# CA336 - Graph Analytics Assignment

Ronan Kelly
21374881

## International Rugby Fixtures

## Dataset and Problem definition

With a keen interest in sport I knew I wanted to do my research project on some sport dataset, and through some research I stumbled upon a dataset on Kaggle containing every international rugby result for each of the Six-Nations and Rugby Championship teams since 1871.

The teams contained in the dataset are Ireland, England, Scotland, Wales, France, Italy, South Africa, New Zealand, Australia and Argentina. The dataset only contains games where these teams are facing each other and not ones containing any teams outside of these big 10 nations.

The dataset contained columns, date which was the date of the match, home team and away team, their respective scores, the competition the match was in, the city and country that the match was played in, whether the game was in a neutral venue, and if the game was in the world cup. This amount of data allowed for lots of analysis on various different topics.

The main problem I'm going to address with this data is using centrality to find out which national team has been most influential in terms of the growth of rugby over the last 140 years. Also finding out which teams have acted as significant connectors or bridges within the international rugby network. This research will not only show me what nations have been most influential but also which ones have been the best at spreading the game around the world. As the data contained dates back to the 1870s historical analysis can be crucial in understanding the nation who was most impactful in the development of the game worldwide.

I also want to use community detection to discover if any communities exist that correspond to geographic regions indicating a rivalry of some sort.

As this dataset only includes games containing the big 10 teams, it does limit our ability to analyse which nations had the most impact on the spread of the game in the more modern era, and also doesn't allow us to understand how much of an influence the world cup had on spreading the game around the globe.

The use of graph analytics will allow me to understand and analyse these links between nations and how they impacted the development of rugby worldwide. The interconnected nature of this graph allows me to possibly uncover hidden patterns in the data.

## Data Preparation

As mentioned previously I got my dataset from kaggle here:
*https://www.kaggle.com/code/lylebegbie/international-rugby-results-data-visualization*

This data initially looked like this:

| | date | home_team | away_team | home_score | away_score | competition | stadium | city | country | neutral | world_cup |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2687 | 2022-08-27 | New Zealand | Argentina | 18 | 25 | 2022 Rugby Championship | Rugby League Park | Christchurch | New Zealand | False | False |
| 2688 | 2022-09-03 | New Zealand | Argentina | 53 | 3 | 2022 Rugby Championship | Waikato Stadium | Hamilton | New Zealand | False | False |
| 2689 | 2022-09-03 | Australia | South Africa | 8 | 24 | 2022 Rugby Championship | Sydney Football Stadium | Sydney | Australia | False | False |

- Original Layout of the dataset

The data was fairly clean with no major pre-processing necessary prior to creating my graph, but I wanted to normalise some of the figures, as well as adding a column saying who won each game. I also wanted to remove the stadium column from my as I didn't think it would be necessary for the analysis I was trying to achieve.

So first of all I used Pandas and Python to add a winner column to my dataset this was the code:

```
import pandas as pd
file_path = '/desktop/results.csv'
rugby_data = pd.read_csv(file_path)
rugby_data['winner'] = rugby_data.apply(lambda row: row['home_team']
if row['home_score'] > row['away_score'] else
(row['away_team'] if row['home_score'] < row['away_score'] else
'Draw'), axis=1)
```

This code created a column titled 'Winner' saying which nation won that specific game. My next aim was to normalise the data column so that it only only said the year this would make analysis a lot easier as we could link together matches by the year they were played in.

This was the code I used to normalise the date column:

```
import pandas as pd
rugby_data = pd.read_csv('results.csv')
rugby_data['date'] = pd.to_datetime(rugby_data['date']).dt.year
rugby_data.to_csv('normalised_rugby_results.csv', index=False)
```

This made my final dataset look like:

| date | home_team | away_team | home_score | away_score | competition | city | country | neutral | world_cup | winner |
|---|---|---|---|---|---|---|---|---|---|---|
| 1871 | Scotland | England | 1 | 0 | 1871 Scotlan | Edinburgh | Scotland | FALSE | FALSE | Scotland |
| 1872 | England | Scotland | 2 | 1 | 1871-72 Hon | London | England | FALSE | FALSE | England |
| 1873 | Scotland | England | 0 | 0 | 1872-73 Hon | Glasgow | Scotland | FALSE | FALSE | Draw |
| 1874 | England | Scotland | 1 | 0 | 1873-74 Hon | London | England | FALSE | FALSE | England |
| 1875 | England | Ireland | 2 | 0 | 1874-75 Hon | London | England | FALSE | FALSE | England |
| 1875 | Scotland | England | 0 | 0 | 1874-75 Hon | Edinburgh | Scotland | FALSE | FALSE | Draw |
| 1875 | Ireland | England | 0 | 1 | 1875-76 Hon | Dublin | Ireland | FALSE | FALSE | England |
| 1876 | England | Scotland | 1 | 0 | 1875-76 Hon | London | England | FALSE | FALSE | England |
| 1877 | England | Ireland | 2 | 0 | 1876-77 Hon | London | England | FALSE | FALSE | England |
| 1877 | Ireland | Scotland | 0 | 6 | 1876-77 Hon | Belfast | Ireland | FALSE | FALSE | Scotland |

- Final Layout of the dataset

The final piece of cleaning I had to do was that when I downloaded the csv file in some cases the hyphen between two years would show up as a string of characters like this, 1924‚Äì25, because of the nature of the rugby season it's rarely spread over two calendar years so it happened very rarely. Therefore I was able to fix it in each individual case by hand fairly quickly. There were also a few occasions when the competition value was blank and on that occasion I called them “Test Match” as most rugby games that aren’t part of a named competition are called a test match.

Now that I had my data all cleaned and normalised so that I was happy with the format of it, it was time to build the graph. I first had to decide what I was going to select my nodes and then also what relationships I was going to create between them.

For the nodes I decided I was going to have one for the match, one for each team in the match, one for the competition that the match took place in, one for the the city the game took place in, one for the country that the city is in, and one for the year the game took place. because there was no match column in the dataset I had to create it using the year, the home and the away team

As for the relationships between the nodes I went for linking each team to the match and also linking the match back to the team that won the game. I also decided to link the match to the competition it was in as well as the city it took place in. I linked the city the game took place in to the country that that city is located in and I also linked the match to the year that it took place in.

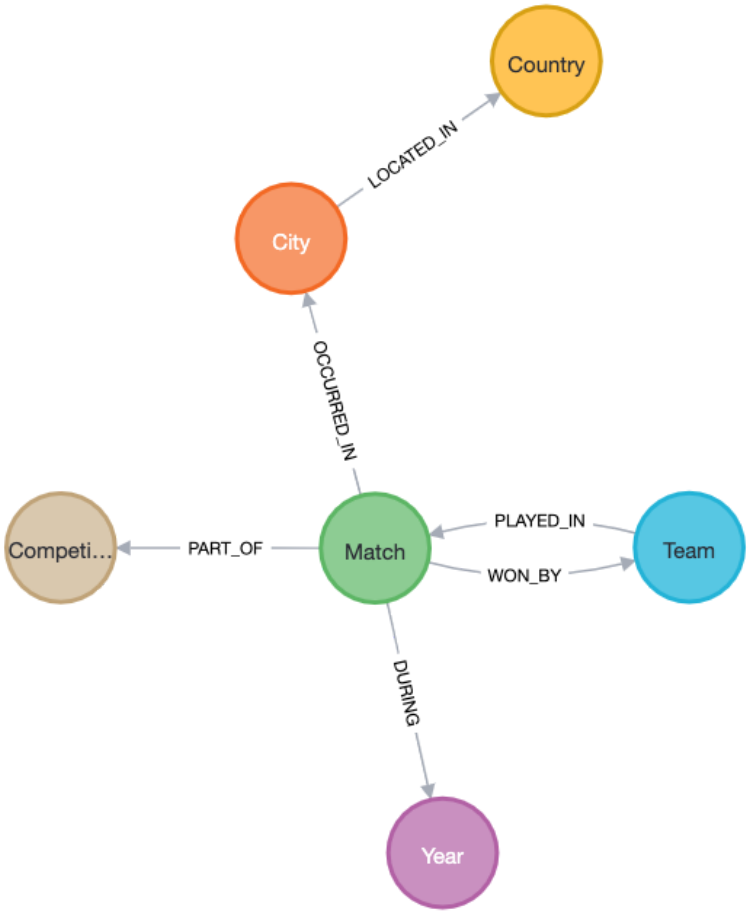
This is the code I used to load the dataset into Neo4j:

```
LOAD CSV WITH HEADERS FROM 'file:///normalized_rugby_results.csv' AS row
MERGE (homeTeam:Team {name: row.home_team})
MERGE (awayTeam:Team {name: row.away_team})
MERGE (winnerTeam:Team {name: row.winner})
MERGE (match:Match {id: row.date + '-' + row.home_team + '-' + row.away_team})
MERGE (competition:Competition {name: row.competition})
MERGE (city:City {name: row.city})
```

```
MERGE (country:Country {name: row.country})
MERGE (year:Year {year: row.date})

MERGE (homeTeam)-[:PLAYED_IN]->(match)
MERGE (awayTeam)-[:PLAYED_IN]->(match)
MERGE (match)-[:WON_BY]->(winnerTeam)
MERGE (match)-[:PART_OF]->(competition)
MERGE (match)-[:OCCURRED_IN]->(city)-[:LOCATED_IN]->(country)
MERGE (match)-[:DURING]->(year)
```

The graph has 3428 nodes and 17377 relationships and this is how the graph is structured:




- Graph Structure

## The Experiments

So to begin I'm going to use centrality to discover which nation has been the best or most influential in the entire history of rugby. I'm going to use a few functions to achieve this, such as using degree centrality to see what nation has played the most matches and also won the most matches. I also think it would be interesting to use degree centrality to see which nation has played the most matches in the World Cup as this is deemed the biggest competition in the sport so a team who has played the most games could be regarded as the biggest and most successful.

So I first used degree centrality for finding out which team had played the most games:

```
MATCH (t:Team)-[:PLAYED_IN]-()
RETURN t.name AS Team, COUNT(*) AS DegreeCentrality
ORDER BY DegreeCentrality DESC
```

I then modified it a small bit to find out which team had won the most matches:

```
MATCH (t:Team)-[:WON_BY]-()
RETURN t.name AS Team, COUNT(*) AS DegreeCentrality
ORDER BY DegreeCentrality DESC
```

Finally I found out which team had played the most World Cup matches, the issue with this first of all which I will talk more thoroughly about in the next section is that this dataset only includes the top 10 teams, and because the World Cup contains teams outside these top 10, the results may be skewed as a better team might have played more games against lower ranked nations, however, the code I used to do this is:

```
MATCH (t:Team)-[:PLAYED_IN]-(m:Match)-[:PART_OF]->(c:Competition)
WHERE c.name CONTAINS 'World Cup'
RETURN t.name AS Team, COUNT(DISTINCT m) AS DegreeCentrality
ORDER BY DegreeCentrality DESC
```

I will also check which team has won the most World cup games might avoid that issue i mentioned earlier, I will adjust the code the same way as before to get this result:

```
MATCH (t:Team)-[:WON_BY]-(m:Match)-[:PART_OF]->(c:Competition)
WHERE c.name CONTAINS 'World Cup'
```

```
RETURN t.name AS Team, COUNT(DISTINCT m) AS DegreeCentrality
ORDER BY DegreeCentrality DESC
```

I’m going to use betweenness centrality to discover which nations have acted as the best connectors between the nodes, this will show me which nation has been the best at growing the game and spreading it around the globe.

So first of all i has to create a graph projection of my code to allow a betweenness centrality algorithm to work and i did that using this code:

```
CALL gds.graph.project(
'rugbyGraph', // Name of the graph projection
['Team', 'Match', 'Competition', 'Year','City','Country'], // Node labels
{
PLAYED_IN: { // Relationship type
type: 'PLAYED_IN',
orientation: 'UNDIRECTED' // Assuming the relationship is undirected
},
WON_BY: { // Relationship type
type: 'WON_BY',
orientation: 'UNDIRECTED'
},
PART_OF: { // Relationship type
type: 'PART_OF',
orientation: 'UNDIRECTED'
},
OCCURRED_IN: { // Relationship type
type: 'OCCURRED_IN',
orientation: 'UNDIRECTED'
},
LOCATED_IN: { // Relationship type
type: 'LOCATED_IN',
orientation: 'UNDIRECTED'
},
DURING: { // Relationship type
type: 'DURING',
orientation: 'UNDIRECTED'
}
}
)
```

This code created a graph projection that allowed me to run various different algorithms on.

Now it was time to run my betweenness algorithm on the graph, I did that using this code:

```
CALL gds.betweenness.stream('rugbyGraph')
YIELD nodeId, score
MATCH (n) WHERE id(n) = nodeId AND 'Team' IN labels(n)
RETURN n.name AS Team, score
ORDER BY score DESC
```

This gave me a score showing which nation had been the best at spreading the game around the world.

For the other end of my project on community detection, I will use it to discover some sort of communities which may shed light on the different rivalries that might exist in the international rugby world.

I first tried to run community detection algorithms on the graph that I created above but the results that came from that weren't useful for analysis as none of the teams were part of any communities with each other. I then decided to create a new graph containing just the team and match nodes, this would keep the usefulness as regards testing for rivalries while also reducing the number of nodes making the algorithm more effective. I achieved this with the following code:

```
CALL gds.graph.project(
'RugbyGraph', // Name of the graph projection
['Team', 'Match'], // Node labels
{
PLAYED_IN: { // Relationship type
type: 'PLAYED_IN',
orientation: 'UNDIRECTED' // Assuming the relationship is undirected
}
}
);
```

I first tried to run the Louvain algorithm to find clusters within the graph and then I moved onto the Label propagation algorithm so that I could compare the results.

```
CALL gds.louvain.stream('RugbyGraph')
YIELD nodeId, communityId
MATCH (n) WHERE id(n) = nodeId AND 'Team' IN labels(n)
RETURN n.name AS Team, communityId
ORDER BY communityId
```

```
CALL gds.labelPropagation.stream('RugbyGraph')
YIELD nodeId, communityId
MATCH (n) WHERE id(n) = nodeId AND 'Team' IN labels(n)
RETURN n.name AS Team, communityId
ORDER BY communityId
```

## Analysis of Results

Finding the most influential nation in rugby can be an interesting task as you can take into account games played, but then wins could be deemed a better indicator of influence as it would show the most successful teams. I decided to try both approaches so that I could compare how these results differed and how they actually relate to the world of rugby that we can see.

| | Team | DegreeCentrality |
|---|---|---|
| 1 | "England" | 685 |
| 2 | "Wales" | 650 |
| 3 | "Ireland" | 639 |

- Nations who have played the most international rugby matches

| | Team | DegreeCentrality |
|---|---|---|
| 1 | "England" | 373 |
| 2 | "France" | 321 |
| 3 | "New Zealand" | 318 |

- Nations who have won the most international rugby matches

As we can see these results differ significantly from each other despite England remaining at the top Wales and Ireland are replaced by France and New Zealand.

This is interesting but also to be expected as although Wales and Ireland have been around for a long time and somewhat successful New Zealand would be regarded as one of the greatest rugby nations of all time and their accolades speak for themselves. Also nations in the southern hemisphere tend to play less matches

annually than northern hemisphere nations, so it's not surprising that we don’t see New Zealand in the top set of results.

I also thought it would be interesting to analyse the successfulness of nations in World Cups, to see if this would mirror the pattern from above.

| | Team | DegreeCentrality |
|---|---|---|
| 1 | "England" | 50 |
| 2 | "France" | 46 |
| 3 | "Wales" | 46 |

- Most games played at a World Cup

| | Team | DegreeCentrality |
|---|---|---|
| 1 | "New Zealand" | 42 |
| 2 | "England" | 33 |
| 3 | "France" | 30 |

- Most wins at a World Cup

It's clear to see from this how superior New Zealand have been since the beginning of the World Cups. I do think it's interesting that South Africa doesn't show up on either of these lists, this is strange as the nation who have won the most World Cups, this could imply easier routes to the final than other countries.

The main issue with this set of results is that the data only includes data from the top 10 teams listed at the start. Meaning that if as a nation you get drawn in an easier group containing lower ranked countries those wins aren’t counted in these results, making it slightly limited in that regard.

However, New Zealand would be deemed as the greatest and likely the most influential rugby nation of the last 40 years which is evident from these results. Both culturally through the ‘Haka’ and through sheer sporting success and dominance every child has grown up hearing the name the “All Blacks”.

Both these sets of results really show the impact England have had on the rugby world, they’ve played and won the most games, while also playing the most number of games in world cups and having the second most number of wins.

Rugby began as a very small sport in England roughly 140 years ago, and for a while it remained contained within the 4 small nations of the United Kingdom of the time. However, for the development of the game the onus was on these few countries to spread the game around the globe. Because of this I thought it would be interesting to use betweenness centrality to view which nation had the most influence when it came to spreading the game worldwide.

| | Team | score |
|---|---|---|
| 1 | "England" | 1176886.5688358685 |
| 2 | "Wales" | 1073142.4618627043 |
| 3 | "France" | 1044479.8241367362 |

- Top 3 Betweenness scores

This shows to me that England, Wales and France have acted as the best connectors to other countries, this would imply that these countries have done the most in growing the game worldwide. Through playing matches against initially smaller rugby nations that have ultimately grown into the powerhouses that we have in the southern hemisphere today.

Even though rugby may be regarded as a gentleman's game, rivalries remain a key part of the game today. Rivalries can be an important characteristic for growing the game as well as developing proper die hard fan bases. I thought it would be interesting to use two community detection algorithms to first of all analyse different rivalries that can be found but also to compare how these algorithms work on this sports data.

| Team | communityId |
|---|---|
| Draw | 10 |
| Scotland | 15 |
| England | 15 |
| Ireland | 15 |
| Wales | 32 |
| Australia | 151 |
| New Zealand | 151 |
| France | 169 |
| South Africa | 177 |
| Italy | 448 |
| Argentina | 500 |

- Louvain Algorithm community detection results

| Team | communityId |
|---|---|
| "Draw" | 10 |
| "Scotland" | 11 |
| "England" | 11 |
| "Ireland" | 11 |
| "Wales" | 11 |
| "Australia" | 11 |
| "New Zealand" | 11 |
| "France" | 11 |
| "South Africa" | 11 |
| "Argentina" | 11 |
| "Italy" | 11 |

- Label Propagation Algorithm community Detection results

As we can see these two sets of results are giving wildly different values. The Louvain seems like it will be much better for analysis as all the values in the Label-Propagation algorithm are the same.

So looking at the results from the Louvain algorithm we can see a few communities that may imply some sort of rivalry or at least a regular encounter. As we look at the first cluster of Ireland, England and Scotland, there are a couple of notably rivalries we can see. The rivalry between Ireland and England speaks for itself, it runs deeper than purely sporting and dates back hundreds of years, its probably most interesting though in terms of rugby as it's the only sport where Ireland play as one whole island. This creates an interesting dynamic as in some ways its the english playing against people who may regard themselves as british. The game is always a tightly contested one and oftentimes can decide such accolades as 'The Six Nations Grand Slam'.

The England and Scotland rivalry is another major one in all sports and especially rugby. The game dates back to the very first rugby back in 1871, and to this day every game between the nations is one for a trophy known as "The Calcutta Cup". Games between these two are always incredibly exciting especially in recent years as Scotland have become a major player in the international rugby scene.

There isn't much of a rivalry between Ireland and Scotland more of a friendship as the two nations see a lot of similarities between each other through their celtic pasts. They are also some of the oldest teams so the fact they have faced each other lots is understandable.

The only other cluster we can find from our Louvain algorithm results is between New Zealand and Australia. This is another major rivalry in the world of rugby, however, more similar to the Ireland and Scotland rivalry, it is not steeped in a cultural hatred of each other and more of a cultural similarity. Both nations share a lot of similar characteristics and it's understandable a friendly rivalry would develop once sport got involved. As two of the most successful nations of the modern era, they often have faced in world cup matches this has only added to the rivalry. They often play test series' against each other where they compete for what is known as "The Bledisloe Cup", this is always a major series in the rugby calendar that can bring about some of the most exciting matches around.

Overall the Louvain algorithm does a good job of finding the major rivalries that exist in the rugby world, however, the Label-Propagation algorithm does a much poorer job of finding these rivalries, not really separating them at all.

## Conclusion

The use of graph analytics algorithms has shown some very insightful results. The centrality analysis identified the nations that have had the greatest impact on the history of rugby as well as spreading the game globally. Nations with high degree centrality highlight their active participation as well as their influence on the sport. Whereas nations with high betweenness centrality show their influence on the spread of the game around the world. The community detection analysis effectively showcased how rivalries in rugby have developed over time.

These findings not only show a historical view on rugby's growth but also provides a nuanced understanding of the sport's current landscape. It shows how interconnected the world of rugby is even throughout history. While the limitations of the dataset need to be factored in, the results still offer an interesting insight into the history and continued development of the game.

## Appendix

How to build the graph:
The CSV:
https://drive.google.com/file/d/1xA1evuTQM7CjxHFTmbn9RtE1t4WuA_rV/view?usp=sharing

The code:

```
LOAD CSV WITH HEADERS FROM 'file:///normalized_rugby_results.csv' AS row
MERGE (homeTeam:Team {name: row.home_team})
MERGE (awayTeam:Team {name: row.away_team})
MERGE (winnerTeam:Team {name: row.winner})
MERGE (match:Match {id: row.date + '-' + row.home_team + '-' + row.away_team})
MERGE (competition:Competition {name: row.competition})
MERGE (city:City {name: row.city})
MERGE (country:Country {name: row.country})
MERGE (year:Year {year: row.date})

MERGE (homeTeam)-[:PLAYED_IN]->(match)
MERGE (awayTeam)-[:PLAYED_IN]->(match)
MERGE (match)-[:WON_BY]->(winnerTeam)
MERGE (match)-[:PART_OF]->(competition)
MERGE (match)-[:OCCURRED_IN]->(city)-[:LOCATED_IN]->(country)
MERGE (match)-[:DURING]->(year)
```