École polytechnique fédérale de Lausanne
Center for Digital Education

Analysing the FIFA World Cup 2018

Network Tour of Data Science

January, 2019

Team 09

Supervisors:

XXX

XXX

1 Introduction

2 Data Analysis

3 Task xy

4 Identifying players in each team

When analysing a world cup only based on the links and number of views from Wikipedia one interesting questions we asked ourself is, if it is possible to identify the players of each team just by using this specific network structure.

In this section we will show two approaches that lead in different ways to the objective of the above-mentioned thesis.

4.1 Heat Signal

In our first approach to identify the players of each team we were using a heat filter. For this method we are using the network of hyperlinks where each player is connected to the page of his team. The aim of this method is to apply a delta impulse as a heat signal on a certain node, which then will be transmitted to the neighbouring nodes. These neighbouring nodes ideally have a close relationship to the initial note, in our case the same national team.

After applying the heat signal on the network we had a continues signal on our network. Because each team is only allowed to have a maximum number of 23 players, we set an upper limit for the heat, in order to filter the 23 most heat transmitted nodes. Those nodes were when our prediction for the different players in the same team.

It turned out that this method worked very good and when comparing the results with the actual true labels, we obtained an accuracy of mostly over 95 percent. In figure 1 we can see the predicted nodes for the Croatian national team in the network. The yellow nodes are nodes that were predicted right, where the green nodes represent nodes that were not assigned even though they belong to the team.

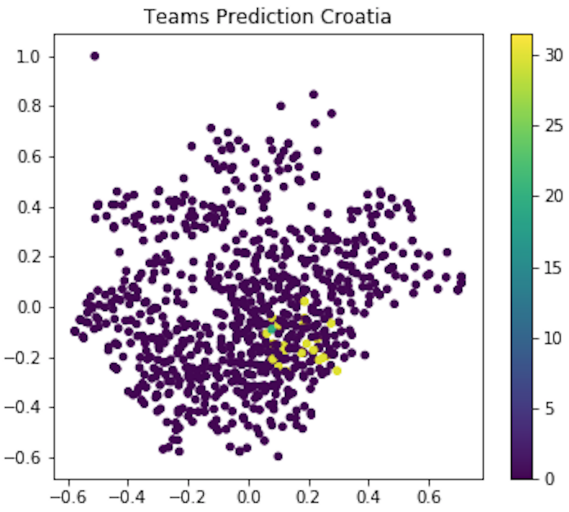

Figure 1: Prediction of Croatian national team in hole network

4.2 Pearson Correlation

In our second approach to solve this problem we were using the pearson correlation between the feature vector of each player. Our presumption was that high correlation is mostly due to similar behaviour exhibited by the team structure. In order to visualise the graph in gephi, we reduced the number of edges to only keep the most correlated ones (removed 97%). The result of the clustering based on the correlation can be seen in figure 2.

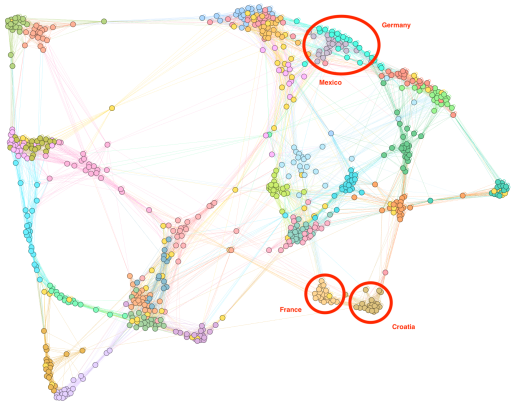

Figure 2: Cluster

At a first glance we can observe that the clustering worked for some cluster better than for others. France and Croatia for example were clustered very well while Germany and Mexico are almost one cluster. This is due to the fact some national teams had more matches than others and therefore more unique matches. Mexico and Germany were eliminated after 3/4 matches while Croatia and France had 7 matches. In addition all group matches of the same team were on the same date. Therefore we can hardly distinguish between two teams that had roughly the same amount of matches.

As a result both the heat signal and the pearson correlation are ways to identify players of each team. However, the first method achieves slightly better results because it is less linked to the data than the second.

5 Data

6 Conclusion