CSC1051 Coursework
2 Person Project
Jared Rushe 22489846
Alex Davis 20449904

Explorative Analysis of the Pittsburgh Steelers NFL Draft History Using Graph Theory

Dataset and Problem Definition:

For our project, we chose to study the National Football League (NFL) draft history of the Pittsburgh Steelers. No team has won more Super Bowls or appeared in more Conference Championship games than the Steelers who are regarded by many as the most successful franchise in the NFL's history. The draft is the primary way to acquire talent in the NFL and an effective selection process is a cornerstone of any winning franchise. We wanted to identify and explore patterns and trends within the draft strategy of the organisation that have allowed them to dominate the league. .

The NFL draft is an annual event where the top college football players are recruited to professional teams. Teams pick from a selection of players in reverse order relative to their performance in the previous season (i.e. Super Bowl winners choose last).

Some of the questions we hoped to address included: Which teams tend to draft from specific colleges? Do teams have a strong focus on certain positions? Do these teams prioritise drafting players in the early or later rounds?

A key element of our research involved the performance based metrics in our dataset. By identifying high-performing players based on their appearances and accolades we could assess which have been the most fruitful draft strategies throughout the franchise's history.

To create our dataset, we scraped data from 'pro-football-reference.com'. We were able to collate information on every draft pick made by the Steelers dating back to their first participation in 1936. In total our data frame had 1367 rows and 9 columns . Our column headings were PlayerID, Player, Team, Year, Rnd (Round), Pos (On-field position), St

(Number of seasons as a primary starter for the franchise), G (Games played for the franchise) and SB (Super Bowls won with the franchise).

The use of graph analytics techniques such as clustering and community detection allowed us to understand the links between players, colleges and teams. We created multiple different graphs from our dataset to gain insights into what attributes were held by ‘successful’ picks under numerous definitions. By identifying inconsistencies between these features and those of the players signed to the team as a whole, we could highlight strategies to recruit players that are likely to prosper.

Data Preparation:
As previously mentioned, we got our data from ‘pro-football-reference.com’. We were able to convert a data frame of the Steelers entire data history in a csv format via a tool on the website and copy and pasted this into a google sheets file. The original data frame had 25 columns including statistics such as passing yards and touchdowns. We reduced the dataframe to the 7 columns needed for our exploratory analysis. We added the Player ID column to give each player a unique identifier. The final column we created ourselves was the SB column. We did this by scraping the rosters from the Steelers 6 Super Bowl winning seasons into a csv file from the same website. We then created a python program to iterate through the rosters and increase the associated SB value by 1 for each time the player's name appeared in the roster of a winning season. The python script can be seen here:

```
import pandas as pd

steelers_file = "C:\\Users\\rushej2\\Downloads\\steeler.csv"
rosters_file = "C:\\Users\\rushej2\\Downloads\\rosters.csv"

steelers_df = pd.read_csv(steelers_file)
rosters_df = pd.read_csv(rosters_file)

roster_players = set(rosters_df['Player'].values)

steelers_df['SB'] = 0

for index, row in steelers_df.iterrows():
    if row['Player'] in roster_players:
        steelers_df.at[index, 'SB'] += 1

updated_steelers_file =
"C:\\Users\\rushej2\\Downloads\\steeler.csv"
steelers_df.to_csv(updated_steelers_file, index=False)
```

The final data cleaning was removing null values from the numerical columns in the data frame. Our source website left all null values as blank cells instead of 0s which we knew would be problematic when transforming our data later on. We created filters to highlight all blank cells in the G, St and PB columns and filled them with “0”.

Here is a sample snippet of our final data frame:

	A	B	C	D	E	F	G	H	I	J
1	Player ID	Year	Rnd	Player	Pos	PB	St	G	SB	College/Univ
2	1	2024	1	Troy Fautanu	OL	0	0	1	0	Washington
3	2	2024	2	Zach Frazier	OL	0	1	8	0	West Virginia
4	3	2024	3	Roman Wilson	WR	0	0	1	0	Michigan
5	4	2024	3	Payton Wilson	LB	0	0	10	0	North Carolina
6	5	2024	4	Mason McCorn	OL	0	1	10	0	South Dakota :
7	6	2024	6	Logan Lee	DL	0	0	0	0	Iowa
8	7	2024	6	Ryan Watts	SAF	0	0	0	0	Texas
9	8	2023	1	Broderick Jone	OL	0	2	27	0	Georgia
10	9	2023	2	Joey Porter Jr.	CB	0	2	27	0	Penn St.

We used the following code to import our dataset into neo4j and create the nodes and relationships for our project.

```
LOAD CSV WITH HEADERS FROM 'file:///steeler.csv' AS row
MERGE (p:Player {id: row.`Player ID`})
SET p.name = row.Player,
    p.year = toInteger(row.Year),
    p.round = toInteger(row.Rnd),
    p.position = row.Pos,
    p.college = row.`College/Univ`,
    p.pb = toInteger(row.PB),
    p.starts = toInteger(row.St),
    p.games = toInteger(row.G),
    p.superbowl = toInteger(row.SB)
MERGE (pos:Position {name: row.Pos});
MERGE (c:College {name: row.`College/Univ`});

MATCH (p:Player {id: row.`Player ID`})
MATCH (c:College {name: row.`College/Univ`})
MERGE (c)-[:PROVIDED]->(p);

MATCH (p:Player {id: row.`Player ID`})
MATCH (c:College {name: row.`College/Univ`})
MERGE (p)-[:ATTENDED]->(c);

MATCH (p:Player), (pos:Position)
WHERE p.position = pos.name
MERGE (p)-[:PLAYS_AS]-(pos);
```

```
MATCH (p:Player), (pos:Position)
WHERE p.position = pos.name
MERGE (pos)-[:PLAYED_BY]->(p);

MATCH (p1:Player)-[:PLAYS_AS]->(pos:Position)<-[:PLAYS_AS]-(p2:Player)
WHERE p1 <> p2
MERGE (p1)-[:SAME_POSITION]-(p2);

MATCH (p1:Player)-[:ATTENDED]->(c:College)<-[:ATTENDED]-(p2:Player)
WHERE p1 <> p2
MERGE (p1)-[:SAME_COLLEGE]-(p2)

MATCH (p1:Player), (p2:Player)
WHERE p1 <> p2 AND p1.round = p2.round
MERGE (p1)-[:SAME_DRAFT_ROUND]-(p2)
```

In total, we had 3 node types of 1673 nodes and 6 relationship types of 124432 relationships.
Nodes:

- Player (10 properties, 1366 nodes)
- Position (1 property, 29 nodes)
- College (1 property, 278 nodes)

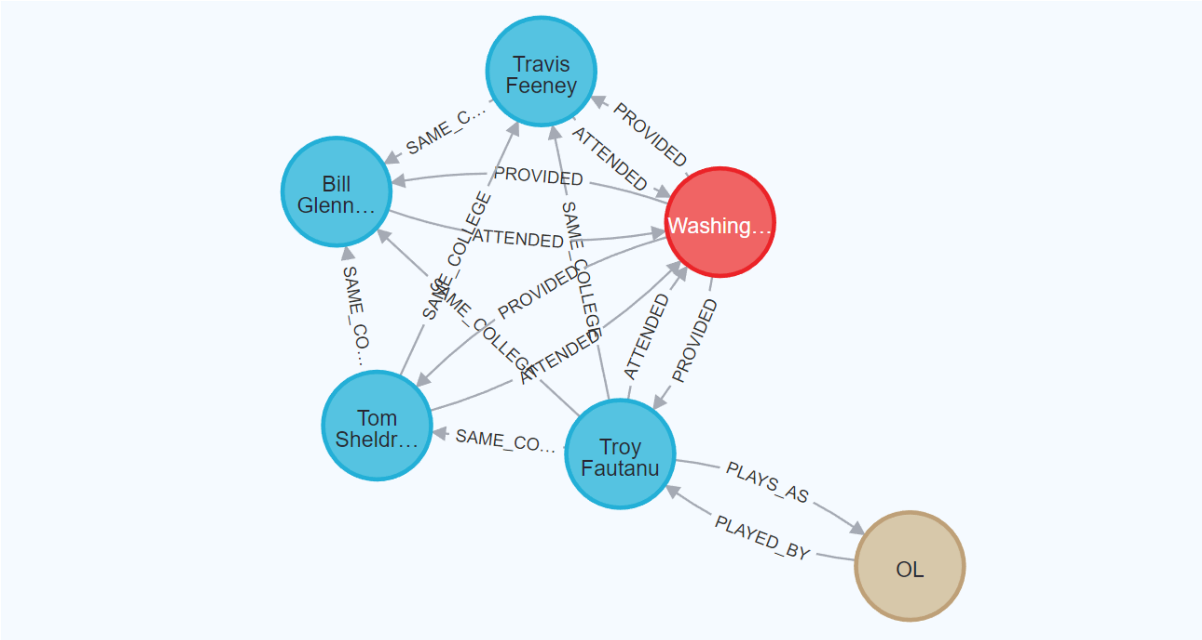
Relationships:

- College - Provided -> Player (1366 relationships)
- Player - Attended -> College (1366 relationships)
- Player - Plays as -> Position (1366 relationships)
- Position - Played by -> Player (1366 relationships)
- Player - Same college - Player (7882 relationships)
- Player - Same draft round - Player (42785 relationships)
- Player - Same position - Player (68767)

I put this data into a graph using gds.graph.project:

```
CALL gds.graph.project(
  'playerGraph',
  ['Player', 'Position', 'College'],
  ['PLAYS_AS', 'ATTENDED', 'SAME_COLLEGE', 'SAME_DRAFT_ROUND',
'SAME_POSITION', 'PLAYED_BY', 'PROVIDED']
) YIELD graphName, nodeCount, relationshipCount;
```

Here is a small sample of the structure of the graph:



Centrality Analysis:

Our primary aim for our project was to employ graph analytics techniques to identify patterns around the schools, positions and rounds that the Steelers prioritise when selecting players in the draft. By then comparing these characteristics to their most successful picks, we expected to find confirmation that these preferences are a major factor behind their elite performance. However, we still endeavoured to detect any discrepancies that could transform their draft recruitment practices in the future.

We began by using degree centrality to analyse which positions are most commonly drafted by the Steelers. Degree centrality tells us which nodes have the highest number of edges connecting with other nodes. By specifically returning the position nodes with the highest degree centrality, we retrieved the positions that had the most players assigned to them. This means they were the positions most drafted by the franchise and the degree was the amount of times the Steelers have drafted that position.

We ran this cypher query to return the 5 positions with the highest degree centrality:

```
CALL gds.degree.stream('playerGraph')
YIELD nodeId, score
WITH nodeId, score
MATCH (pos:Position) WHERE id(pos) = nodeId
RETURN pos.name AS Position, score as degree
ORDER BY degree DESC
LIMIT 5;
```

	Position	degree
1	"T"	176.0
2	"B"	153.0
3	"G"	119.0
4	"DB"	115.0
5	"LB"	112.0

The position with the highest degree centrality was T (Tackle: 176). The other positions in the top 5 were B (Back: 153), G (Guard: 119), DB (Defensive Back: 115) and LB (Line Backer:112). These nodes being so highly involved in the graph suggests that they have been identified as key positions for the team and a priority for the team. The high connectedness is also intuitive as these are positions that take up multiple spots on a roster so the team is likely to recruit more of them.

To consider whether the signings of these positions has contributed proportionately to the teams titles, we created a new graph to only include relationships involving players who have won a Super Bowl with the franchise. We created this graph using this cypher query:

```
CALL gds.graph.project.cypher(
  'sbGraph',
  'MATCH (p:Player) WHERE p.superbowl > 0 RETURN id(p) AS id UNION MATCH
(c:College) RETURN id(c) AS id UNION MATCH (pos:Position) RETURN id(pos) AS id',
  'MATCH (c:College)-[:PROVIDED]->(p:Player) WHERE p.superbowl > 0 RETURN id(c) AS
source, id(p) AS target UNION
   MATCH (p:Player)-[:ATTENDED]->(c:College) WHERE p.superbowl > 0 RETURN id(p) AS
source, id(c) AS target UNION
   MATCH (p1:Player)-[:PLAYS_AS]->(pos:Position)<-[:PLAYS_AS]-(p2:Player) WHERE
p1.superbowl > 0 AND p2.superbowl > 0 RETURN id(p1) AS source, id(p2) AS target
UNION
   MATCH (p1:Player)-[:SAME_DRAFT_ROUND]->(p2:Player) WHERE p1.superbowl > 0
AND p2.superbowl > 0 RETURN id(p1) AS source, id(p2) AS target UNION
   MATCH (p1:Player)-[:SAME_COLLEGE]->(p2:Player) WHERE p1.superbowl > 0 AND
p2.superbowl > 0 RETURN id(p1) AS source, id(p2) AS target UNION
   MATCH (pos:Position)-[:PLAYED_BY]->(p:Player) WHERE p.superbowl > 0 RETURN
id(pos) AS source, id(p) AS target'
) YIELD graphName, nodeCount, relationshipCount;
```

The positions with the highest degree centrality within this new graph were:

	Position	score
1	"DB"	17.0
2	"LB"	13.0
3	"WR"	10.0
4	"T"	10.0
5	"RB"	10.0

As expected, three of the positions, DB (17), LB (13) and T (10), were ranked highest for degree centrality amongst both graphs. The fact that these nodes still have high levels of incoming and outgoing relationships amongst championship winners validates the franchises tendency to favour those positions. The presence of new positions WR (Wide Receiver: 10) and RB (Running Back: 10) suggests that perhaps these are areas that the team should place more importance on. Conversely, it could be argued that the high conversion rate of recruits to champions shows that they have strong talent identification in that area.

Next, we investigated the colleges that have the strongest relationship to the Steelers within the graph. By extracting the colleges with the highest measure of betweenness, we found the nodes central to the network of connections between players and colleges. This might indicate that certain colleges are "preferred" or more connected to the Steelers' team strategy. This is the Cypher query I used to return the 5 colleges with the highest betweenness centrality:

```
CALL gds.betweenness.stream('playerGraph')
YIELD nodeId, score
WITH nodeId, score
MATCH (c:College) WHERE id(c) = nodeId
RETURN c.name AS College, score
ORDER BY score DESC
LIMIT 5;
```

	College	score
1	"Texas"	3184.2531158875577
2	"Notre Dame"	960.179490800362
3	"Pittsburgh"	846.8926407658173
4	"West Virginia"	693.8772544161945
5	"Maryland"	579.936383637506

Far and away the school with the highest betweenness centrality score was the University of Texas. As a school with a proud footballing history and a reputation for producing talenting athletes, it is unsurprising to learn it is the most influential College node in the graph. To inspect if these colleges are serving the franchise well by providing players who make a meaningful impact , we created another subgraph limited to relationships involving player nodes who had played 3 or more seasons as a primary starter.

```
CALL gds.graph.project.cypher(
  'startingGraph',
  'MATCH (p:Player) WHERE p.starts >= 3 RETURN id(p) AS id UNION MATCH (c:College)
RETURN id(c) AS id UNION MATCH (pos:Position) RETURN id(pos) AS id',
  'MATCH (c:College)-[:PROVIDED]->(p:Player) WHERE p.starts >= 3 RETURN id(c) AS
source, id(p) AS target UNION
   MATCH (p:Player)-[:ATTENDED]->(c:College) WHERE p.starts >= 3 RETURN id(p) AS
source, id(c) AS target UNION
   MATCH (p1:Player)-[:PLAYS_AS]->(pos:Position)<-[:PLAYS_AS]-(p2:Player) WHERE
p1.starts >= 3 AND p2.starts >= 3 RETURN id(p1) AS source, id(p2) AS target UNION
   MATCH (p1:Player)-[:SAME_DRAFT_ROUND]->(p2:Player) WHERE p1.starts >= 3 AND
p2.starts >= 3 RETURN id(p1) AS source, id(p2) AS target UNION
   MATCH (p1:Player)-[:SAME_COLLEGE]->(p2:Player) WHERE p1.starts >= 3 AND
p2.starts >= 3 RETURN id(p1) AS source, id(p2) AS target UNION
MATCH (pos:Position)-[:PLAYED_BY]->(p:Player) WHERE p.starts >= 3 RETURN id(pos)
AS source, id(p) AS target'
) YIELD graphName, nodeCount, relationshipCount;
```

1	"Georgia"	559.9596623870307
2	"Baylor"	371.12102224473017
3	"Indiana"	115.53894081007417
4	"Penn St."	90.54185566596227
5	"Clemson"	63.60793729749363

Surprisingly, none of the 5 colleges with the highest betweenness centrality score for the new graph were the same as the original. For players that went on to nail down a place in the starting line-up, the colleges that act as intermediaries linking these nodes together are completely different. There is an underlying imbalance between the colleges linked to the players contributing to the Steelers success and the colleges where they appear tend to draft most of their players from. This could be cause for the organisation to reevaluate the strategic importance of undervalued schools like Georgia and Baylor within their overall draft approach.

The final centrality measure we used to gain insight into the Steelers recruiting habits was PageRank. PageRank assigns a rank to each node based on the number and quality of relationships they have with other players, colleges, and positions. Players with higher PageRank scores are likely more central to the graph because they are connected to a lot of other players that went to the same college, were selected in the same round or play the same position. Therefore, it is likely that the schools, positions and rounds of these players are those that are heavily sought-after by the team and the players themselves provide examples of what could be described as the Steelers 'archetypal' candidate.

We retrieved the 5 players with the highest PageRank score via the following code:

```
CALL gds.pageRank.stream('playerGraph')
YIELD nodeId, score
WITH nodeId, score
MATCH (p:Player) WHERE id(p) = nodeId
RETURN p.name as Player, p.college as College, p.round as Round, p.position as Position,
score
ORDER BY score DESC
LIMIT 5;
```

1	"Kendrick Green"	"Illinois"	3	"OL"	4.395857780802358
2	"Logan Lee"	"Iowa"	6	"DL"	4.1466414969553
3	"Devin Bush Jr."	"Michigan"	1	"LB"	4.1030972427841395
4	"Spencer Anderson"	"Maryland"	7	"OL"	4.042143633793101
5	"Ryan Shazier"	"Ohio St."	1	"LB"	4.0152800251161285

Based on our assumptions about how PageRank would be implemented in the context of our project, we inferred that the positions, rounds and colleges of these players would represent those that are frequently targeted within prospective signings. Therefore, it was no surprise to see two of the top ranked players being Linebackers, one of the positions that had the highest degree centrality as well. The spread of draft picks from two in the first round to two in rounds 6 and 7 suggests that the team generally prioritise trading up their mid round picks and are happy to select a few top talents in exchange for having to enlist some less desirable graduates. To check if these trends of characteristics are reflected among players that go on to have lasting careers, we ran the query on a graph created only using player nodes who played over 50 games for the franchise.

```
CALL gds.graph.project.cypher(
  'played50GamesGraph',
  'MATCH (p:Player) WHERE p.games >= 50 RETURN id(p) AS id UNION MATCH
(pos:Position) RETURN id(pos) AS id UNION MATCH (c:College) RETURN id(c) AS id',
  'MATCH (p:Player)-[:PLAYS_AS]->(pos:Position) WHERE p.games >= 50 RETURN id(p)
AS source, id(pos) AS target UNION '
  'MATCH (p:Player)-[:ATTENDED]->(c:College) WHERE p.games >= 50 RETURN id(p) AS
source, id(c) AS target UNION '
  'MATCH (p1:Player)-[:SAME_COLLEGE]->(p2:Player) WHERE p1.games >= 50 AND
p2.games >= 50 RETURN id(p1) AS source, id(p2) AS target UNION '
  'MATCH (p1:Player)-[:SAME_DRAFT_ROUND]->(p2:Player) WHERE p1.games >= 50
AND p2.games >= 50 RETURN id(p1) AS source, id(p2) AS target'
) YIELD graphName, nodeCount, relationshipCount;
```

	Player	College	Round	Position	score
1	"Najee Harris"	"Alabama"	1	"RB"	2.0575643781109854
2	"Pat Freiermuth"	"Penn St."	2	"TE"	1.8625169956994565
3	"Alex Highsmith"	"Charlotte"	3	"LB"	1.762997572579968
4	"Devin Bush Jr."	"Michigan"	1	"LB"	1.6903794370667775
5	"Isaiah Buggs"	"Alabama"	6	"DT"	1.5167823314634394

From reviewing the top PageRanked nodes from players who played over 50 games, we picked up on a few key details. The recurring presence of LB throughout our exploration as an influential node cements the positive impact of the Steelers evidently prioritising it as an area of recruitment. RB (Running Back) being a highly connected position within degree centrality and also being linked to the most important player by this metric implies that the draft holds great potential to yield a future superstar in this role. Two first round picks again having such high connections shows that, although it may seem obvious, using first round picks to take on premium prospects is reaping its rewards and there is value in keeping these picks as opposed to trading them for other players. However, once again these results highlight the conflict between the colleges that the team tends to recruit from compared to where their most valuable assets actually originate. Michigan is the only school on both lists for PageRank and Alabama featuring twice amongst the top 5 for players with over 50 games played means that it must be extremely highly connected amongst stalwarts of the organisation but it is not being reflected in the overall recruitment policy

Community Detection Analysis

Our goal for this analysis was to identify clusters of players who share similar attributes such as college attended, position played and draft round. This can help us to understand the characteristics of the most successful and least successful player clusters and therefore provide insights into the Steelers draft strategies.
For community detection, all relationships in the graph must be bidirectional. To do this we created a subgraph that contains only player nodes and the relationships between them, such as SAME_COLLEGE, SAME_POSITION, and SAME_DRAFT_ROUND. These relationships were modeled to represent key connections between players that might influence their performance or clustering.

```
CALL gds.graph.project.cypher(
  'triangleGraph',
  'MATCH (p:Player) RETURN id(p) AS id',
  'MATCH (p1:Player)-[:SAME_COLLEGE]-(p2:Player) RETURN id(p1) AS source, id(p2) AS
target
   UNION
   MATCH (p1:Player)-[:SAME_POSITION]-(p2:Player) RETURN id(p1) AS source, id(p2) AS
target
   UNION
   MATCH (p1:Player)-[:SAME_DRAFT_ROUND]-(p2:Player) RETURN id(p1) AS source,
id(p2) AS target'
) YIELD graphName, nodeCount, relationshipCount;
```

We applied the louvain modularity and used it to detect communities of players based on shared relationships. The communities were added as a property of each player.

```
CALL gds.louvain.stream('triangleGraph')
YIELD nodeId, communityId
MATCH (p:Player) WHERE id(p) = nodeId
SET p.communityId = communityId;
```

We ran the below code to find which communities had the best and worst performing players based on their average SuperBowl Wins, average Games Played, average ProBowl Selections and average Starts and found the results.

```
MATCH (p:Player)
WITH p.communityId AS communityId, COUNT(p) AS numPlayers,
    AVG(p.games) AS avgGames,
    AVG(p.superbowl) AS avgSuperbowlWins,
    AVG(CASE WHEN p.pb > 0 THEN 1 ELSE 0 END) AS avgProBowlSelections,
    AVG(p.starts) AS avgStarts
ORDER BY avgSuperbowlWins DESC, avgGames DESC, avgStarts DESC
LIMIT 10
RETURN communityId, numPlayers, avgGames, avgSuperbowlWins,
avgProBowlSelections, avgStarts;
```

Best performing communities:

	communityId	numPlayers	avgGames	avgSuperbowlWins	avgProBowlSelections	avgStarts
1	763	98	73.9795918367347	0.2244897959183674	0.28571428571428575	3.8469387755102047
2	1324	459	52.278867102396546	0.17647058823529405	0.12200435729847489	1.8692810457516347
3	871	176	22.77840909090908	0.1250000000000001	0.04545454545454546	1.0624999999999996
4	642	247	24.283400809716603	0.10526315789473688	0.0647773279352227	0.7449392712550604
5	1205	120	21.208333333333325	0.08333333333333327	0.04999999999999999	1.0416666666666676

Worst performing communities:

	communityId	numPlayers	avgGames	avgSuperbowlWins	avgProBowlSelections	avgStarts
1	1086	154	0.9155844155844156	0.0	0.0	0.01948051948051948
2	838	112	8.276785714285712	0.0	0.035714285714285705	0.41964285714285715
3	1205	120	21.208333333333325	0.08333333333333327	0.04999999999999999	1.0416666666666676
4	871	176	22.77840909090908	0.1250000000000001	0.04545454545454546	1.0624999999999996

We then grouped the most common draft rounds, colleges and positions among players in the most successful, second most successful and least successful communities for analysis:

```
CALL {
  MATCH (p:Player)
  WHERE p.communityId = 763
  RETURN p.round AS value, COUNT(*) AS count, "DraftRound" AS type
  ORDER BY count DESC
  LIMIT 5
```

```
    UNION
    MATCH (p:Player)
    WHERE p.communityId = 763
    RETURN p.college AS value, COUNT(*) AS count, "College" AS type
    ORDER BY count DESC
    LIMIT 5
    UNION
    MATCH (p:Player)
    WHERE p.communityId = 763
    RETURN p.position AS value, COUNT(*) AS count, "Position" AS type
    ORDER BY count DESC
    LIMIT 5
}
WITH value, count, type
RETURN
    COLLECT(CASE WHEN type = "DraftRound" THEN value END) AS DraftRounds,
    COLLECT(CASE WHEN type = "College" THEN value END) AS Colleges,
    COLLECT(CASE WHEN type = "Position" THEN value END) AS Positions;
```

Most successful community:

DraftRounds	Colleges	Positions
[1, 17, 13, 15, 12]	["Pittsburgh", "Ohio St.", "Georgia", "USC", "Michigan St."]	["QB", "FB", "DB", "WR", "RB"]

Second most successful community:

DraftRounds	Colleges	Positions
[3, 6, 4, 5, 2]	["Notre Dame", "Penn St.", "Pittsburgh", "Georgia", "Michigan"]	["LB", "DB", "WR", "DE", "RB"]

Least successful community:

DraftRounds	Colleges	Positions
[12, 15, 19, 20, 22]	["Pittsburgh", "Pennsylvania", "Florida", "Notre Dame", "Duquesne"]	["B", "BB"]

We can now use this to analyse these communities. The most successful community is made up of 98 players with an average of 74 games played between them. Considering the NFL season consists of only 17 regular season games plus up to a maximum of 6 postseason games in the playoffs, this is a very high average number, which shows the group have been consistently active in the NFL.This metric is a strong indicator of durability, reliability, and career longevity, indicating that these players have established themselves as key contributors to the Steelers. They have an average of 0.22 average SuperBowl wins and 0.28 average ProBowl selections, showing that this group contains some extremely talented players.

Unsurprisingly, the most successful group contains a high number of players from round 1 in the draft, which shows that the most desirable talents in the draft often tend to perform well. However, surprisingly, rounds 17, 13, 15 and 12 also appear, indicating that these players massively outperformed expectations. This suggests effective scouting, development programs, or a strong work ethic and determination among these players. Players from very late rounds may have been overlooked during the draft for various reasons, such as coming from smaller colleges, lacking flashy statistics, or not performing well at combines and the ability for the Steelers to capitalise on these late picks may be a key component to being such a successful franchise. This was very surprising to find, as we assumed the first four or five rounds would have had the biggest impact for them.

If we compare the draft rounds to the second most successful community, we can see that they are a lot more predictable, being 3, 6, 4, 5 and 2. This shows consistency in the Steelers mid round picks, who are usually considered to be solid contributors to teams but not necessarily the standout stars expected from the first round. The success of this community demonstrates that the Steelers excel at identifying reliable talent in the middle rounds, which are crucial for building depth and filling specific needs on a roster. An average of 52 games played across 459 players in this community shows that these players can be seen as successful picks, being durable and reliable. They contributed to an average of 0.17 SuperBowls which is not far from the 0.22 of the most successful community, however, they only have an average of 0.12 ProBowl selections, confirming that they tend to be solid contributors but not the stars that can be found in the most successful group.

It's not surprising to find that the least successful community consists of low round draft picks - 12, 15, 19, 20 and 22. The low number of games of 0.91, 0.0 SuperBowl wins and 0.0 ProBowl selections shows that these players did not have illustrious careers in the NFL, implying that the Steelers may have been using these late rounds to fill roster gaps as opposed to looking for high impact contributors.

We found that the most successful community was made up of positions QuarterBack, FullBack, Defensive Back, Wide Receiver and Running Back. The inclusion of Quarterback, Wide Receiver, and Running Back highlights the importance of skill positions in determining a player's success. These roles are often central to a team's offensive output and tend to receive significant attention during the draft. Quarterbacks, in particular, are pivotal for team performance, and their success is closely tied to the success of the team. The presence of Defensive Backs reflects the importance of defensive contributions. Fullbacks are a declining position in modern NFL teams, with fewer teams utilizing them in their offenses. Their inclusion in the most successful community shows these players found ways to remain impactful, despite the position's reduced prominence. This was surprising to see in the most successful group.

The second most successful group consisted of LineBacker, Defensive Back, Wide Receiver, Defensive End and Running Back. This group has a strong emphasis on defense, with Linebackers, Defensive Backs, and Defensive Ends forming a majority. While the most successful group leaned heavily on offensive positions like Quarterback and Fullback, this group shifts the focus toward defense and a balanced mix. This emphasises the mid-round draft strategy of getting reliable players as opposed to superstars.

Lastly, the college that players were drafted from show us a lot about the Steelers draft strategy. The most successful group tended to be drafted from Pittsburgh, Ohio State, Georgia, USC and Michigan State. Colleges like Ohio State, Georgia, and Michigan are consistently among the top-ranked football programs in the NCAA and have some of the

highest number of players in the NFL currently. These schools are known for producing NFL-ready talent, thanks to their robust coaching staff, high-quality training facilities, and ability to attract elite athletes. The Steelers' focus on drafting players from these programs suggests a strategy aimed at securing talent with a proven track record of success at a high level of competition.
The inclusion of the University of Pittsburgh highlights a local connection. Drafting locally may offer benefits, such as detailed knowledge of the players and potentially stronger motivation for these players to succeed for their home team. This could indicate a strategy of leveraging regional familiarity while also gaining community support.

The second most successful community also contains Georgia, Michigan and Pittsburgh. This confirms the success that the Steelers have had drafting from these colleges. The addition of two more highly proven football colleges such as Notre Dame and Penn state shows the Steelers strategy of drafting from highly rated and proven colleges.

The least successful group includes a mix of highly prestigious programs (e.g., Notre Dame and Florida) alongside smaller or less football-dominant schools (e.g., Pennsylvania and Duquesne) and also includes local college Pittsburgh. This diversity emphasises that while the Steelers have usually targeted top-tier programs, they have also taken chances on players from less competitive or less football-focused colleges. We were surprised to see colleges like Notre Dame and Florida which are historically strong programs, yet their representation in this group suggests underachievement, particularly Florida having not appeared in the two most successful groups.

Conclusion

Our analysis of the Pittsburgh Steelers' NFL Draft history shows a clear draft strategy to which their success can be accredited. Our centrality investigation highlighted an emphasis on prioritising certain positions, considering draft round value and utilising connections with specific colleges. By comparing the overall data to subgraphs of top performers, we were able to identify the most productive policies along with those that should be amended. The Steelers show a clear preference for drafting Tackles, Backs, Guards, Defensive Backs, and Linebackers. This focus aligns with the positions that have contributed most significantly to their Super Bowl victories. The team's success in identifying talented players in later draft rounds (rounds 12-17) is noteworthy and indicates an excellent scouting system. However, there's a disconnect between the colleges they frequently draft from and those that have produced their most impactful players. This underlines a potential area for improvement in their draft strategy.

Community detection analysis further illuminates the Steelers' approach - The most successful player community is characterised by a high proportion of first-round picks, but also includes players from surprisingly late rounds, reinforcing the team's ability to find hidden gems. This group also features a mix of offensive and defensive positions, highlighting the importance of a balanced roster. The second most successful community, mainly made up of mid-round picks, shows the team's consistent success in finding reliable contributors.

The Steelers' sustained success over decades can be attributed, in part, to their effective draft strategy. However, the analysis also reveals potential areas for improvement and change. Re-evaluating their approach to scouting players from specific colleges or capitalising on the potential of running backs emerging from later draft rounds could lead to even greater success in identifying future stars.

By continuing to analyse and adapt their draft strategy, the Pittsburgh Steelers can aim to maintain their position as the NFL's most successful franchises.

Appendix:

Pittsburgh Steelers Draft History Dataset: https://www.pro-football-reference.com/teams/pit/draft.htm
Pittsburgh Steelers Roster Dataset: https://www.pro-football-reference.com/teams/pit/1974_roster.htm
Final Data Frame:
https://docs.google.com/spreadsheets/d/154DBiu84s0cSPIWgnsFxksAG9Ws__wwct07ncoNq_1g/edit?usp=sharing
Full Neo4J Code:
https://docs.google.com/document/d/1e-cO9tz2rj3XwQD4wqJFUVSpM8Xs1bzJIVpb31BrcMs/edit?usp=sharing