Analysing Football Passing Networks Using Network Science

Dinesh Adhithya
Department of EECS
IISER Bhopal
hdinesh18@iiserb.ac.in

Kiran D
Department of EECS
IISER Bhopal
kiran18@iiserb.ac.in

Rohit Taeja
Department of EECS
IISER Bhopal
rohit18@iiserb.ac.in

Mehul Paithane
Department of EECS
IISER Bhopal
mehul18@iiserb.ac.in

***Abstract*—The application of network science to real life problems has introduced new methodologies to analyse statistical problems. In previous decades, team sport decisions often were subjective, trusting in the intuition and personal experience of decision makers. Generally, football outcomes at the team level can be expressed in terms of results, goals scored, differences in the goals or probability of winning. Thus, statistical methods can be used to detect determinants and predict football outcomes. The main innovative contribution of this paper is to discuss whether, if and how passing network properties and performance indicators derived from network summary measures are crucial for the match outcome. Such measures may be able to improve statistical models where the football outcome is defined as the probability of winning the game. Here we have shown that network science allows addressing different aspects of the team organization and performance, that was earlier not captured by classical analyses and at a descriptive level, we have provided useful graphic visualizations to compare teams and their individual level of connections. This paper also aims to make 2 models, one where we consider the players as nodes and the other one considering the football field into grids and making them nodes.**

***Index Terms*—Passing networks, Football, Centrality, Machine Learning**

I. Introduction

The power of Network Science relies on the fact that, once a system is composed of a series of interacting units, it can be projected into a network, i.e., a set of N nodes connected through L links. Passing systems are in fact dynamical systems themselves. From the diversity of applications of Network Science, in this Opinion paper we are concerned about its potential to analyze one of the most extended group sports: Football. The passing networks are one example more of the bunch of new method, metrics and visualizations which have arisen in the last years trying to analyze what happens inside the pitch along a football game. These networks aim to provide useful information for trainers and football experts. They can also be used as performance indicators for football teams.
The passing networks are based on a basic approach to graph theory and analysis, where we consider the existence of:

1) **Individual entities (nodes).**
2) **Connection between them (edges).**

This can be translated to the game of football as, the players can be the nodes each pass to other player an edge. this passing is then further used to analyse a team's passes and their strategic play. [1]

Summary statistics of any football match consists of values like: ball possession and number of complete passes. These passing networks aim to provide useful graphical visualizations to compare teams. therefore we directly compute and discuss network properties, such as centralization, clustering and cliques. We can model the probability of the match winner. However, not all methods consider the strength of the relationship between players, the type of interactions among them in both at the micro-level (players) and macro-level (team) and the global team organization. In fact, passes and the related network are undoubtedly representative of game style, even compared to goals, shots and other summary statistics, representing more than 80% of the events in football.
Analyzing passing networks has some advantages: It provides an easy way to detect patterns or strong and weak ties among players and their positions in the lineup, can provide useful evidence of players' skills, teams tactics and connections between positions. Although information retrieved from this structure is not directly able to provide spatial information about the pitch, passing networks can be plotted clearly, helping to understand team cohesion from the micro- and macro-perspective. Finally, data are generally open access at the individual level for the main international and European tournaments [2] [3].

II. Materials and Methods

Data acquisition:
The Stats-bomb football match event dataset was used for this work. The stats-bomb dataset had been used in the form of a python package called statsbombpy. The data is provided as JSON files exported from the stats-bomb Data API, in the following structure:

1) Competition and seasons stored in competitions.json.
2) Matches for each competition and season, stored in matches. Each folder within is named for a competition ID, each file is named for a season ID within that competition.
3) Events and lineups for each match, stored in events and lineups respectively. Each file is named for a match ID.
4) StatsBomb 360 data for selected matches, stored in three-sixty. Each file is named for a match ID.

For dynamic python programming we shall use jupyter notebook.

III. Results

Our work was mainly focused on studying the dynamics of passing networks during football matches. Pass happens to the most frequent event in a football match and we had devised two approaches to study passing networks:

1) Passing networks considering each player as a node. In this approach we constructed multi-edge passing networks with the player as a node , with passes considered only between players of the same team and those who belong to the starting 11.
   a) The images below shows the passing network created among the players from France and Portugal from the UEFA EURO 2020 tournament:

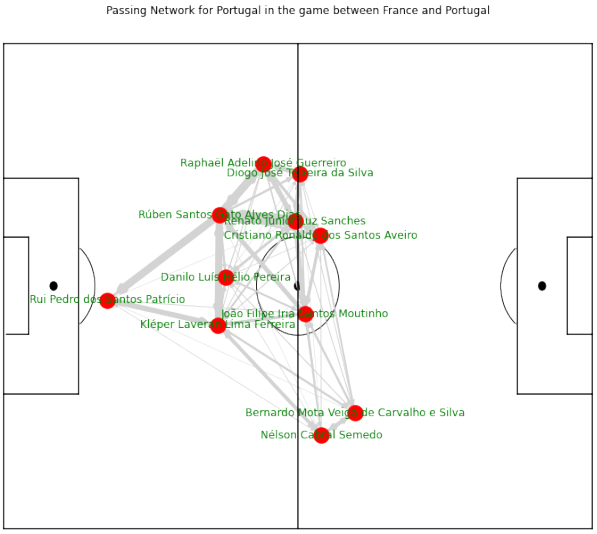


Fig. 1. Passing network for Portugal in France vs Portugal

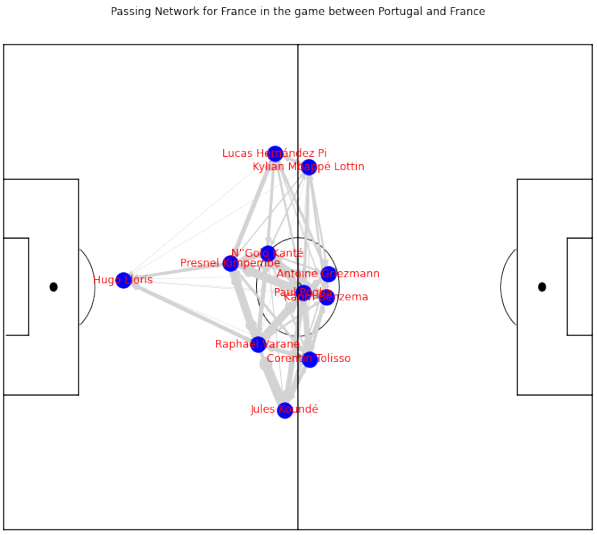


Fig. 2. Passing network for France in France vs Portugal

   b) The total degree distribution of the players over the entirety of EURO 2020 tournament:

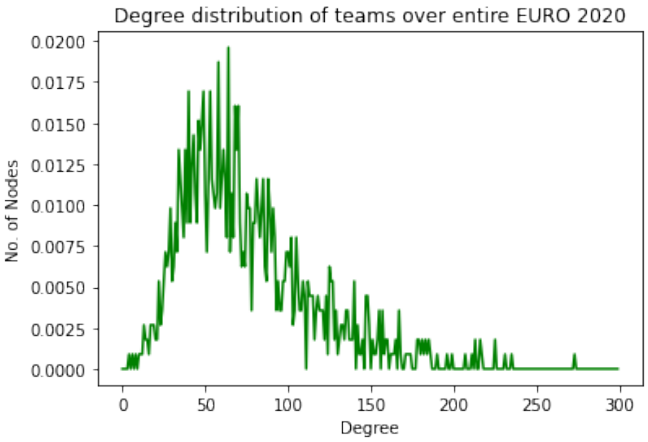


Fig. 3. Degree distribution of teams over entire Euro 2020.

   c) This distribution was compared with a configuration model G(n,m) , where n is no of nodes and m is no of edges. A football match has 11 players, therefore 11 nodes and an average football match has 828 passes per match. We compare the degree distribution of the configuration model with the actual degree distribution obtained from the dataset:

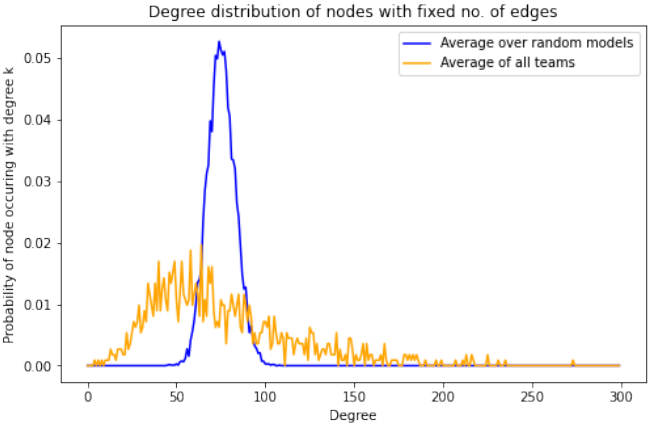


Fig. 4. Degree distribution of nodes with fixed edges

A machine learning based approach was used to predict the match result based on various centrality measures. The input features based to the models were average clustering, transitivity, average shortest path length, average degree centrality, average pagerank centrality, average betweenness centrality and average closeness centrality and match result (win, lose or draw) was taken as output and a SVC was fit for the data. The model had an accuracy of 60%. Similarly, when done for regression with output as goal difference, a poor R2 score of 0.2 was obtained. It was also found that average clustering,

transitivity and average degree centrality had 0.4 positive pearson's correlation with the output. Average shortest path length, average betweenness centrality and average closeness centrality had -0.4 negative pearson's correlation with the output. But, average pagerank centrality had no correlation with the output because the average pagerank centrality was nearly constant for all the networks.

	precision	recall	f1-score	support
0	0.56	0.69	0.62	408
1	1.00	0.01	0.01	190
2	0.64	0.77	0.70	498
accuracy			0.60	1096
macro avg	0.73	0.49	0.44	1096
weighted avg	0.67	0.60	0.55	1096

Fig. 5. Classification report for random forest classifier

This indicated that a larger clustering coefficient, transitivity and average degree Centrality ensured that a team had better chances of winning, a larger clustering coefficient meant that the ball could be retained in well by a team, larger degree centrality signifies a player being connected to a larger number of players, helping them move the ball to different locations easily and better retain the ball. The betweenness centrality shows a negative correlation as a player being part of multiple shortest paths between various nodes in a network, means that player could be targeted and could be man-marked to cut that player from the network, which would make the network less robust.

2) Passing networks, making grids on the football and each grid an individual node.
In this approach the football pitch was divided into 24 equal parts and each square was considered as a node. A multi-edge directed graph was constructed with each square as a node, and a pass from one box to another was considered as an edge between the two boxes.
   a) The images below show the 24-grid passing networks reated among the players from france and portugal from the UEFA EURO 2020 tournament

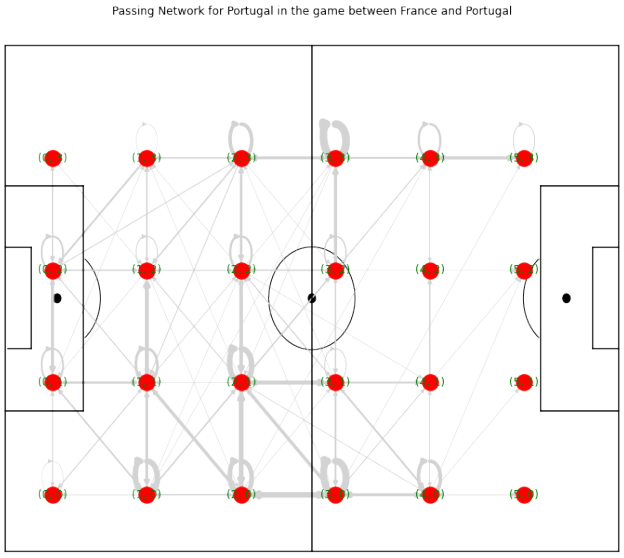


Fig. 6. Grid passing network for Portugal

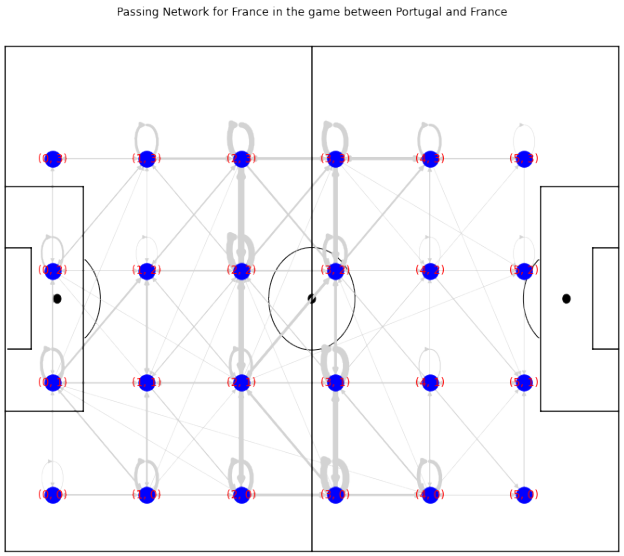


Fig. 7. Grid passing network for France

   b) The total degree distribution of these grids over the entirety of EURO 2020 tournament are given below (Figure 8):
   c) This distribution was compared with a configuration model G(n,m) , where n is no of nodes and m is no of edges. A football match has 24 grids, therefore 24 nodes and an average football match has 828 passes per match. We compare the degree distribution of the configuration model with the actual degree distribution obtained from the dataset (figure 9):

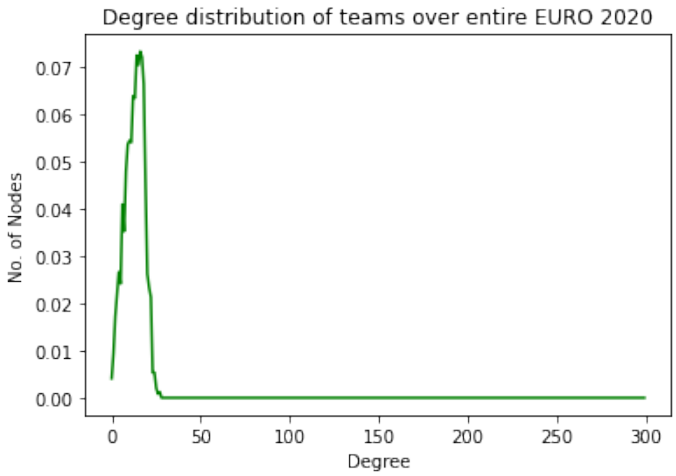


Fig. 8. Degree distribution of teams based on 24-grid networks

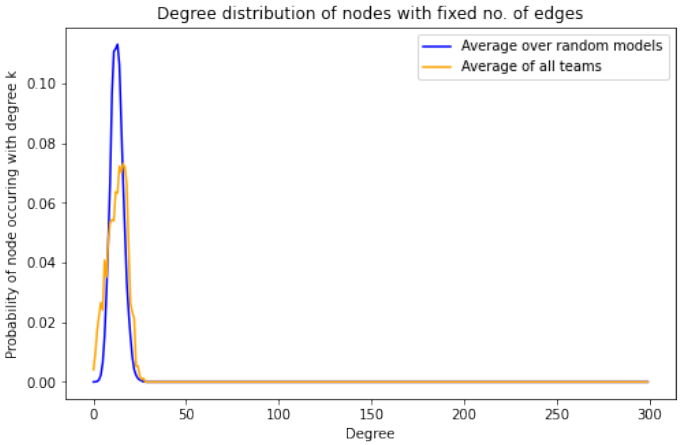
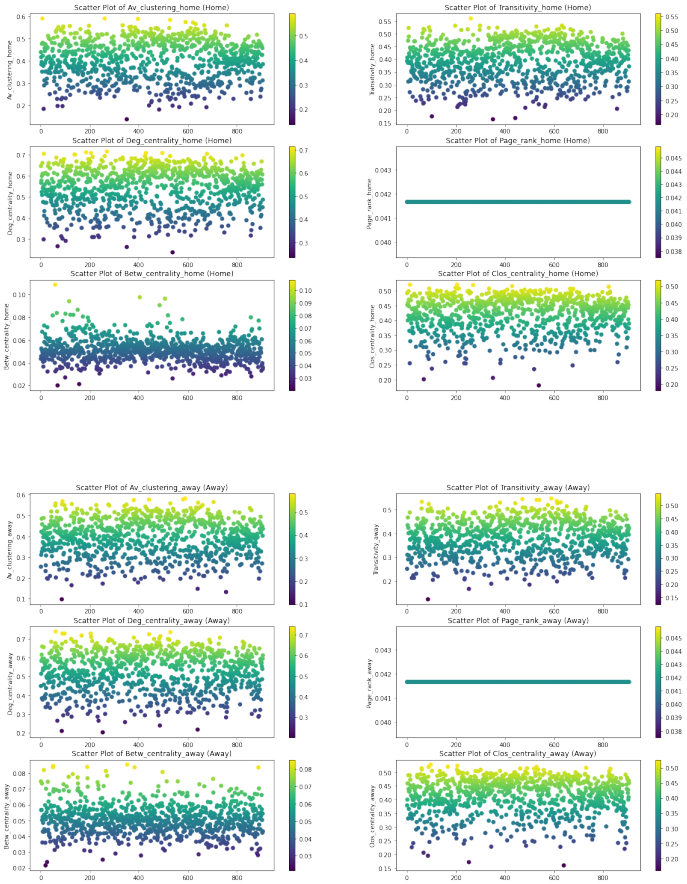


Fig. 9. Degree distribution of nodes with fixed edges

3) Machine learning using spaces as nodes: Using the constructed Network using the spaces in the field as Nodes, a Machine Learning model was constructed to predict the outcome of the matches using various centrality measures as features. The centrality measures used in this study were average clustering, transitivity, average degree centrality, average page rank, average between-ness centrality and average closeness centrality. The attributes were extracted from the data, and the following figures given below are the scatter plots of the said data, for the home team and the away team.



Various algorithms like Linear Regression, Support Vector Machines, K-Nearest Neighbors, Random Forests and Neural Networks were run. Hyper-parameter tuning was done in order to ascertain which parameters worked best with the models, and the best performing model is chosen to be the representative of the most accurate predictor. In the case of using spaces as nodes, the Random Forest algorithm gave the best accuracy score of 0.75. The heat map of the confusion matrix is shown in the figure below. The others, while performing decently, did

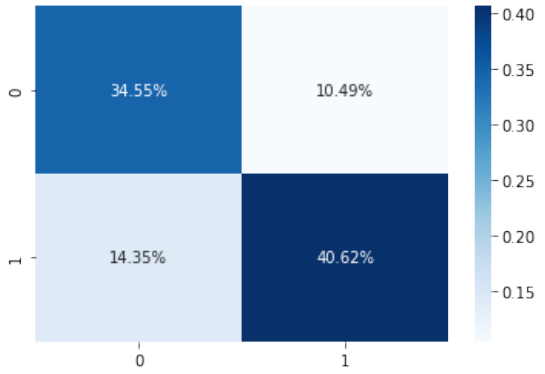


Fig. 10. The heat map of the confusion matrix.

not have as good of an accuracy. The feature importances (centrality measures), were measured with respect to the best performing model. The results are shown below as a Bar Plot in the figure given below.

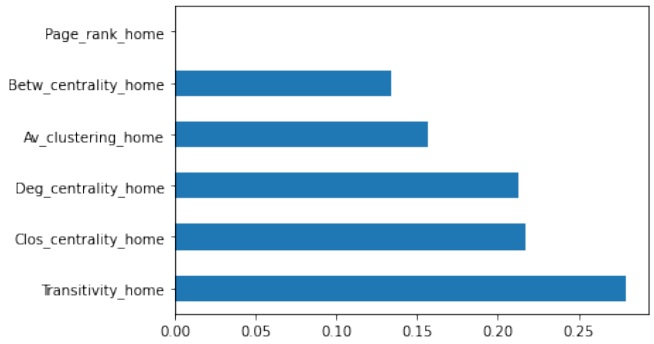


Fig. 11. Feature importances for centrality measures based on the best performing model.

The codes that reproduces the results can be found on the following GitHub link: https://github.com/Dinesh-Adhithya-H/Analyzing-Football-Passing-Networks.

IV. Discussions

A. General match statistics:

To understand the relationship between the passing network structure, network measures and football outcomes, we use quantitative methods that consider different probability assumptions. The real data applications for this paper are too many like [4]:

1) It provides an easy way to detect patterns or strong and weak ties among players and their positions in the lineup.
2) It can give us an insight about the relationship among the players.
3) It also can tell us how the network properties increase and decrease the chances of winning a match.

V. Conclusion

By considering passes during a football match as a network , of two types one as players as nodes and another by considering position on the pitch as a node we measured various centrality measures and how they affect match result for a particular team . This same exercise can be carried out for other sports dominated by passing events such as basketball. Our study shows that passing networks have a large extent of clustering compared to configuration model's of similar parameters. Maximizing degree centrality and closeness centrality and minimizing betweenness centrality increases chances of winning . Such metrics could in future be used for live match statistics which can help a sports viewer appreciate and engage in the game better.

References

[1] J. M. Buldú, J. Busquets, J. H. Martínez, J. L. Herrera-Diestra, I. Echegoyen, J. Galeano, and J. Luque, "Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game," *Frontiers in Psychology*, vol. 9, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2018.01900

[2] E. Arriaza-Ardiles, J. Martín-González, M. Zuniga, J. Sánchez-Flores, Y. de Saa, and J. García-Manso, "Applying graphs and complex networks to football metric interpretation," *Human Movement Science*, vol. 57, pp. 236–243, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167945717306280

[3] X. Du, W. Cai, J. Liu, D. Yu, K. Xu, and W. Li, "Basketball player's value evaluation by a networks-based variant parameter hidden markov model," 12 2020.

[4] P. Cintia, S. Rinzivillo, and L. Pappalardo, "A network-based approach to evaluate the performance of football teams," 09 2015.