SI 608 Project Midpoint Report

Project Title: Network Dynamics and Team Success in the NBA (2015–2021)

Team Members: Mohammad Farhat, Lamonte Nunn, Mustafa Al-Ogaili, & Ali Al-Ogaili

1. Research Questions and Motivation

We model teams as passing networks: players = nodes; directed, weighted edges = playmaking interactions. Because wins emerge from coordinated actions, network structure should relate to outcomes (win/loss, efficiency). Networks let us quantify cohesion, centralization, and redundancy in ball movement beyond individual box-score metrics. Our main research questions are:

1. Do teams with more cohesive passing networks higher density, clustering, and reciprocity exhibit higher win probability and offensive efficiency
2. How does the departure of a key player affect the structure and cohesion of the team's network?
3. Which network properties best predict team success?

Feedback from our proposal has been addressed by clarifying overlapping questions and adding an evaluation plan.

2. Related Work

Studies by Fewell et al. (2012) and Passos et al. (2017) show that sports teams display small-world characteristics—high clustering and short paths that enhance coordination. Gyarmati et al. (2014) introduced passing networks to evaluate teamwork efficiency. Our study expands this by examining six NBA seasons (2015–2021) and incorporating the effects of player departures with predictive modeling.

3. Dataset

We use the NBA Play-by-Play dataset from Kaggle (schmadam97), covering 2015–2021. It records every play event (passes, assists, rebounds, etc.) for ~11,000 games. This rich dataset enables construction of directed, weighted passing networks. While it provides extensive data, missing tracking data for some games introduces limitations.

4. Data Collection and Processing

- **Source.** Kaggle NBA Play-by-Play (schmadam97), seasons 2015–2021; ~11k regular-season games.
- **Key fields (used):** game_id, season, period, clock, event_type (shot, assist, rebound, substitution, turnover), player and team IDs, score updates. We rely on assist and substitution events for assist networks and on-court reconstruction.
- **Suitability.** Event granularity enables directed, weighted assist-based networks and possession estimates for ORtg.

Limitations/Biases:

Pass coverage: Play-by-play typically logs assists but not all passes → networks reflect shot-ending interactions (positive bias toward finishers).

Scorer subjectivity for assists; injury reasons not always coded; schedule/pace differences across teams.

Mitigations: per-48 shared-minutes normalization; opponent, pace, and lineup controls; robustness checks with alternative star thresholds (top 5% / 15%); sensitivity using degree binarization.

Tools: Python (pandas, numpy), NetworkX, scikit-learn; Kaggle API for retrieval; reproducible scripts with fixed seeds.

Pipeline:

Ingest & Filter. Pull seasons 2015–2021; keep regular season; harmonize player/team IDs across seasons.

Lineups (on-court windows). Reconstruct on-court players by applying substitutions in event order. For each pair (i, j), compute shared minutes per window.

Edges. For each window and team: weight
$wi \to j$= assists from i to j / shared minutes × 48. Build directed weighted graphs; restrict to players with ≥10 min in the window to stabilize small-sample noise.

Metrics. Node: weighted out-degree, betweenness, closeness, eigenvector. Team: density, transitivity, reciprocity, weighted out-degree centralization (Freeman), Gini of out-degree, position assortativity.

Targets/Controls. Outcomes: Win (0/1), ORtg. Controls: opponent fixed effects (or Elo proxy), pace, back-to-back, rest days, injuries count.

Star & Departure Detection. Compute team-season percentiles for minutes/usage; detect trade via team_id change; injury/absence as ≥3 consecutive DNPs after a ≥5-game high-minute run; define event time t relative to first missed game (or trade date).

Quality & Reproducibility. Scripted end-to-end; unit checks (edge count ≤ N(N-1), minutes balance, possessions sanity); versioned data artifacts.

Compute constraints. ~11k games × rolling windows; parallelize by team-season; cache lineup states; optionally subsample windows for model selection, then refit on full set.

5. Analysis Completed So Far

Descriptives: Across team-windows, networks are sparse but cohesive; degree distributions are right-skewed with moderate centralization. Reciprocity varies with scheme (iso-heavy vs motion offenses).

Associations (preliminary, controlled for pace/opponent):

- Lower out-degree centralization and higher average betweenness (shared playmaking) associate with higher win odds.
- Reciprocity shows a positive association with ORtg, consistent with two-way ball flow.
- Event-study around star departures shows an immediate drop in density and increase in centralization at t=0, with partial reversion by t+10.

6. Plans for Further Analysis and Evaluation Plan

Upcoming work: finalize temporal event analysis and train predictive models (logistic regression, random forest). Cross-validation (70/30 temporal split) and SHAP-based feature importance analysis will be used.

Evaluation Plan:
• Predictive accuracy (AUC, RMSE, R^2) for win/loss and point differential.
• Difference-in-differences analysis for player departures.
• Baseline comparison against team averages.
• Bootstrapped confidence intervals for feature significance.

Risks include computational load from 11,000+ graphs; mitigated via parallelization and subsampling.

7. Reflection on Feedback

We addressed feedback by adding a clear evaluation plan, using directed weighted graphs, introducing a sliding-window analysis for temporal depth, and distinguishing between overlapping research questions. Feasibility of Q2 was improved by defining and quantifying 'star player' events.

8. References

*Fewell, J. H. et al. (2012). *Basketball teams as strategic networks.* PNAS.*
*Gyarmati, L., Kwak, H., & Rodriguez, P. (2014). *Understanding Team Performance in the NBA.**
Guo, T., Cui, Y., Min, W., Shen, Y., & Mi, J. (2022). ***Exploring the relationship between basketball rotation and competitive performance using substitution network analysis.*** Journal of Sports Sciences, 40*(7), 776–784*