Analysis of the football passing networks

Erazem Pušnik[a] **and Rok Miklavčič**[a]

[a]University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia

The manuscript was compiled on April 4, 2023

In this project we present the analysis of passing during a football match. Using social network analysis, we can construct a passing network, in which players are represented as nodes and passes between then as weighted edges. Once the network in constructed, we can focus on studying the relations between the nodes in order to extract meaningful information about their performance, as well as visualizing the network. We included measures such as degree, closeness, clustering centrality, PageRank score, edge connectivity, density, transitivity, modularity and others. We concluded that midfielders form the most important nodes of the network.

In sports, the analysis of each teams tactics, their style of play based on the results and their next opponent has played a crucial part in each teams preparation and focus. Specifically, in football, the world's most played sport, the need for pre and post game analysis has increased significantly.

Within the past decade, network analysis has emerged as a method for analyzing intra-team passing networks. A team is considered as a complex network whose nodes are its players (entities) and the connections linking the nodes of the network are the passes between the players. The goal of the observed team's network is to overcome the opponents network.

Once a network is constructed we can apply selected metrics from network science. For example, we can focus on the importance of specific players where the centrality metric would be used. There are various kinds of centralities such as degree centrality, betweenness centrality, flow centrality, closeness centrality, and eigenvector centrality.

Using network analysis to analyzing passing networks provides a deeper insight into the team's interactive behaviour compared to methods such as passing frequency and percentages. It allows us to gather more specific details which make each team unique, such as connectivity, clustering coefficient, shortest path length, degree counts, and so on.

Related work

In the paper (1) different studies on the topic are examined and assessed where we can search for key findings based on several samples. The (2) tells us that the passing networks in football matches are a dynamic system and shows dimension between teams and their respected networks. The main focus of the paper comes from the article (3) which studies the goal scoring passing networks. Our idea is derived from this paper since we will be researching the passing networks of the winning teams and comparing them to their opponents. Article also studies the degree centrality of these networks, network density, cohesion, connections, and duration. Another article which builds and studies the football passing networks with the *SNA* is (4).

Results

Player analysis. We calculated the data throughout the 7 matches of the World Cup tournament and averaged their values for all players involved. This is shown in the table 1. We notice that the majority of the maximum values are occupied by *P. Pogba* and the majority of the minimum are occupied by *O. Giroud*. This is somewhat expected as *P. Pogba* is a midfielder and his role in the team is to bring the ball up the pitch from defenders towards forwards.

O. Giroud on the other hand, is the most forward attacker of the team but not the primary goal scorer. His role is to create space for his wingers and take on as many defenders as possible without the ball at his feet. He is also tall, strong but not very fast (compared to wingers) and is therefore an areal rather than a ground threat. This results in the lowest degree centralities and lowest *PageRank* as the majority of the passes towards *O. Giroud* are either aerial passes, which usually aren't completed as they get intercepted by the opponent, or miss the player. Furthermore, since he is an attacker, he is surrounded by opponent's defenders and pressed a lot more compared to French defenders. The *simpler* passes are also more likely to be intercepted. As we can see in the table 1, *O. Girouds* clustering coefficient is the highest. This means that his neighbourhood is fairly connected, as the midfielders such as *N. Kante*, which has the lowest clustering coefficient, are doing the connecting.

Lastly, the *PageRank* winner is *B. Pavard* who is a right back (defender) of the team. The *PageRank* score is determined as a recursive notion of "popularity". The basic idea is that "a player is popular if he gets passes from other popular players". It is important to note that *P. Pogba* was a close second at *PageRank* but since *B. Pavard* had fewer connections and those connections were between most important nodes, he beats *P. Pogba* as he connects nearly everyone.

Measure	Results
Max/min degree centrality	**0.929 (P. Pogba)**/0.055 (O. Giroud)
Max/min in-degree centrality	**0.458 (N. Kante)**/0.033 (O. Giroud)
Max/min out-degree centrality	**0.371 (N. Kante)**/0.022 (O. Giroud)
Max/min closeness centrality	**0.469 (P. Pogba)**/0.081 (O. Giroud)
Max/min betweenness centrality	**0.056 (P. Pogba)**/0.0 (O. Giroud)
Max/min clustering coefficient	**0.143 (O. Giroud)**/0.089 (N. Kante)
Max/min PageRank	**0.039 (B. Pavard)**/0.006 (O. Giroud)

Table 1. Average results of measures for French players throughout the tournament (7 matches)

To visualize the distribution of centralities for each player, we used *seaborn* distribution plot (5). The plot displays a kernel density estimate and histogram with bin size determined automatically. On x axis, we display our centrality values, and on y axis, we display the frequency of the centrality value. Both closeness and betweenness centralities of the French team, can be seen on figure 1.

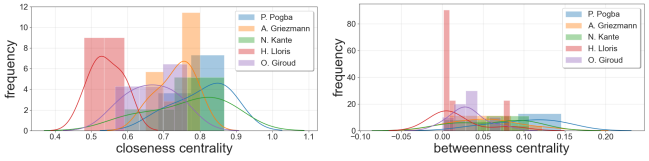

Fig. 1. Distribution plot of closeness (left) and betweenness (right) centralities for each player

Team analysis. The French national team averaged values shows in the table 2 throughout the 7 matches. We compare the results

All authors contributed equally to this work.

[1]To whom correspondence should be addressed. E-mail: fine.author@email.com.

to the opposing team Croatia, which also reached the final, but from the other side of the brackets. What we notice is that Croatia actually beat France on every measure except for edge connectivity and of course the most important one: the final match.

The results are again somewhat expected as the Croatia on paper played against worse opponents compared to the opponents that the France faced. For example the France faced *Belgium (3), Argentina (5), Denmark (12)* and Croatia faced *Russia (70), Denmark (12) England (8)* where the numbers in the brackets show teams official FIFA ranking for that year (6).

Measure	Results (France)	Result (Croatia)
Average degree	13.667	14.945
Average density	0.537	0.568
Average clustering coefficient	0.693	0.713
Average edge connectivity	1.571	1.571
Average max clique number	12.857	13.571
Average transitivity	0.755	0.767
Average triangle count	358.286	430.714

Table 2. Average results of measures for the French National team throughout the tournament (7 matches)

For each match, we have generated an image of the graph with both *NetworkX* library and *Gephi* graph. With *NetworkX*, we used spring layout, which positions nodes using Fruchterman-Reingold force-directed algorithm (7). The only visualization setting we used is node direction and scaling by their degree, as can be seen in figure 2. This visualization served us a simple and quick overview of the network.

With *Gephi* library, we used more advance methods of visualization. To position the nodes, we used Force Atlas 2 layout algorithm. Nodes were ranked using *PageRank* algorithm, with higher scoring ones being colored with darker shade of green. Links between the nodes are directional, with their thickness and color being based on the weight of the link. For labels, we scaled them based on the node ranking. An example of the visualization can be seen on figure 2.

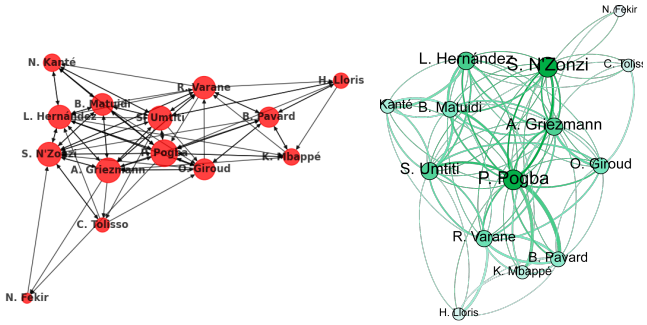

Fig. 2. Visualization of the French team network in a match against Croatia using *NetworkX* (left) and *Gephi* (right) library.

Match analysis. When calculating the measures for the French players in the final, we notice an already seen theme. The table 3 shows that *P. Pogba* is again the main French player passing and connecting the team together. At the metric *max out-degree centrality*, he is joined by *A. Griezmann*, who has a more attacking role. From this we can determine that in this match *A. Griezmann* dropped lower in the field to get the ball and based on his high degree count 18 compared to *P. Pogba's* 21 played similarly.

We also notice that the minimum values are occupied by *H. Lloirs* and not *O. Giroud* as seen on the average data in the table 1. Since *H. Lloris* is a goalkeeper, we assume that in this match the French were attacking the majority of the game, thus leaving the goalkeeper with little to no passing duties. But when comparing it to other data available for this match, we see that this is not the case, and that the passing network analysis is in this case misleading.

The ball possession in the game was in favour of Croatia 66% to 32%, total number of shots were also in Croatia's favour 14 to 7 and so were the number of passes with 528 to 289 and the passing accuracy with 83% for Croatia and only 68% for France. This data shows that Croatia were attacking most of the time. Why are the French's goalkeepers centralites values so low? Surely he must've had the ball quite a couple of times if Croatia was attacking the majority of the time. Most of his passes were not completed as he kicked the ball far from his goal and also was not included in the build-ups of the French attacking plays (8).

Measure	Results
Nodes	14
Edges	88
Max/min degree centrality	**1.615 (P. Pogba)**/0.538 (H. Lloris)
Max/min in-degree centrality	**0.846 (P. Pogba)**/0.231 (H. Lloris)
Max/min out-degree centrality	**0.769 (A. G. & P. P.)**/0.308 (H. L. & N. K. & K. M.)
Max/min closeness centrality	**0.867 (P. Pogba)**/0.52 (H. Lloris)
Max/min betweenness centrality	**0.145 (P. Pogba)**/0.002 (N. Kante)
Max/min clustering coefficient	**0.844 (N. Kante)**/0.5 (P. Pogba)
Max/min PageRank	**0.131 (P. Pogba)**/0.0397 (H. Lloris)
Density	0.484
Clustering coefficient	0.681
Edge connectivity	1
Max clique number	9
Transitivity	0.716
Modularity	0,17
Diameter of the network	3
Average path length	1,555

Table 3. Results of measures for french players in the final match between France - Croatia, 4 - 2

Discussion

The results show that midfielders of the football teams form the most important nodes, if we perform the social network analysis based on completed passes between players. This is expected, since midfielders are literally in the middle of their team and thus connect defense with offense. Their goal is to bring the ball forward and create chances for their attackers to finish.

Results also show that the attackers will usually have the lowest centrality values as they either don't get the ball or they don't actually pass since their goal is to shoot, not pass.

Goalkeepers are usually the ones starting the attack for each team and therefore can have quite high centrality values but when put under pressure, their passing accuracy drops significantly, and so does the centrality values.

Overall, the passing network gives us a lot of insight into who are the go to players when passing the ball and which players have other duties. The stats can be interpreted in a number of ways but can be misleading when trying to reason the actual game result and should be used in addition to other stats to avoid assumptions.

Further work could include more matches or even competitions in the analysis for the selected time frame. Another option would be to merge the data with additional research, such as situations leading to goals or even situations leading to goal errors on the defensive end.

The main contribution of this work is showing that social networks can be used to effectively analyze football passing networks of football teams as well as insights into individual players of the team.

Methods

Data gathering and parsing. In order to properly analyze passing networks in football, we needed to find a good source of football matches data sets. The data set we decided on, was the "Soccer match event data set" version 5 (9). It contains all the spatio-temporal events (passes, shots, fouls, etc.) that occurred in an entire 2017/2018 season of five European competitions (La Liga, Serie A, Bundesliga, Premier League, Ligue 1), as well as data from FIFA World Cup 2018 and UEFA Euro Cup 2016 events. A match event contains information about its position,

time, outcome, player and characteristics. The data set also contains the following collections: coaches, competitions, player rank, players, referees and teams. For our use case, we will only make use of the events, matches, players, competitions and teams collections, since the other ones do not contain the data we need for construction and analysis of the passing network. While the data set isn't exactly up-to-date, it should still present us with a wide range of useful and meaningful raw data.

The data is provided to us in the *CSV* format and for the purpose of parsing it properly, we needed to convert it to *JSON*. Some of the data is also hard-coded as a *JSON* string, so we needed to phrase it properly. This process can be seen in figure 3, as the *preprocess* step.

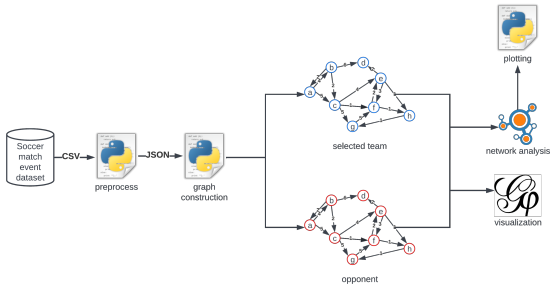

Fig. 3. Steps in process of social network analysis

Transforming collections to a network. To properly represent the passing network of a team, we needed to decide on the structure that makes the most sense. We ended up using a directed network, in which we represented players of the team as nodes, with edges corresponding to the passages between the players. To showcase the number of passes between the players, we used weights on the edges. When generating the network, we only used data from the World Cup tournament. We had to extract events for each match and filter only accurate passes, since we are not interested into the other types (e.g. head, high...). Once we had all the raw data, we could easily add the nodes to our network (*graph construction* phase in figure 3). A simple example of our passing network can be seen on figure 2, that we have used in our results section.

Network analysis and tools used. The passing network analysis was carried out in the following order:

1. **Player analysis**: The main goal is to examine player's individual contribution. We can achieve that by calculation of the following measures:
   - **Degree centrality**: Degree is a simple measure of centrality that counts how many neighbors a node has. In directed network, there are two versions of the measure. In-degree is the number of in-coming links, while the out-degree is the number of out-going links.
   - **Closeness centrality**: It measures node's average distance from all other nodes. The distances between nodes with a high closeness score are the shortest.
   - **Betweenness centrality**: It is a way determining how much of an impact a node has over the flow of information in a graph. It is frequently used to find nodes that serve as a bridge between two parts of a graph. A score is assigned to each node based on the number of shortest paths that pass through it. Nodes that more frequently lie on the shortest paths between other nodes will have higher betweenness centrality scores.
   - **Clustering coefficient**: For a node, it describes the likelihood that the neighbours of the node are also connected. Two versions of this measure exist: the global and the local. The global version gives an overall indication of clustering in the network, whereas the local version gives an indication of the embeddedness of single nodes.
   - **PageRank**: Based on the number of incoming relationships and the importance of the corresponding source nodes, the algorithm calculates the importance of each node in the graph. We could say that a node is only as valuable as the nodes it connects to.
   - **Triangle count**: For each node in the graph, it counts the number of triangles. A triangle is made up of three nodes, each of which is connected to the other two.

   Based on the results, we can decide which player contributed the most in the match.
2. **Team analysis**: The purpose is to understand how the connectivity of the team affects its performance and style of play. The measures useful in this case are *clustering coefficient*, which we have already described, as well as the following ones:
   - **Edge connectivity**: The minimum number of edges that need to be deleted, such that the graph gets disconnected, is called edge connectivity.
   - **Max clique number**: A clique is a subset of undirected graph vertices in which every two distinct vertices are adjacent. A graph's maximum clique is a clique with no more vertices than the graph's maximum clique.
   - **Transitivity**: It is the overall probability for the network to have adjacent nodes interconnected, thus revealing the existence of tightly connected communities.
3. **Match analysis**: Serves as a fusion of both player and team analysis. We can determine the contribution of individuals in relation to the team. The helpful measures are all the previous ones as well as ones such as:
   - **Modularity**: The density of connections within a module or community is measured by modularity, which is a measure of the structure of a graph. Graphs with a high modularity score will have many connections within a community, but only a few pointing outwards to other communities.
   - **Diameter of the network**: We can define it as the longest of all the calculated shortest paths in a network. It is the shortest distance between the two most distant nodes in the network.
   - **Average path length**: Between two nodes in a graph, shortest path, it is the one with the minimum number of edges. If the graph is weighted, it is a path with the minimum sum of edge weights. The average path length is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. It is a measure of how efficient information or mass transportation is in a network.

   In order to confirm and prove our analysis, we can compare the results with an analysis of a match done by an outside source, written after the event has occurred. We also need to take into account substitutions of players that occurred during the match.

Network analysis of all described metrics was primarily be done with *NetworkX* library (1) (*network analysis* phase in figure 3), while graphical visualization of the social networks was primarily done with *Gephi* (10) (*visualization* step in figure 3). We decided to use it, as it allows us to customize the visualization of the network to a higher degree. Once we had completed our network analysis, we could use the data to draw the plots of different measures, as is represented with *plotting* step in figure 3.

1. Sergio Caicedo-Parada, Carlos Lago-Peñas, and Enrique Ortega-Toro. Passing networks and tactical action in football: a systematic review. *International Journal of Environmental Research and Public Health*, 17(18):6649, 2020.
2. Javier M. Buldú, Javier Busquets, Johann H. Martínez, José L. Herrera-Diestra, Ignacio Echegoyen, Javier Galeano, and Jordi Luque. Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. *Frontiers in Psychology*, 9, 2018. ISSN 1664-1078. . URL https://www.frontiersin.org/article/10.3389/fpsyg.2018.01900.
3. Scott Mclean, Paul M Salmon, Adam D Gorman, Nicholas J Stevens, and Colin Solomon. A social network analysis of the goal scoring passing networks of the 2016 european football championships. *Human movement science*, 57:400–408, 2018.
4. Scott Mclean, Paul M Salmon, Adam D Gorman, Karl Dodd, and Colin Solomon. Integrating communication and passing networks in football using social network analysis. *Science and Medicine in Football*, 3(1):29–35, 2019. . URL https://doi.org/10.1080/24733938.2018.1478122.
5. Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. . URL https://doi.org/10.21105/joss.03021.
6. FIFA. Men's Ranking, 2018. URL https://www.fifa.com/fifa-world-ranking/men?dateId=id13603.
7. spring_layout NetworkX 2.8.2 documentation, 2022. URL https://networkx.org/documentation/stable/reference/generated/networkx.drawing.layout.spring_layout.html.
8. France 4-2 Croatia - FIFA World Cup 2018 - Match Report, 2018. URL https://www.whoscored.com/Matches/1307192/MatchReport/International-FIFA-World-Cup-2018-France-Croatia.
9. Luca Pappalardo and Emanuele Massucco. Soccer match event dataset, Feb 2019. URL https://figshare.com/collections/Soccer_match_event_dataset/4415000/5.
10. Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*, 2009.