Introduction

Sports betting networks provide a rich landscape for studying social interactions and behavioral patterns. The goal of this project is to identify sports bettors who share similar betting behaviors. To analyze these betting behaviors, we will employ a combination of clustering and graph-based techniques. K-means clustering will be applied to model betting behavior distributions and detect common betting archetypes. Further, community detection algorithms, such as BigCLAM, will be utilized to construct communities based on playing interests. The hypothesis guiding this study is that betters exhibiting similar transaction and betting behaviors will form distinct clusters, and that graph-based community detection will uncover overlapping group memberships indicative of more complex social structures within the betting network.

Ultimately, by integrating insights from both clustering and community detection, this project aims to provide a deeper understanding of the sports betting users and networks, combining behavioral similarity with structural connectivity. This hybrid approach enhances interpretability and supports more robust interventions in domains like responsible gaming, fraud prevention, and targeted outreach.

Data

The dataset used in this analysis originates from the Transparency Project, a research initiative under the Division on Addiction at the Cambridge Health Alliance, a Harvard Medical School teaching hospital[1]. This study presents the first large-scale descriptive analysis of actual Daily Fantasy Sports (DFS) behavior, specifically focusing on National Football League (NFL) contests.

The data includes 10,385 players who registered with DraftKings, one of the most popular DFS platforms, between August 1, 2014, and September 30, 2015. All participants made an initial deposit and entered at least one paid NFL contest during that period. Two datasets derived from this study that form the basis of our analysis are the TacklingData5All.csv and TacklingData6Play.csv datasets.

First, TacklingData5All.csv contains aggregated metrics across all DFS contests each player participated in, including NFL and non-NFL events. The metrics in TacklingData5All.csv include Date1st, the start date for the first contest the user entered; DateLst, the start date of the player's last contest; nDays, the number of unique contest start dates; nCont; the number of contests the player entered; nEntries, the total number of entries over the contests during the 2014 NFL season; nLineups, the total number of lineups over the contests during the 2014 NFL season; TotFees, the total amount spent on entry fees, AvgBuyIn is the average entry fee over the contests the player entered; TotWinnings, the player's total winnings; nUserUp, the number of contests where the player made a profit (i.e., the amount won in prizes exceeds the amount spent on entry fees); and UserID, the ID number that links this data to the demographic data in the user demographics data file.

Second, TacklingData6Play.csv captures more granular information on the types of contests each player entered, particularly focusing on sport and contest type. The metrics in TacklingData6Play.csv include UserID, the ID number that links this data to the demographic data in the user demographics data file; DidNFL, is TRUE if the player played in at least one NFL-based contest; DidNBA, is TRUE if the player played in at least one NBA-based contest; DidOth, is TRUE if the player played in at least one contest based on some other league; Cnt1, the numbers of 50/50 contests; Cnt2, the numbers of head-to-head contests; Cnt3, the numbers of

[1] http://www.thetransparencyproject.org/index.htm

multiplier contests; Cnt4, the numbers of league contests; Cnt5, the numbers of "move your way up" contests; and Cnt6, the numbers of other contests.

Together, these datasets provide a detailed behavioral profile for each player and the foundation to build both feature based clusters and similarity based network graphs. By integrating these perspectives, we can effectively identify player archetypes and uncover the underlying structure of the sports betting community.

Real-world

The real-world application of this problem is twofold. First, daily fantasy sports (DFS) companies can leverage community and cluster information derived from user behavior to recommend personalized promotions. These targeted promotions are designed to increase user engagement and ultimately maximize profits. Second, regulatory agencies, or the DFS companies themselves, can use network based analyses to detect coordinated fraudulent behavior, such as rings of bettors using insider information or manipulating betting lines.

Sportsbooks have long leveraged data science techniques to set and adjust betting lines in their favor. Historically, these calculations were based mainly on sports-related data, such as player performance and team statistics. However, the rise of online DFS platforms has introduced a powerful new data source: user behavior. By analyzing individual betting patterns, DFS companies can target users with highly specific promotions, nudging them toward contests they are likely to enter. This targeted marketing has become a key driver of the sports betting industry's explosive growth over the past decade.[2]

In fact, the strategic use of user data is now a standard practice across the industry, often powered by advanced machine learning techniques such as neural networks.[3] These systems allow companies to model user preferences and dynamically adapt offerings to maintain user engagement. Understanding the behavioral archetypes and relationships between users is a massive advantage, making clustering and community detection especially relevant tools in a commercial setting.

Alongside the commercial benefits, there is growing concern over the rise of fraud within DFS and the broader sports betting ecosystem. Identifying such fraud is essential not only to protect the financial interests of platforms, but also to uphold the integrity of the industry and safeguard bettors. Data science methods, particularly those focused on graph analysis and community detection, offer promising tools for uncovering suspicious activity.

One prominent example involves the federal investigation of a betting ring tied to two NBA betting cases in 2024. Authorities discovered the ring by identifying patterns among accounts placing large, similar wagers on prop bets involving former Toronto Raptors player Jontay Porter and Miami Heat guard Terry Rozier.[4] This case highlights how community detection algorithms, when applied to betting behavior, can play a pivotal role in identifying fraudulent activity early.

In summary, both industry leaders and regulators have strong incentives to adopt clustering and community detection techniques. Whether for profit maximization or fraud prevention, these tools are shaping the landscape of modern sports betting.

[2] https://www.espn.com/espn/betting/story/_/id/43922129/us-sports-betting-industry-posts-record-137b-revenue-24

[3] https://www.researchgate.net/publication/385353406_A_Systematic_Review_of_Machine_Learning_in_Sports_Betting_Techniques_Challenges_and_Future_Directions

[4] https://www.espn.com/mens-college-basketball/story/_/id/43680104/sources-gambling-ring-tied-college-basketball-games

Methods

This analysis employs a combination of clustering and network based community detection techniques to uncover behavioral archetypes and latent social structures among DFS users. Specifically, we apply K-means clustering, Principal Component Analysis (PCA) for dimension reduction, and the BigCLAM (Big Cluster Affiliation Model) algorithm for community detection. Each technique operates on a different representation of the user, providing complementary insights into both behavior and structure.

We first perform K-means clustering on player aggregated behavioral features from TacklingData5All.csv, including metrics such as total entry fees (TotFees), total winnings (TotWinnings), number of contests (nCont), number of days active (nDays), number of profitable contests (nUserUp), and other similar statistics. These metrics capture users' betting volume, success, and frequency. K-means partitions the users into k disjoint clusters (hard clustering) minimizing cluster variance until convergence defined by:

$$min\,(\sum_{i=1}^{k}\sum_{x\in C_i}\left|\left|x-u_i\right|\right|^2)$$

where u is the centroid of the cluster C, and x is a user's feature vector.

We experiment with different values of k and use the elbow method to determine the optimal number of clusters based on the total within-cluster sum of squares (inertia). This allows us to identify distinct DFS betting archetypes, such as casual, consistent, or high-stakes players.

To aid in interpreting and visualizing the clusters, we apply Principal Component Analysis (PCA) to reduce the high-dimensional feature space to two dimensions. This does not affect clustering but helps highlight patterns in user behavior across clusters and overall aid in viewing results in visual formats like graphs.

While K-means groups users based on individual features, it does not capture relationships or shared affinities between users. To model this network structure, we apply BigCLAM to a graph derived from TacklingData6Play.csv.

We construct an undirected, weighted user-user graph where edges are formed based on similarity in contest participation patterns (e.g., betting on the same leagues or entering the same types of contests). Each user is a node; edges connect users who show similar distributions across contest types (e.g., both heavily favoring NBA contests and 50/50 games). Edge weights reflect the degree of behavioral overlap.

BigCLAM models overlapping communities by learning a non-negative affiliation matrix $F\in R^{nxc}$, where n is the number of users and c is the number of latent communities. Each entry F_{uc} reflects the strength of user u's affiliation to community c. The probability that two users u and v share an edge is given by the probabilistic model:

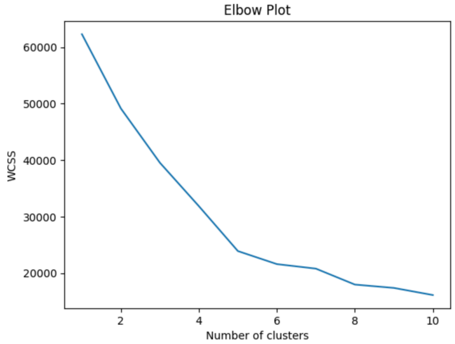
$$P(u\sim v)\;=\;1\;-\;exp(-\sum_{c=1}^{C}F_{uc}\bullet F_{vc})$$

This formula allows users to belong to multiple communities, which is crucial for modeling complex or hybrid betting behaviors (e.g., a user who bets both on the NBA and in head-to-head contests). The number of communities c is a hyperparameter and chosen based on modularity and coherence of community assignments. The probabilistic model used by BigCLAM here is the Poisson edge generation model, where the model assumes edges between nodes arise independently with probabilities determined by the community affiliation strengths.

To integrate the clustering and community detection results, we conduct a cross-analysis between K-means clusters and BigCLAM communities. This involves two main techniques: alignment analysis and community profile analysis. Alignment analysis involves checking the degree of overlap between clusters and communities to see whether behavioral archetypes tend to co-occur with specific network-based communities. Further, Community profile analysis involves characterizing each community based on its member users' K-means archetypes and contest preferences. This approach allows us to move beyond one-dimensional user modeling and uncover structures in the DFS betting ecosystem to derive deeper insights.

Results

K-means was ran over a range of cluster values k using scaled 'nDays', 'nCont', 'nEntries', 'TotFees', 'AvgBuyIn', and 'TotWinnings' values for each user from the TacklingData5All.csv dataset. The resulting elbow plot was constructed to determine the k which produced a reasonably low within-cluster sum of squares, a metric used to evaluate the quality of clusters:



Since the 'elbow' for the plot occurs at 5 for the number of clusters and steadily decreases for the next couple values, the ideal chosen k was determined to be 6.

Thus, the k-means algorithm was run with the hyperparameter 6 for the number of clusters. Betting archetype labels were determined by analyzing the mean values of the clustered groups.

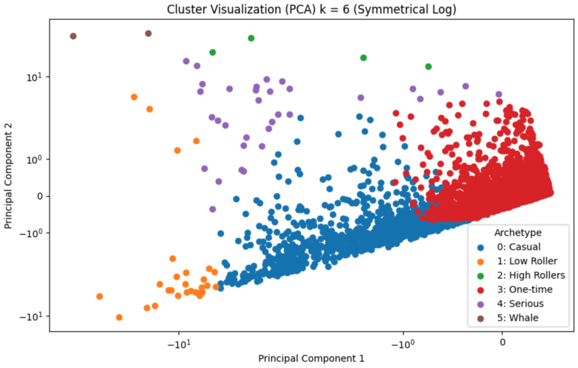
Derived from the cluster groups means values the following labels we assigned:

- Casual Players (0) - These users participate moderately (64 days active, 225 contests on average), placing low-stakes bets (AvgBuyIn: $5.22, TotFees: $1,444). They typically enter around 327 total lineups, with total winnings ($1,165) close to their spend, indicating relatively balanced outcomes. Likely bettors during the regular season or popular events, they exhibit mild engagement without high risk. More successful than one-time betters but not highly profitable.
- Low Rollers (1) - Highly active users (93 days, 1,945 contests), betting small amounts (AvgBuyIn: $5.51) with moderate total spend ($17,122). Their high number of entries (~3,375) and consistent participation suggests a strategic, disciplined approach. Despite low buy-ins, their total winnings ($16,492) are nearly breakeven, suggesting a high skill-to-stakes ratio. These are likely analytical, experienced grinders maximizing ROI from small stakes.
- Problematic High-Stakes Bettors (2) - These users bet very infrequently (12 days, 37 contests) but with extremely high stakes (AvgBuyIn: $785, TotFees: $23,912). Despite low participation, they win large amounts ($16,949), implying high volatility. This pattern aligns with sporadic, high-risk behavior, possibly by wealthy individuals seeking

thrill bets on major events, or by problem betters during relapse phases. Risk-reward profile is extreme.

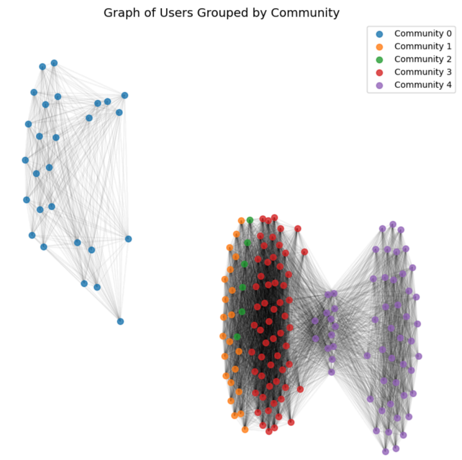
- One-Time Bettors (3) - Users in this cluster are mostly event-driven participants, only 11 days active, with 27 contests and very low spending (TotFees: $206, AvgBuyIn: $6.84). Their winnings are modest ($137), suggesting casual, one-off participation likely triggered by promotional offers or major televised sports events like the Super Bowl. Represent the least engaged and least successful segment.
- Serious Regulars (4) - This group is quite active (74 days, 327 contests), with substantial total spend ($26,362) and a moderate average buy-in ($102). Their total winnings ($26,842) slightly exceed fees, suggesting competent and committed DFS players. With ~591 entries, they balance volume and stakes, likely representing season-long players or semi-professional users. They win enough to sustain interest, without massive profits.
- High Rollers (5) - An elite cluster defined by extremely high stakes and volume: over 2,000 contests, $173,131 in entry fees, and $675,040 in total winnings. With an average buy-in of $66 and 40 days of activity, these users likely operate at near-professional or syndicate levels, placing many lineups (~2,183) and winning heavily. Represent the most successful and highest-value users in the dataset, rare but very impactful.

Next, PCA was used to reduce dimensionality to two dimensions for the purpose of plotting the clusters. The plotted clusters can be seen below with corresponding betting archetype labels:



In general, the Cluster Visualization plot shows users further to the left on the x-axis had higher winnings and played more frequently and users further to the right had less winnings and played less frequently. Further, users higher up the y-axis bet significantly more money and had larger buy-ins than users lower on the y-axis.

The values used for the community detection algorithm BigCLAM were ‘DidNBA’ and ‘DidOth’ from the TacklingData6Play.csv dataset and ‘Cnt1_Norm’, ‘Cnt2_Norm’, ‘Cnt3_Norm’, ‘Cnt4_Norm’, ‘Cnt5_Norm’, and ‘Cnt6_Norm’ calculated from each users contest number value and normalized for the total number of contests played in for that user. BigCLAM was then run with hyperparameters number of communities of 5, learning rate of 5e-4, iterations of 500, and community threshold of .01 after tuning to achieve lowest possible log-likelihood. Using the python package networkx, the community graph was visualized as:

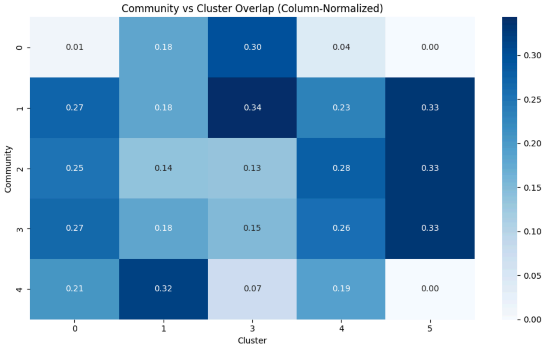


Further by analyzing the summary statistics for each community, the general community descriptions were determined to be:

- Community 0 - Users who almost exclusively play in contests. About 39% of users participated in NBA league bets; minimal participation in other sports league bets. No overlap between this community and others.
- Community 1 - NBA participation ~90%, Conservative participation in contests with > 90% in Cnt6. Substantial overlap between this community and 2 and 3.
- Community 2 - All league participation. NBA participation ~100%, Conservative participation in contests with > 90% in Cnt6. Substantial overlap between this community and 1 and 3.
- Community 3 - All league participation. Moderate contributions across Cnt1–Cnt6, especially Cluster 6 (~80%). Semi-diverse contest participation. Substantial overlap between this community and 1, 2, and 4.
- Community 4 - All league participation. Most diverse contest participation. Heavy participation in Cnt1 (50/50 contests) at ~23%, much higher than other communities. And only ~50 participated in Cnt6 (other contests). Moderate overlap between this community and community 3.

In the general, the community description reveals a few key patterns in participation and connection amongst users. First, users are overwhelmingly likely to participate in Cnt6 (other contests), unless in community 3, where contest participation is diversified among the other contests. Second, league participation is overwhelmingly common, except in community 0 who play in only exclusively contests and community 4 who play in NBA slightly less.

The cross-analysis of the clustered groups and communities revealed the following heat-map normalized by clusters:



This heat map reveals unique insights between clusters and communities. Such insights include:

- Cluster 0 almost never having membership in community 0
- Cluster 1 having relatively strong membership in community 4
- Cluster 2 having no community membership due to no meeting the community membership strength threshold
- Cluster 3 having relatively strong membership in community 1 and 0 as well as very weak membership in community 4
- Cluster 4 almost never having membership in community 0
- Cluster 5 having membership spread equally among communities 1, 2, and 3

From the previous cluster-betting archetype assignments and community descriptions the cross analysis reveals:

- Casual betters almost always bet on the NBA and other leagues and almost never are contest exclusive betters.
- Low rollers have diverse contest participation, more often playing more in 50/50 contests.
- Problem betters have little to no discernible betting pattern, likely due to their impulsive betting nature.
- One-time betters are all over the map. Split into distinct groups either not betting on leagues at all or other times betting on leagues frequently while consistently playing Cnt6 (other contests). This is likely due to one-time betters being targeted strongly by promotions and not betting long enough to exhibit patterns or congregate to a social community.
- Serious betters almost always bet on the NBA and other leagues and almost never are contest exclusive betters. Very similar to casual betters.
- High rollers are difficult to derive insights from due to this archetype being incredibly rare to begin with. In general, it can be said high rollers often bet on leagues and on the NBA and other leagues and almost never are contest exclusive betters. Similar casual and serious betters.

Conclusion

Two main sources of improvement for our results would be better data and utilization of machine learning techniques. First, this dataset is from a decade ago and consists at times of limited data. Using a bigger dataset with the growth of DFS and legalization of sports betting in the last decade would no doubt produce better results. Further, increasing the granularity of data could absolutely improve clustering and would greatly improve community assignment, as the TacklingData6Play.csv dataset left much to be desired in regards to specificity. Second, as discussed prior in this write-up, modern DFS companies leverage machine learning techniques such as neural networks to derive their insights from user data, which due to neural networks powerful computation capabilities and modern infrastructure would likely also result in better results.

In conclusion, data science techniques used to answer the proposed hypothesis were mostly successful. It was discovered that betters exhibiting similar transaction and betting behaviors formed distinct clusters which were used to classify betting into archetypes. Further, graph-based community detection uncovered overlapping group memberships indicative but the complex social structures within the betting network weren't as apparent as originally hypothesized.