D. Etapa 4: Indexação Espacial (Ramificação B)

Paralelamente à criação do grafo, os mesmos vetores normalizados foram processados para otimização de busca. Instanciou-se uma estrutura de dados *Ball Tree* (Árvore de Esferas).

Diferente do grafo, que é estático e foca na estrutura global, a *Ball Tree* organiza os dados hierarquicamente em hiperesferas. Sua função no sistema é permitir consultas de vizinhança (*queries*) com alta performance, viabilizando a etapa seguinte de validação.

E. Etapa 5: Motor de Validação e Scouting

A convergência das duas ramificações ocorre no Motor de Validação. Este módulo opera cruzando as informações obtidas:

- **Input 1 (Do Agrupamento):** O rótulo da comunidade (ID) atribuído a cada jogador.
- **Input 2 (Da Indexação):** A lista dos vizinhos mais próximos recuperados pela *Ball Tree*.

Para cada jogador alvo, o motor calcula a interseção entre sua comunidade e a comunidade de seus vizinhos. Se a maioria dos vizinhos recuperados pela *Ball Tree* pertence à mesma comunidade detectada pelo Grafo, o jogador é classificado como "Puro". Caso contrário, é classificado como "Híbrido", gerando os resultados discutidos na seção seguinte.

V. Resultados e Discussão

Os experimentos foram realizados utilizando o *dataset* da temporada regular 2024-2025, totalizando $N = 213$ jogadores após a filtragem de relevância. A análise dos resultados divide-se em: otimização de hiperparâmetros, caracterização das comunidades detectadas e validação individual via sistema de *scouting*.

A. Otimização de Parâmetros Topológicos

A fidelidade da detecção de comunidades depende intrinsecamente da conectividade do grafo k-NN. A escolha do número de vizinhos k não foi arbitrária, mas guiada por um critério empírico.

Empiricamente, realizou-se uma análise de sensibilidade da Modularidade (Q) variando k no intervalo $[2, 15]$. Conforme ilustrado na Fig. 5, embora valores de $k < 5$ apresentem modularidade elevada ($Q_{max} \approx 0.77$ para $k = 2$), estes resultam em grafos desconexos (muitas comunidades diferentes com poucos jogadores), inúteis para a análise global da liga.

Observou-se um máximo local de estabilidade em $k = 6$ ($Q \approx 0.58$). Este valor representa o ponto de equilíbrio onde os arquétipos estão bem definidos antes que a introdução de ruído (excesso de arestas) comece a degradar a separação dos grupos.

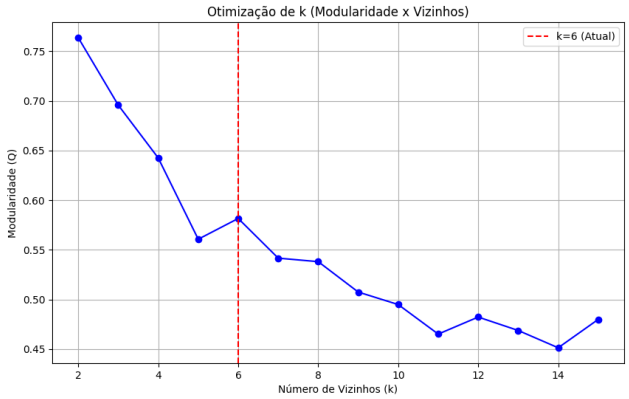

Fig. 5. **Análise de Sensibilidade da Modularidade.** A linha tracejada vermelha marca $k = 6$, escolhido por ser o primeiro pico local que satisfaz o critério de conectividade global ($\ln N$).

B. Caracterização das Comunidades (Arquétipos)

Aplicando-se o algoritmo de Clauset-Newman-Moore com $k = 6$, a rede convergiu para uma divisão em 5 comunidades principais. A visualização espacial (Fig. 6) demonstra a separação topológica dos grupos.

Para interpretar o significado tático de cada cluster, analisamos os centróides (médias) das estatísticas normalizadas. A Tabela II resume os perfis identificados:

TABLE II
Centróides das Comunidades (Transposta)

Métrica	**C0**	**C1**	**C2**	**C3**	**C4**
JG (Jogadores)	58	52	51	39	10
PpG (Pontos)	15.09	14.78	**23.69**	19.22	10.71
ApG (Assist.)	3.50	3.42	**6.09**	3.24	2.10
RpG (Rebotes)	6.39	4.92	5.52	**10.56**	4.42
BpG (Tocos)	0.68	0.43	0.45	**1.50**	0.50
SpG (Roubos)	**1.53**	1.00	1.19	0.99	1.14
$3PA$ (Vol. 3pts)	5.01	7.43	**7.44**	2.96	5.84
$2PA$ (Vol. 2pts)	7.12	4.73	10.63	**10.99**	3.16
$USG\%$ (Uso)	17.96	17.20	**27.58**	21.28	12.63
FTA (Lances L.)	2.67	1.85	**5.24**	4.42	1.05

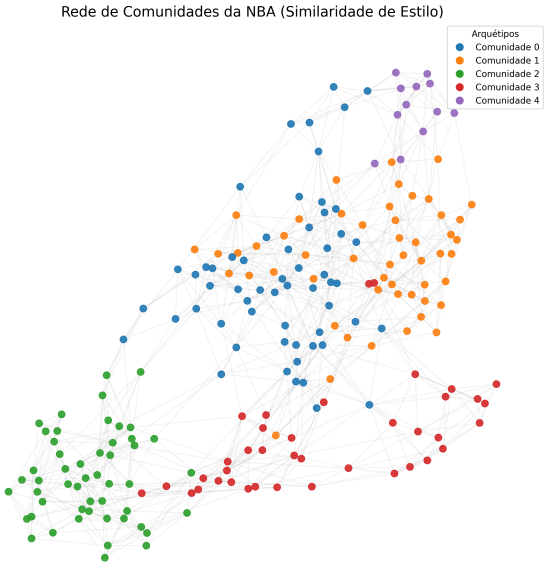

Fig. 6. **Grafo de Similaridade da NBA.** A espacialização (Force-Directed Layout) revela que os jogadores se agrupam organicamente por função, e não por posição nominal em quadra.

- **Comunidade 0 (*Versatile Defenders/Two-Way*):** Este grupo se destaca pela maior média de Roubos de Bola (1.53) da liga, combinada com uma contribuição sólida em rebotes (6.4). São peças de equilíbrio que contribuem positivamente em estatísticas de ataque e defesa sem demandar o volume de bola (USG%) das estrelas.
- **Comunidade 1 (Espaçadores de Quadra):** Especialistas arremessos (principalmente arremessos de 3 pontos), com menor responsabilidade defensiva no garrafão (baixo índice de rebotes e tocos).
- **Comunidade 2: (*Ball Carriers/Handlers*):** Alta taxa de assistências, alto USG% (porcentagem de posses de bola de um time que o jogador utiliza enquanto está em quadra), maior taxa de arremessos entre todas as comunidades (18,07 arremessos tentados somando 2PA e 3PA). São jogadores que controlam o ritmo ofensivo.
- **Comunidade 3 (Dominant Bigs):** Pivôs dominantes no garrafão. Lideram isoladamente em Rebotes (10.6), Tocos (1.50) e volume de jogo interno (11.0 tentativas de 2 pontos a cada 36 minutos), sendo a única comunidade com média de tentativas de 3 pontos abaixo de 3.0. São os âncoras defensivos e finalizadores de curta distância.
- **Comunidade 4 (*Low-Usage Rotation*):** Um grupo pequeno (apenas 12 jogadores) caracterizado por métricas de volume reduzidas em todas as categorias principais (10.7 Pontos e menor USG% entre todas as comunidades). Representam jogadores que entram em momentos específicos ou jogadores em desenvolvimento, que atuam com responsabilidades ofensivas limitadas.

C. Validação de Scouting e Jogadores Híbridos

A aplicação da *Ball Tree* permitiu a validação cruzada dos perfis, testando a robustez das fronteiras entre as comunidades detectadas.

1) Estudo de Caso 1: Ball Carrier (Anthony Edwards): Utilizando o jogador **Anthony Edwards** (Comunidade 2) como pivô de busca, o sistema identificou uma vizinhança homogênea composta exclusivamente por outros criadores de jogadas de elite.

A Tabela III apresenta os 5 vizinhos mais próximos, demonstrando a alta coesão do agrupamento (100% de correspondência com a C2).

TABLE III
VIZINHOS MAIS PRÓXIMOS DE ANTHONY EDWARDS

Rank	Jogador	Distância (d)	Comunidade
1	Donovan Mitchell	1.7228	C2 (Match)
2	Jayson Tatum	1.7663	C2 (Match)
3	Tyler Herro	2.2202	C2 (Match)
4	Jordan Poole	2.3580	C2 (Match)
5	Stephen Curry	2.4045	C2 (Match)

A validação visual é apresentada no gráfico de radar (Fig. 7) e no recorte topológico da rede (Fig. 12), confirmando Edwards como um representante central (”Puro”) do seu arquétipo.

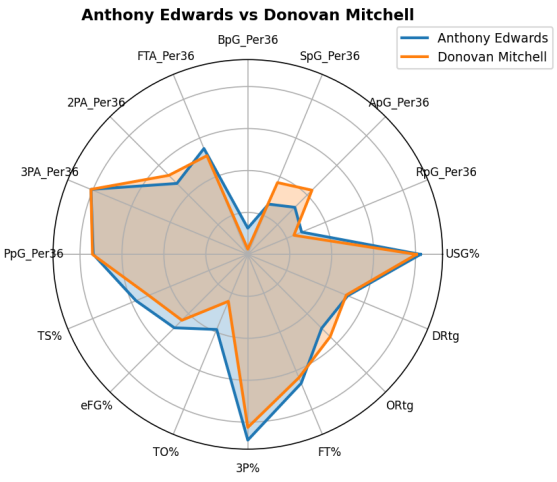

Fig. 7. **Scouting: Edwards vs Mitchell.** A sobreposição quase perfeita das áreas no radar valida a similaridade vetorial, confirmando que ambos desempenham a mesma função tática de volume ofensivo.

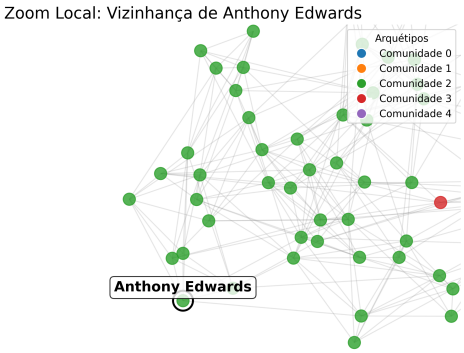

Fig. 8. **Topologia Local (Edwards).** O atleta ocupa uma posição central em um cluster denso da Comunidade 2.

2) Estudo de Caso 2: Espaçador de Quadra (Sam Hauser): Para validar a identificação de especialistas em perímetro, analisou-se **Sam Hauser**, classificado na Comunidade 1. A busca retornou vizinhos com perfil estatístico de alto volume de tentativas de 3 pontos e baixo uso de bola para criação própria (Tabela IV).

TABLE IV
VIZINHOS MAIS PRÓXIMOS DE SAM HAUSER

Rank	Jogador	Distância (d)	Comunidade
1	Isaiah Joe	2.6521	C1 (Match)
2	Aaron Nesmith	2.9592	**C0 (Híbrido)**
3	A.J. Green	2.9745	**C4 (Híbrido)**
4	Payton Pritchard	3.0199	C1 (Match)
5	Amir Coffey	3.2993	C1 (Match)

A análise dos vizinhos híbridos revela a complexidade da função de Hauser:

- **Conexão com C0 (Aaron Nesmith):** Indica que Hauser possui métricas defensivas (*SpG* e *RpG*) ou posicionais semelhantes às de um jogador da comunidade 0, não sendo unidimensional.
- **Conexão com C4 (A.J. Green):** Esta intersecção é crucial. A Comunidade 4 agrupa jogadores de rotação com baixo *Usage Rate*. A proximidade vetorial sugere que Hauser, embora seja um arremessador de elite (C1), opera taticamente como uma peça complementar de finalização, sem a responsabilidade de criação de jogadas, assemelhando-se funcionalmente aos jogadores de rotação da C4.

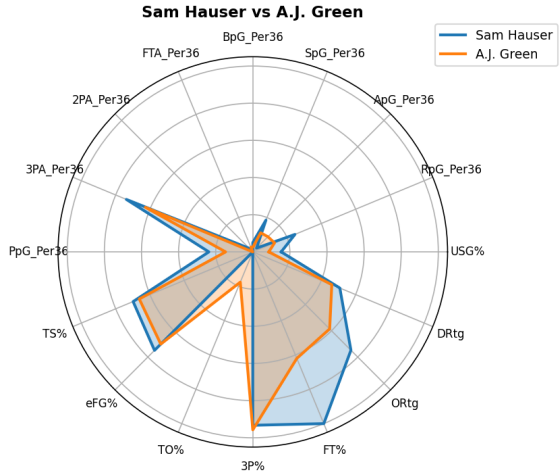

Fig. 9. **Scouting: Hauser vs A.J. Green.** O gráfico de radar evidencia a especialização em arremessos de 3 pontos (pico em 3PA) com baixa criação de jogadas (baixo AST e USG), característica que o aproxima tanto dos especialistas (C1) quanto dos jogadores de rotação (C4).

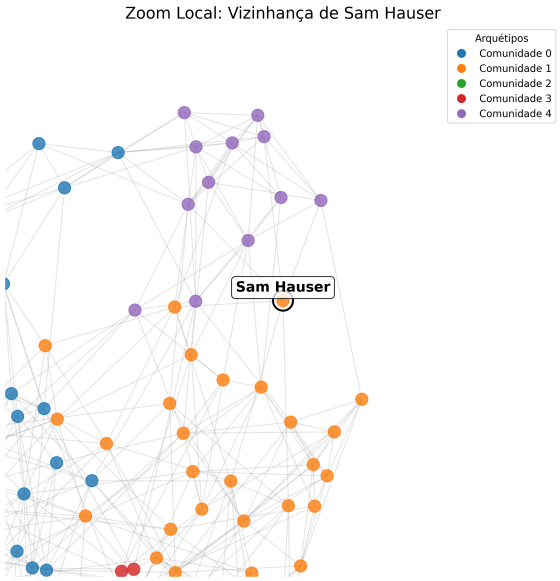

Fig. 10. **Topologia Local (Sam Hauser).** O atleta ocupa uma posição central em um cluster que se divide entre Comunidade 1, comunidade 0 e Comunidade 4

3) Estudo de Caso 3: Jogador Híbrido (Giannis Antetokounmpo): O caso de **Giannis Antetokounmpo** ilustra a capacidade do sistema de lidar com "*outliers*" táticos. Alocado na **Comunidade 3** (*Dominant Bigs*) devido à dominância em rebotes, pontos no garrafão e tocos, sua vizinhança na *Ball Tree* revela uma natureza híbrida complexa (Tabela V).

Diferente dos casos anteriores, onde a distância (d) do primeiro vizinho girava em torno de 1.5 a 2.0, o vizinho mais próximo de Giannis está a uma distância elevada ($d \approx 4.41$). Isso quantifica a singularidade do atleta: não há ninguém "muito perto" dele no espaço vetorial.

A lista de vizinhos confirma o hibridismo funcional:

TABLE V
VIZINHOS MAIS PRÓXIMOS DE GIANNIS ANTETOKOUNMPO

Rank	Jogador	Distância (d)	Comunidade
1	Anthony Davis	4.4178	C3 (Match)
2	Alperen Sengun	5.0834	C3 (Match)
3	Paolo Banchero	5.2120	**C2 (Híbrido)**
4	Shai Gilgeous-Alexander	5.7790	**C2 (Híbrido)**
5	Nikola Jokic	5.8385	**C2 (Híbrido)**

- **Matches (C3):** Anthony Davis e Alperen Sengun validam a base "Big" de Giannis (proteção de aro e finalização interna).
- **Híbridos (C2):** A conexão com Shai Gilgeous-Alexander (um armador da C2), Paolo Banchero (C2) e Nikola Jokic (C2) evidencia que Giannis carrega uma responsabilidade de criação e volume ofensivo ($USG\%$) típica de *Ball Carriers*, transcendendo a função tradicional de um *Big*.

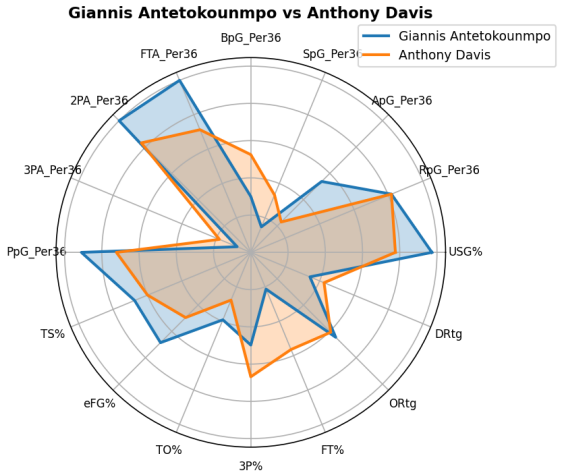

Fig. 11. **Hibridismo Tático.** O perfil de Giannis cobre áreas de força física (C3) e criação de jogadas (C2), conectando-o a armadores e pivôs simultaneamente.

D. Discussão dos Resultados

A segmentação obtida desafia a taxonomia tradicional da NBA. Enquanto a classificação clássica se baseia em atributos físicos (altura) e localização em quadra (armador vs pivô), os agrupamentos detectados via Modularidade revelaram uma organização baseada em *volume* e *função*.

Destaca-se a emergência da Comunidade 4 (*Low-Usage Rotation*) e da Comunidade 2 (*Ball Carriers*). A distinção entre estes dois grupos não é a posição nominal (ambos podem conter armadores ou alas), mas sim a responsabilidade ofensiva ($USG\%$). Isso demonstra que algoritmos não-supervisionados capturam a hierarquia do elenco, separando estrelas de operários, algo que a posição nominal "Ala-Armador" não consegue distinguir.

Adicionalmente, a validação via *Ball Tree* provou que a "posição" de um jogador não é um ponto fixo, mas uma região de probabilidade. Jogadores como Giannis Antetokounmpo

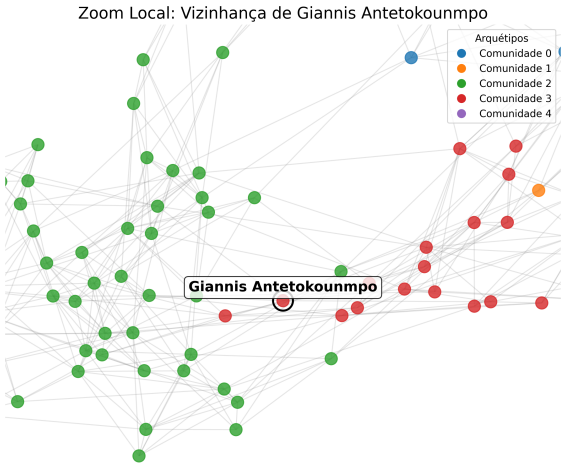

Fig. 12. **Topologia Local (Giannis Antetokounmpo).** O atleta ocupa uma posição central em um cluster que se divide entre Comunidade 3 e Comunidade 2.

(C3 $\rightarrow$ C2) operam como "pontes" na topologia da rede, conectando o jogo físico dos pivôs à dinâmica de criação dos armadores. A identificação desses híbridos é vital para o *scouting*, pois aponta atletas que oferecem versatilidade tática e vantagens de *mismatch*.

VI. CONCLUSÃO

Este trabalho apresentou uma arquitetura computacional para a redefinição de arquétipos na NBA, superando as limitações das posições tradicionais através de uma abordagem baseada puramente em dados. A combinação da análise de redes complexas com indexação espacial mostrou-se eficaz tanto na macro-segmentação quanto na micro-análise de atletas.

A otimização do hiperparâmetro topológico fixou $k = 6$ como o limiar ideal de conectividade, garantindo a detecção de 5 comunidades funcionais robustas. A análise dos centróides revelou que a liga moderna se organiza prioritariamente em torno do volume de arremessos, volume de posse e responsabilidade defensiva, e não apenas pela estatura dos atletas.

Por fim, o motor de validação baseado em *Ball Tree* introduziu uma métrica quantitativa para a "pureza" posicional. A capacidade de distinguir matematicamente jogadores especialistas (como Sam Hauser) de *outliers* híbridos (como Giannis Antetokounmpo) oferece às equipes uma ferramenta analítica poderosa para a composição de elencos e identificação de talentos subvalorizados.

Trabalhos futuros poderão expandir este modelo incorporando dados de rastreamento espacial (*player tracking*), permitindo que a similaridade considere não apenas o resultado estatístico, mas a movimentação dos atletas em quadra.

REFERENCES

[1] M. E. J. Newman, "Modularity and community structure in networks," *PNAS*, 2006.

[2] A. Clauset, M. E. J. Newman, C. Moore, "Finding community structure in very large networks," *Physical Review E*, 2004.
[3] C. M. Keshri, S. Ghosh, "Player Performance Prediction in the NBA Using Machine Learning," *IEEE ICCCNT*, 2019.
[4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA: Morgan Kaufmann, 2011.
[5] D. Oliver, *Basketball on Paper: Rules and Tools for Performance Analysis*. Washington, D.C.: Brassey's, 2004.
[6] M. Dolatshah, A. Hadian, B. Minaei-Bidgoli, "Ball*-tree: Efficient spatial indexing for constrained nearest-neighbor search in metric spaces," *arXiv preprint arXiv:1511.00628*, 2015.
[7] M. Alagappan, "Redefining NBA basketball positions," *MIT Sloan Sports Analytics Conference*, 2012.
[8] d J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.