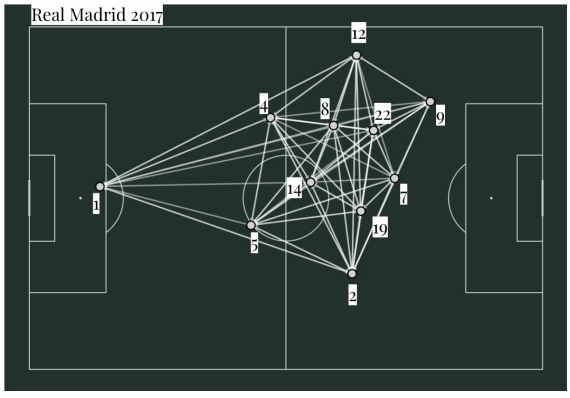
Alex Johnson
April 13, 2022
Section 1

Passing Networks: A Mathematical Perspective

Part 1: Finding and Creating the Network

Given my general interest in soccer, I have frequently sought forms by which I could better understand and analyze the beautiful game. Combining this with my passion for mathematics, I was naturally led to the book *Soccermatics* by the applied mathematician David Sumpter. Dr. Sumpter uses fundamental techniques of mathematical frameworks to model diverse phenomena that occur in the sport, expanding on a wide range of analytical tools, including geometry and statistics. However, of particular interest to me, chapter seven of this book discusses in detail the idea of passing networks in soccer, networks that are made up of individual players on a given team as the nodes and directed edges representing a pass made from player *i* to player *j* over the course of a match. Though I thought the concept of a passing network to be interesting at the time of reading this book, I never possessed the tools necessary to analyze, understand, and evaluate a passing network. This project gave me the opportunity to learn more about said networks, and, therefore, about soccer tactics, player importance, and the sport in general. The rest of this report will serve as a discussion about my findings in analyzing such networks, leading to a final discussion about the real-world features of these networks I am analyzing.

I will first briefly go over how I created the various networks I analyzed. A company called StatsBomb has created a package in Python that makes public detailed statistics of events that occur in various soccer matches. Conveniently, the match-level data includes information about all passes made over the course of the match, detailing which players were involved in the action, along with their location on the pitch when the pass was played. I developed a program that organizes this data in such a format that it may be fed into the NetworkX package to then be mathematically analyzed according to the tools developed in this class. I will not go over the specifics of the data preparation process in this report, but I of course include the program written so it may be referenced (function name get_pass_network). I now show a plot of such a network that will serve as a guiding example for much of the following analysis:



Deleted: with

Deleted: further

Part 2: Network Analysis

Basic Structure: As can be seen by the plot above, the passing network itself isn't entirely large; it is composed of only 11 nodes (one node for each player on the team) along with a varying number of edges, 93 in this network. Note that this network is a directed network, with an outgoing edge representing a pass being made by a player and an incoming edge representing a player receiving a pass. It should be noted that if a player *i* passes the ball to player *j* multiple times, *i* and *j* are still only connected by one edge in the graphical representation. To overcome this, I weight the edges according to the number of passes that are outgoing from one player to the player being passed to, making this network weighted. Last, I plotted the players according to their average position on the field over the course of the game, introducing a spatial component to these networks.

Centralities: Finding the common centralities of the network plotted above leads to interesting results, since they, give or take, may be interpreted in giving the most important players in the passing network. However, first I give more context into this specific network. Being the son of a mother from Madrid, I have followed the family tradition of supporting the soccer team Real Madrid, and as such I created a passing network of one of their greatest modern triumphs, their Champions League final victory against Juventus in 2017. The results of the centrality measures at first surprised me as they did not fall in line with my initial analysis of who the most important players in the match were. Yet, going over the match again, I realized that the results yielded by the centrality measures were nonetheless accurate and thus incredibly interesting. According to degree, eigenvector, betweenness, and closeness centrality, the most important player for Real Madrid was number 2, Dani Carvajal (refer to the code to see specific values). Among other things, this implies that Carvajal both received and passed the ball the most, shared passes with other important players, has the shortest geodesic distance to all other players, and is, relatively, on the highest number of shortest paths between any two players in the network. These findings are simple in nature, but they do add immensely to match analysis, and, at the very least, highlight the impact Carvajal had on this game.

After being encouraged to seek new centrality measures that combine my expert knowledge in soccer using mathematical tools, I referenced *Soccermatics* to find such a centrality measure. In his discussion, Dr. Sumpter references a study performed by the mathematical sociologist Thomas Grund in which Dr. Grund developed a network centrality measure specific to passing networks. Mathematically, this network centrality measure, which I informally call "pass centrality," is defined as follows:

$$PassCentrality = \frac{\sum_{i=1}^{11}(p^* - p_i)}{(n-1)\sum_{i=1}^{11} p_i}$$

where p^* refers to the maximum number of passes received by any player and p_i refers to the number of passes received by player *i*. Note that pass centrality gives one number for the entire passing network and it isn't specific to any one player. Note further that pass centrality will

Deleted: where this network has

Deleted: itself

Deleted: received and passed the ball

Deleted: to

Deleted: that

always give a number between 0 and 1, with the implication that a 1 means that all the passes in the match had been made to the player who received the most passes, while the number 0 means that all players received the ball the same number of times. In other words, pass centrality tells us how focused, or localized, a passing network may be on just one player (the player who receives the most passes). It isn't hard to believe that teams with a high pass centrality tend to struggle in comparison to teams with low pass centrality, since teams that focus their passing on only a few players aren't able to move the ball effectively over the entire pitch, whereas teams who involve many players in their passing generally are able to move the ball into dangerous positions. Dr. Grund's study echoed this intuition; teams with lower pass centralities have an 8% advantage in terms of scoring more goals than the opposition, a significant enough advantage to win matches.

After implementing this idea as a program in Python (function name pass_centrality), I calculated the pass centrality for the network plotted above for Real Madrid, yielding approximately 5%. Real Madrid's opposition on the day, Juventus, had a pass centrality of around 7%. Given this context, it isn't too surprising to see that Madrid's more spread-out passing network passed the ball effectively across the pitch, scoring 4 goals against Juventus en route to victory. I calculated several other pass centralities, including those in the World Cup match between Spain and Portugal in 2018, along with the pass centrality of Barcelona in the 2011 Champion's League final. The example of Barcelona serves as an important counterexample: the 2010-2011 Barcelona team is considered by many pundits as the greatest club team ever assembled, but in the match in which I analyzed them, they had a pass centrality of around 11%, a figure that was much higher than their opposition. However, Barcelona were dominant on the day, winning the match 3-1, showing that pass centrality, though incredibly informative, isn't perfectly predictive.

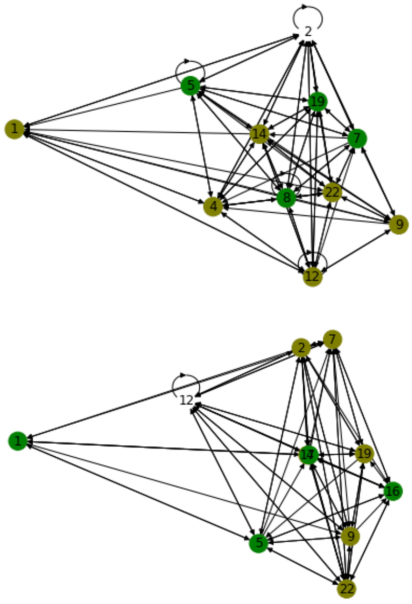
Network Measures: In measuring network features, I first remind the reader of the importance of the triangle in soccer. A triangle allows for players to make quick, interchanging passes to escape the pressure of the opposition's defense and thus maintain possession of the ball, building attacks that may lead to a goal. Note that the global clustering coefficient of a network is defined as the proportion of triangles in a network over the number of connected triples. Since connected triples themselves aren't inherently useful in a passing network, then the higher the transitivity of a network, the easier it is for a team to escape pressure and maintain possession since an increasing amount of their connected triples are triangles.

In analyzing two passing networks, I calculate the global clustering coefficient of the Portuguese national team in their World Cup match against Spain in 2018, and of the same Barcelona team referenced earlier in their match in 2011. For Portugal, the coefficient ended up being approximately 0.628. Note that this result is much higher than many other types of networks but happens to be lower than the coefficient of many other passing networks, since this indicates that only 63% of its connected triples are triangles, thus making it much harder for Portugal to escape the opposition's pressure. For Barcelona, the coefficient was around 0.848, the highest such figure I could find. This result made sense in the context of this Barcelona team, as this team was known for its quick passing and moving, using such triangles expertly to escape

relentless opposition pressure, building attacking scenarios from almost nothing. Their relatively high number of triangles compared to connected triples seem to verify this tactical analysis.

Community Structure: For detecting communities, I wanted to see how effective a simple modularity maximization community detection algorithm would be in predicting team formation (such as a 4-3-3, for example). If community detection can do this, then it would be easy to measure how true a team is able to stay to their desired formation over the course of a game, showing us the adaptations that were necessary to win a match. Likewise, we could then use community detection algorithms to show how formations change over the course of a season for a single team.

In an attempt to observe this, I observed the passing networks of two games of Real Madrid in 2017. After using the simple community detection algorithm, I observed the following communities in the passing networks, the first for the game in the plot showed above, the second for a game earlier on in the season against Barcelona:

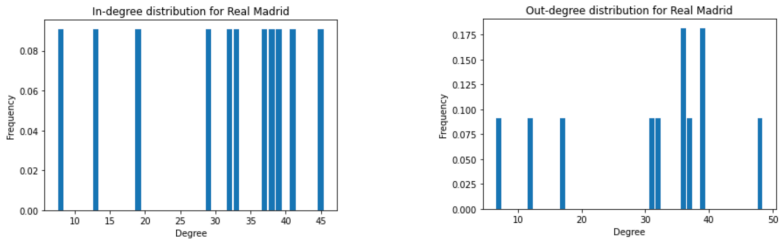


Notice first that the networks themselves are flipped horizontally from what they should have been. Second, though the passing networks seem somewhat similar, the players don't seem to be paired in the same communities too frequently. This isn't too worrisome on its own since the players in the team over both games are different, and so this may be accounting for the different players, though it is true that different players who are playing the same position aren't included in the same communities. Additionally, note that the community detection algorithm doesn't give a clear team formation structure, instead creating what seem to be arbitrary connections. Overall,

it doesn't seem that a standard community detection algorithm gives clear formation predictions, and neither does it give too much in way of determining formation changes between games. A community detection algorithm specific to passing networks seems useful, since we could use it to give communities like what we should be expecting, with any such algorithm likely making use of the spatial features and the weighted aspect of the passing network.

Real World Features: When observing the global clustering coefficient, we observed that the result given for passing networks is generally high, the lowest observation being around 0.63. If indeed most passing networks are above this figure (for the scope of this project, this seems a reasonable assumption), then this indicates that, at least in relation to transitivity, passing networks don't seem to be very realistic.

To analyze further if passing networks model real networks well, I now check the in-degree and out-degree distributions to see if they follow a power law. I plot them as follows, using the first 2017 Real Madrid team as a reference:



As we can see, these distributions clearly do not follow a power law distribution, with the distributions also not changing when plotted on a log-log scale. Again, in this way, passing networks don't seem to share the same features of real networks.

Part 3: Conclusion and Real Network Features

I first note that there are potential sources of error in the data collection process, as is evident by the self-edges in all the passing networks analyzed. Since a pass in soccer, by definition, includes two players, it seems that self-loops should not exist. I chose to leave them in since I am not completely aware of the data collecting process and there may be a reason these self-edges exist.

Though these passing networks have been created from data from the real-world, that does not imply that they must possess features associated with real networks. In fact, through this analysis, I have found that passing networks don't typically have features associated with real networks as they have abnormally high global clustering coefficients along with in and out-degree distributions that do not follow a power law. In addition to this, passing networks are generally much smaller than many real-world networks. Also, note that passing networks don't grow over time; they are limited to only 11 nodes. As such, it seems that most models we discussed in class seem to be bad models for passing networks. For example, when considering the Barabási-Albert model, we may build the model such that it satisfies the clustering coefficient, yet, as a dynamic model, we will not satisfy the static nature of passing networks, and since players most likely do not preferentially choose who to pass to. Other models, such as the configuration or the Price model, cannot be constructed in such a way that they match the degree distributions of passing networks, since they give power-law distributions. Any network generation model would necessarily have to account for only generating 11 nodes, while also ensuring a high global clustering coefficient, while not giving power-law distributions.

Deleted: Erdos-Renyi

Deleted: .

In conclusion, passing networks are very indicative of team performance in matches, giving key insights when analyzed mathematically. However, I note that not everything that happens in a match can be represented on a passing network (such as goals!). In this way, there is always something to love and marvel at in the context of the beautiful game.

Deleted: Though I didn't observe other real-world features, such as the giant component, it isn't hard to conclude that passing networks aren't associated with real networks.