ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

A Network Analysis of the 2018 FIFA World Cup

A Network Tour of Data Science

Project report

Authors:

Maxence DRAGUET
Robert INJAC
Yannick KLOSE
Manana LORTKIPANIDZE

Team 09

Professors:
Pierre VANDERGHEYNST
Pascal FROSSARD

January, 2019

1 Introduction

This project aims to exploit the intensely hyped and connected month that a Football World Cup constitutes. The last one, that happened in 2018 though precising surely is unnecessary, saw a record **3.5 billion people** following its events, according to the association organising it, FIFA. Naturally, a peak in Web searches can be expected in connection with this massive phenomenon. We therefore hope that the networks we could form from the specific set of Wikipedia pages connected to this context could offer insights into who played, when and with whom. We will see that these questions can indeed find their answers from a network approach.

The entire code for this project is accessible online[1].

2 Data Analysis

2.1 Data Gathering

The data used in this project was gathered from Wikipedia both manually and by using the eponym API. Collecting the names of the page of the 736 players, the 32 national football teams and the 32 countries associated, a set of CSV file was generated to answer predefined specific questions. A signal corresponding to the number of views on a time window centred around the world cup, from the 10th June 2018 until the 20th July 2018, was collected on a per day basis. A notable problem occurring at this step was that some young and less famous players did not have a Wikipedia page by the time of the world cup (their values in the signal were set to 0). The signal is stored both in absolute and normalised values. Finally, hyperlinks, in order to build graphs based on these on each page, were sampled for the 800 nodes consisting of the three type of elements considered in this project.

2.2 The Hyperlink Network

This first model is built using the 800 nodes gathering our three categories and the hyperlinks connecting them. An important remark at this stage is to notice the network thus formed will inherently be directed with binary weight and, in this case, was seen to be connected. How does each category of nodes fit inside this network? Observing the out-degree (= leaving) and the in-degree (arriving) distributions on figure 1, we can clearly distinguish the three types of node.

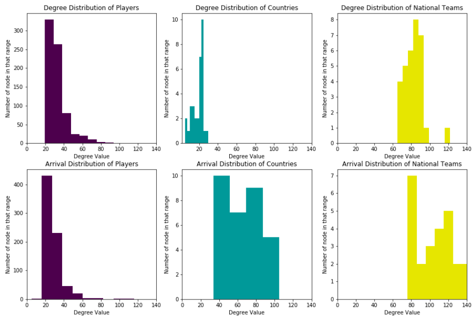

Figure 1: Above: out-degree distribution; below: in-degree (arrival) distribution. Purple: layers; blue: countries; yellow: national football teams.

Players tend to connect less to the other nodes and are even less pointed to. *Countries* are highly referred too but point less to other nodes. It was observed that they in fact mostly point towards each other, as shown in the appendix, figure 11. This behaviour is expected as players are unlikely to be mentioned in the Wikipedia page of a country while the contrary is very plausible. Finally, *national teams* form the centres, the hubs of our network, being the most predominantly connected type of nodes.

2.3 Player popularity

The hyperlink network's structure is suitable for identifying the most important

[1] https://github.com/ProjectNTDS/Network_World_Cup_Analysis

and/or popular players of the World Cup. The easiest way to measure this is checking which players have a high degree (in + out degree). In table 1, the top 5 players by degree are shown. To a football fan, these will be known names: they are currently among the most famous players in the world.

Rank	Player	Degree
1	Lionel Messi	157
2	Eden Hazard	149
3	Cristiano Ronaldo	145
4	Luis Suárez	140
5	Neymar	127

Table 1: Most connected players

Another approach of determining importance/popularity of a player is to examine the number of shortest paths passing though a player's node. We shall exploit betweenness centrality: a measure of centrality in a graph based on shortest paths[2]. Large value of this metric relates to the importance players. In table 2, the top 5 players by betweenness value are shown. Again, they are some of the most famous players, with some of them being common from the previous table but, interestingly, not all of them. This is of course due to the different notions of "importance" centred around each metric.

Rank	Player	Degree
1	Keylor Navas	0.053
2	Luis Suárez	0.052
3	Lionel Messi	0.046
4	Mohamed Salah	0.046
5	Son Heung-min	0.034

Table 2: Players with highest betweenness values.

On figure 2 the betweenness values of the whole network can be seen. Most of the nodes have a very low betweenness centrality with just few of them above 0.03.

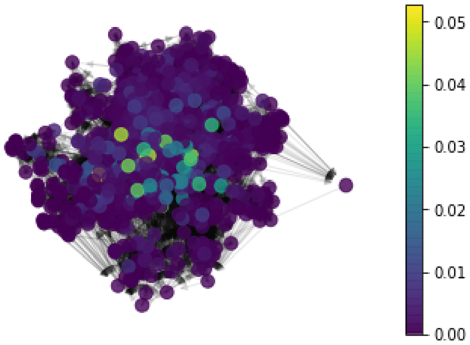

Figure 2: Betweenness values for the whole network

3 Identifying Teams

Analysing the hyperlinks network form by nodes related to the World Cup, as introduced in subsection 2.2, we set to identify the players of each team. This section explores three different approaches to perform this objective.

3.1 Heat Transfer

The first method to identify the players of each team is based on **heat filtering**. By applying a delta impulse on a certain node, the signal can then transmit to the neighbouring nodes, in good analogy with heat transfer. These neighbouring nodes should statically have a close relationship to the source node, such as sharing the same team if the impulse is put on a *player* or a *national team*.

Thus applying the "heat" signal on the *national team* nodes of the network returns a distributed signal over the nodes. As each football team is only allowed to have a total number of 23 players, performing a **selection** of the 23 most intense "heat" signals on *player* nodes seems like a reasonable approach to gather a team. Sequentially setting the source signal on each *national team* returns a clustering prediction of player nodes.

This method was observed to offer **satisfying results** when comparing the predicted team label to the true ones: the accuracy of the method is close to 85 %

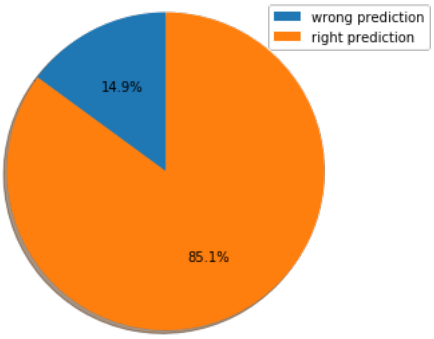

Figure 3: Heat signal team assignment predictions for all players.

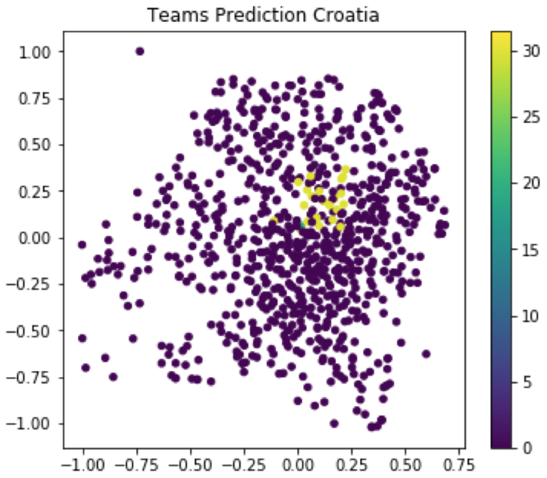

Figure 4: Prediction of the Croatian national team.

as shown in figure 3). As an example, figure 4 displays the quality of the label prediction for the Croatian national team. Nodes are coloured yellow if the label was correctly predicted, green if they were not assigned even though they should have and purple if they were correctly considered to be unrelated to the label under study.

3.2 Pearson Correlation

This second approach is centred around the use of the **Pearson correlation** between the feature vector (entries corresponding to the normalised signal of visits) of each player, thus restricting the nodes to the *player* type. The assumption was that high correlation would be due to similar behaviour exhibited by the underlying team structure. After performing the correlation for each node, a symmetric weighted adjacency matrix is obtained, having as entries the correlation coefficients and being thus naturally undirected. For each *player* node, each row of the matrix, the 21 most important correlation values are **selected** and **binarised**. In this way, each player should connect to most of its team, in the ideal scenario.

Note that a team is composed of 23 players so requiring 22 connections should in principle have each player linked to the entire team. The value proposed here, 21, is such that, in the case of a player not being correlated to its team members, due for example to the bug mentioned in the introduction (players that did not have a Wikipedia page at the time of the World Cup have a signal set to 0), no undesired links are established towards out-of-team players. The entire team is still expected to cluster, since some player of the team may be more closely correlated to a subset of its players and could thus jointly form a cluster. There is some freedom regarding the number of connections we impose but 21 was observed to be a sweet spot with the displayed network offering the most compelling behaviour. In the appendix, subsection 6.2, a thresholding method is proposed and the associated correlation network is displayed. The resulting correlation network is shown in figure 5.

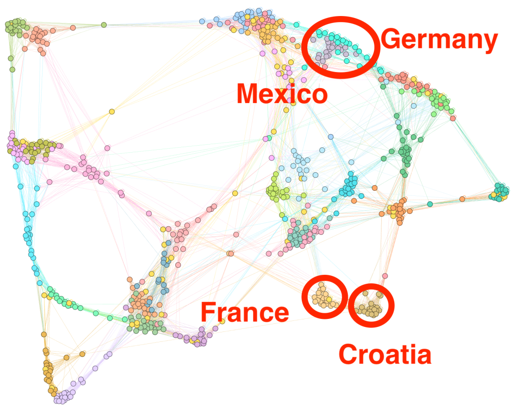

Figure 5: Correlation network using selection.

The apparent quality of the clustering depends on the team analysed. France and Croatia, for example, are very well clustered while Germany and Mexico tend to form a single entity. This disappointing behaviour can be connected to the fact teams played a number of matches depending on their successive victories. Mexico and Germany were respectively eliminated after 4 and 3 matches while Croatia and France both had 7 matches. In this context, correlation on player is naturally more efficient at separating teams playing a high number of matches at seemingly random dates. Unfortunately, World Cup matches are assigned a schedule based on group of teams, thus confusing the correlation approach even more. Therefore, this method does not distinguish between teams that had roughly the same number of matches around the same dates. This approach is expected to perform vastly better under a championship format where each team plays against all other foes.

3.3 Louvain Method

The Louvain method for community detection is an iterative algorithm to **extract communities**, in this context a synonym for cluster, from large networks[1]. The algorithm optimises a measure called modularity. A value between -1 and 1, it compares the density of links inside the assumed cluster compared to links between it and the other communities. It starts by considering each node as its own single cluster. A cluster is then only joined with a neighbouring one if this operation yields an improved modularity. Hopefully, the communities are going to match the *national teams*.

On figure 6 we can see the results of the Louvain method on our network. Unfortunately, the method converges to 18 different clusters, while we were expecting 32 different teams. However, a closer look to the proposed communities does indicate a certain structure. Performing some analysis on the communities, we first observe there are **no teams split** over different communities: all the players of a given team appear in the same cluster. While some clusters correspond to single teams (for example Sweden and Senegal), some others result in several countries joining. For example, Tunisia and Egypt are in the same cluster, Japan and South Korea in another one. This structure is not surprising: Tunisians and Egyptians are likely to have a high proportion of players in the same regional clubs. Once again, a hidden information lies behind the structure: most, if not every, players are part of a regional football club. The inter-player connections resulting from this aspect have a notable impact on the quality of the method.

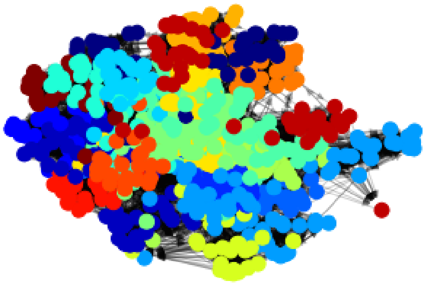

Figure 6: Clusters established by the Louvain method

4 Finding matches

A total of 64 matches occurred during the Word Cup. We set here to detect which countries played a match at a certain date using the *national teams* nodes and the number-of-views-per-day table. Logically, this signal should peak for a team during the days it played.

On figure 7 is displayed the graph of the normalised number of views per day for the Croatian national football team. Clearly, 7 peaks stand out at the following successive dates: 16/06, 21/06, 26/06, 01/07, 07/07, 11/07, and 15/07. These match **exactly** the days that Croatia played in the 2018 World Cup. Naturally, the largest peaks - 11/07 and 15/07 - correspond to the semi-final against England and final against France.

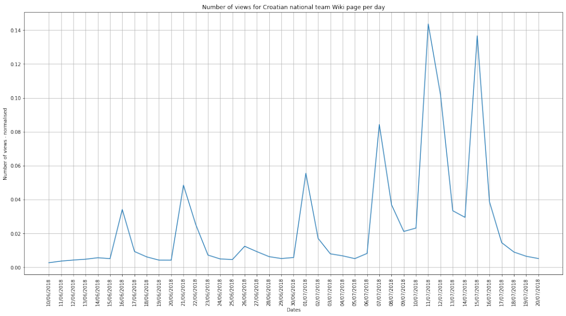

Figure 7: Normalised number of views of Croatian national team.

To find out which countries played a match at a certain date, an evolving graph for each day of the competition can be formed. It always has the same nodes, the 32 national teams, and, for each day, connections are made between nodes if both associated teams had a peak in normalised number of page views on that day. The notion of "peak" was computed to be a certain signal value above neighbouring values by a certain amount (a threshold). Therefore, by looking at this graph, connected nodes should be teams that played on that day.

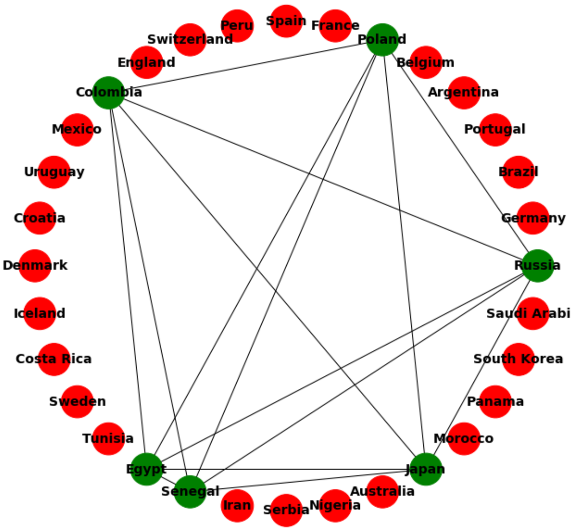

Figure 8: Dynamic graph: 19/06/2018.

An example of this for the 19th of June 19 is proposed on figure 8. Looking at the official FIFA page, 6 countries played during that day: Colombia, Japan, Poland, Senegal, Russia and Egypt. These exactly match the countries connected on the graph. This approach was shown to be efficient for any dates.

5 Conclusion

To conclude this study, we saw that different interesting networks can be formed to explain and visualise the dynamics of the large-scale event the Football World Cup is. Limiting the approach to the information available through Wikipedia, we built a hyperlink and a correlation network as well as a dynamic visualisation one . The properties of these graphs and the use of various methods offer great insights into the subject studied here, such as different ways to cluster players into teams, a link between fame of a player and the way it connects to others, the power of the number-of-view signal analysis and so on.

References

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[2] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

6 Appendix

6.1 The Hyperlink Network

A 40% sample of the hyperlink network is displayed in figure 9 with the *players*, *countries* and *national teams* respectively being assigned the values 0, 5 and 10.

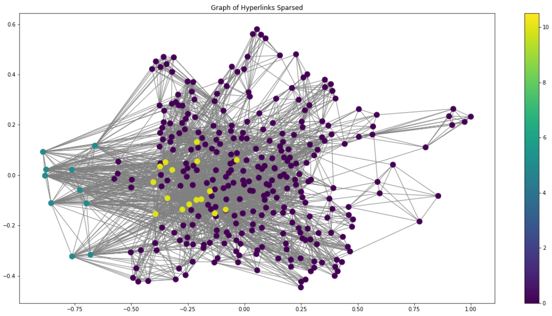

Figure 9: A 40% sample of the hyperlink network. Purple: layers; blue: countries; yellow: national football teams.

This does confirm the behaviour mentioned in subsection 2.2. Figure 10 displays the associated adjacency matrix. As the graph is directed, it is not symmetric. The 736 first entries are the *players*, followed by the 32 *countries* and finally the 32 *national teams.*

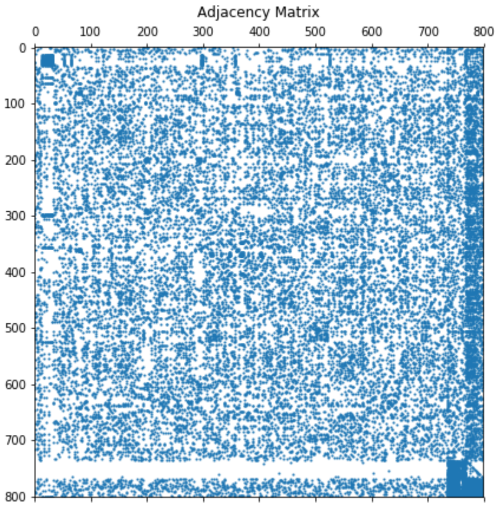

Figure 10: Hyperlinks network associated adjacency matrix.

Again, the behaviour mentioned above is also apparent in this visualisation, with a strong right band due to every types of nodes being linked to the *national team* ones. Note the blank space in the *countries* toward player region, with only some very rare connections corresponding to very famous players.

Inspecting the last nodes, as displayed in figure 11, one can notice that *countries* only connect with their respective *national team* and between themselves, making a characteristics square pattern.

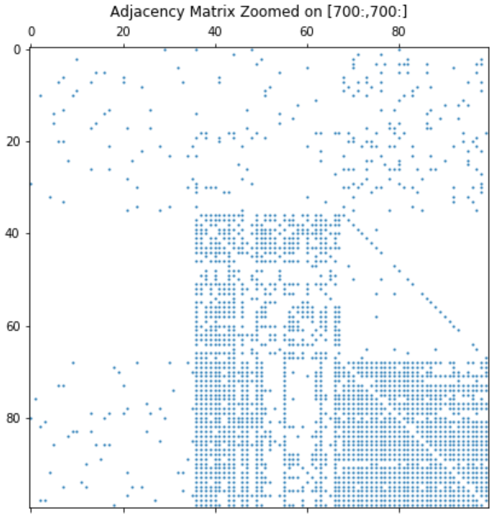

Figure 11: 100 × 100 *last nodes of the hyperlinks network adjacency matrix.*

6.2 Pearson Correlation

In this section, a different method to threshold the correlation-weighted adjacency matrix is proposed. Here the weighted adjacency matrix undergoes an overall thresholding procedure to keep the most correlated values in a binary fashion, thus removing about 97% of the edges.

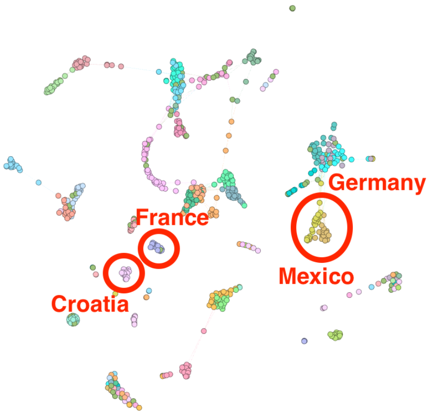

Figure 12: Correlation network with colours for team labels - global threshold

The resulting graph is displayed on figure 12. The result is similar to that of figure 5, though in this case both Croatia and France appear as isolated clusters.