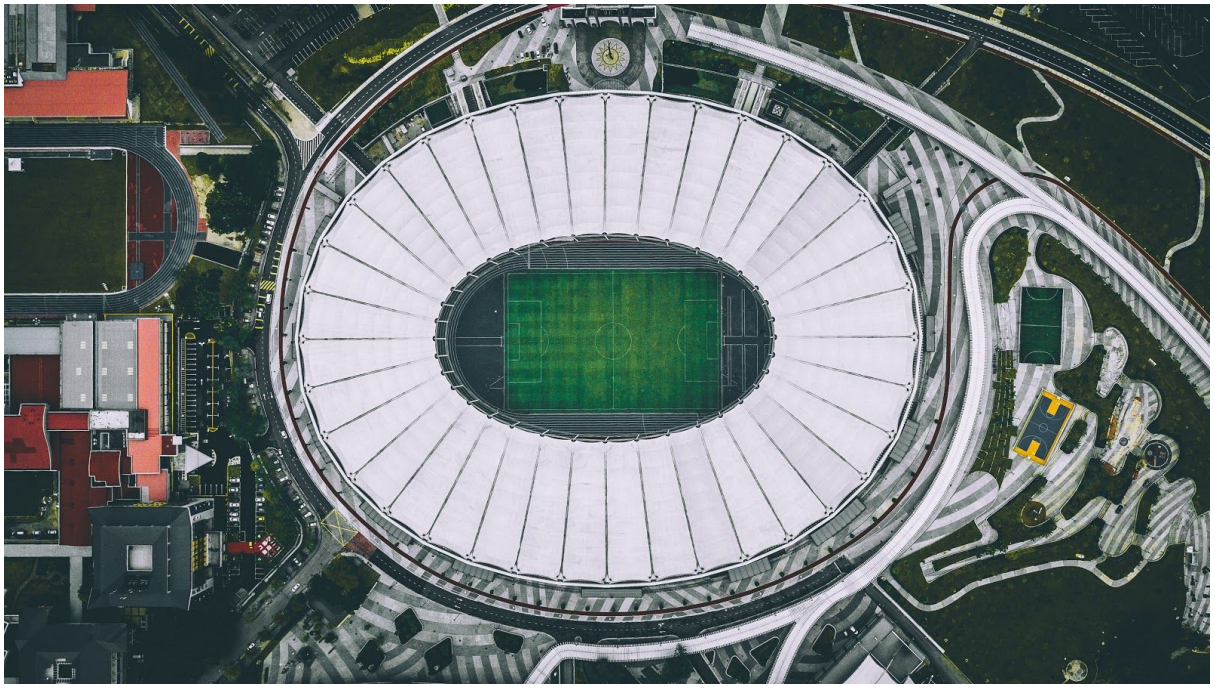
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF BUSINESS
ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

Social Network Analysis

International Football Matches (1920-2020)



Konstantinos Babetas, 8160078, kbabetas@gmail.com

Professor: Dimitrios Pournarakis

Department of Management Science and Technology 2019-2020

Table of Contents

1. Idea

Which is the most successful national team in football? Is it Brazil as it is the only national team that has won the FIFA World Cup five times? Is that enough to declare them as the best national team that has ever existed?

How about which national team is the most important one? This is a completely different question because importance does not equal success. Would the world of football be any different if a specific national team did not exist?

Is a national team important because it has won many matches but has failed to win any titles? These teams are usually labeled as “Losers” and we can all think about some National Teams that have played in finals, but always lose, like the Netherlands the last couple of decades, or Spain before they started dominating in international football. What is more, it is important to remember that in football, like every other sport, there is a winner and a loser. And sometimes, the importance of the win heavily depends on the success and the strength of the team that lost.

All of these questions will be attempted to be answered through this network analysis which can also be found, along with all the notebooks and photos, at GitHub.

2. Dataset

2.1. Source

The Dataset that was selected for the Network Analysis is the International football results from 1872 to 2019 which was found on Kaggle.

2.2. Overview

The Dataset includes 41.540 results from the national football teams starting from the very first official match in 1872 up to 2019. The matches are strictly men’s full internationals and they do not include Olympic Games or matches where at least one of the teams was the nation’s B Team, Under-23 or a league selected team.

The Dataset includes nine columns:

- Date: The date of the match
- Home_team: The name of the home team

- Away_team: The name of the away team
- Home_score: The score of the home team, not including penalty shootouts
- Away_score: The score of the away team, not including penalty shootouts
- Tournament: The name of the tournament in which the match was played
- City: The name of the city or town in which the match took place
- Country: The name of the country in which the match was played
- Neutral: A boolean value indicating whether the match took place at a neutral venue

3. Data Preprocessing

The Dataset has already been cleaned so no invalid or null values exist and, therefore, no cleaning was necessary.

The first change that was implemented is that the information regarding the home and the away team does not offer any insight to us in our analysis. Therefore, we changed the dataset to contain a column with the team names that won and lost, along with their respective scores. After we did that we went on to create the Nodes and Edges Table.

3.1. Nodes Table

From the original Dataset all the unique countries names were extracted and were sorted alphabetically whilst being given a unique Identification Number (ID) to help with the network creation later.

For every country the total matches were calculated as well as the total goals scored for and against that specific country.
The same procedure was followed to calculate the total wins and total losses of the country and the percentage of wins and losses.

At the end of the data preprocessing the Nodes Tables looks like this:

	id	label	matches	goals_for	goals_against	wins	losses	per_win	per_loss
0	0	Albania	188	146	365	46	142	0.244681	0.755319
1	1	Algeria	207	352	229	126	81	0.608696	0.391304
2	2	Andorra	111	26	339	3	108	0.027027	0.972973
3	3	Angola	150	217	176	83	67	0.553333	0.446667
4	4	Antigua and Barbuda	111	167	226	41	70	0.369369	0.630631
...	...	...	...	...	...	...	...	...	...
162	162	Wales	320	434	517	140	180	0.437500	0.562500
163	163	Yemen	125	127	321	31	94	0.248000	0.752000
164	164	Yugoslavia	178	408	235	118	60	0.662921	0.337079
165	165	Zambia	305	530	301	196	109	0.642623	0.357377
166	166	Zimbabwe	195	302	212	119	76	0.610256	0.389744

167 rows × 9 columns

3.2. Edges Table

For the edges table the Winning team was put as the Target and the Losing team as the Source. Initially the Source and Target were the other way around but the decision was made to switch them as due to the high number of edges in the graph they were not providing any coherent information. Whereas now the node-country with the most arrows pointing towards it, has the most wins.

What is more, every edge has a weight that corresponds to the total wins of the Target Country in that connection. Therefore, the Edges Table looks like this:

	Source	Target	Weight
0	0	6	1
1	0	8	6
2	0	10	1
3	0	13	2
4	0	14	1
...	...	...	...
4673	166	146	2
4674	166	148	1
4675	166	150	1
4676	166	152	2
4677	166	165	6

4678 rows × 3 columns

4. Assumptions

Some assumptions were made both before and after the first round of data preprocessing in order to make better sense of the results in the graph visualization and in its analysis.

First of all, the matches that were kept for the analysis took place after 1920. The reason behind this decision is twofold. First, the first matches within the first years were mainly between the UK Countries and as a result they had a significantly higher amount of matches than the rest of the countries as football started mainly there before they other countries caught up. The second reason was that by limiting the matches from 1920 to 2020 we could focus on what happened in international football in a whole century, in the past 100 years, which is more eye-pleasing.

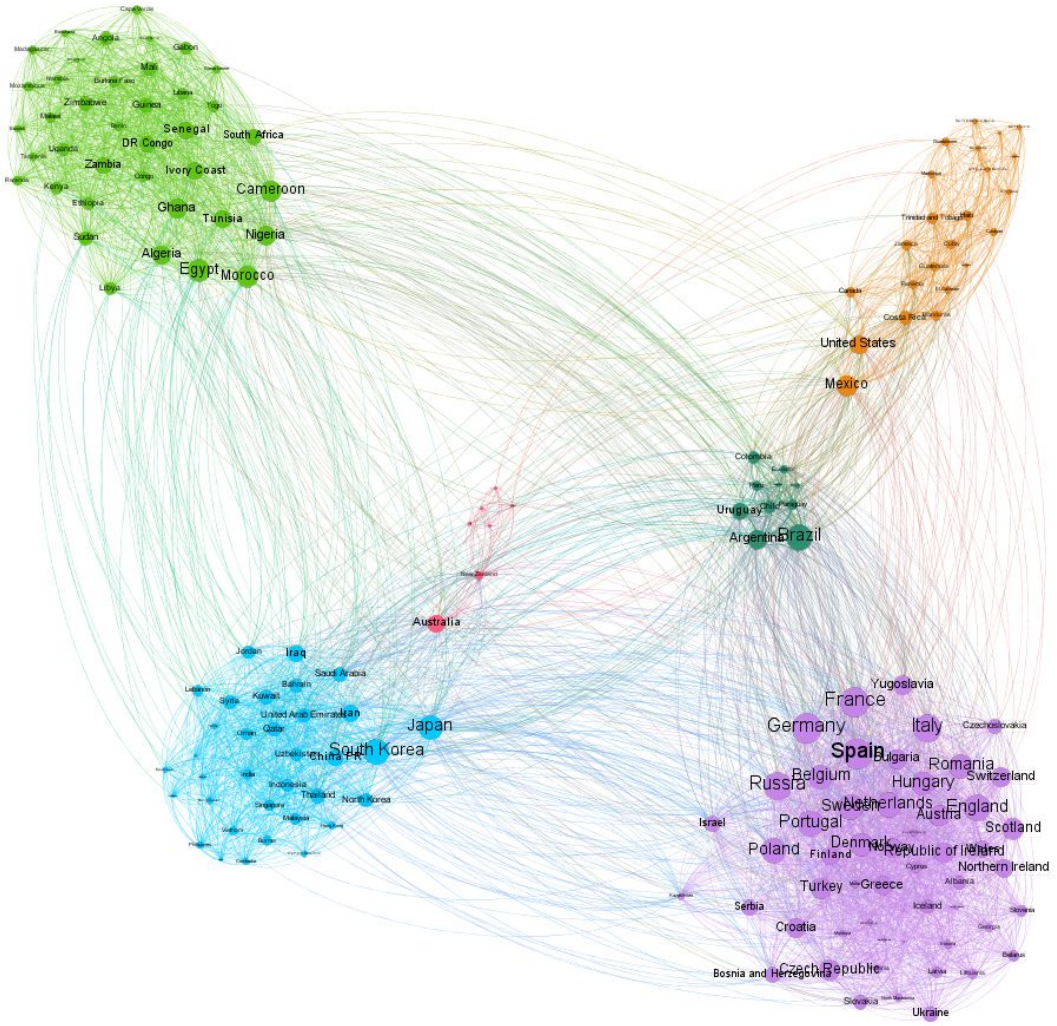
Then the decision was made to include only the results from official matches in tournaments and not friendly matches as the purpose of this analysis was to see how countries perform in the official matches. Then, mainly for the sake of simplicity, ties were eliminated from the entirety of this study as they would not offer any significant insight for our analysis and would only cause more complexion in the graph representation, which is already quite complex. Last but not least, again for the sake of simplicity the countries which had less than 100 matches in the past 100 years were disregarded from the analysis as they did not help us in the analysis of the network.

Having made all these assumptions we can move on to our graphical representation of the network.

5. Graph Representation

For the Graph Representation no further filter was used and the nodes were partitioned based on their modularity class while the size of each node depends on the ranking of the in-degree, which means that the bigger the node the more incoming edges it has, hence the more wins that country has.

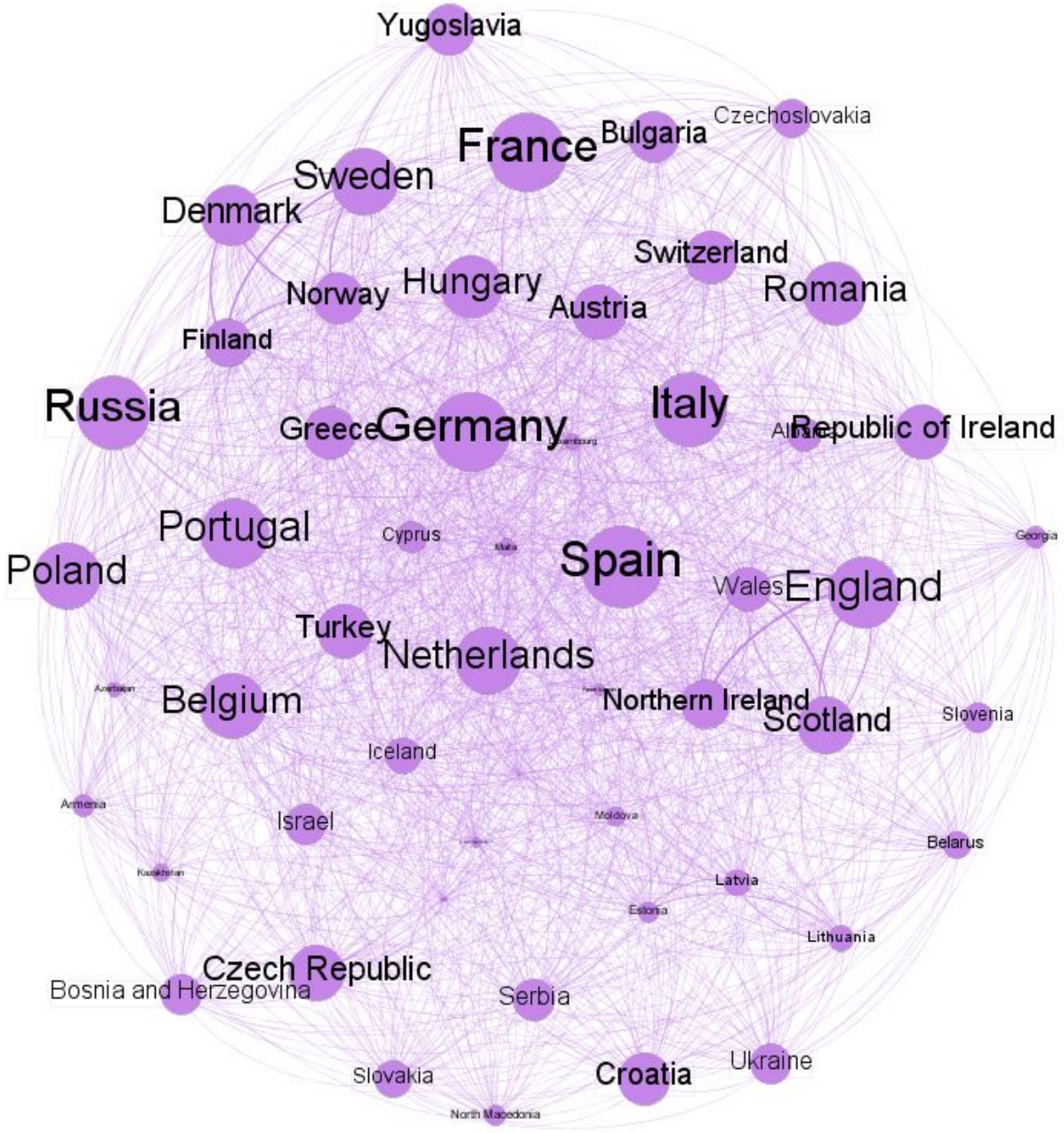
5.1. Full Graph



As we can see there have been a lot of matches played in the last 100 years and as a result, there is some chaos in the graph. Still, we can see that some conventionally successful countries like Spain, France and Germany seem to be the bigger nodes, which actually makes perfect sense,

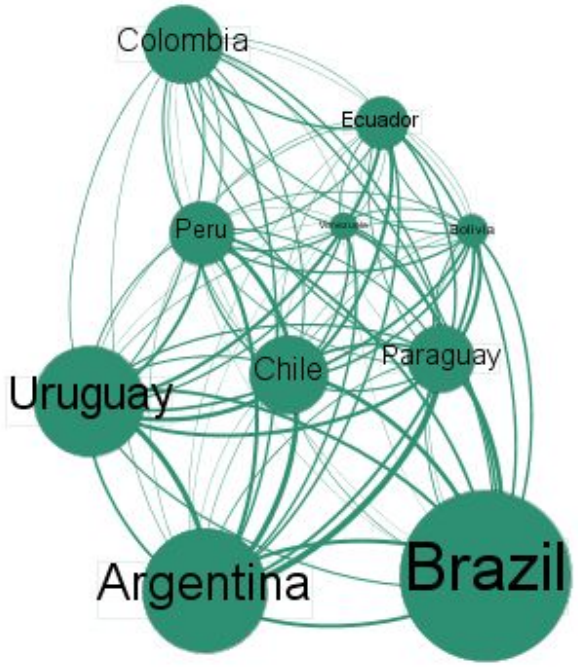
However, it would be beneficial for our analysis to filter based on the modularity class so that we can observe what is happening in each class and if what we are seeing makes sense.

5.2. Europe



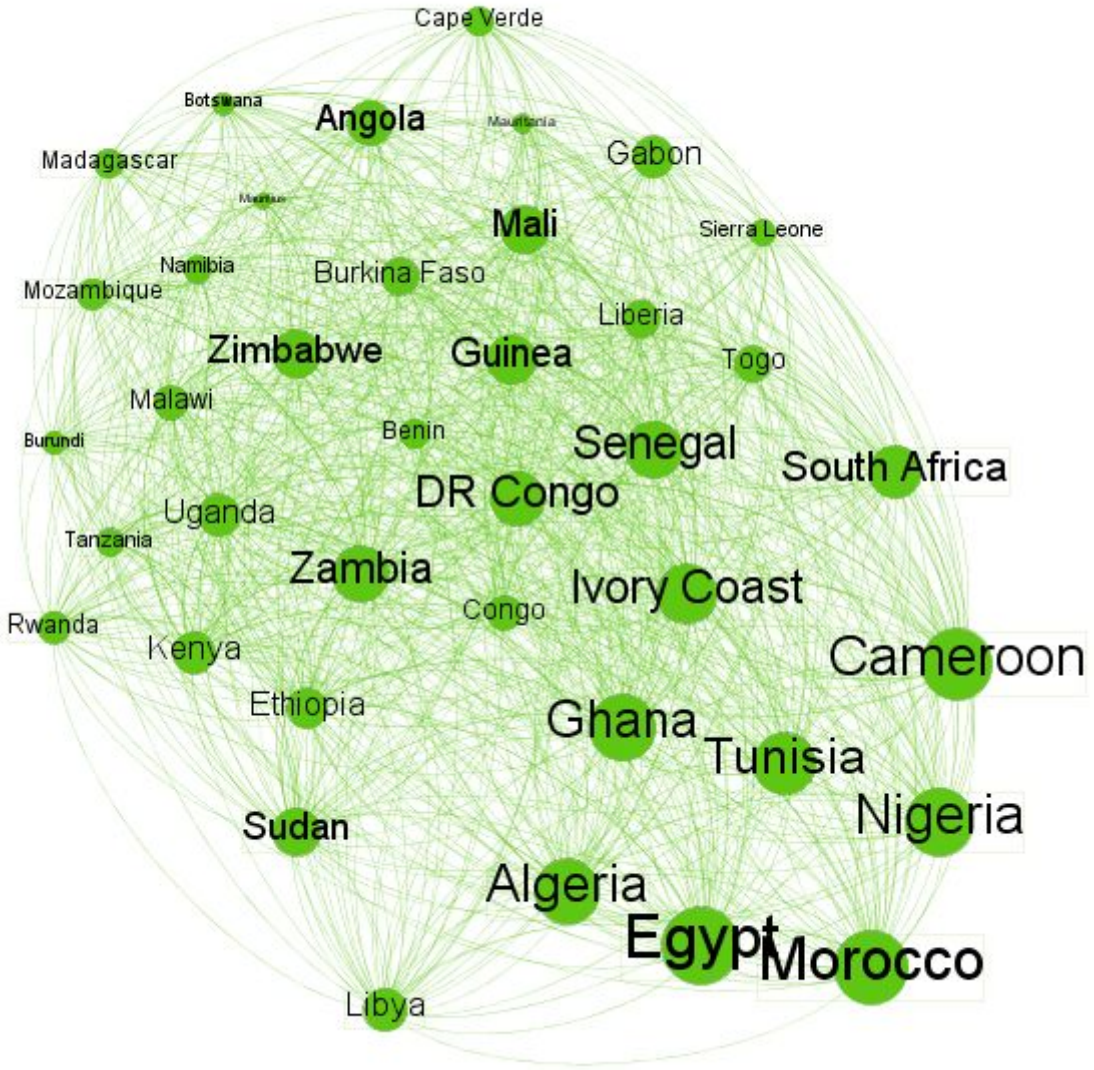
As we can see Spain, Germany and France have the highest in-degree, 69, 67 & 66 respectively, which makes sense as they are the most successful european national teams. What is quite interesting to observe in this graph is the fact that Yugoslavia and Czechoslovakia, two countries that have ceased to exist are located in the northern boundary of the graph whereas the countries in which these two were split into are located in the southern boundary, something that logically makes a lot of sense and it is correctly depicted in our graph.

5.3. South America



Again in South America, we see that Brazil is by far the most successful one with an in-degree of 59 followed by Argentina and Uruguay. This representation again makes a lot of sense as these three are undoubtedly the most successful teams with the most wins and titles.

5.4. Africa

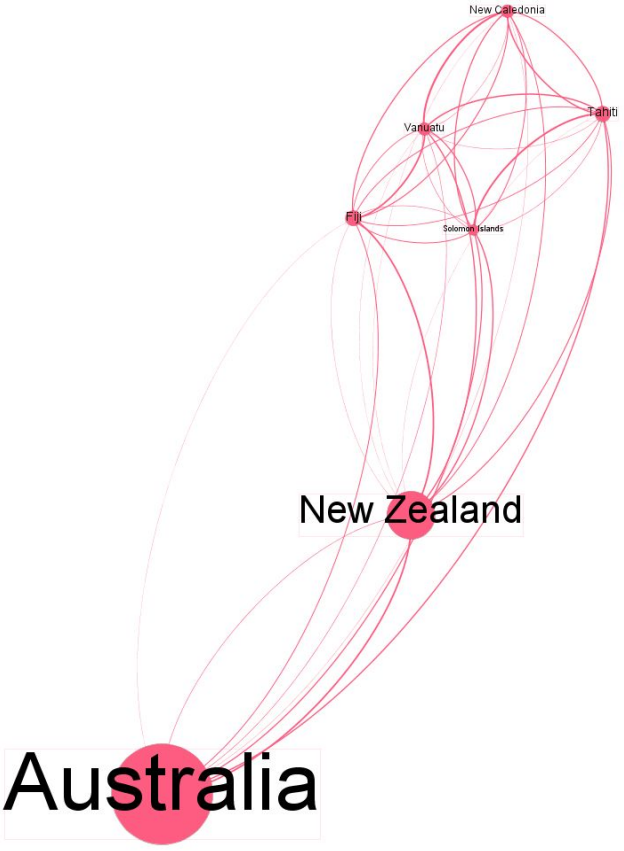


In Africa, Egypt takes the lead with an in-degree of 50 followed by Morocco, Cameroon and Nigeria and Ghana, all of which are the teams that have won the Africa Cup of Nations the most times.

5.5. North America

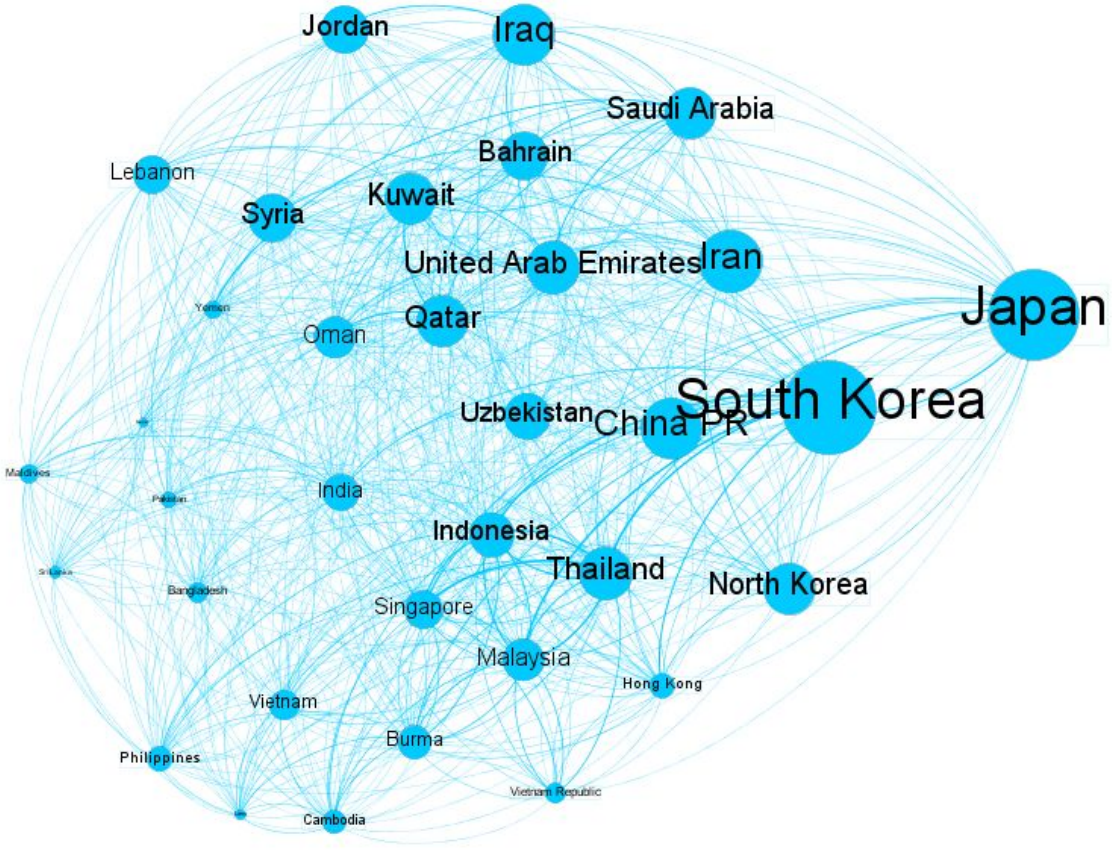
In North America, Mexico has the highest in-degree of 47 and then the United States with 41. Mexico has been traditionally considered to be a very good team and has also managed to play in every Round of 16 in the World Cup since the World Cup of 1994.

5.6. Oceania



As expected in Oceania, Australia is by far the most successful country with New Zealand following and the other countries struggling to win any matches.

5.7. Asia



Last but not least, in Asia we can see that South Korea and Japan are the most successful countries.

6. Basic Topological Properties

Before we go more in depth, no pun intended, let's first calculate some basic topological properties of our network so as to better understand it. First of all, we have 167 nodes-countries and 4.678 Edges.

The network diameter is 5 which means that the longest shortest path connecting two countries is 5.

The average of the shortest paths for all pairs of nodes, the average path length, is 2.23 which means that on average two countries have a distance of 2.23 between them.

7. Component Measures

According to the Connected Components Report in Gephi, since the network is directed, we have 1 Weakly Connected Component and 2 Strongly

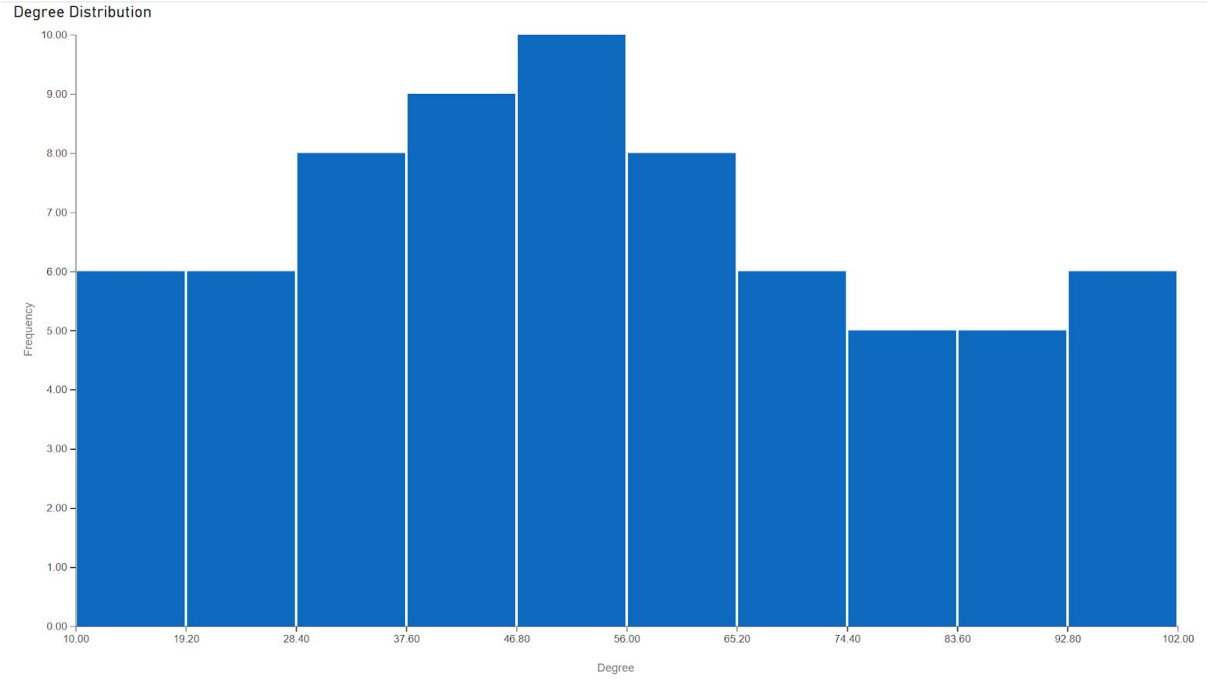
Connected Components. San Marino is the reason behind this as San Marino has never won a single official match so it has an in-degree of 0.

Other than that we do not have a single giant component as is the case in many other networks. Maybe if we had left all the countries and had not limited our analysis to the countries that have less than 100 games in the past 100 years we maybe would have had some countries that have only played with each other, although the chances for that are still pretty slim. Due to that, we do not have a giant component and other smaller ones, but our component size distribution just shows that all of our 167 countries are in one component.

8. Degree Measures

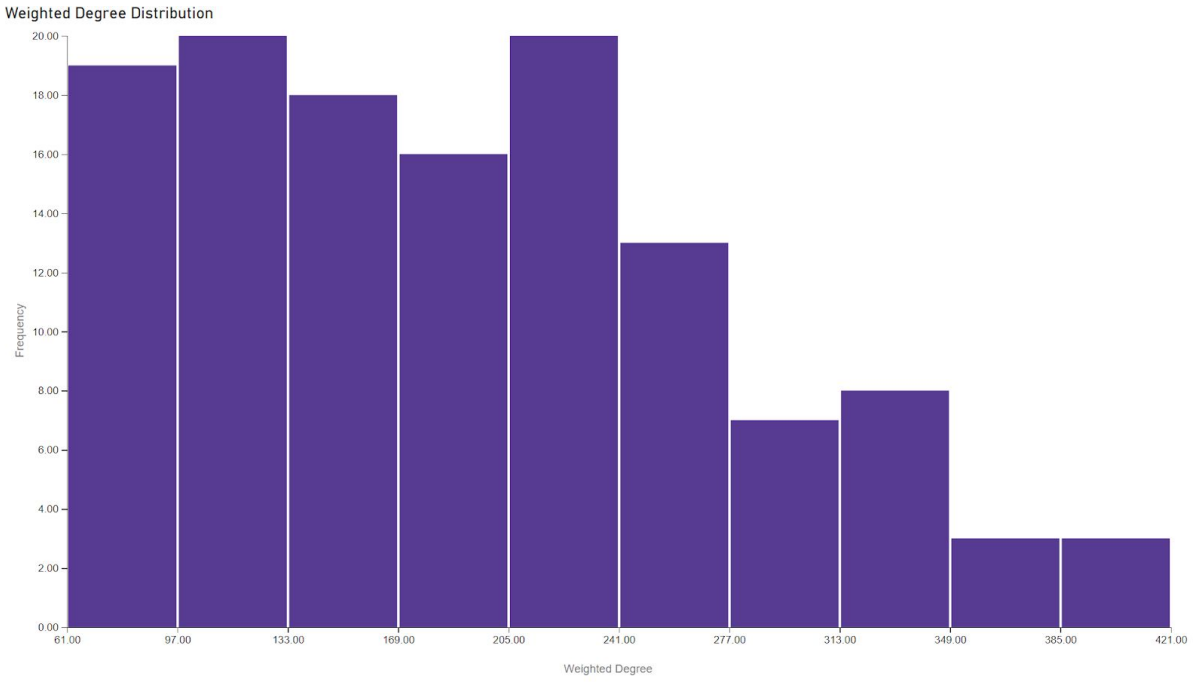
The average degree of the network is 28.012, the country with the highest degree is South Korea with 102. The highest indegree, the most wins against different opponents, belongs to Spain (69).

The Degree Distribution shows us that most national teams have a degree between 28 and 65 as you can see from the histogram below:



The Average Weighted Degree is 92.443, which is significantly higher than the Average Degree as it takes into account the weights of every edge. The national team with the highest weighted degree is Argentina and Brazil with 421. The highest weighted indegree belongs to Brazil (324).

The Weighted Degree Distribution shows us that most national teams have a weighted degree between 61 and 241 as you can see from the histogram below:

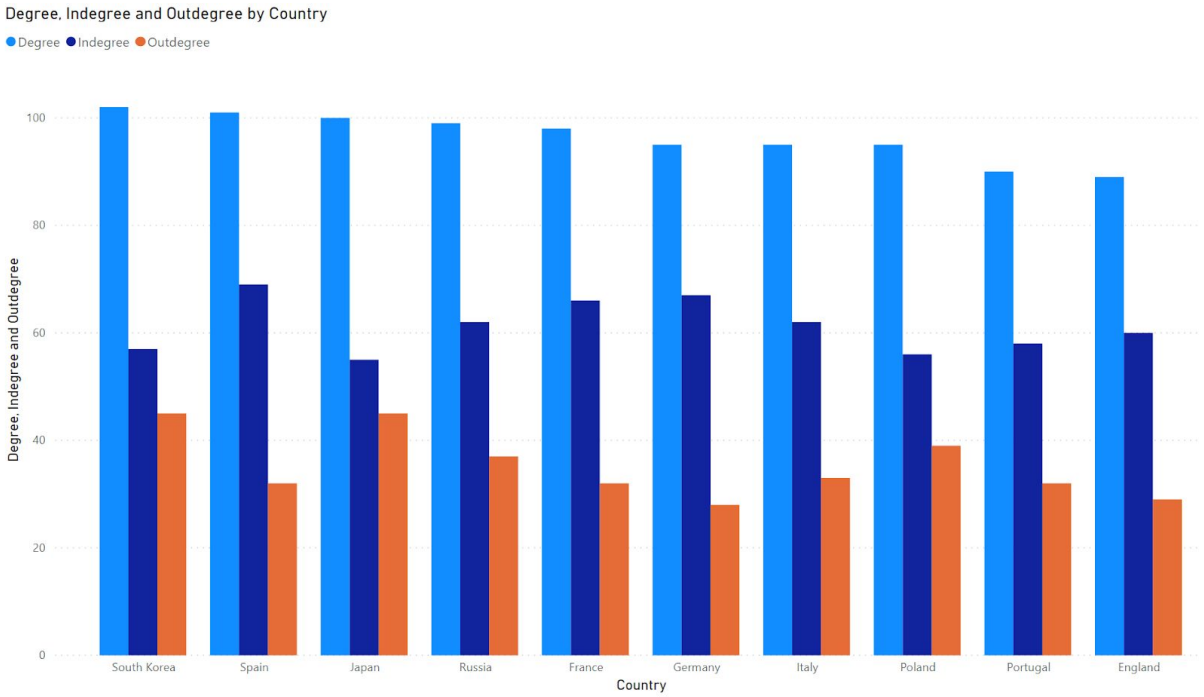


9. Centrality Measures

9.1. Degree Centrality

The Degree Centrality shows how connected a node is. In our case the Degree Centrality shows the distinct matches that were won or lost with different opponents.

Here we can see the top 10 countries based on their degree centrality:

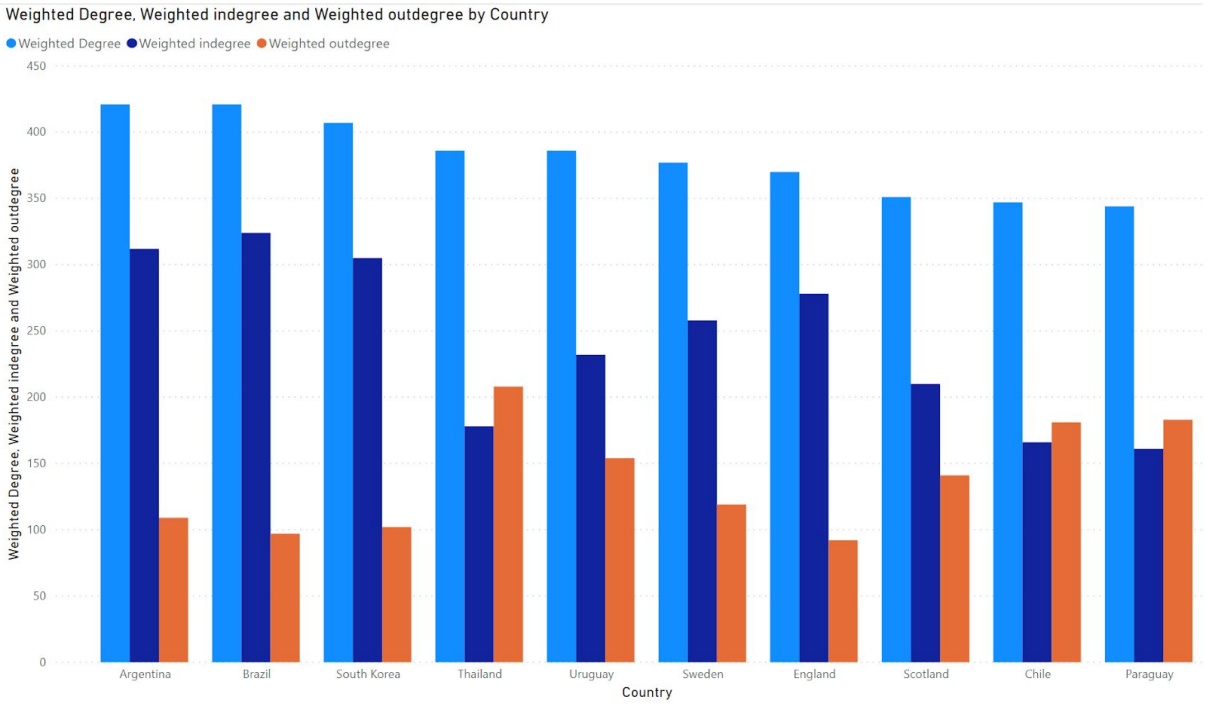


South Korea and Spain take the lead in the Degree Centrality, although we can see that Spain has a higher indegree from South Korea meaning that Spain has more wins than Korea.

9.2. Weighted Degree Centrality

However, the degree centrality does not necessarily help us understand which team has played the most matches because a team that has both won and lost to the same opponent will have two edges, whereas a team that has only won that opponent, even if they have won a thousand times, will have only one edge. As a result, we can calculate the Weighted Degree Centrality which actually shows the weighted number of edges so that is more close to real life than the simple degree centrality.

Here we can see the top 10 countries based on their weighted degree centrality:



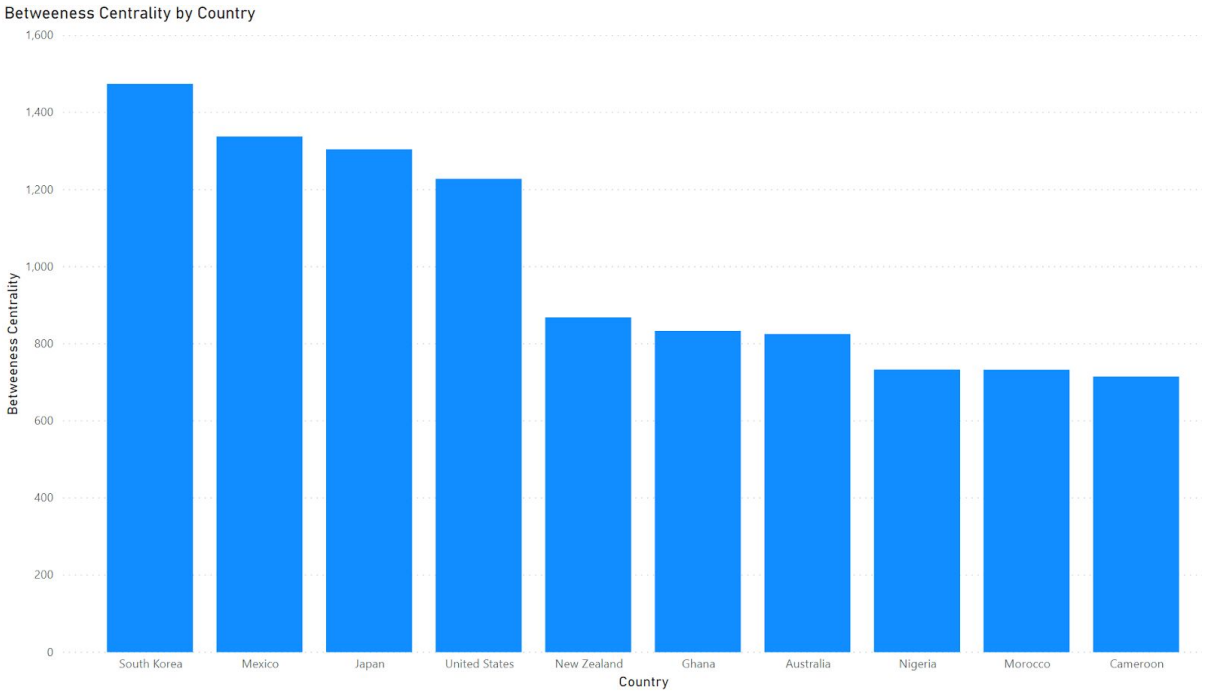
As we can see Argentina and Brazil take the 1st place as they both have a weighted degree of 421 games, although Brazil has won more. The big surprises are South Korea and Thailand in the 3rd and 4th place respectively. So by relying on the Weighted Degree Centrality we would have that South Korea and Thailand are considered to be among the best teams of the century, a statement that even the least interested in football person could tell you that it is not true.

Let's see if there is another centrality measure that can provide us with a better insight.

9.3. Betweenness Centrality

Betweenness Centrality shows how important a node is in terms of connecting other nodes. In other words which countries lie on shortest paths between two other countries.

Here we can see the top 10 countries based on their betweenness centrality:



Again the results are surprising. We see that South Korea, Mexico, Japan and the United States take the first places. That is due to the fact that betweenness centrality takes into account the nodes-countries that other countries need to go through in order to reach other countries. What does that mean?

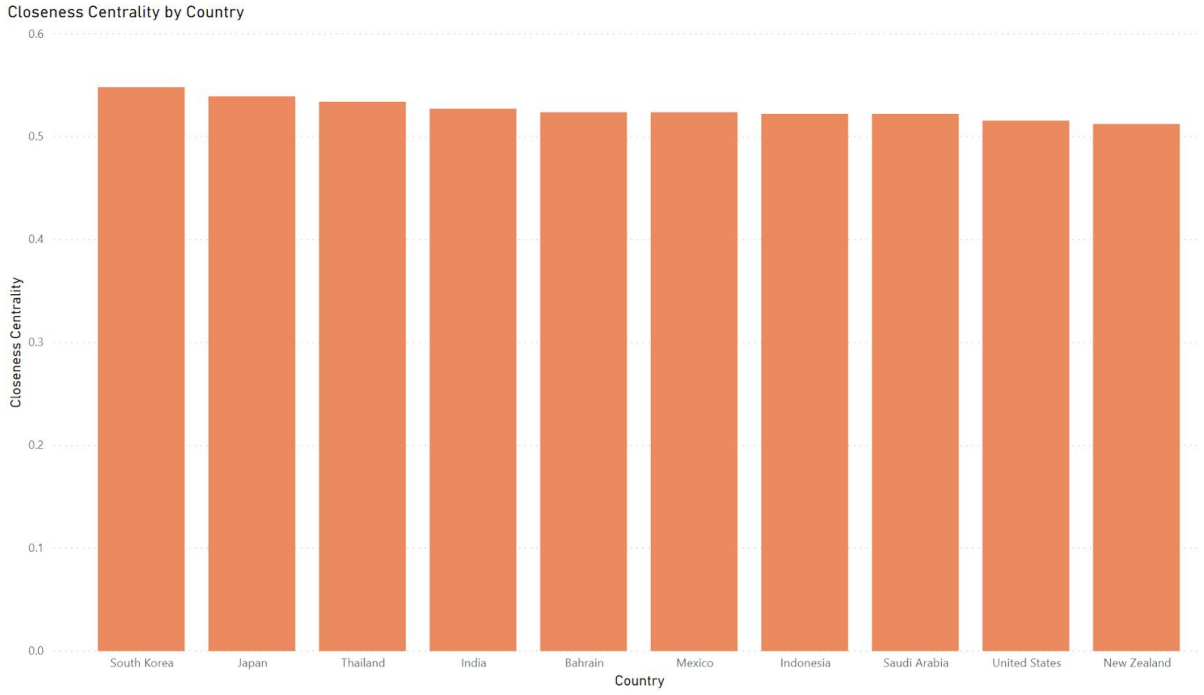
If we look closer we will see that all countries from the top 10 are pretty significant countries and belong to all the continents apart from the two most successful in football (Europe and South America). These countries have most successfully represented their continents in other international tournaments, such as the World Cup and hence they have connections with the European and South American countries.

On the other hand, the other countries in continents like Asia and Africa have almost never played with the “major players”. As a result, the only countries that connect them to those countries are the top ten in betweenness centrality.

Therefore, if we were to take out these countries many smaller countries in Asia, Africa and North America would not have a connection with many countries in Europe and South America. So the top countries in the betweenness centrality are very important for football as they are the countries which connect the other countries.

9.4. Closeness Centrality

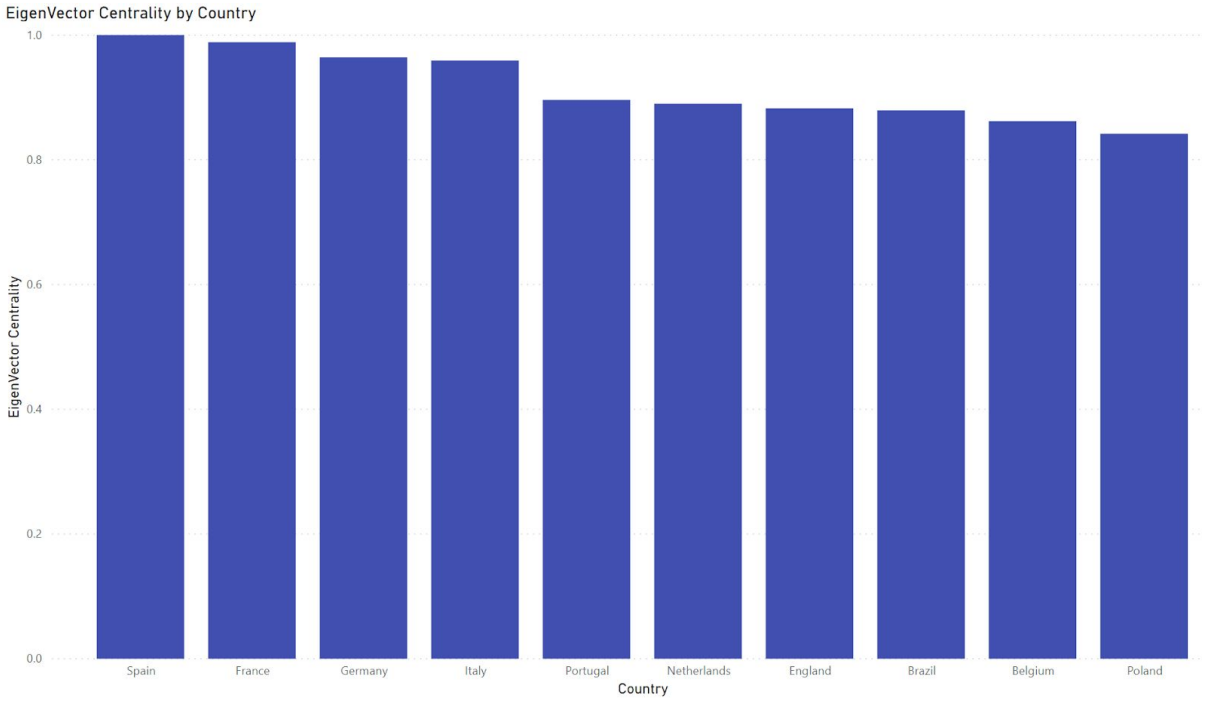
The Closeness Centrality shows how easily a node can reach other nodes or, in other words, how close a node is to the center of the network. Closeness is based on the length of the average shortest path between a country and all other countries in the network.
Here we can see the top 10 countries based on their Closeness centrality:



Again we see that countries not from the main two continents are dominating in the Closeness Centrality measure.

9.5. Eigenvector Centrality

Eigenvector centrality shows how much a node is connected to other important nodes in the network. It is a weighted degree centrality with a feedback boost for playing with other important countries.
Here we can see the top 10 countries based on their Eigenvector centrality:

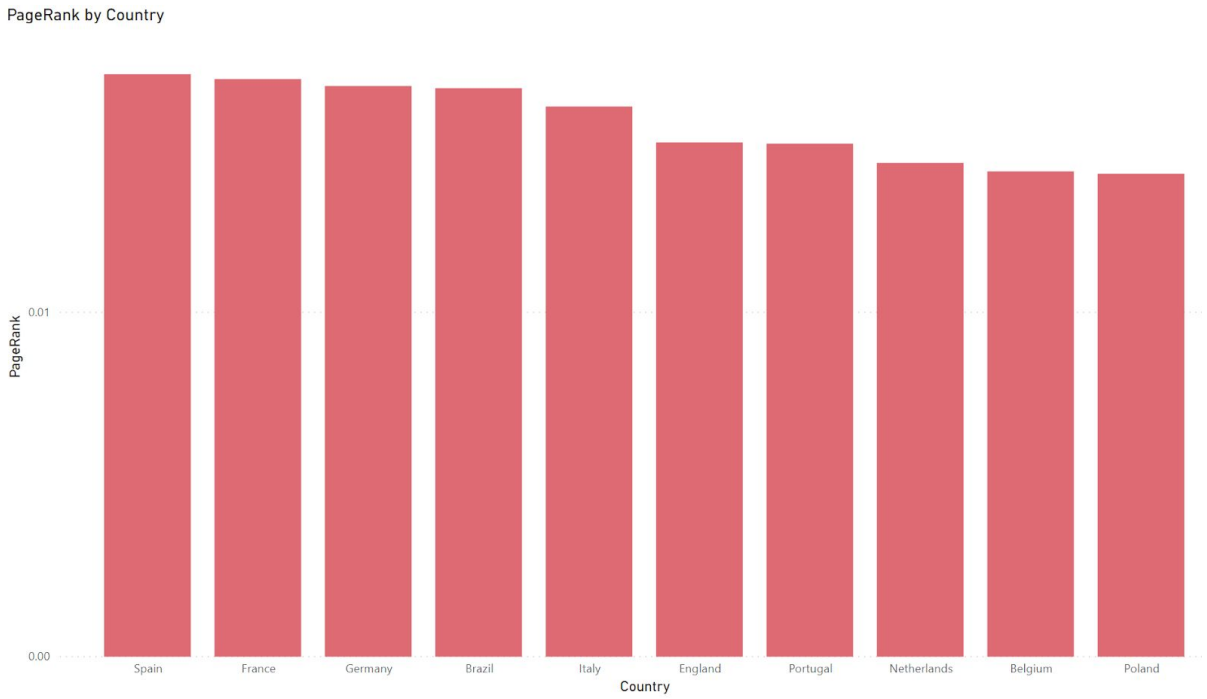


We can now see that some of the more traditional important countries are the leading countries in regards to the Eigenvector centrality as they have played, and thus they have connections with other important countries. So Spain, France, Germany and Italy are dominating in this measure and are the more traditional successful and important national teams.

9.6. PageRank Centrality

PageRank is a variant of Eigenvector Centrality and its main difference is that it accounts for link direction. Each country in the network is assigned a score based on its number of incoming links (its indegree). These links are also weighted depending on the relative score of its originating node.

Here we can see the top 10 countries based on their PageRank centrality:



We can now say that most probably the PageRank has done the best job in outlining the best national teams in regards to what a football fan would say. Spain, France, Germany and Brazil are close together at the top and they are, undoubtedly, the most top-performing national teams of the century.

10. Clustering Effects

The average clustering coefficient in our network is 0.519 which means that the average probability that two past opponents of a country X have played against each other is 51.9%.

Transitivity is the ratio of all triangles over all possible triangles. So transitivity, like density, expresses how interconnected a graph is in terms of a ratio of actual over possible triangle connections. In Gephi we were unable to easily calculate transitivity and, therefore, we loaded our graph into pandas and with the assistance of the networkx library we calculated that the transitivity in our network equals 0.614.

```
with open('final_node_table.csv', 'r') as nodecsv:
    nodereader = csv.reader(nodecsv)
    nodes = [n for n in nodereader][1:]

node_ids = [n[0] for n in nodes]
node_names = [n[1] for n in nodes]
data_filename = 'final_edge_table.csv'
edges = pd.read_csv(data_filename)

G = nx.from_pandas_edgelist(edges, 'Source', 'Target', 'Weight')
triadic_closure = nx.transitivity(G)
print("Triadic closure:", triadic_closure)
```

```
Triadic closure: 0.6144751896266791
```

Can we also calculate how many triangles we have in our graph? Again with the Networkx library we found that we have 29.002 triangles. Since we are computing triangles for the entire graph each triangle is counted three times, once at each node. As a result, we need to divide by 3 to get the actual number without having duplicates.

```
triangles = nx.triangles(G)
print(triangles)
```

```
{0: 831, 6: 526, 8: 1053, 10: 407, 13: 617, 14: 1024, 17: 657, 20: 922, 38: 553, 40: 1026, 4
5: 1154, 50: 1135, 51: 1199, 53: 531, 54: 1258, 56: 1039, 64: 1038, 65: 808, 70: 690, 71: 123
0, 85: 571, 86: 903, 102: 1060, 109: 941, 110: 907, 117: 1102, 118: 1104, 120: 952, 121: 120
0, 122: 1241, 128: 1061, 130: 756, 134: 583, 138: 1305, 142: 1044, 143: 904, 151: 914, 153: 5
65, 162: 838, 164: 624, 1: 598, 3: 420, 15: 304, 19: 807, 21: 508, 25: 730, 43: 742, 47: 491,
52: 392, 55: 673, 60: 479, 67: 483, 72: 557, 74: 877, 75: 337, 77: 552, 83: 362, 88: 410, 91:
454, 98: 746, 106: 684, 129: 523, 131: 271, 136: 600, 140: 480, 146: 371, 148: 390, 150: 668,
155: 523, 165: 580, 166: 513, 2: 398, 33: 771, 36: 917, 37: 978, 46: 682, 48: 661, 76: 480, 8
0: 676, 97: 637, 108: 545, 133: 756, 18: 309, 27: 256, 39: 571, 94: 331, 95: 304, 99: 353, 15
2: 480, 4: 93, 12: 126, 34: 175, 35: 145, 41: 86, 57: 101, 58: 156, 59: 135, 61: 229, 73: 21
6, 93: 157, 124: 82, 125: 114, 141: 153, 149: 205, 5: 655, 16: 164, 28: 333, 29: 523, 30: 28
5, 42: 194, 96: 720, 114: 268, 115: 237, 156: 482, 159: 60, 92: 927, 7: 498, 26: 210, 49: 15,
68: 629, 69: 473, 78: 400, 104: 285, 107: 412, 111: 378, 127: 543, 137: 977, 154: 428, 9: 41
6, 84: 436, 63: 385, 81: 337, 89: 424, 116: 297, 119: 408, 144: 404, 147: 554, 157: 365, 163:
325, 11: 363, 22: 297, 66: 510, 79: 215, 90: 221, 101: 246, 112: 246, 139: 280, 160: 297, 32:
306, 44: 151, 105: 91, 100: 432, 123: 351, 62: 229, 113: 198, 23: 332, 87: 304, 31: 396, 82:
450, 24: 309, 132: 468, 161: 104, 103: 18, 135: 15, 145: 24, 158: 15, 126: 615}
```

```
sum(triangles.values())/3
```

```
29002.0
```

11. Bridges

A bridge is a connection between two nodes that if it were to be deleted it would cause the component to get divided into two different components. According to Networkx we do not have any bridges, but we do have two local bridges.

```
nx.has_bridges(G)
```

```
False
```

```
list(nx.local_bridges(G))
```

```
[(165, 42, 3), (80, 111, 3)]
```

The first local bridge connects Ecuador with Zambia, with Ecuador being the winner and Zambia being the losing team. If that local bridge was to be removed then the shortest path length between Zambia and Ecuador would rise to 3 (span).

The second local bridge connects Oman to Latvia and, again, the removal of the local bridge would increase the shortest path length between these two countries to 3 (span).

12. Community Structure

Since we are analysing all the national football teams we have already created some communities or clusters in our mind. The different continents were the clusters that we had in mind before we started the analysis, but wanted to see if these communities would show up in the network analysis through Gephi. And it actually showed. We have a modularity of 0.647 and we have 6 communities that correspond to the 6 continents. The communities have a higher density relative to other nodes within their module, but low density with those outside.

So it makes sense that the modularity classes that were produced coincide with the geographical continents as countries within the same continent are more likely to have played against each other through the course of 100 years. Maybe if we had limited our analysis in the past 10 or 15 years our results would have been different.

There was an initial doubt that the network would manage to put every country in the correct class as some countries, such as Australia which belongs to the Oceania Modularity Class has participated in the AFC Asia Cup 4 times, and even won one time. Therefore, the graph shows that Australia is significantly closer to the Modularity Class of Asia, but still managed to correctly place it in the Oceania Class.

13. Graph Density

The Graph Density of our network is 0.169. The graph density is simply the ratio of actual edges in the network to all possible edges in the network. Therefore, our network is not a very dense one.

14. Homophily

Assortativity measures the similarity of connections in the graph with respect to the node degree. According to the NetworkX library the assortativity coefficient in our graph equals 0.32.

```
r=nx.degree_assortativity_coefficient(G)
print("%3.4f"%r)
```

```
0.3214
```

A saying exists that says that “birds of a feather flock together” and we can see that something quite similar happens in our network. We have already seen that the national teams from the same continent are positioned closer together, as they interact (play matches) with each other more often due to their geographical connection.

We previously calculated that the average clustering coefficient of our network is 0.519. Now we decided to focus only in Europe and see how homophily affects several of our measures. We calculated that the average clustering coefficient in the Europe-only graph is 0.606, which means that the average probability that two countries have played against each other if they have both played with another country X is 60%.

We also calculated the triadic closure in Europe and found that it is 0.82 which is significantly higher than the triadic closure in the whole graph. What is more, the density of our graph increased a lot as it is 0.577 and now our graph is pretty dense.

As a result, we see that homophily plays a very important role in our network. And that makes a lot of sense. Had we also included the friendly matches it probably would have played an even more important role as the neighboring countries are more likely to play against each other in friendly matches as they usually share similar languages, time zones and transportation is easier.

But let's conduct a homophily test. We will continue focusing on Europe and let's suppose that we in the network we have there is a *p* fraction of all nodes being in the modularity class of Europe and a *q* fraction of all nodes being outside of Europe. If we independently assign each node the modularity class of Europe with probability *p* and the non-Europe class with probability *q*, then both ends of the edge will be between European national teams with probability p^2, and both will be between non-European national teams with probability q^2. On the other hand, if the first end of the edge is a European National Team and the second is a non-European, or vice versa, then we have a cross-gender edge, so this happens with probability $2pq$.

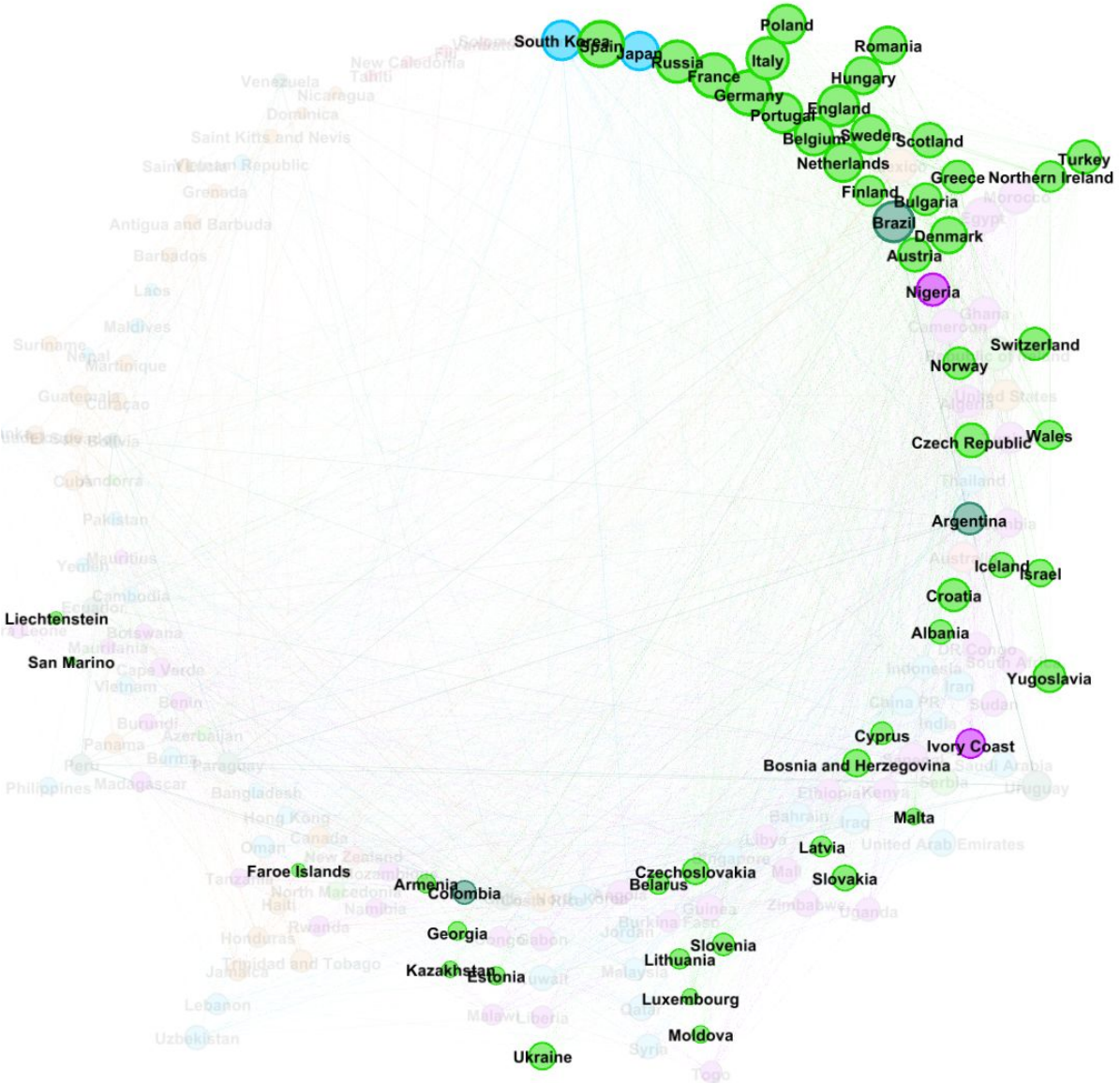
So if the fraction of cross-gender edges is significantly less than $2pq$, then there is evidence for homophily.

So we have 54 nodes in Europe, so $p = \frac{54}{167}$ and we have 113 non-European nodes so $q = \frac{113}{167}$. The total edges are 4.678 and we have 1.652 edges strictly between European national teams and 492 edges between a European and a non-European national team. So we have 492 cross-gender edges in our graph and if we had no homophily we would expect $2pq = 2 \cdot \frac{54}{167} \cdot \frac{113}{167} = 0,43 > \frac{492}{1652} = 0,29$. As a result, we have some more evidence of homophily.

15. Greece

We can not finish our analysis without taking a closer look in Greece and its position in the international football this past century.

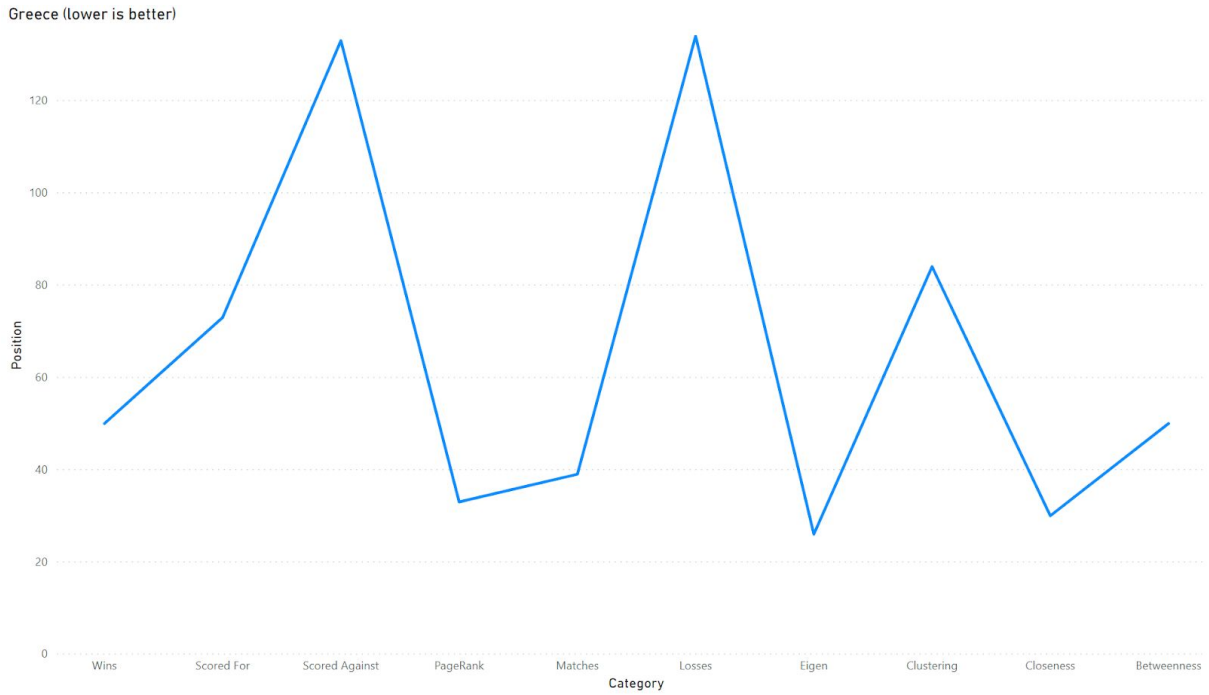
Let's first take a look at what the graph would look like with only Greece and its connections:



As we can clearly see Greece has mainly played with other European National Teams, as also shown in our Homophily analysis. We can also see that Greece has played against some non-European National Teams such as Nigeria (World Cup 1994, World Cup 2010) and Ivory Coast (World Cup 2014) from Africa, Japan (World Cup 2014) and South Korea (World Cup 2010) from Asia and Colombia (World Cup 2014), Brazil (Confederations Cup 2005) and Argentina (World Cup 2010) from South America. Greece has not won or lost in an official match against National Teams from North America and Oceania.

Let's see how Greece does in regards to the other countries. Greece has played 243 matches in total (39th), has scored 342 goals (73rd) and has conceded 378 goals (133rd). It has 125 wins (50th) and 118 losses (134th). It has a Closeness Centrality of 0.48 (30th), a Betweenness Centrality of 224.55 (50th), a PageRank of 0.01 (33rd) and an Eigen Centrality of 0.63 (26th).

Finally, its clustering coefficient is 0.5 (84th). The line chart with all the positions can be seen here (lower is better):



16. Conclusion

Through our Analysis we confirmed that the traditionally successful National Teams such as Spain, Germany, France, Italy and Brazil have had a huge impact in what we know as International Football. In addition to that, we also came to the realization that some countries like South Korea, Mexico and Japan are what make the whole network stay connected and without them many National Teams would have no connections to other continents.

Moreover, we understood how much of an important role homophily plays, as nodes are more likely to connect with their friends-neighbors and connect with their friends-neighbors. Which, as stated in the analysis, is something to be expected but it became even more apparent through the network analysis.

Last but not least, we saw that the geographical borders tend to matter less and less nowadays. Many tournaments, in order to boost their popularity and their competitiveness have started including countries that, traditionally, would not have participated, like Australia in the Asian Cup. That is something that did not happen 100 years ago, when countries and National Teams were a lot more confined to their continental borders. As a result, as this occurrence happens more and more it is very likely that if we were to conduct the same analysis in 20 years we would see completely different results both in the

importance of some National Teams and to their community structure and homophily.

17. Tools Used

- Pandas & Jupyter Notebook: Data Preprocessing
- Gephi: Network Representation & Analysis
- NetworkX: Python Library used for calculation of some measures in the graph.
- PowerBI: Visualization of Centrality measures & Degree Distributions
- GitHub: Network Analysis Repository - https://github.com/kbabetas/International-Football-Network-Analysis

18. Sources

- Barabási A.L. (2012). Network Science. Cambridge, United Kingdom: Cambridge University Press.
- Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets.* New York, NY: Cambridge University Press.
- https://unsplash.com/photos/vKSIwGivUuA
- https://programminghistorian.org/en/lessons/exploring-and-analyzing-network-data-with-python
- https://www.futurelearn.com/courses/social-media/0/steps/16055
- http://www2.unb.ca/~ddu/6634/Lecture_notes/Lecture_6_homophily_handout.pdf
- https://faculty.nps.edu/rgera/MA4404/Winter2018/16-Homophily.pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4267571/
- https://www.nature.com/articles/s41598-018-29405-7