Summary

In our report, we seek to find a concrete method to analyze a team's passing network in order to identify areas of improvement and possible changing of players' positions, in an effort to improve the overall success of the team. We initially began by making a network consisting of nodes representing the players (with their respective position given by the average of all the passes they gave and received) and edges constituting passes between players. However, as we will describe in further detail in Section 3.1, we decided this was a sub-optimal method for pointing to areas of improvement, even if it could point to the success of the team to a degree. In place of this, we tracked all the passing triangles that occurred throughout the games. We then created a two-dimensional histogram consisting of the coordinates where these triangles occurred, and identified five significant zones we wanted to track in which more passing triangles occurred than anywhere else. We then created another graph using these five zones as nodes to see how interconnected these high density passing zones were, knowing that the occurrence of passing triangles indicated high passing density and more time spent. This lead us to conclude that while the team may be connected at a person to person level, the heaviest passing zone (Highlighted in yellow in Figure 2) was not well connected to the defensive passing zones (shown by the thin lines indicating passing frequency). By linking the defense better to the yellow node, the Huskies team can accelerate their play and improve their overall connection.

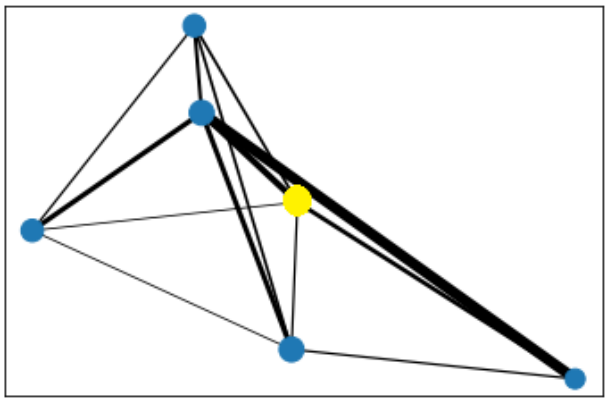

Figure 1: Connections between High Density Passing Zones

Decomposing Passing Networks

Team No. 2020452

Problem D

Contents

1 Introduction

Our goal was to quantify and formalize the structural and dynamical features of the Huskies passing network in order to find weaknesses and portions for improvement. In the passing network we were specifically looking for dyadic and triadic configurations.

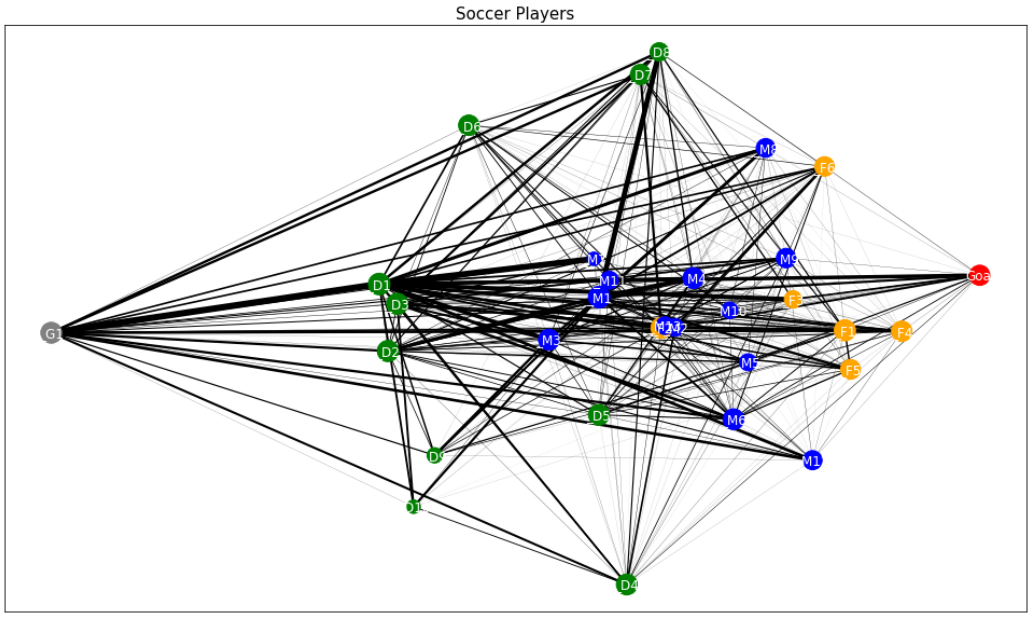

Figure 2: Node Map

Information about Figure 2:
Number of nodes: 31 (Players)
Number of edges: 390 (Passing Lanes)
PageRank S.D: 0.0068

Our first step was to create a model of the passing network of the Huskies. Each node in Figure 2 represents the average position recorded from each player the instant passed or received a pass. This figure, and all of our figures are oriented with the Huskies goal on the left, and facing the opponent (a left defender would be in the top left corner). The node size represents

each players PageRank value (described further in section 3.1). In our instance however this was not particularly interesting as almost all the players had similar degrees, shown by the standard deviation below figure 2. This standard deviation would be an important metric to track had we the time to compare the different possible formations, combinations, and substitutions possible for the coach. The node map highlights the position of each player on the field in addition to the players who have the highest pass rates in between one another. The model also indicates the position of the goal as a node to show the pass chains that are successful in scoring. We did not enlarge the nodes in Figure 1 because it is self evident which players had the greatest amount of interaction.

2 Statement of Problem

Our goal was to analyze the data of the competitive soccer team, the "Huskies", and provide a meaningful model of a network tracking the passes between players. This network leads to the analysis of the team's success and how well the team plays together. Other factors such as skill sets, individual versus collective performance, and passing chain links are possible areas for research in order to meet the goal. Our solution should be able to be generalized and applicable to other teams across various sectors.

2.1 Simplifications

In approaching the creation of our model we made some simplifications and excluded some of the sample data given to us, an a effort to meet the time constraints given to us by the coach.

2.1.1 Macro Player Analysis

When analyzing the given data, we made the simplification of including all players, including substitute players, on the same node map, instead of making separate maps for each game and only eleven players. As shown in figure 2 above, This allowed us to analyze overall trends in order to present a more generalized solution instead of only being able to produce solutions to each game. The trends associated with this could be very pertinent to the team, but would require a larger sample size, as we only had 38 games, and the

formations did not vary that much. In order to predict wins, or even just predict formations, another seasons results would be necessary to produce statistically significant metrics.

Furthermore, there was data given on which coach was active for which games. However having three different coaches divided the data set significantly, and thus we decided it wasn't of immediate value due to the lack of data, and difficulty.

2.1.2 Data Types

Another way that we simplified the data was including passing events and not the full events log, which include the kicks, goals, and launches. We made this choice partly due to the nature of a node map, in that its purpose is to display the passing activity, and though the information could be integrated into our model to provide more information about the different patterns which could be useful, our focus is on passing.
When dealing with our data, all the coordinates came in a 0-100 range. Using this we scaled the coordinates to represent a soccer field, which is a rectangle rather than a square. In doing so we lost slight accuracy, but achieved more readable metrics.

2.1.3 Probability

The edges in our model represented the frequency a pass occurred between the two nodes. However, this did not take into account failed passes. Using a model with the pass completion percentage is an area of further study, as it leads to many more complications. Having a low pass completion is obviously sub-optimal, but sometimes what the coach may want from the players. If the central midfielder has 100% pass completion with lead forward it would likely be indicative that the forward is not moving into aggressive positions, or it could be indicative of the teams overall strategy. If the forward was constantly in aggressive positions that had only a 30% chance of pass completion, but a 40% of a goal had he gotten the goal, that is clearly better than a 90% pass completion with a .1 chance of a goal (.3*.4 ¿ .1*.9)

3 Analysis

3.1 Why Move Smaller

When analyzing large networks such as ours, the PageRank algorithm is the most widely used metric. Developed originally to rank the importance of different websites, it is calculated by observing the number of links that go from site to site. It has been used in various academia the context of soccer to explain the trends and the uniqueness of certain teams. [1] [2] [3]

Our initial approach to this problem was to make the network of players, and calculate various metrics such as PageRank and Betweenness Centrality. However, after doing this, we realized the results were not objectively valuable. The players with the highest page rank were unsurprisingly the midfielders, and those who centrally placed in the field. It was largely a metric of their position. Average PageRank consequently was a measure of how much passing occurred. Telling a coach to "pass more" is common sense and thus, not particularly valuable.

Furthermore the standard deviation of the PageRank we calculated from the standard PageRank algorithm was approximately 0.0068. Had we computed passing density into this metric we might of had more deviation, but with such a low amount it appeared discouraging.

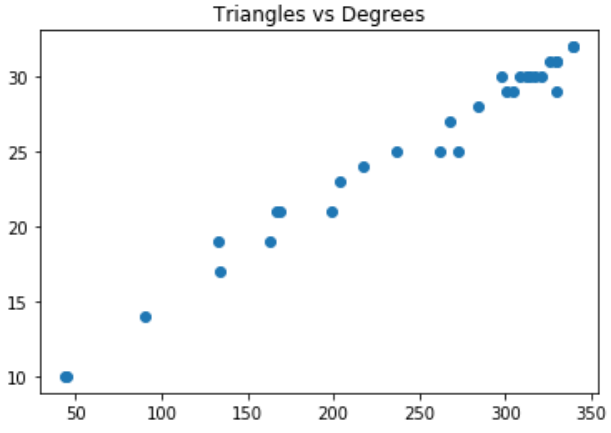

Figure 3: Possible Triangles vs Degrees

The calculation of passing triangles for each player also proved to be a dead end, as almost all players were connected to some degree, and thus the triangles had a heavy correlation to the degree (the number of edges attached to a node) of each player, as shown in Figure 3.

3.2 Network Reduction Process

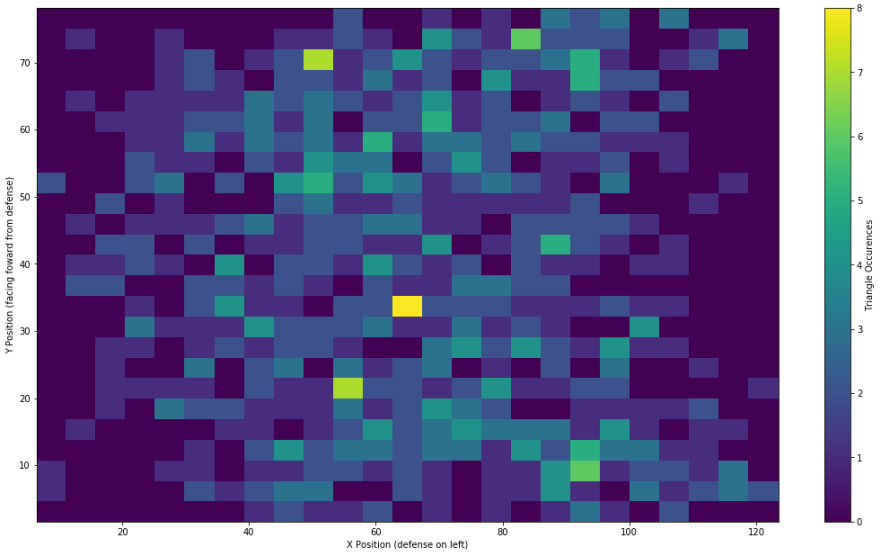

Figure 4: 2D Coordinate Histogram of Passing Triangles

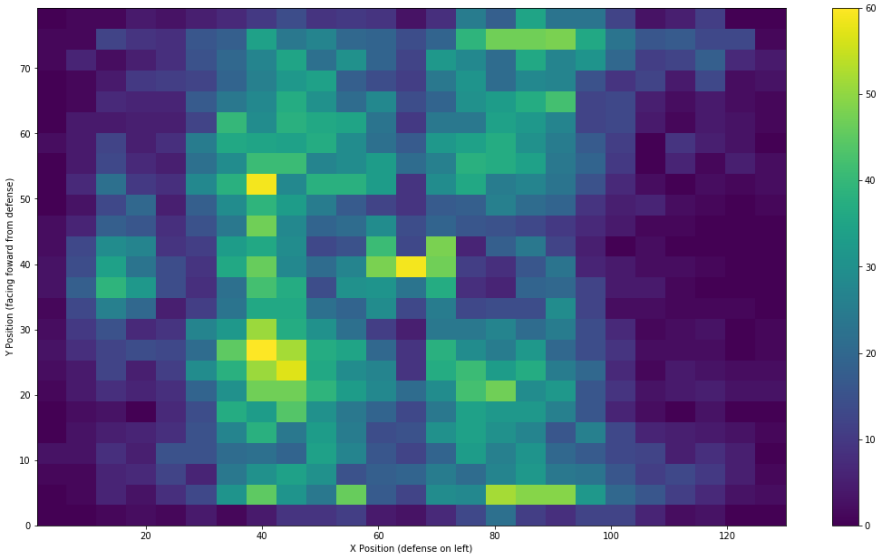

Figure 5: 2D Coordinate Histogram of All Passes

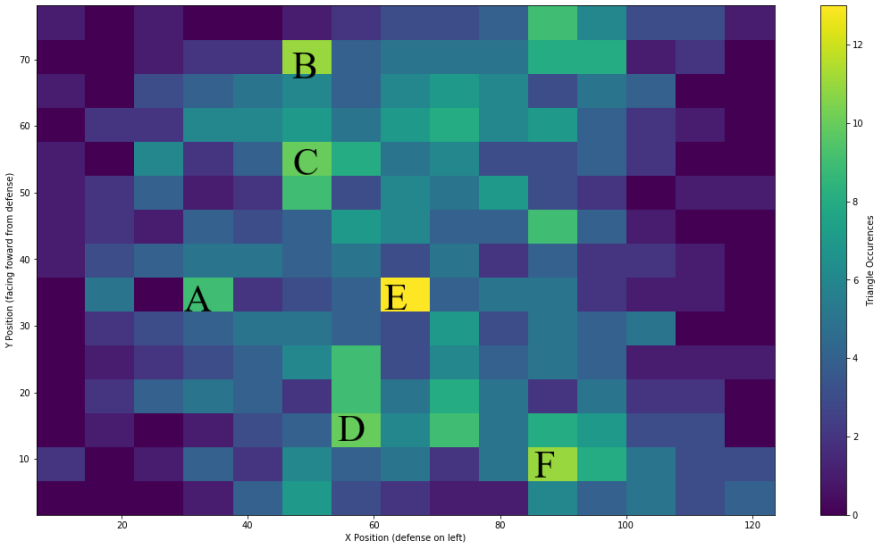

Figure 6: 2D Coordinate Histogram (Larger Bin Size)

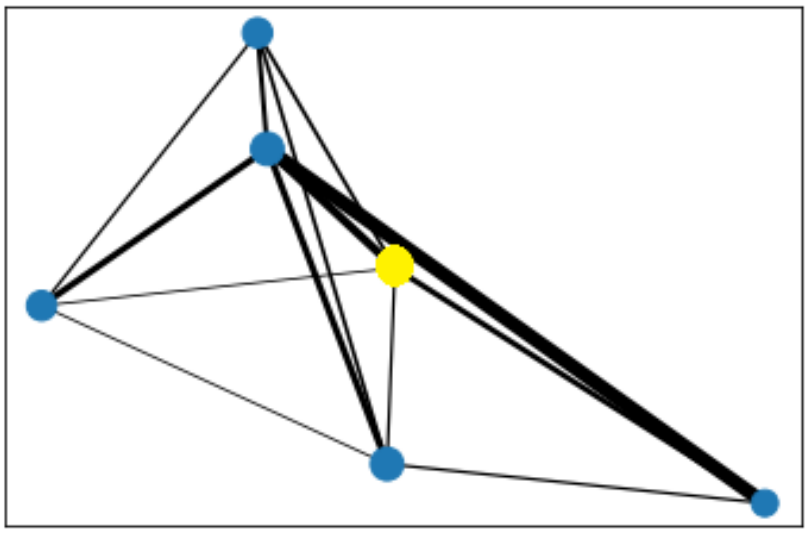

Figure 7: Connections Between High Density Passing Zones

3.2.1 Network Reduction Process

We built the initial two-dimensional histogram by calculating all the passing triangles between the players (Figure 4). We defined a passing triangle as a series of three passes that ended at the initial passer. While there are many ways to calculate high density passing zones, the occurrence of triangles is a common indicator in soccer of not only frequent passing, but chemistry between players and passing the ball in order get around the opposing team and find openings rather than just moving the ball up field.

We can tell this method of triangle calculations is accurate due to the two dimensional histogram of all the passes (calculated from the origin). This showed us approximately the same results as our triangle method, which is a good thing.
Although the zones in the initial heat map (Figure 4) have lower occurrences than one might expect, this is due to our method of taking the center of the triangle as a single point rather than the overall area of the triangle when making the overall heat map. However, this led to a similar result once the bin size was increased (shown in Figure 6), we could pinpoint the most common location of passing triangles easier.
After calculating the most frequent area passing triangles were centered around. We made a node map using the center of the high density areas (labeled in Figure 6), and calculated the passes that came out of that zone, into another dense passing zone. This is represented in Figure 7. Our conclusion was largely based off this final figure.

3.3 Conclusion

Our approach to solving this problem was to represent and analyze the given data in two ways: in a graph of the passing networks for the Huskies soccer team throughout the season and in a graph of the passing networks between high density passing zones. The results from Figure 7 (links between high density passing networks) were surprising, as it showed that the connections specifically coming out of defensive passing zone to the highest density passing zone in midfield (the yellow dot in Figure 7) were infrequent. This midfield zone was where the Huskies controlled the ball the most, and where most of their offensive plays started from. While typical soccer wisdom advises clearing the ball to the sides out of defense to minimize danger, these points were higher up on the field, and thus more involved in the transition to offense. Telling the defenders to find the midfielders in this central zone rather than looking for the players on the wings would aid in the generation of more offensive opportunities, and cutout wasted time in moving from the defensive middle to the wings and then back to the middle.

4 Broader Implications

The largest takeaway from this study was that the majority of passes, and connections were in very concentrated specific zones. This is likely very similar to human networks where subsections of people in similar groups interact and account for a vast majority of the passes, emails, or interactions across a team, business, or country. Strengthening the connections between these dense zones is something difficult but worthwhile to pursue.

Additionally, creating networks that link to the most connected player (or heaviest passing zone in our case) can be an easier method to increase overall connection rather than trying to create individual links to each player. This is a point of failure for the PageRank algorithm in its application to social networks, as reporting to a central figure is common practice, and less chaotic then trying to encourage connections between everyone, in order to boost the average PageRank score.

4.1 Further Research

In the process of computing our network, and reduction process, we identified several key areas of relevant study listed below.

4.1.1 Overlapping Networks

One of the avenues of further research we identified specifically relating to soccer, is in comparing passing networks between two teams to optimize one's network in relationship to the opposition. If a manager knows in detail the other team's network, he could attempt to maneuver his players as to avoid the other network. For example, given the average location of the opposing players, and weights for their interception frequency, it would possible to manipulate the player formation such that his most important links do not intersect the opposing players, or the zones in which the opposing team steals the ball most frequently. However moving players around is typically not done in soccer due to a variety of reasons so constraints on how dramatically the formation could be shifted would need to be set.

4.1.2 Performance Predicting and Network Building

Predicting the overall success of the network by various algorithms (PageRank, Between,Shortest Chain) in addition to quantifiable results such as goals, individual happiness, and shots is another worthwhile endeavor. If one could come up with an combination of these algorithms or significantly accurate predictions through machine learning, the ability to manipulate your own network to find the "optimal" orientation would be extremely valuable. This was not possible given our smaller data set.

References

[1] Markus Brandt and Ulf Brefeld. Graph-based approaches for analyzing team interaction on the example of soccer. 2015.

[2] Emerging Technology from the arXiv. Pagerank algorithm reveals soccer teams' strategies. 2012.

[3] I. Echegoyen F. Seirul.lo J.M. Buldu, J. Busquets. Defining a historic football team: Using network science to analyze guardiola's f.c. barcelona. 2019.