Player Archetypes and Weighted Network Clustering in the NBA

By Ziyad Gaffar

Motivation

- NBA shifting to *data-driven evaluation* over traditional scouting
- Positional roles becoming **fluid** → need for performance-based archetypes
- Challenge: *How to identify roles and synergy algorithmically?*
- Clustering identifies **similar players**, however
- Does **not** capture *how well they play together*
- Objective:
  - Detect **player archetypes**
  - Model **role-based compatibility**

Framework Overview

- PCA: reduce dimensionality
- k-means++: player clustering
- Build weighted co-occurrence network
- Louvain: detect communities
- Choose best k using modularity frontier

Clustering (k-means++)

- Improves center initialization
- Reduces risk of local minima
- **O(log k)-competitive** cost
- Produces statistically stable archetypes

k-Means++ Algorithm

1) Choose an initial cluster center c_1 uniformly at random from X.
2) **Randomization Step:** Choose the next center c_i by selecting a point $x \in X$ with probability
$$\frac{D(x)^2}{\sum_{y \in X} D(y)^2}.$$
3) Repeat Step 2 until k centers are chosen.
4) Execute Lloyd's heuristic using $c_1, \ldots, c_k$ as the initial centers.

Theorem 3.1. If C is constructed with k-means++, then the corresponding potential function ϕ satisfies::

$$\mathbb{E}[\phi] \leq 8(\ln k + 2)\,\phi_{\mathrm{OPT}}.$$

*Lemma 3.1: If a center is picked uniformly at random from an optimal cluster **A**, the expected clustering cost on **A** is at most twice the optimal cost:*

$$\mathbb{E}[\phi(A)] \leq 2\,\phi_{\mathrm{OPT}}(A)$$

Lemma 3.2: Let A be an arbitrary cluster in C_{OPT}, and let C be an arbitrary clustering. If we add a random center to C from A, chosen with D^2 weighting, then:

$$\mathbb{E}[\phi(A)] \leq 8\,\phi_{\mathrm{OPT}}(A).$$

Lemma 3.3: Let C be an arbitrary clustering. Choose $u > 0$ "uncovered" clusters from C_{OPT}, and let X_u denote the set of points in these clusters. Also let $X_c = X - X_u$. Now suppose we add $t \leq u$ random centers to C, chosen with D^2 weighting. Let C' denote the resulting clustering, and let ϕ' denote the corresponding potential. Then $\mathbb{E}[\phi']$ is at most :

$$\phi(X_c) \;+\; 8\,\phi_{\mathrm{OPT}}(X_u) \cdot (1 + H_t) \;+\; \frac{u - t}{u}\,\phi(X_u).$$

Weighted Networks

$$Q = \frac{1}{2\tilde{m}} \sum_{i,j} \left[w_{ij} - \frac{\tilde{d}(i)\,\tilde{d}(j)}{2\tilde{m}} \right] \delta(c_i, c_j),$$

- Nodes = players, edges = co-occurrence strength
- **Modularity:** measure of community quality
- Higher *Q* = *densely connected with same group and sparse with different groups*

- w_{ij} is the edge weight between nodes i and j.
- The weighted degree of node i is:

$$\tilde{d}(i) = \sum_j w_{ij}.$$

- The total edge weight is:

$$\tilde{m} = \frac{1}{2} \sum_{i,j} w_{ij}.$$

- The community indicator function is

$$\delta(c_i, c_j) = \begin{cases} 1, & \text{if nodes } i \text{ and } j \text{ are in the same community,} \\ 0, & \text{otherwise.} \end{cases}$$

Louvain

$$\Delta Q = \left[\frac{\sum_{\text{in}} + 2w_{i,\text{in}}}{2m} - \left(\frac{\sum_{\text{tot}} + w_i}{2m}\right)^2\right] - \left[\frac{\sum_{\text{in}}}{2m} - \left(\frac{\sum_{\text{tot}}}{2m}\right)^2 - \left(\frac{w_i}{2m}\right)^2\right]$$

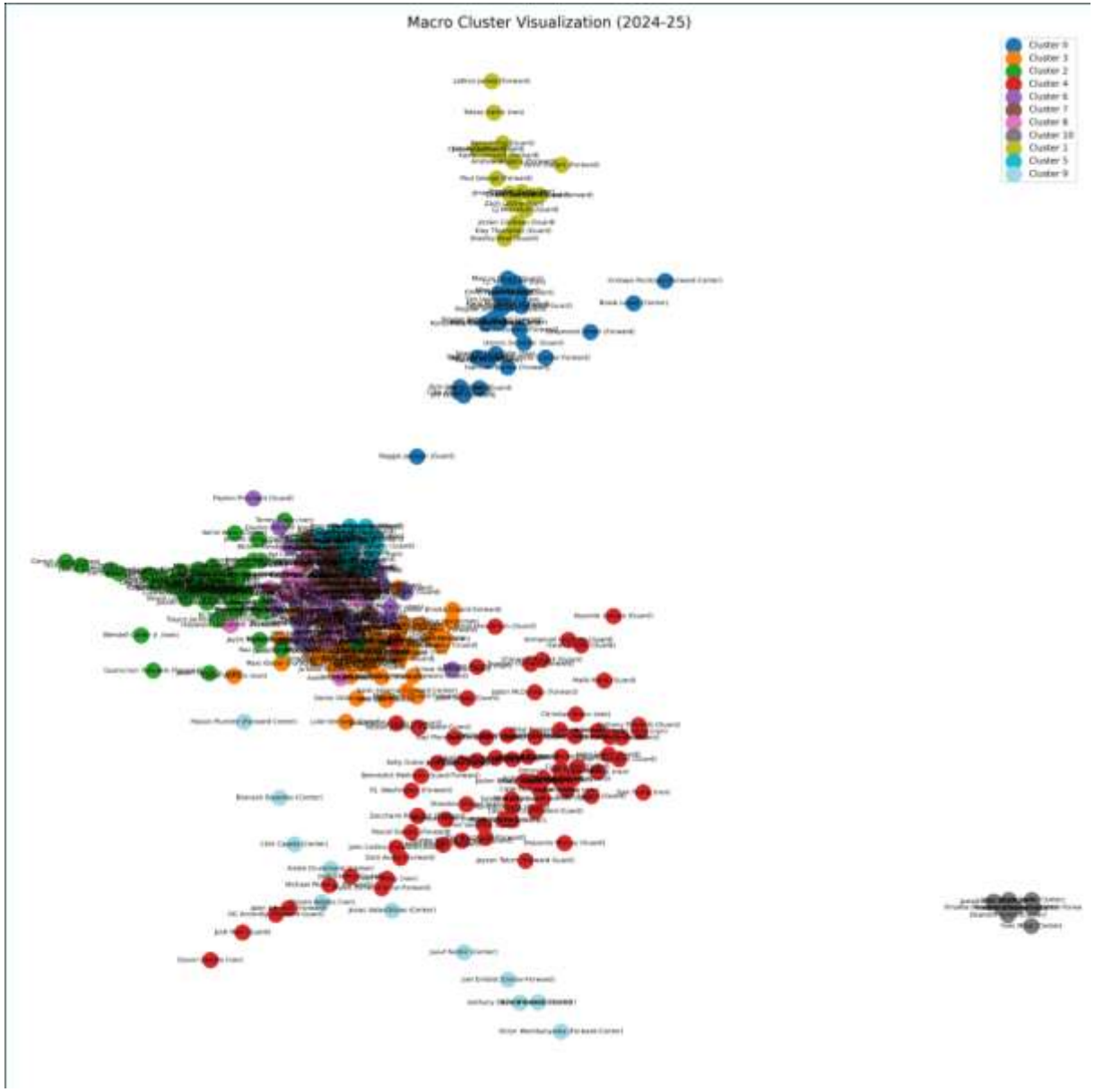
- Louvain algorithm: Two phases.
  1. Local modularity optimization
  2. Community aggregation
- Guarantees *non-decreasing modularity*

Experiment Setup

- 2024–2025 NBA season data
- Split into 6 datasets:
  - Scoring, Passing, Rebounding, Defense, Hustle, Clutch
- Tested multiple values of *k*
- Built weighted networks
- Applied Louvain and computed modularity

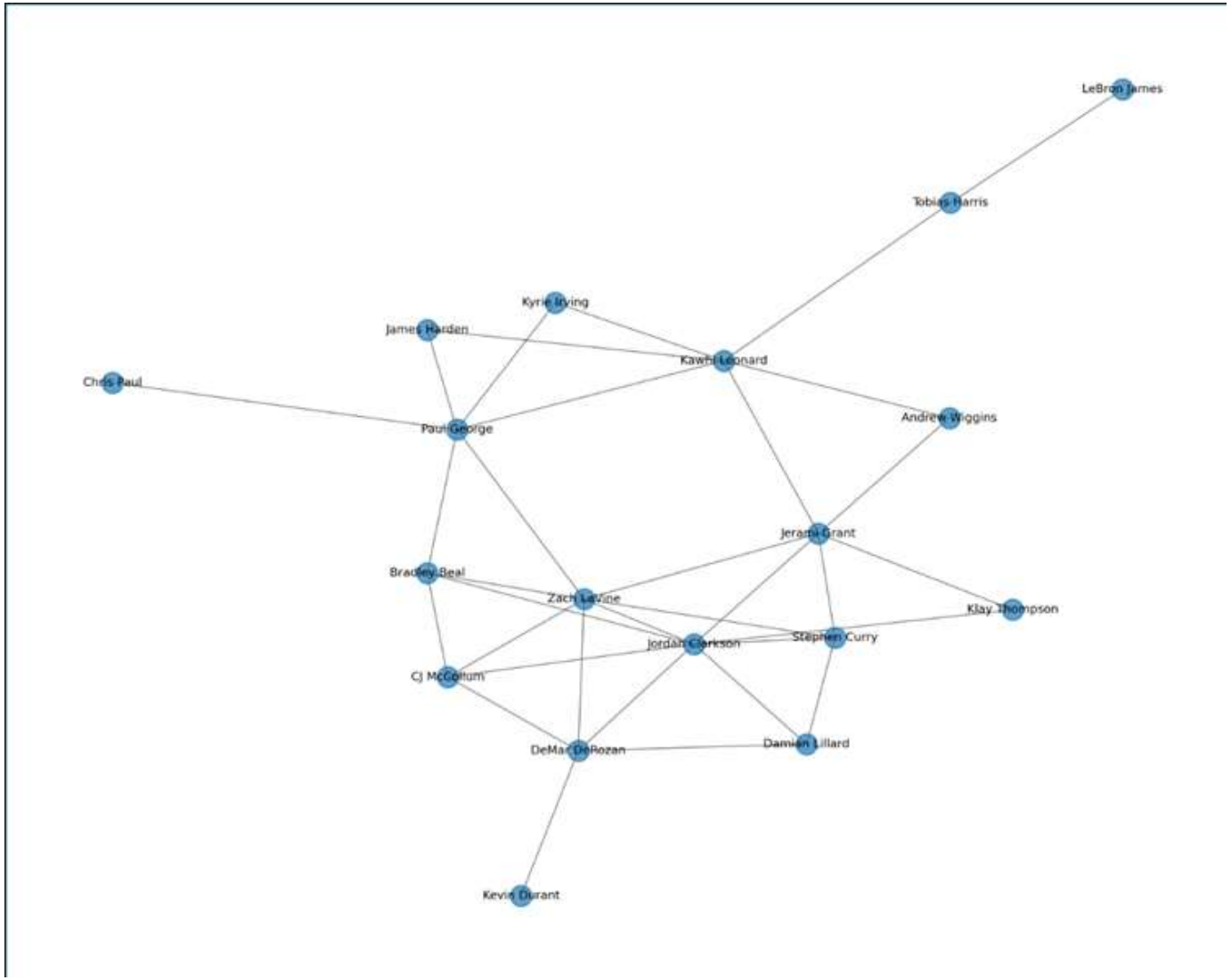
Results

- Original paper: $Q > 0.65$
- My implementation: $Q \approx 0.54$
- Still above $0.50 \rightarrow$ meaningful clusters
- Possible reasons:
  - No publicly available implementation
  - Less dataset Consistency



Micro-Cluster Network

- Players clustered based on:
  - Elite offensive role similarity detected
  - High scoring volume
  - Ball-handling + creation ability



Conclusion

- Clustering + network analysis = powerful roster construction tool
- Can help teams identify optimal lineups based on:
  - Player Style
  - Complementary roles
- Promising for future sports analytics applications
  - Example: hockey players share positions but differ tactically

References

- Arthur, D., & Vassilvitskii, S. (2007). k-means: the advantages of careful seeding. Proceedings of the Eigh teenth Annual ACM-SIAM Symposium on Discrete Algorithms, 1027–1035. https://doi.org/10.5555/1283383.1283494
- Arratia, A., & Renedo Mirambell, M. (2021). Clustering assessment in weighted networks. PeerJ. Computer Science, 7, Article e600. https://doi.org/10.7717/peerj-cs.600
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics, 2008, P10008-12. https://doi.org/10.1088/1742-5468/2008/10/P10008
- Muniz, M., & Flamand, T. (2022). A weighted network clustering approach in the NBA. Journal of Sports Analytics, 8(4), 251–275. https://doi.org/10.3233/JSA-220584