TERM PROJECT

Layup Laboratory

A Causal Inference and Network Science Approach to March Madness

Anthony Gallante

Northwestern University
School of Professional Studies

February 15, 2026

Abstract

This project focuses on using Network Science and Causal Inference concepts included in the MSDS-452-DL curriculum for the purpose of creating a set of diversified, competetive March Madness brackets. The first goal of this effort is creating a model that measures the causal effect of in-game events in terms of their influence on the outcome on the game. Such a model can be used as the basis for Monte Carlo simulation to determine how often one team is likely to defeat another. In recognition that some online competitions allow the user to submit dozens of brackets, the second goal is developing a strategy for building the March Madness bracket itself.

Author Note

The code for this project is available at the LayupLaboratory repository at the author's personal GitHub page: https://github.com/AnthonyGallante/LayupLaboratory

Modifications to this report in support of Checkpoint B are shown in red.

Introduction

The NCAA Division I basketball tournament, commonly referred to as March Madness, is a 63-game tournament that decides the national champion each year. As one of the most widely recognized sporting events in the United States, its arrival signals the annual tradition in which millions of fans submit their tournament predictions to casual competition pools, both in person and online. While technically consisting of 67 games, I will be excluding the play-in games from this discussion, since a majority of the popular March Madness competitions begin with the Round of 64.

The high-stakes single elimination format creates a chaotic environment in which upsets, games in which a weaker team eliminates a stronger team, are commonplace. Failure to recognize these upsets in the early rounds of the tournament tend to invalidate entire branches of bracket predictions. In recognition of the astronomically low odds that any bracket prediction is perfectly correct, generous rewards are promised to those who take on the challenge: Warren Buffet, former CEO of Berkshire Hathaway, famously offered $1 billion in 2014[1] and the SpaceX aerospace company announced that the 2025 grand prize winner would receive a trip to Mars[2].

While the author has no hopes of claiming such rewards, he has always been fascinated with the technical challenges that are present with event forecasting. Bracketology, the term coined specifically to describe forecasting for the NCAA basketball championship, requires the analyst to make predictions at the individual game level as well as the bracket level.

Literature Review

With millions of users entering free March Madness competitions every year, the barrier for entry is low. It is a topic that both the casual sports fan and academic are willing to discuss. One may easily find herself stumbling upon interesting bracket prediction techniques from miscellaneous blogs. In one example, a blogger with the screen name *BioPhysEngr* wrote about his experience of building a bracket based on the concept of eigenvector centrality, only to be beaten by his mother picking randomly[3].

Traversal of binary tree data structures, on the other hand, is a richly researched topic and I anticipate finding useful academic sources for the purpose of bracket optimization.

While I continue to review the literature on both game level prediction and bracket optimization, I will briefly summarize several ideas that have caught my eye thus far.

Entropy-Based Strategies for Multi-Bracket Pools[4]

Brill et. al propose an entropy-based strategy when given the opportunity to submit multiple brackets to a competition. When allowed few brackets, optimal greedy strategies should be

[1] Cory Mitchell, "What Are the Odds of Getting a Perfect Bracket?," October 8, 2024, https://www.investopedia.com/ask/answers/082714/what-are-odds-getting-perfect-bracket-warren-buffetts-1-billion-march-madness-bracket-challenge.asp.

[2] X Business, "X Launches X Bracket Challenge with a Trip to Mars on the Line," March 13, 2025, https://x.com/XBusiness/status/1900293498177552601.

[3] BioPhysEngr, "Eigenbracket 2012: Using Graph Theory to Predict NCAA March Madness Basketball," March 13, 2012, http://blog.biophysengr.net/2012/03/eigenbracket-2012-using-graph-theory-to.html.

[4] Robert S. Brill et al., "Entropy-Based Strategies for Multi-Bracket Pools," *Entropy* 26, no. 8 (2024): 615, https://doi.org/10.3390/e26080615.

used, but contestants should choose more late-round upsets in order to *increase the amount of entropy in their brackets* as more are permitted, essentially diversifying their submissions.

Statistics Slam Dunk[5]

Uses causal inference frameworks to refute the idea of the "Hot Hand." Does not appear to delve deeply into either causal inference or network science.

SMOGS: Social Network Metrics of Game Success[6] *and Social Network Analysis of College and Professional Basketball*[7]

A network can be formed by modeling the players as nodes and the number of passes between each player as directed, weighted edges. Both Xu et al. and Bu et al. suggest that teams who pass the ball across the entire team have an advantage over teams that do not. The exception is in the Men's National Basketball Association (NBA), which rely on star players to serve as a distribution node, making a majority of the passes. The data sources currently being used (later discussed in Table 1) do not have data to this resolution, making this resource difficult to use at the current moment. The dataset used by Bu et al. does not appear to available for all NCAA Division I basketball teams, as SportsVu optical sensors are not installed in all facilities, despite being standard in theNBA. I am hopeful that player-to-player passing data is available at the college level.

Methods

This project is divided into two distinct lines of effort: game level prediction with causal inference and bracket optimization with various network science techniques.

Game Level Prediction

A causal model is used to understand how each event captured in game-level box scores affect the final scores of each team. Figure 1 below shows a preliminary model, which is based on the idea that team possessions lead to scoring opportunities. Here, I model the assumption that increasing events like rebounds and steals leads to more changes in possession. Table 2 shows the preliminary average treatment effect (ATE) for each of the variables shown in the diagram. Many games can be simulated with Monte Carlo methods by randomizing the players participating on each team and the number of possessions that each team gets in each game. Because our simulation is based on possession-level statistics, key variables in our causal model are those that influence team possessions and scoring opportunities. These include rebounds (ORB and DRB), blocks (BLK), steals (STL) and turnovers (TOV). Additionally, because none of these actually contribute directly to a team's score, their effects must be mediated through direct scoring attempts (2PA, 3PA, FTA). The refutation of the DAG below is currently in progress.

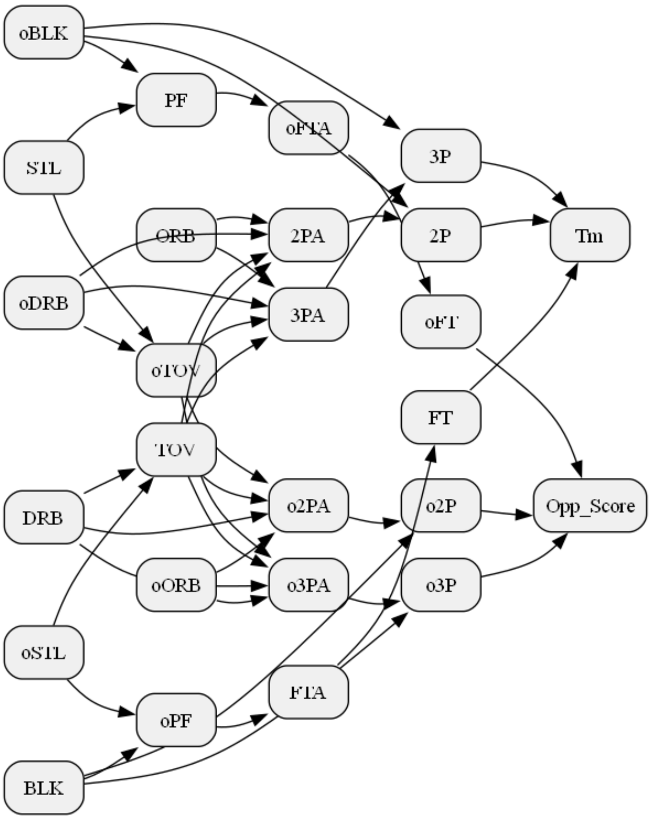
[5] Gary Sutton, *Statistics Slam Dunk* (Manning Publications, 2024).

[6] Fan Bu et al., "SMOGS: Social Network Metrics of Game Success," in "Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (Aistats)," special issue, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)* (Naha, Okinawa, Japan), PMLR, vol. 89 (2019).

[7] Carol Xu, "Social Network Analysis of College and Professional Basketball" (Ph.D. dissertation, 2018).

Figure 1

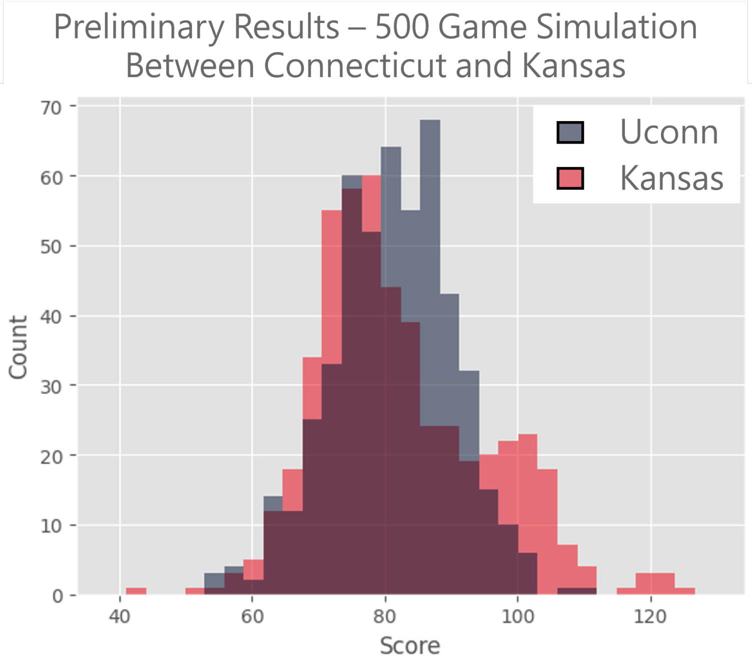
A preliminary DAG that can be used to estimate the point value for events that happen during a game of basketball. A minor adjustment was made, removing duplicate arrows.



Games are built around a randomly generated number of possessions and a weighted sample of players at the start of each match. Each team is given approximately the same number of possessions per game, though it varies according to each team's pace rating. Using per-possession player-level statistics, we estimate the number of events all players on the team are responsible for the decided number of possessions. This simple method will continue to be fine tuned over the course of this project, using past games as validation sets. In the mean time, it at least appears to showcase potential for teams with a considerable amount of data.

Figure 2

An overlapping histogram showing preliminary simulation score distributions between the University of Connecticut and the University of Kansas (500 games).



A benefit of the causal inference framework used in this project is the ease of explainability for each simulation. For the game above in Figure 2 above is the ability to feed the simulated game logs into a large language model and ask for a short analysis. For the 500 simulations above, gpt-5-nano[8] returns the following analysis:

"Connecticut edges Kansas in the simulation, riding superior efficiency, balanced scoring, plus stronger rebounding and turnover discipline; Kansas pushes tempo but falters in efficiency and defensive consistency overall."

While this annotation process has no influence on the outcome of a round of simulations, it at least provides a fun attempt at explaining why one team might have defeated another, which might be useful in upset-situations. While Figure 2 shows the outcome of 500 games, fewer simulated games would likely be desired in high-entropy brackets, where chance plays a higher role. Thus, one way to increase the randomness in a bracket is to reduce the number of simulalted games played between each team.

Bracket Optimization

The second line of effort in this project is optimizing the bracket performance itself. A greedy algorithm, always choosing the team with the higher chance of winning, is likely impractical when generating many brackets. Such a strategy might never account for certain upsets. One solution would be to create one greedy version of the bracket, and create a new graph wherever our game-level simulations are uncertain of the winner. This method will

[8] OpenAI, "Chatgpt," 2024, https://chat.openai.com/.

hopefully prevent us from creating too many brackets where 16 seeds beat 1 seeds, but will maximize diversity in close games.

Brill et al. provide a great starting point for this problem, though I am interested in learning about different approaches as well.

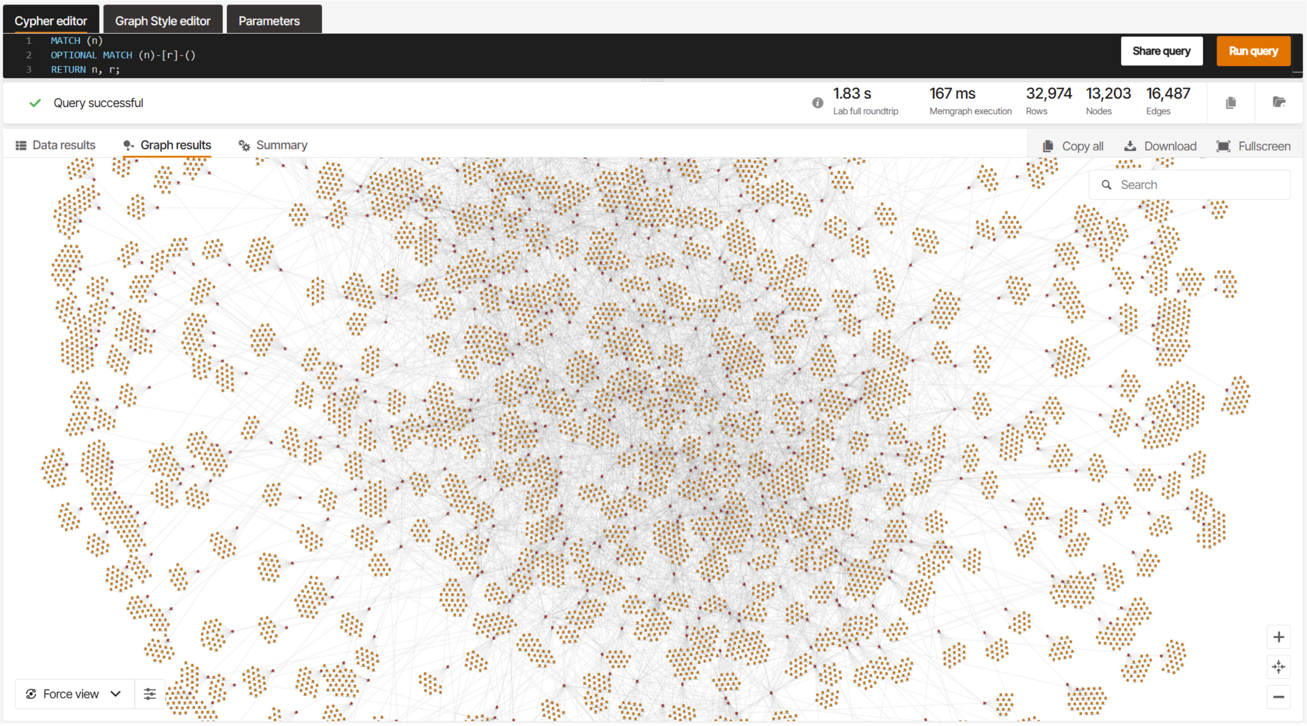
Results

Two graph databases have been created for this project: one to visualize the relationships between players and teams that happened during the basketball season, and one to query pre-calculated simulations for all teams. There are 727 unique teams in the 2026 HoopR dataset, including many schools that, while not division I themselves, play division I teams during the regular season. 12,476 athletes are included in this dataset. It is worth noting that athletes can transfer schools and play for multiple teams over the course of the season.

In Figure 3 below, we visualize each basketball team, the athletes that play on each team, and the outcome of all games played thus far. All edges are directed, and edges between two teams are directed and weighted according to the final difference in score.

Figure 3

A sample of the 13,203 nodes and 16,487 relationships formed by teams and players that have participated in the NCAA basketball season.



```
MATCH (n)
OPTIONAL MATCH (n)-[r]-()
RETURN n, r;
```

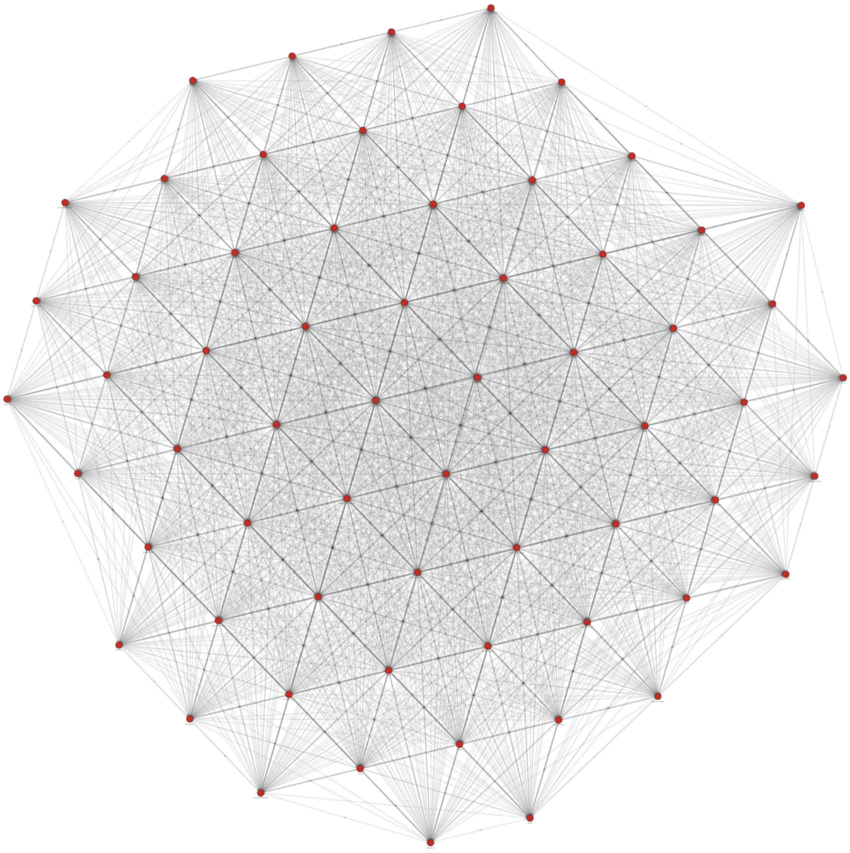
✓ Query successful

```
1.83 s              167 ms               32,974   13,203   16,487
Lab full roundtrip  Memgraph execution   Rows     Nodes    Edges
```

In Figure 4 below, we see the results of a simulated round-robin tournament of 64 teams, to be replaced with the actual teams taking place in the NCAA Basketball Tournament once announced. This graph is created in order to drastically reduce the computations needed to produce brackets; all simulations can be performed one time each ahead of time, and only queries of the outcome will need to be performed in any given bracket permutation. While more work needs to be done on the simulations, simulations in their current state seem to favor teams with a higher ELO rating[9], which at least indicate forward progress.

Figure 4

Simulated round-robin tournament with 64 NCAA basketball teams.



Data Sources

See Table 1 for a description of data sources used thus far. This table will be updated as the project progresses.

[9] Warren Nolan, "ELO Ranking - 2026 Men's College Basketball," 2026, https://www.warrennolan.com/basketball/2026/elo.

Table 1

Data Sources

Source	Description
sports-reference.com[10]	Team and Individual Player level data
warrennolan.com[11]	Team ELO ratings
hoopr.sportsdataverse.org[12]	Play-by-play data, schedules, and box scores

Data from the sources in the table above will be used to populate causal models and a memgraph database. It may be reasonable to pre-compute a large number of games between all possible combinations of teams and store those in a graph database prior to the bracket generation stage. Computing these outcomes beforehand will allow the bracket-building program to simply query results, rather than performing simulation calculations at runtime.

Conclusions

Preliminary average treatment effects (ATE) from the DAG shown in Figure 1 are shown in Table 2. I hope to collect more data as the season progresses and calculate conditional average treatment effects (CATE) at the team or player level in order to account for differences in team and player behavior. At this point, I have no reason to believe that one team's block should have the same impact as another team's. Switching to a CATE approach will hopefully capture differences in team's offensive and defensive strategies.

Because CATE is more computationally expensive than ATE, DAG validation will be done using ATE; however, the current structure of this project assumes a single common causal model.

Table 2

Preliminary Average Treatment Effects for Player-Level Events

Event	Team Score Effect	Opponent Score Effect
2-Point Field Goal	1.998643	0.0
2-Point Attempt	1.217853	0.0
3-Point Field Goal	3.000105	0.0
3-Point Attempt	0.947277	0.0
Free Throw	1.100958	0.0
Free Throw Attempt	0.850505	0.0
Steal	1.046184	−0.779294

[10] Sports Reference LLC, "College Basketball at Sports-Reference.com," 2026, https://www.sports-reference.com/cbb/.

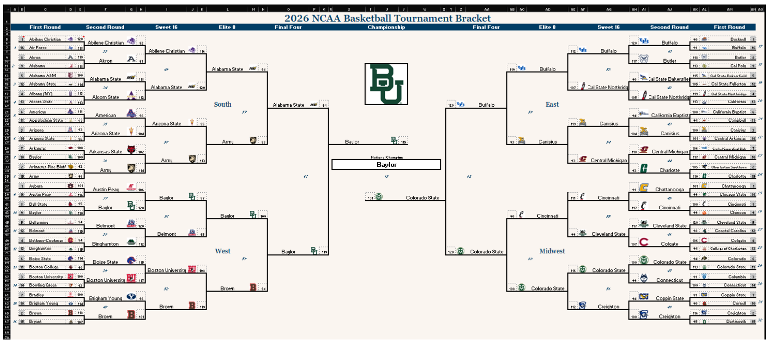
[11] Nolan, "ELO Ranking - 2026 Men's College Basketball".

[12] Saiem Gilani, "Hoopr: The Sportsdataverse's R Package for Men's Basketball Data.," 2021, https://hoopr.sportsdataverse.org/.

Event	Team Score Effect	Opponent Score Effect
Block	0.069190	−0.560639
Turnover	−0.883766	0.814006
Personal Foul	0.0	1.112083
Offensive Rebound	0.182142	0.0
Defensive Rebound	0.851554	−0.848281
Allow 2-Point Field Goal	0.0	2.002056
Allow 2-Point Attempt	0.0	0.982315
Allow 3-Point Field Goal	0.0	2.976299
Allow 3-Point Attempt	0.0	0.888978
Allow Free Throw	0.0	0.985768
Allow Free Throw Attempt	0.0	0.752256
Allow Steal	−0.901049	0.875331
Allow Block	−0.708807	0.069735
Allow Turnover	0.976682	−0.870936
Allow Personal Foul	1.004520	0.0
Allow Offensive Rebound	0.0	0.143493
Allow Defensive Rebound	−1.090965	0.688674

Results from this project can be visualized in a spreadsheet. I've put together an .xlsx file that can be used to display a bracket from a .csv file, though it is not yet hooked up to the rest of the Python code. See Figure 5 for an example output, using randomly selected teams.

Figure 5
An example output visualization spreadsheet.



I am hopeful that the objectives defined above are attainable with the data available. If time permits, this effort can be validated with data from past tournaments. While I will be entering the brackets created by this project into the official March Madness competition (details yet to be released), my only requirement for success is that this approach is better than random.

References

BioPhysEngr. "Eigenbracket 2012: Using Graph Theory to Predict NCAA March Madness Basketball." March 13, 2012. http://blog.biophysengr.net/2012/03/eigenbracket-2012-using-graph-theory-to.html.

Brill, Robert S., Abraham J. Wyner, and Ian J. Barnett. "Entropy-Based Strategies for Multi-Bracket Pools." *Entropy* 26, no. 8 (2024): 615. https://doi.org/10.3390/e26080615.

Bu, Fan, Sonia Xu, Katherine Heller, and Alexander Volfovsky. "SMOGS: Social Network Metrics of Game Success." In "Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (Aistats)." Special issue, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)* (Naha, Okinawa, Japan), PMLR, vol. 89 (2019).

Gilani, Saiem. "Hoopr: The Sportsdataverse's R Package for Men's Basketball Data.." 2021. https://hoopr.sportsdataverse.org/.

Mitchell, Cory. "What Are the Odds of Getting a Perfect Bracket?." October 8, 2024. https://www.investopedia.com/ask/answers/082714/what-are-odds-getting-perfect-bracket-warren-buffetts-1-billion-march-madness-bracket-challenge.asp.

Nolan, Warren. "ELO Ranking - 2026 Men's College Basketball." 2026. https://www.warrennolan.com/basketball/2026/elo.

OpenAI. "Chatgpt." 2024. https://chat.openai.com/.

Sports Reference LLC. "College Basketball at Sports-Reference.com." 2026. https://www.sports-reference.com/cbb/.

Sutton, Gary. *Statistics Slam Dunk*. Manning Publications, 2024.

X Business. "X Launches X Bracket Challenge with a Trip to Mars on the Line." March 13, 2025. https://x.com/XBusiness/status/1900293498177552601.

Xu, Carol. "Social Network Analysis of College and Professional Basketball." Ph.D. dissertation, 2018.