Showing the Coach What He Can't See: Graphical Passing Networks as a Method of Soccer Team Analysis

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364740">https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364740</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Showing the Coach What He Can’t See: Graphical Passing Networks as a Method of Soccer Team Analysis

Suriya Kandaswamy

A thesis presented to the Computer Science department in partial fulfillment of a degree of Bachelor of Arts with Honors

Harvard University
Cambridge, Massachusetts
April 3, 2020
Abstract

Despite being the most popular sport in the world, soccer has a relatively under-developed analysis space compared to some other sports. The low-scoring and passing-based nature of soccer makes it incredibly complex, but also makes it possible to characterize a team’s match activity as a passing network in the form of a mathematical graph. We use passing networks and metrics developed for characterizing graphs in order to describe a soccer team’s performance overall and with respect to individual players by drawing analogies between graphs and passing relationships. Many of these metrics make it possible to compare a team’s performance between two different networks with respect to ball movement and strength of possession in terms of visual representations and quantifiable relationships. Contextualizing graphical differences between passing networks allows us to determine the effect of opponent and match features on team decision-making during the game. The most significant finding from this research is that a team’s ability to control ball flow and possession is significantly affected by an opposing team’s tactical choices, such as formation or halftime changes to formation, and these changes are both able to be visualized and measurable using a graph theoretical approach.
Acknowledgements

I am incredibly grateful for the guidance of my advisor, Laurie Shaw, not only for the weekly meetings and endless suggestions for how to take this project further, but also for opening up the opportunity for me to combine a passion for sports with a love for computer science in order to develop something meaningful.

I also can’t not thank my parents for always encouraging me to take the next step so that I can always strive to turn something good into something great. And also for not yelling at me when I dented the fence from kicking a soccer ball into it one too many times.
1 Introduction 3

2 Literature Review 6
   2.1 Network Representation and Visualization 6
   2.2 Centrality 8
   2.3 Connectedness 10

3 Graph Theory and Networks 12
   3.1 Graphs 12
      3.1.1 Adjacency Matrix 13
      3.1.2 Paths and Cycles 13
      3.1.3 Subgraphs 14
      3.1.4 Random Walks 15
   3.2 Metrics 15
      3.2.1 Eigenvalues 15
      3.2.2 Centrality 16
      3.2.3 PageRank 19
      3.2.4 Clustering 20

4 Passing Network Visualizations 23
   4.1 Soccer Event Data 23
   4.2 Network Visualization Methods 24
      4.2.1 Average Player Location 25
      4.2.2 Closeness by Relationship Strength 26
      4.2.3 Fixed Location by Offensive Formation 28
      4.2.4 Fixed Formation Grouped by Role 30

5 Full Season Analysis 31
   5.1 Network Visualization 31
   5.2 Half-to-Half Network Difference 31
   5.3 Full-Network Connectedness 33
   5.4 Adjacency Heatmap 35
   5.5 Centrality Measures 36
   5.6 Highest-Betweenness Edges 38
6 Match Comparison

6.1 Visual Network Comparison
6.2 Differencing Networks
6.3 Full-Network Connectedness
6.4 Centrality Measures
6.5 Highest-Betweenness Edges
6.6 Top Triplets
6.7 Summary

7 In-Match Tactical Changes

7.1 Visual Network Comparison
7.2 Differencing Networks
7.3 Full-Network Connectedness
7.4 Centrality Measures
7.5 Highest-Betweenness Edges
7.6 Top Triplets
7.7 Summary

8 Conclusion

8.1 Future Work
Chapter 1

Introduction

With nearly 4 billion worldwide fans, soccer is the most popular sport in the world, without a doubt. Despite this, the analysis space for soccer is relatively underdeveloped and it has been difficult to put a numerical spin on individual and team performance. Soccer, unlike a number of sports that have thus far been more deeply explored in the realm of sports analytics, is heavily passing-based and contains many more complexities in the game that make it hard to decide not only how to analyze, but even what results to look for in the overall analysis.

Baseball and basketball, which are among the sports that are most heavily analyzed have clearer metrics of game-time “success” that make target-oriented analysis much easier to approach. Baseball can be split into statistics such as runs and batting averages in order to evaluate player quality. For example, one can simply divide the number of hits a player has over the number of times the player is at bat and then, if this batting average falls around .300, decide that this is an excellent hitter. Without a doubt, a team wants a better hitter! Basketball teams are small, at five players per side, and is such a high-scoring sport that each play can be evaluated as successful depending on whether it results in points scored or not. Soccer, on the other hand, is extremely low-scoring with constant changes of possession, and a team does not move the ball in a single direction over the course of the match, as teams will pass the ball backwards and sideways almost as often as they pass the ball forward toward the opposite goal. The issue at hand then becomes whether we can quantify certain aspects of a soccer team in order to characterize them in a way that allows comparison to other teams and even against themselves in different matches.

Ball movement in soccer is hugely important in determining the strength of a team, and generally (although not always), the team that maintains higher possession of the ball during the match tends to win, or is at least deemed stronger. From a statistical standpoint, there is a positive correlation between better ball possession and higher winning percentage in teams, as Parziale and Yates [18] found in a study performed on English Premier League data. Similarly, the number of completed passes completed in addition to possession can also be a significant indicator of team dominance. Incredibly well-known and historically powerful teams like Bayern
Munich, Paris Saint-Germain, and Barcelona rank at the top of a computed ranking of the “best passing teams”, indicating that team-wide ball control is an important factor in determining which teams are likely to perform well [15].

Although possession may be a decent indicator of overall team performance over the course of a season, there is incredible variance both between individual games and between halves, and there are many more factors that play into whether a team is successful or not. Simply passing around the ball for the majority of the game will not result in goals (and thus wins) or having a poor defense yet having a strong passing structure is also not useful. Teams need to be able to develop both predictive and reactive strategies by focusing on multiple features of a match, many of which hinge on the “who”, “what”, and “why” of ball movement during games. Of particular interest are passing networks as a method of characterizing the behavior of soccer teams. Soccer teams have eleven players on the field at once, all of whom must interact with each other in different ways throughout the match. Much of the activity during a soccer game can be categorized into ball movement and passing, and thus if we reduce the activity in a game into a network, or graph, then we may be able to glean further insights into the behavior of the team based on metrics that can be derived from a network using a graph-theoretical approach. Passing networks are able to capture an abundance of information about the “who” and “what” of ball movement in a game by indicating which players pass to whom and how often these relationships are utilized, so investigating the “why” will enable teams to positively adjust their own strategies both on the whole and with respect to the patterns of game play that other teams employ. By developing and exploring passing networks for individual teams, we can determine what factors and behaviors of a certain team influence the performance of not only themselves but also of their opponents.

The aim of this research is to be able to narrow down, and perhaps even model, what factors and metrics contribute to a team’s success in different matchups. By modeling a team’s behavior with respect to quantifiable metrics, it becomes easier to determine what makes a team “good” in general or at least what behaviors of one team best address the choices that another team makes in a given match. There are a handful of characteristics of soccer gameplay, such as triangles and formations, that are taught to even the youngest of players, and this paper will focus on using passing networks to build stories around how these features affect and reflect team performance. We begin by representing a team in a match using a passing network and deriving quantifiable metrics from this network. Then we compare the passing network and metrics of the same team from a different match in order to see what differences between the two matches could have caused differences in the network and metrics. Finally, we can use these comparisons to derive predictions for what kinds of different match details will cause a team to behave a certain way in order to draw deeper, useful insights about a team’s performance.

Note: While all of the data being used for this analysis is sourced from a real season of a major professional European league, all team names in this paper have been replaced, in order to preserve anonymity, with the names of childhood teams
that the author has played on.

This thesis will begin with a review of related work done in the realm of soccer analytics in Chapter 2. Then Chapter 3 will cover an explanation of graphs and networks and the metrics that we’ll be using to characterize the passing networks, as well as the context by which each of these metrics relate to soccer matches. Next, Chapter 4 will discuss different methods for visualizing a passing network in order to display as much relevant information as possible in a single image. Chapter 5 analyzes a team’s “average” network over the course of a season, Chapter 6 proceeds to lay out the methodology for comparing the passing networks of two matches against each other, and then Chapter 7 tries to explain the impact of in-match tactical shifts by comparing the passing networks from both halves of the same match. Finally, Chapter 8 wraps up with a summary of findings and potential avenues for future exploration of soccer analytics through the lens of passing networks.
Chapter 2

Literature Review

A number of papers have explored different methods and metrics for quantifying soccer team behavior via passing networks. Passing networks in particular have become a common and useful method of representing activity of soccer teams in a comprehensive and digestible manner. Additionally, the graphical representation of passing networks allows a more straightforward approach to deriving summary metrics that can allow us to characterize and compare team performance.

2.1 Network Representation and Visualization

All pursue related avenues of team analysis by first beginning with a graphical passing network representation of each team, specifically with each node representing a player on the team and each edge weight reflecting the number of passes completed between the two players on either end of the edge during a specified time period. Variations of this network design have also been explored in order to express additional or different information about a team’s configuration during a match. A paper survey conducted by Buldú et al. discusses three different comprehensive structures for passing networks: one in which every node represents a single player (as described above), one in which every node represents a location on a soccer pitch that is divided into positional regions, and a combination of the two in which every node is a player-location pair.

The first method, employed by Peña and Touchette, Gonçalves et al., and Barghi, of representing each player as its own node is the most straightforward method, as the representation is limited to 11 distinct nodes - one for each player. It is possible that this representation can have more than 11 nodes if substitutes during the match are also included as their own nodes, but the number of nodes will not far exceed the standard number of 11. Additionally, it is simple to identify the strength of particular passing relationships simply by looking at the edge connecting two players. An example of a passing network that uses this style is taken from Peña and Touchette and shown in Figure 2.1.

The second method assigns a node to each “zone” on the pitch, and edges are
Figure 2.1: Spain’s 2010 FIFA World Cup Final passing network where each node is an individual player and each edge weight corresponds to the number of passes completed between two players, taken from Figure 1 in [19].

drawn between zones, providing information about how often the ball traveled between two zones on the pitch. Cintia et al. [10] use this method to visualize where the ball is most likely to travel physically on the pitch, regardless of which players are actually involved in the movement. This provides a way of visualizing pass length, but does not necessarily provide any information about which players are involved in the interactions, and whether certain player-to-player relationships are important to the network. Figure 2.2 provides an example of such a zone-based passing network.

The third method combines both the first and the second, allowing us to visualize nodes as player-zone pairs, and making it possible to visualize not only who is passing, but also from where they are passing on the pitch. This is a visualization method set forward by Cotta et al. [11], where the pitch is split into 9 zones, meaning that there can be up to $11 \times 9 = 99$ nodes in the passing network. Figure 2.3 demonstrates how a player can cross between zones and interact with other players on the team from some zones but not others. This representation does make it difficult to summarize the overall interaction between two players since those interactions will be split into zones, and certain relationships between pairs of players that are strong may end up appearing weaker due to being split over these multiple zones.

The flexibility of a graph representation allows passing networks to express a variety of information through adjustments in the significance of nodes, edges, and weights. Chapter 4 explores a variety of visualization methods that express different information about the passing network and identifies benefits and drawbacks in terms
Figure 2.2: Argentina’s 2014 FIFA World Cup Semifinal passing network where each node is a pitch zone and each edge weight corresponds to the number of passes completed between the two zones, taken from Figure 3 in \[10\]

of expressiveness and clarity of information about team interactions and player roles.

2.2 Centrality

Varied metrics can expose different information about the performance of a team and its individual players for each match. One common category of metrics to examine within a team is “centrality” \[19\] \[6\] \[14\] \[5\], which can give insight into the role that each player plays in the behavior of a passing network. Centrality measures can be shortest-path-based or random-walk-based, depending on whether higher values should be assigned to players that are present in shorter present paths or to players that are more likely to be present in any given path between all pairs of players in the network, respectively.

Gonçalves et al. \[14\] take a look at path-based closeness and betweenness centralities of players on youth association soccer teams in order to derive information about how important each player is to ball movement in the network. The authors of this paper were able to use centrality measures, among other network details, to compare the performance of competing teams in two different matches and found that teams with high closeness and low betweenness measures were more likely to have stronger ball flow. They posited that having a few players with high betweenness indicates a team’s reliance on these few players, making it easier for an opposing team to block a handful of players and disrupt the team’s entire passing flow.

Barghi \[5\] uses PageRank as a metric for identifying team strategy through passing network analysis. Highly and strongly connected players tend to have higher PageRank values and are thus hubs of ball travel within the network - they are more
Figure 2.3: Spain’s 2010 FIFA World Cup Semifinal passing network where each node is a player-zone pair, sized by degree centrality, and colored by zone region (defense is zones 1-4, midfield is zone 5, and offense is zones 6-9), taken from Figure 2 in [11] likely to receive the ball during the game. In his analysis of the 2010 FIFA World Cup final match between Spain and the Netherlands, Barghi finds that the Dutch goalie had the highest PageRank of the entire team, indicating that the Netherlands played a more defensive game, while offensive midfield players had the highest PageRank in the Spanish team, demonstrating that the Spanish team was able to progress the ball forward towards the Dutch goal more often.

Centrality measures can thus be used as a means of characterizing team strategy by identifying strength of and reliance on particular players in the network. It is worth noting that Peña and Touchette [19] found that the players in the role of forwards on the team tend to have the lowest closeness, betweenness, and PageRank scores due to their low involvement in ball movement. Forwards are generally uninvolved until it is time to press towards the opposing goal and they receive the ball to make an advance. Using details like this, we can compare the centralities of different players both within and between networks in order to notice strategy differences as well as the relative importance of all the players on a team.
2.3 Connectedness

Soccer in particular values the presence of “triplets” - strongly connected subgroups of three players - in the passing structure of a team. Passing within a triangle shape allows each player to position himself to have two options to dish the ball to when being defended by an opponent, making it easier to maintain possession for longer periods of time. Peña and Touchette [19] explore the use of a clustering coefficient, which exposes the number and strength of triplets that each individual player is involved in. They use clustering scores as a measure of determining the connectedness of a team; by calculating the coefficient for each player, we are able to get a picture of the volume of triplets that each player is a member of. Ultimately, by analyzing two different matches between four different teams, Peña and Touchette are able to use the clustering coefficients of a team to determine that teams with high coefficients are well-connected while teams with lower coefficients tend to have trouble balancing ball movement throughout the network. In general, higher clustering coefficients across all nodes indicate more connectedness throughout the network.

Information about the team as a whole can also be diluted down into a handful of metrics, such as network strength and algebraic connectivity, which can be derived from the eigenvalues of the full network adjacency matrix and the network’s Laplacian matrix, respectively. Buldú et al. [8] use network strength and algebraic connectivity as measures of “robustness” and “cohesion”, respectively. More nodes and stronger edges (edges with heavier weights) lead to higher network strength, and therefore a high network strength value is reflective of a team that passes frequently. Higher algebraic connectivity is correlated with more edges in the network, meaning that a team that has more pairs of players connected via a passing relationship will have a higher algebraic connectivity value. In their study, Buldú et al. find that under coach Guardiola, FC Barcelona, a team very well known for its passing-heavy style of play, have network strength and algebraic connectivity values that are far higher than those of their rivals, indicating that players pass quite frequently during a match while also passing to a wide variety of teammates. Having a higher algebraic connectivity makes it harder for an opponent to defend and block a few passing relationships in order to disconnect one part of the network from another.

Many of the papers discussed in this section utilize graph theory and quantifiable metrics in order to compare teams to each other. This was done to compare two teams that played against each other in a single match as well to compare a team to its rivals within its entire league. Ideas surrounding visualization, centrality, and connectedness are vital in pursuing a graphical understanding of a team’s strategy and the roles that individual players have in a passing network. This thesis aims to build on the findings of the literature discussed here in order to compare a team to itself across a variety of matches. While it is true that two different teams may have concretely different strategies within the same match, it is also true that a single
team may change its strategy based on different match conditions. By summarizing and building stories around a team’s performance across different matches, we can observe how a single team reacts to the changes that an opposing team makes either between matches or during the game.
Chapter 3
Graph Theory and Networks

Graphs play a primary role in the analysis of soccer teams through the lens of player-to-player interactions. They provide us with the ability to compress the activity of a team during a given match into single entity that can then be manipulated and have calculations performed on it in order to derive a number of concise metrics representing its different features and qualities.

3.1 Graphs

We begin by defining a graph in the context of networks. Graphs contain nodes and edges, often represented as the sets \( V \) (i.e., vertices) and \( E \), respectively. The full graph is then given the notation \( G = (V, E) \). Each edge \( e \in E \) can be represented as \((u,v) = e\), where \( u, v \in V \) are the nodes that edge \( e \) connects. In a directed graph, we differentiate between the edges \((u,v)\) and \((v,u)\) because they have opposing directions. It is possible for a directed graph to contain an edge \((u,v)\) without containing \((v,u)\). Undirected graphs, on the other hand, do not differentiate between \((u,v)\) and \((v,u)\). There is no directional component to edges, and there can only be one connecting feature between any two nodes. The best graph type depends on the importance of direction in the application at hand.

Edges can also be weighted or unweighted. In an unweighted graph, the existence or non-existence of an edge is all that is relevant to the graph. There is no variation between the edges as entities in and of themselves. In a weighted graph, each edge can be assigned its own weight. This weight can represent anything in the context of the problem that a graph is representing. In some cases, the weight represents the “distance” between the two nodes that the edge connects. In other cases, the weight represents the “strength” of the connection.

The degree of a node \( v \) is the number of nodes that are connected to node \( v \) by an edge, regardless of edge direction. The out-degree of a node \( v \) is the number of nodes that are connected to node \( v \) via outgoing edges while the in-degree of a node \( v \) is the number of nodes that are connected to node \( v \) via incoming edges. Out-degree and in-degree are only relevant in directed graphs while degree can be applied to nodes.
in both directed and undirected graphs.

In a soccer context, we use graphs to represent passing networks. A passing network is essentially a graph where each node represents a player on the team and each edge represents some kind of relationship between the two players it connects. Typically, the edge will represent a “passing relationship", where the existence of an edge \((p_1, p_2)\) means that players \(p_1\) and \(p_2\) pass to each other in the time period represented by the network. The weight of the edge, though, can vary based on what types of details about each relationship we want to be able to visualize through the network representation. A directed edge would represent the passing relationship from player \(p_1\) to player \(p_2\) while an undirected edge indicates that there is a passing relationship between \(p_1\) and \(p_2\), which could possibly be in just one direction or both directions. The degree of a node \(p_i\) represents the number of players that player \(p_i\) directly interacts with; the out-degree would be the number of players that \(p_i\) directly passes to in the time span while the in-degree would be the number of players that directly pass to \(p_i\) within the time span of the network.

### 3.1.1 Adjacency Matrix

A graph can be represented as an adjacency matrix. Given a graph \(G = (V, E)\), we represent the node set \(V\) as an ordered set, indexing each node with its own number, meaning that \(V = \{v_1, v_2, ..., v_n\}\), where the number of nodes \(|V| = n\). The adjacency matrix \(A\) is then constructed as an \(n \times n\) matrix where each entry \(A_{ij}\) represents the edge \((v_i, v_j)\).

In an unweighted adjacency matrix, each entry \(A_{ij}\) represents whether the edge \((v_i, v_j)\) exists in the edge set \(E\). If edge \((v_i, v_j)\) is present in the graph, then \(A_{ij} = 1\), otherwise \(A_{ij} = 0\). In a weighted adjacency matrix, each entry \(A_{ij}\) represents the weight \(w(e)\) of the edge \(e = (v_i, v_j)\). If edge \((v_i, v_j) \notin E\), then \(A_{ij} = 0\), equivalent to a zero-weight edge.

In an undirected graph, the entry \(A_{ij}\) is equal to the entry \(A_{ji}\), while in a directed graph, the entry \(A_{ij}\) corresponds to the value for directed edge \((v_i, v_j)\) and the entry \(A_{ji}\) corresponds to the value for directed edge \((v_j, v_i)\).

Regular matrix functions, such as calculating eigenvalues and eigenvectors, can be executed on a graph by executing them on the graph’s adjacency matrix.

### 3.1.2 Paths and Cycles

There exists a path between nodes \(v_1\) and \(v_n\) if there exists a sequence of edges \(e_1, e_2, ..., e_{n-1} \in E\) such that for \(i = 1, 2, ..., n\), \(e_i = (v_i, v_{i+1})\) for some sequence \(v_1, v_2, ..., v_n \in V\). \(v_1, v_2, ..., v_n\) need not be distinct to form a valid path. A cycle is a path such that \(v_1 = v_n\) in the sequence of vertices that form this path. In other words, a cycle is a path that begins and ends on the same vertex.

Adjacency matrices can be used to identify the existence of paths and cycles involving a specific number of edges. For the desired number of involved edges \(d\),
we can use the unweighted adjacency matrix $A$ to compute $A^d$, the product of $A$ multiplied by itself $d$ times. The value $A^d_{ij}$ indicates how many paths with $d$ edges exist from node $v_i$ to node $v_j$. Similarly, the value $A^d_{ii}$ indicates how many cycles with $d$ edges exist involving node $v_i$. We can use the weighted adjacency matrix instead of the unweighted adjacency matrix in order to determine whether $d$-edge paths exist, but will not be able to derive meaningful information about the exact number of paths or cycles.

A graph is connected if every node is reachable from every other node in the graph. A graph is disconnected if there exists at least one node in the graph that is not reachable by every other node.

In a soccer passing network, if there exists a path between nodes $p_1$ and $p_n$, this means that there exists a set of passing relationships present in the network that connect player $p_1$ to player $p_n$. In a very literal translation, a “path” during a game of soccer between players $p_1$ and $p_n$ in a passing network would exist if there was an actual series of consecutive, un-intercepted passes starting from player $p_1$ and ending at player $p_n$. But, due to the limitations of a graph representation, we cannot visualize every single actual path taken by the ball during the match. Thus, a path from player $p_1$ to player $p_n$ in the context of a passing network indicates that there exist passing relationships, or edges, $(p_i, p_{i+1})$ for $i = 1, 2, \ldots, n$ for some sequence of players $p_1, p_2, \ldots, p_n$. In this way, paths and cycles in passing networks provide a method of succinctly summarizing player-to-player relationships while also expressing the means by which a ball could have traveled around the team in a specified match.

### 3.1.3 Subgraphs

Properties of subgraphs can provide information about individual nodes and the network as a whole. Subgraphs are subsets of the nodes and edges of the full graph. Consider a graph $G = (V, E)$. Given a subgraph that contains a subset $S_V \subseteq V$ of the nodes in the original graph, an edge $(u, v)$ can only be present in the subgraph if $u, v \in S_V$ and $(u, v) \in E$. By analyzing subgraphs - for example, triplets of nodes and their edges - we can derive the relative “importance” of specific subnetworks within the larger network.

The subnetworks that a single node is involved in can provide information about the connectedness of the node in question. Clustering, which will be discussed in Section 3.2.4, is a feature of a network that is generally centered around individual nodes. Some nodes, which are involved in more tight subnetworks, would have higher clustering scores than nodes that are only linked into the network by one or two edges. Involvement in multiple subnetworks can provide information about a node’s role in maintaining the cohesiveness of a network, giving insight into the impact on the whole network when the single node is removed. Subnetworks can provide insight into the influence that interactions among smaller subgroups of nodes within a network have on the characterization of the entire network.
3.1.4 Random Walks

Random walks are another useful tool for analyzing the structure of a network. In highly connected networks, i.e. networks that have edges present between most pairs of nodes, measures based on shortest paths between pairs are not as useful in learning about the significance of each node within a network, as the majority of pairs will be connected by a single edge, albeit perhaps with different weights. Random walks allow us to use probabilistic measures of connectedness in order to determine node importance. Each edge is assigned a probability of being selected for traversal and then, starting from any one node in the network, the next edge to traverse for the “walk” is selected according to this probability. At each node, the next edge is selected according to this probability distribution. By this methodology, random walks provide a heuristic for determining the frequency with which we would trace a path from one node to another in a network. Using this heuristic measure, we can determine the likelihood of a path between any pair of nodes \(u, v\) in a network involving a specific node \(w\) and use this to determine node \(w\)’s “importance” in the network. This idea will be further discussed in Section 3.2.2.

3.2 Metrics

While representing a passing network as a graph is the first step towards quantifying the interactions between players in a match, it is important to find metrics that can reflect qualities about both pairwise and subnetwork interactions. Several metrics exist for determining different information about the roles that nodes and edges play in a network with respect to each other. In particular, one node or edge may have more relevance to the characterization of an entire network than the rest. Each of these metrics apply to either the entire graph, individual nodes, or individual edges. For soccer in particular, we want metrics that can characterize the strength of the overall team, strength and relevance of pairwise relationships, importance of particular players, and group involvement of each player. By using the metrics that will be discussed in this section, qualitative descriptions of certain players and the team as a whole (e.g., “he is a central player on the team”) can be reduced to numbers that can be more accurately compared and analyzed.

3.2.1 Eigenvalues

Network strength represents the strength of connections within a network. If there are more and stronger edges in a network, then the network strength is higher. The largest eigenvalue, \(\lambda_1\), of the weighted adjacency matrix of a graph, is analogous with the strength of the network [1]. A network with more strong connections and nodes yields a higher \(\lambda_1\).

Algebraic connectivity measures a similar metric to network strength, reflecting how interconnected a network is. A network with more paths between pairs of
nodes will have a higher algebraic connectivity, although these connections do not necessarily need to be strong, as in network strength. The second smallest eigenvalue, $\lambda_2$, of a graph’s Laplacian matrix is analogous with the graph’s algebraic connectivity, meaning that a larger $\lambda_2$ represents a network that is tougher to disconnect. The Laplacian matrix $L$ of a network is given by $L = D - A$, where $D$ is the graph’s degree matrix (a matrix where each diagonal entry $D_{ii}$ is the degree of node $i$ and each non-diagonal entry is 0) and $A$ is the unweighted adjacency matrix of the network. Thus, the Laplacian matrix contains information both about connectivity and level of connectivity of each node, as an entry $L_{ij} = 1$ if there exists an edge $(i, j)$ and an entry $L_{ii}$ contains information about how many nodes $i$ is directly connected to by an edge. We select the second-smallest eigenvalue of the Laplacian matrix because the smallest eigenvalue will be 0 if the network is connected [1], and thus the smallest eigenvalue provides no useful quantifiable information about the network. An additional property of $\lambda_2$ is that $\lambda_2 > 0$ if and only if the graph $G$ is connected [22].

Eigenvalues can provide further insights into the behavior of a passing network as a whole. A network with more strong connections and nodes yields a higher $\lambda_1$. Since the number of nodes across networks will be constant due to the fact that every soccer team will have 11 players on the field at a time, higher values for $\lambda_1$, or network strength, imply networks with more and stronger links (passing relationships between pairs of players). The adjacency matrix used to calculate $\lambda_1$ is not normalized over the number of total passes completed during the game. Network strength is highly dependent on the volume of passes completed in a network and therefore a network with more and stronger connections should have this reflected in its $\lambda_1$. This way, network strength also becomes a measure of the volume of passes completed in a network.

A network that is “less divisible” yields a higher $\lambda_2$. This means that a team that is more easily split into two groups - a team that has weaker links between two subnetworks of players - will have a smaller $\lambda_2$. Naively, smaller $\lambda_2$ values are worse because an opposing team has fewer player-to-player relationships that it needs to concern itself with disrupting in order to break the overall connectivity of a team’s passing network. When a team’s passing mechanism is broken into two disjoint networks as such, it is possible for an opposing team to block off an entire half of a team from interacting with the ball and maintaining possession.

### 3.2.2 Centrality

Centrality measures provide indicators for how “important” an element is within a graph [16], according to different standards of measure. Both edges and nodes can be assigned centrality measures and, in general, the higher the centrality value, the more “important” the component is to the graph. Two of the most common measures of centrality are closeness and betweenness, and their most common applications are defined by shortest paths - in graphs whose edges are weighted by distance, the
shortest path between nodes $s$ and $t$ would be the path from $s$ to $t$ with the smallest sum of edge weights. Meanwhile, in graphs whose edges are weighted by strength, edge values would have to be adjusted to reflect higher strength as smaller weight instead in order to properly reflect the desirability of a “shortest path”.

Closeness centrality is a measure of how “close” a node is to other nodes in the graph. The shortest-path closeness centrality of a node $v$ inversely reflects the lengths of the shortest paths from $v$ to every other node in the graph. Shortest-path closeness centrality is formally defined as

$$c_C(v) = \frac{|V| - 1}{\sum_{t \neq v} d_G(v, t)}$$

where $|V|$ is the number of nodes in the graph and $d_G(v, t)$ is the length of the shortest path between nodes $v$ and $t$ [7].

Betweenness centrality is a measure of how often an edge or node is crossed when taking the most efficient path between all pairs of nodes. The shortest-path betweenness centrality of a node $v$ reflects the number of shortest paths between all pairs of nodes in the graph that include $v$. The formalized definition for shortest path-based betweenness centrality is defined as

$$c_B(v) = \frac{1}{(|V| - 1)(|V| - 2)} \sum_{s, t \in V} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$$

where $\sigma_{s,t}$ is the number of shortest paths between nodes $s$ and $t$, as there may be more than 1 shortest path between $s$ and $t$ if the shortest $n$ paths all have the same length. $\sigma_{s,t}(v)$ is the number of shortest paths from $s$ to $t$ that contain $v$ in the path. Betweenness centrality is a measure of the likelihood that a player is in the middle of a path between any two other players in the network.

Similarly, edge betweenness centrality of an edge $e$ reflects the number of shortest paths between all pairs of nodes in the graph that include $e$. The formalized definition for shortest path-based edge betweenness centrality is defined as

$$c_{EB}(v) = \frac{1}{(|V| - 1)(|V| - 2)} \sum_{s, t \in V} \frac{\sigma_{s,t}^{edge}(e)}{\sigma_{s,t}}$$

where $\sigma_{s,t}^{edge}(e)$ is the number of shortest paths from $s$ to $t$ that contain $e$ [7].

Random-walk betweenness, unlike shortest path betweenness, does not assume that the path taken from one node $s$ to another node $t$ will always be the shortest path. Shortest path-based centrality measures are useful in graphs for which we can be confident that the shortest path is always going to be the path taken between two nodes. In the context of a graph, a random walk from one node to another uses probabilities assigned to each edge to decide which edge to include next in the path. A random walk-based centrality measure can be used in graphs for which we are unsure which path will be taken between two nodes. These centrality measures
take into account the likelihood with which certain nodes and edges are crossed in a graph traversal from one node to another.

Random-walk betweenness applies a probability to each edge in the graph and assigns higher betweenness scores to nodes that are more likely to be crossed via random walks between all pairs of nodes. This type of betweenness is also referred to as current-flow betweenness because a graph can be represented as an “electrical network”, with a node’s betweenness rating reflecting its average throughput in the network and an edge’s “conductivity” reflecting the strength of an edge, or the probability with which an edge will be traversed. Brandes and Fleischer [7] provide the framework through which we can calculate current-flow betweenness of each node in a graph. Random-walk betweenness, or current-flow betweenness, is defined as

\[ C_{CB}(v) = \frac{1}{(|V| - 1)(|V| - 2)} \sum_{s,t \in V} \tau_{s,t}(v) \]

where \( \tau_{s,t}(v) \) is the throughput of node \( v \), defined as

\[ \tau_{s,t}(v) = \frac{1}{2} \left( -|b(v)| + \sum_{e \in e} |x(e)| \right) \]

where, in a random walk from \( s \) to \( t \), \( |x(e)| \) is approximately a measure of how likely edge \( e = (u, v) \) is to be traversed. \( b(v) \) indicates whether we are at the start, end, or elsewhere in the random walk, with \( |b(s)| = |b(t)| = 1 \) and \( |b(v)| = 0 \) if \( v \notin \{s, t\} \). In the analogy of the “electrical network” mentioned earlier, \( s \) is the network’s “current source”, \( t \) is the network’s “current sink”, \( |x(e)| \) is the “electrical current” carried over edge \( e \), and \( b(v) \) represents how much current is supplied by an “electrical node” \( v \). A player will have a high betweenness centrality if he is able to receive the ball by many strong paths from several players and then is able to similarly distribute the ball by many strong paths to several players on the team.

Random-walk edge betweenness, or current-flow edge betweenness is calculated in a similar manner, except that

\[ C_{CBE} = \frac{1}{(|V| - 1)(|V| - 2)} \sum_{s,t \in V} \tau_{s,t}(e) \]

where \( \tau_{s,t}(e) \) is the throughput of edge \( e \), defined as \( \tau_{s,t}(e) = |x(e)| \). Edge betweenness centrality in a passing network follows similar logic to node betweenness centrality, except that it is the measure of the likelihood that the interaction between two specific players is included in a passing path between any two randomly selected players. If a certain direct passing connection between two players is used frequently to get the ball between any two other players on the team, then the edge betweenness of that direct passing connection will be high.

Current-flow closeness centrality employs a similar analogy by treating nodes as points in the network with calculable “electric potentials”. In determining the
current-flow closeness centrality of a node \( s \), we treat \( s \) as the “current source” of the network and then take the inverse average of the potential difference between \( s \) and every node \( t \) in the network, treating \( t \) as the “current sink” of the network for each difference calculation. Thus, Brandes and Fleischer [7] define current-flow closeness centrality as

\[
CC_{\text{c}}(s) = \frac{1 - |V|}{\sum_{t \neq s} p_{s,t}(s) - p_{s,t}(t)}
\]

where \( p_{s,t}(v) \) is the absolute potential of node \( v \) in the network setup, \( s \) is the current source, and \( t \) is the current sink. The idea is that the “effective resistance” \( p_{s,t}(s) - p_{s,t}(t) \) is lower for two nodes \( s \) and \( t \) that have more connecting paths of higher conductance (stronger edges). In other words, \( t \) is the destination node to be reached from node \( s \) and \( p_{s,t}(s) - p_{s,t}(t) \) is a measure of the difficulty to reach \( t \) from \( s \). In a passing network, closeness centrality is a measure of the likelihood that a player starts with the ball and is able to distribute it to other players in the network via any path. A player will have a high closeness centrality if he is connected by many strong paths to several players in the team.

Ultimately, closeness centrality is a measure of the importance of a player with respect to distributing the ball around the network, betweenness centrality is a measure of the importance of a player with respect to being a conduit for ball flow throughout the network, and edge betweenness centrality is a measure of the importance of the relationship between two players to ball movement in the network.

### 3.2.3 PageRank

PageRank [17] is actually, in essence, another centrality measure, as nodes with higher PageRank are deemed “more important” to the network. Essentially, a node that is linked to by a popular node becomes more popular by relation. The strategy implemented in determining these rankings is analogous to executing a random walk through the network. The PageRank of a node \( v \) is defined as

\[
x_v = \frac{1 - \alpha}{|V|} + \alpha \sum_{u: (u,v) \in E} \frac{x_u}{d_{\text{out}}(u)}
\]

where \( d_{\text{out}}(v) \) is the out-degree of node \( v \) and \( \alpha \) is the “damping factor”. Say that, in terms of a random walk, we are currently at a node \( u \). \( \alpha \) is the probability that the next step taken in the random walk will be a node connected to \( u \) via an outgoing edge. With probability \( 1 - \alpha \), the next node is instead selected at random from the entire network. The first term in the PageRank formula for a node \( v \), \( \frac{1-\alpha}{|V|} \), represents the probability with which node \( v \) is selected at random. The second term, \( \alpha \sum_{u: (u,v) \in E} \frac{x_u}{d_{\text{out}}(u)} \), represents the probability with which node \( v \) is reached via a random walk.

In the context of passing networks, PageRank can help determine which players are popular targets for passing sequences. Ultimately, if we can determine which
players are the most popular, we can also determine which players make or, when removed through substitution or injury or blocked by the opposing team, break the effectiveness of a team’s passing network.

For PageRank in the context of soccer passing networks, we define it as given by Barghi [5]:

\[ x_j = q + p \sum_{t \rightarrow j} \frac{1}{d_{out}(t)} x_t \]

where \( p \) “is a heuristic parameter indicating the probability that a player will pass the ball away rather than keep it and go for a shot by himself” and \( q = \frac{1-p}{n} \), with \( n = 11 \), for the number of players on the team. The original PageRank algorithm includes a parameter \( \alpha \), which indicates the probability with which we select a new node in the random walk uniformly at random. In this new soccer-specific definition, \( \alpha = 1 - p \) because we want to select a new completely random player to jump to in the PageRank calculation with the same probability that any given player will not pass the ball to another player.

In terms of event data, which is discussed more in-depth in Section 4.1, we adjust the definition of \( p \) to be defined as a heuristic representing the probability with which a player will attempt to make a legal pass rather than lose the ball, take on another player with the ball, or attempt a shot. There are 7 different event types to take into account here: pass, offside pass, take on, shot missed, shot hit post, shot saved, and shot made goal. Of these 7 event types, only the “pass” event is counted as a proper pass attempt for the \( p \) heuristic. We apply this empirically determined \( p \) value towards the alpha considered in the NetworkX pagerank_numpy [3] algorithm in order to calculate the PageRank of each player in a team’s passing network.

3.2.4 Clustering

Smaller groups of nodes may be more strongly connected within their subnetwork than to other nodes in the larger graph, forming clusters. While a graph may be connected, it is possible that certain groups of nodes are more strongly connected than others. In the context of weighted graphs, this means that a subset of nodes can form a cluster within a network if they have particularly strong edges connecting them to each other. In this sense, certain groups of nodes can be “more clustered” than other groups. In soccer, the “magic number” of players involved in a passing subnetwork is 3. Triplets, usually present as on-field triangles of players, are particularly strong passing structures and thus are important to the strength of a team’s passing network and possession.

The clustering coefficient of a node \( u \) is the probability that a pair of nodes adjacent to \( u \) is also connected by an edge [16]. In other words, it is the weight of triplets of nodes that are formed with \( u \) over the total number of triplets that could possibly be formed with \( u \). The clustering coefficient [12] of node \( u \) is given as
Figure 3.1: Possible Triplet Formats Including Node $u$

follows:

$$C^W_u = \frac{(\hat{W} + \hat{W}^T)^3_{uu}}{2[d_u(d_u - 1) - 2d^{\leftrightarrow}]}$$

where $\hat{W} = (w_{uv})$, with $w_{uv}$ being the weight of the edge from node $u$ to node $v$ in the network, and

$$d^{\leftrightarrow} = \sum_{u \neq v \in V} A_{uv}A_{vu}$$

where $A$ is the unweighted adjacency matrix, meaning that each element $d^{\leftrightarrow}$ indicates how many other nodes $u$ has a reciprocal edge relationship with (a node $u$ has a reciprocal edge relationship with node $v$ if there exists an edge from $u$ to $v$ as well as an edge from $v$ to $u$), and finally $d_u$ is the total (in + out) degree of node $u$. The numerator of the clustering coefficient formula, $\frac{1}{2}(\hat{W} + \hat{W}^T)^3_{uu}$, represents the weighting of all triplets including node $u$ that are formed in the graph. The denominator of the formula $d_u(d_u - 1) - 2d^{\leftrightarrow}$ gives the total number of possible triplets that could be formed with $u$. $d_u(d_u - 1)$ gives the number of triplets that can be formed with the neighbors of node $u$ since there exist $\frac{1}{2}d_u(d_u - 1)$ possible groups of three nodes involving $u$ and two neighbors of node $u$, which can be connected to each other in two different directions. In a triplet subgraph, each node must have degree of two in order to be counted, and thus we must subtract $2d^{\leftrightarrow}$ from the denominator in order to discount triplets in which two nodes point back at each other and cause the third node to only have degree 1. Figure 3.1 provides graphical representations of all eight types of triplets that are counted towards the clustering coefficient of
Figure 3.2: Disallowed Triplet Formats Including Node $u$

node $u$ while Figure 3.2 provides a graphical representation of three-node subgraphs that would not count as a triplet.

Fagiolo [12] notes that the above definition of a clustering coefficient conflates the many effects of a cluster in a directed network and thus splits the effects into four more specific clustering coefficients. We only define one of these coefficients here – namely the cyclic clustering coefficient, defined as

$$ C_{cyc}^{u} = \frac{(\hat{W})_{uu}^{3}}{d_{u}^{in}d_{u}^{out} - d_{u}^{out}} $$

where $d_{u}^{in}$ and $d_{u}^{out}$ are the in and out degrees of node $u$, respectively. The cyclic clustering coefficient reflects the likelihood that a node will be involved in a three-node directed cycle, which is visualized in graphs (1) and (2) of Figure 3.1.
Chapter 4

Passing Network Visualizations

A passing network is a tool that we can use to visualize a team and its activity during a match. While a simple graph representation can convey certain pieces of information about a passing network such as pass frequency and players involved, it is difficult to represent any further information using a strict mathematical graph formulation. An appropriate visualization will be able to demonstrate to a reader further details about the passing network that are relevant to determining information about team dynamics. In a visualization, these qualities can be represented as functions of location, node size, edge thickness, and color. There are infinite options for what team and match details to include and exclude from a passing network, and so this chapter will cover the benefits and drawbacks of different passing network representations that we explored, all based on real event data.

4.1 Soccer Event Data

Opta [21] is a service that tags event data in sports matches. In the context of soccer, this includes passes, throw-ins, shots on goal, and interceptions, among other events. Additionally, these events are given further qualifiers; for example, passes are given qualifiers that describe whether a pass was completed, onsides, or made by throw-in, among other descriptors. For the purposes of this paper, we are concerned with pass and shot events as well as descriptors regarding completion, location, and players involved.

We have data from 25 matches played by one team in a single season. While this paper keeps the team and league names anonymous by assigning pseudonyms, it is worth noting that all of the data comes from matches played by a major team in a major European league in the 2018-19 season. Each game has been manually annotated to contain details about every single event including, but not limited to, passes, shots, substitutions, and formation changes.

A match is split into a long series of “events”, each of which is assigned a unique (within the match) event ID. Each event is assigned one of 65 types, such as “pass”, “player off”, “player on”, “start”, or “goal”. An event is also associated with one
of the two teams on the pitch, so a team ID attribute is also assigned. \( x \) and \( y \) attributes determine where on the pitch the event occurred while the outcome attribute indicates the event type-specific results of the event. For a pass event, for example, an outcome of 0 means that the pass was unsuccessful (somehow did not reach another player of the same team - perhaps via interception or kicking the ball out) while an outcome of 1 means that the pass successfully reached another member of the team. The \( \text{min} \) and \( \text{sec} \) attributes give the game clock time at which an event occurred while the period attribute indicates which game period the event occurred during, such as first half, second half, or first half extra time, among other period designations. A player ID can also indicate which player was responsible for the execution of the event.

Individual events, in addition to attributes, can have any number of “qualifiers” that provide even more details. For example, a pass event could have qualifiers indicating that the pass is a “long pass” or a “cross”, among other types of passes. A shot event could have qualifiers indicating whether it was classified as “strong” or “weak” and “rising” or “dipping”. There are dozens of qualifiers for each event type that provide very detailed descriptions of each event in a match. For the purposes of this paper, qualifiers are not of particular interest because they provide information that is too detailed for the development of passing networks and calculation of metrics.

4.2 Network Visualization Methods

There is an abundance of information present in soccer matches that can be used to develop insights into a team’s behavior in a given match or over the course of multiple matches. In order to best consolidate this information, it is important to represent as much relevant detail as possible without cluttering the representation with too many features. One of the main issues to consider when developing an appropriate graphical representation is where to position each node on the diagram. Different positioning techniques can expose different information about the individual players and the team overall, and thus play a big role in the differences between the methods of network visualization explored in this section. Additionally, for the sake of comparison between two different passing networks, pass frequencies are not normalized by the number of passes in a match, as this would make it difficult to see whether a relationship in one network is actually stronger than a relationship in another network.

Passing networks can also be time-adjusted within a match, representing activity over either chunks of time or the entire match period. Choices made in representing different periods of time within a match enable varying comparisons of the behavior of a single team over time or between matches. Halftime is a particularly significant time because teams will sometimes decide to recalibrate and adjust their formation from the first half to the second half. This choice can reflect changes - both in the
team changing its formation and the opposing team - that may not be represented in a cumulative passing network for the entire match. While this section gives examples of full match networks, we will see in future sections that the time span for a passing network can reveal different information about a team’s dynamic.

We explore four network visualization methods that expose different information about the team’s passing network in a given match. Event data from the same match is used to generate each of the graphics in this chapter.

4.2.1 Average Player Location

The first method of representation expresses the average location of each player on the pitch, the distribution of each player’s movement during the game, and the number of directed passes between each pair of players over the course of the match. An example of this representation for the main team we are analyzing in this paper, which we are calling the “Sharks”, in one of their matches is given in Figure 4.1.

![Average Player Location](image)

**Figure 4.1:** Average Player Location Representation. Each player node here is located at its average event position over the course of the game, calculated by taking the average of each player’s location when making or receiving a pass. Different colored edges indicate opposing passing directions while the shape of a node indicates the approximate directional movement of the player over the course of the match.

Each player node on the map is represented by a bivariate normal distribution, where player positions are approximated by a series of points given by a player’s location when sending or receiving a pass. This type of representation helps account
for the variability in a player’s movement on the field without reducing his position
down to the single average location at which his node appears. The actual distribu-
tion is shrunk down to a quarter of the size in order to provide a cleaner visualization
without displaying large, overlapping ellipses.

Each edge is given weight corresponding to the number of passes occurring be-
tween the two involved players. A heavier edge represents more passes made. The
red edges are those pointing from one player to another who is approximately fur-
ther back (i.e., closer to the defended goal) on the pitch and the orange edges are
those pointing in the opposite direction. The different colors exist mainly to help the
viewer distinguish between the directionally different relationships between a pair of
players.

The shape of each node reflects the direction of movement of each player, indicat-
ing which parts of the pitch each player controlled during the match represented by
the passing network. A reader will want to pay attention to thicker edges in order to
identify the most frequently used passing relationship in the network. Additionally,
players with many strong connections are of particular interest as hubs of network
activity. The density of edges displayed on this network do make this representation
difficult to process, though.

4.2.2 Closeness by Relationship Strength

From a quantitative standpoint, it is possible that the actual physical location of
each player on the pitch is not particularly relevant, so this second method of repre-
sentation ignores the actual physical location of each player on the pitch and instead
attempts to focus on sub-group interactions within the team. Using NetworkX’s
spring_layout [2], we attempted to place player nodes closer to each other if they
passed to each other more often, while moving nodes further apart for players who
did not pass to each other as frequently. This distancing is calculated via force-
directed placement, as defined by Fruchterman and Reingold [13]. The idea behind
this layout method is that each pair of nodes is assigned an attractive force defined
by edge weight and an equivalent repulsive force assigned to all nodes. In this con-
text, the larger the edge weight (in other words, the more frequently that two players
passed to each other), the more attractive the force is between the two.

Nodes are colored based on player position: the black node is the goalkeeper, red
nodes are defenders, yellow nodes are midfielders, and green nodes are strikers. The
goalkeeper is anchored at (0,0) and the other nodes are generated around it for every
network. Node sizes reflect the clustering coefficient of each player. Effectively, a
larger node indicates a player that is involved in more passing sub-networks within
the team. The thickness of each edge represents the pass distance: a thicker edge
indicates a passing relationship that involves shorter passes on average while a thinner
edge indicates the opposite. Finally, the color of the edge reflects the “laterality” of
a passing relationship. A bluer edge means that the passes made between the two
involved nodes tended to be more longitudinal on average, meaning that the passes
Figure 4.2: Closeness by Relationship Strength Representation. Different colored nodes indicate the positional role that a player has on the team, with green for offense, yellow for midfield, red for defense, and black for the goalkeeper. Location is not determined by on-pitch position of individual players, but rather by the strength of relationships between players given by passing frequency. Edge color indicates whether passes are more lateral or more longitudinal on average between a pair of players while the edge thickness reflects the average pass distance between the pair.

tended to go up and down the pitch. A redder edge means that the passes tended to be more lateral, across the width of the pitch. The laterality itself is calculated by taking the angle of the pass with respect to one side of the pitch.

This method of visualizing a passing network more intuitively expresses how “tight” relationships are between groups of players. Players who pass to each other more frequently are closer to each other, and thus makes it easier for a reader to understand that the relationships between closer players are stronger without having to decide which edges are thicker as an indication of relationship strength, as in the first representation explored. Additionally, the clustering coefficients given by node size are useful for determining which players are most integral to ball movement within the network by comparing node sizes. Edge thickness as a means of describing the average distance over which two players pass to each other provides a straightforward way for the reader to decide whether the relationship between two players is strong in terms of frequency, proximity, or both. Finally, the color of the edges gives information about whether a passing relationship is more possessive or advancing - lateral passes are generally used to help the team maintain possession.
without necessarily progressing towards the opposing goal, while longitudinal passes on average indicate offensive advancement.

Unfortunately, upon closer inspection we discovered that the functionality of \texttt{spring\_layout} was not quite as we had expected, and the actual displayed node locations actually varied in relative pair-wise distance when generating the network for the same match multiple times. Additionally, this network representation itself seems to be too information-rich, cluttering the view with many features that may not necessarily be important to the comparison of networks of the same team across different matches.

4.2.3 Fixed Location by Offensive Formation

In soccer, a team will usually organize itself into a \textit{formation} that generally dictates the locations of individual players in a team relative to each other at any point during a match. A formation will be described as a series of numbers, with each number describing the number of players organized approximately linearly starting with the line of last defense, after the goalie. For example, a 5-3-2 formation would be one with 5 defenders, 3 midfielders, and 2 forwards. Note that all the numbers add up to 10 to represent all members of the team, except the goalie.

While a team may define a specific formation at the beginning of a match, this formation may take a different form in execution, and may even change at different points throughout the match. These formation definitions and changes are taken from the data-driven match analysis approach employed by Shaw and Glickman [20]. With continuous (not event-based) tracking data for each player on the pitch, we are able to measure, at different points in the match, where each player is on the pitch relative to his teammates at all times - not just when he is controlling the ball. These relative positions are then clustered in order to identify the overall formation that a team employed during a match, or over a stretch of time during a match. This strategy allows us to see what formation a team relied on during the match, regardless of what type of formation they may have intended to use based on their positioning at the start of the game, before the whistle was blown. These defined formations are useful both from a network classification standpoint and a visualization standpoint.

The third representation method is a half-way meeting point between the first and second representations we explored. In this representation, tracking data has been used to determine the average location of each particular player with respect to the team’s approximate offensive formation. The player nodes are then fixed based on the player locations according to the given formation and the relative distance between the players is representative of their positions when the team is in possession of the ball. Like in the second representation, the size of each node reflects the clustering coefficient of each player, so a larger node indicates that a player is involved as a part of more triplets in the network. The edge weights indicate how many passes were made in total between a pair of players (in either direction) over the course of
Figure 4.3: Fixed Location by Offensive Formation Representation. For the sake of clarity, this network visualization diagram has been superimposed on a soccer pitch and player ID numbers have been removed. This helps to picture what on-field role each player has as well as the direction of play indicated by the diagram. The node on the far left is the goalie and the team is attacking the right-side goal. ID numbers are helpful for network analysis, but have been omitted here. Substitutes are on the outside of the pitch.

the match and represent the overall bidirectional strength of a pairwise relationship.

Note that Figure 4.3 contains more than 11 nodes even though a soccer team has only 11 players. This is because substitutes must be represented on the network alongside the starting 11 players in the match, and so the nodes on the fringes of the network, outside of the formation, are the substitutes that are brought in later in the match. Each substitute is lined up vertically with the player whom they substituted out during the match.

This representation allows us to visualize the information about a team’s behavior in a match that seems to be the most significant in deriving relevant information about a team’s passing network. By positioning each player by his formation role, the reader is able to derive information about the strength of a team in terms of its defense, midfield, and offense, as well as the left and right sides of the field. Edge weights allow us to compare the strength of pair-wise relationships of players while node sizes help us to determine the relative importance of individual players to the sub-team (cluster) activity within a team. With the formation structure, we can see whether groups of nearby players tended to interact in a specific way. In Figure 4.3 for example, we see that the bulk of ball movement occurs on the edges of the network. The central players have larger nodes than the rest, though not by much,
indicating that much of the clustering occurs in the center of the network, although there is overall not much clustering occurring in the network as a whole.

### 4.2.4 Fixed Formation Grouped by Role

![Figure 4.4: Fixed Location by Offensive Formation Role-Grouped Representation.](image)

As in Figure [4.3], this diagram has been superimposed on a pitch with role numbers removed. For analysis purposes, the pitch is not necessary as long as it is clear that the goalie is on the far left and the team is attacking towards the right. Additionally, labeling nodes with role numbers makes comparison and analysis easier, but labels have been omitted for clarity in this figure.

This representation method is the same as the previous one, except that each node represents a collective “role” rather than an individual player. While the majority of players in a match are not substituted, those who are substituted will have their network statistics combined with the player whom they substituted out instead of being assigned their own node. This way, we are characterizing the team truly by its formation and the role that each position plays in a team’s performance in a match. Figure [4.4] takes the substitutes from the edge of the network in [4.3] and combines them into their respective role nodes.

This is the visualization method utilized for the remainder of this thesis as it provides enough positional information about the players grouped by their individual roles on the team via the formation while also providing information about each player’s distributive role in the network via the nodes sized by clustering coefficient. While directional relationships are lost in this representation, we are more clearly able to see the mutual strength of relationships within the network by looking at edge thickness.
Chapter 5

Full Season Analysis

We begin by generating “average” season metrics for the team we are looking at - the Sharks - as well as an “average” season passing network. Analyzing the team’s performance on average will provide a basis for contextualizing findings in other network analysis steps. Additionally, we will look at metrics over the course of the season so that any outliers or trends may stand out. By generating an average team network, it is possible to average out the effects of outlying properties of individual match passing networks. We can then generate a description of the team’s average strategy over the course of the 25-game season using the metrics and visualizations generated.

5.1 Network Visualization

In order to generate an average network, for each pair of player roles \((p_1, p_2)\), we average the number of passes from \(p_1\) to \(p_2\) over the course of the 25 games and assign that average pass number to the edge \((p_1, p_2)\). We then use this new network to generate a visualization and match metrics.

We notice in Figure 5.1 that the two center midfielders (players 4 and 8) on the Sharks have the highest clustering coefficients (see Section 3.2.4), as indicated by the size of their nodes. The strongest passing relationships are among the defenders (players 3, 6, 5, and 2), moving the ball along the outside of the network, up and down the sides of the pitch. The edges on the inside of the network passing through the center midfielders (players 4 and 8) are also relatively strong and, in addition to their high clustering coefficients, indicate that they are primary ball movers due to the high number of heavy edges that they are connected to.

5.2 Half-to-Half Network Difference

We can generate a difference network by subtracting the features of one network from the other. Let us say that we are comparing two networks: network A and
**Figure 5.1**: Average Full Match Network. This network visualization is to be read left to right: the team goalie is player 1 and the team attacks towards the right. Defenders are players 3, 6, 5, and 2; midfielders are players 11, 8, 4, and 7; forwards are players 9 and 10. Node size is reflective of the clustering coefficient of the player and edge thickness represents the number of passes completed between a pair of players, in both directions.

network B. We can generate two opposing difference networks, A - B or B - A, which reflect identical information about the differences between both networks. To generate network A - B, we begin by subtracting the clustering coefficient of each node in network B from the corresponding nodes in network A and then assign the absolute value of these differences to each node in the difference network A - B. If the clustering coefficient of the node in network A is larger, then we color the difference node blue, otherwise we color the difference node orange. We then repeat this process for the individual edges of the network, using passing weights instead of clustering coefficients as the values to subtract. We use this methodology on the individual halves of the Sharks’ average passing network to produce a half-to-half difference network in Figure 5.2.

By looking at the difference network between the two average half networks for Sharks, we find that every player except the two strikers (roles 9 and 10) have higher clustering measures in the first half than in the second, because their nodes are blue. Additionally, the popularity of passes involving these two players also increases in the second half while central, tight passes involving midfielders and defenders are more common in the first half. There almost appears to be a “separation” in edge colors from back to front - there are more and stronger blue edges in the back of the network while orange edges are more prevalent in the front of the network. This
Figure 5.2: Difference Network Between Average Half 1 and Half 2. In this difference network, the node size reflects the *difference* in clustering coefficient between the first half and the second half. A larger node indicates a larger difference. Similarly, the thickness of an edge \((p_1, p_2)\) reflects the *difference* in the number of passes completed between \(p_1\) and \(p_2\) between the two halves. Blue features indicate that the value if higher in half 1 while orange features indicate that the value is higher in half 2.

indicates an average increase in offensive activity in the second half, likely due to the team deciding to implement more of an offensive strategy as the game clock runs down.

It is worth noting that the node sizes are incredibly small in this network, meaning that clustering has barely changed from the first half to the second for any player in the team, reflecting that the reliance on individual players to move the ball around the network is approximately same from the first half to the second. Consistent with the change in passing frequency from the back to the front of the network, some of the largest differences in clustering from the first half to the second half, as shown in Table 5.1, are in defending players such as 2 and 5, although curiously there is also a large difference in the clustering for player 11, who is an offensive midfielder.

### 5.3 Full-Network Connectedness

There are two details about connectivity measures over the course of the average match that are important in understanding the usefulness of network strength, \(\lambda_1\), and algebraic connectivity, \(\lambda_2\) (see Section 3.2.1). As shown in Table 5.2, \(\lambda_1\) for the full match is higher than \(\lambda_1\) for both the first half and the second half. Since
network strength is reliant on the number of passes completed, the higher network strength for the full match indicates that the full match has more strong connections throughout the entire network compared to just the first half or just the second half - a result of the full network covering more passing events than just a single half. Additionally, this average passing network has a higher network strength in the first half than in the second, indicating that Sharks generally passed more frequently and across wider ranges of the network in the first half than in the second. This could theoretically be linked to player fatigue, as players getting tired over the course of the match generally (at least from qualitative match observations), leads to fewer completed passes and less player-to-player ball movement.

Table 5.2 also conveys useful information about the difference in algebraic connectivity ($\lambda_2$) between the full match and each half. Note that $\lambda_2$ for the full average match is much larger than for each half, but that $\lambda_2$ is approximately the same between halves. A higher algebraic connectivity is indicative of a team that is able to distribute the ball around the network via a variety of different passing paths, so it is understandable that the full match has a higher $\lambda_2$ than each half because the full match is made up of more passes, and thus likely has more unique and strong connections than each half. The algebraic connectivity being approximately the same between halves is an indication that there was no significant change in the number of channels of distribution in the network. While the same number of connections may have been utilized in each half, it is possible that the connections that were heavily traversed - or even present - changed from the first half to the second. This would
lead to a larger variety of connections in the full network, making the full network much more difficult to disconnect, and thus leading to the significantly larger $\lambda_2$ observed in Table 5.2.

While useful to look at in the context of this single match, network strength and algebraic connectivity are more interesting in the context of comparing one network to another, as will be shown in Chapter 6. The values of $\lambda_1$ and $\lambda_2$ explored here will prove useful if one chooses to compare the Sharks’ performance in one match against their average performance over the course of the entire season.

5.4 Adjacency Heatmap

Figure 5.3: Average Full Match Adjacency Heatmap. Each box indicates the relative strength of a passing relationship from the player on the y-axis to the player on the x-axis. Lighter colors indicate more passes.

Since every role-based network will have the same number of nodes, we can use the adjacency matrix of a passing network and turn it into an 11 by 11 heatmap, where each point is colored based on the intensity of the passing relationship between the two players represented. While not particularly useful for drawing any concrete insights into the dynamics of a team, it is an interesting method of visualizing the relative strength of directional relationships within a passing network. We notice in Figure 5.3 that the connection from player 5 to player 6, given by the yellowest box in the map, is the strongest one in the network. Connections from player 2 to
player 5, player 6 to player 5, player 5 to player 2, and player 6 to player 3 are also particularly noticeable in this average match network. Looking back at Figure 5.1 we see that these player all correspond to the defenders on the team and correlate with the visible strength of edges in the network visualization.

The average adjacency heatmap is included in this section as a demonstration of the variety of methods for visualizing a team’s passing network, but will not be used as a method of network comparison in later chapters.

### 5.5 Centrality Measures

<table>
<thead>
<tr>
<th></th>
<th>Closeness</th>
<th>Betweenness</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.29</td>
<td>0.25</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>2.41</td>
<td>0.46</td>
<td>0.09</td>
</tr>
<tr>
<td>3</td>
<td>2.23</td>
<td>0.40</td>
<td>0.09</td>
</tr>
<tr>
<td>4</td>
<td>2.38</td>
<td>0.42</td>
<td><strong>0.09</strong></td>
</tr>
<tr>
<td>5</td>
<td><strong>2.43</strong></td>
<td><strong>0.46</strong></td>
<td>0.09</td>
</tr>
<tr>
<td>6</td>
<td>2.35</td>
<td>0.43</td>
<td>0.08</td>
</tr>
<tr>
<td>7</td>
<td>1.96</td>
<td>0.28</td>
<td><strong>0.09</strong></td>
</tr>
<tr>
<td>8</td>
<td><strong>2.44</strong></td>
<td><strong>0.45</strong></td>
<td><strong>0.09</strong></td>
</tr>
<tr>
<td>9</td>
<td>1.65</td>
<td>0.23</td>
<td>0.08</td>
</tr>
<tr>
<td>10</td>
<td>1.57</td>
<td>0.22</td>
<td>0.08</td>
</tr>
<tr>
<td>11</td>
<td>1.97</td>
<td>0.30</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 5.3: Average Centrality Measures

Centrality measures give information about the “most important” players in a passing network (see Sections 3.2.2 and 3.2.3). Identifying the nodes with the highest centrality measures helps us to identify which players were most integral to contributing to passing paths and directing ball flow in the network. In Table 5.3 we notice that player 8 has one of the highest centrality measures across closeness, betweenness, and PageRank, suggesting that it is likely the player that is most involved with ball movement in the network.

When we order the nodes of a network by their centrality measures as in Table 5.4 we obtain information about the relative importance of all the players on the team. For instance, we find in the Sharks’ average network, players 8, 2, 5, 4, 3, and 6 (defenders and center midfielders) tend to top out the centrality rankings while the offensive players and the goalkeeper are generally, as expected, less involved in ball distribution. Unfortunately, using the high-to-low ordering as a summary of a network’s relative centralities does not capture the magnitude by which the centralities of two nodes may differ within the network. It is possible that two nodes have nearly identical centralities, but the fact that player A has a marginally larger centrality than player B places player A higher in the ordering than player B. If
we instead plot these centralities on a scatter plot, we are able to identify patterns among players based on the relative magnitudes of their centralities.

The first feature to notice about the plot in Figure 5.4 is that the relationship between closeness and betweenness is approximately linear. Players with high betweenness will generally also have high closeness - a player that is frequently involved in passing paths between random pairs of players (high betweenness) will likely also have a strong ability to distribute the ball to multiple players around the network (high closeness). The one player that seems to stray from this pattern is player 1 - the goalkeeper - as he has a higher than linear expectation for betweenness. This outlier makes sense, though, since a goalie does not have the ability to pass the ball to a lot of players (low closeness), but will sometimes receive the ball from and distribute the ball to defenders when the team is in a defensive position (higher betweenness).

Additionally, we look at clusters of players that have similar centralities. These clusters convey some information about the level of involvement that certain players have in a network and how players are grouped by similar levels of involvement in a network. There appear to be three (or four, if we count the goalkeeper) distinct clusters of players: \((2, 3, 4, 5, 6, 8), (7, 11),\) and \((9, 10)\). The first cluster corresponds to the defenders and center midfielders on the team. As expected, these players have the most important roles in the network with respect to ball distribution and are all highly involved in passing paths and in delivery of the ball to all parts of the network. The second cluster \((7, 11)\) corresponds to the wide, offensive midfielders who are generally responsible for moving the ball from the midfield to an attacking position for the team, but are not quite as involved in the distribution of the ball to all corners of the network. The final, lowest cluster \((9, 10)\) corresponds to the team strikers. Corroborating the findings of Peña and Touchette [19], the forwards of the Sharks are the least involved in ball distribution, as their role is usually to wait to receive the ball and then advance towards the goal. They are not particularly involved in passing the ball around to other players on the team.

<table>
<thead>
<tr>
<th>Closeness</th>
<th>Betweenness</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.4: Ordering of Average Centrality Measures
5.6 Highest-Betweenness Edges

Section 5.5 discussed the importance that individual players had with respect to ball distribution, but only looking at each individual player ignores the relative importance that certain pairwise relationships may have within the passing network. Recall that edge betweenness is a measure for how likely a passing relationship is to be used in order to distribute the ball between any two players in the network. Both Table 5.5 and Figure 5.5 give us a picture of which relationships are most integral to the distribution of the ball throughout the network by listing and highlighting the edges with the top 10 edge betweenness scores in the average Sharks match. Notice that none of these edges involve the two strikers (players 9 and 10) because passing relationships involving forwards are generally used at the end of a team’s possession in an attempt to advance the ball up the field for a shot at the goal. We find that the edges connecting the defenders are among those with the highest betweenness, but only those that connect a defender to his most proximal defensive teammate. This indicates a defensive passing strategy that on average involves passing to the nearest teammate rather than attempting to cross the ball across the field.

Another curious detail about the top 10 edge list is that no edge involves either of the center midfielders (players 4 and 8) until we reach the tenth-highest edge betweenness value. We may have deigned to predict that players 4 and 8, given their
Table 5.5: Top 10 Edge Betweenness Scores for Average Network

<table>
<thead>
<tr>
<th>#</th>
<th>Edge</th>
<th>Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(5, 6)</td>
<td>$8.8 \times 10^{-4}$</td>
</tr>
<tr>
<td>2</td>
<td>(6, 3)</td>
<td>$7.5 \times 10^{-4}$</td>
</tr>
<tr>
<td>3</td>
<td>(2, 5)</td>
<td>$7.1 \times 10^{-4}$</td>
</tr>
<tr>
<td>4</td>
<td>(5, 2)</td>
<td>$7.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>5</td>
<td>(6, 5)</td>
<td>$6.9 \times 10^{-4}$</td>
</tr>
<tr>
<td>6</td>
<td>(7, 2)</td>
<td>$6.5 \times 10^{-4}$</td>
</tr>
<tr>
<td>7</td>
<td>(5, 1)</td>
<td>$6.3 \times 10^{-4}$</td>
</tr>
<tr>
<td>8</td>
<td>(2, 7)</td>
<td>$6.3 \times 10^{-4}$</td>
</tr>
<tr>
<td>9</td>
<td>(3, 6)</td>
<td>$5.4 \times 10^{-4}$</td>
</tr>
<tr>
<td>10</td>
<td>(3, 8)</td>
<td>$5.3 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

high clustering and individual centrality scores, would have also been involved with some high-betweenness edges. The evidence to the contrary indicates that players 4 and 8 are the most important players in the network in terms of distributing the ball to all players in the network, but don’t particularly rely on single relationships to make this distribution happen. The edges that are the most “important” are situated on the edges of the network, connecting the defenders and outside midfielders (particularly player 7), as progressing the ball via outer players is a common strategy for offensive playmaking in soccer. An opposing team would seek to disrupt some of these high-betweenness relationships in order to prevent the Sharks from using their usual strategy to get the ball from one player to any other player on the team.

5.7 Top Triplets

Triangles are a basic structural component of soccer strategy and are thus integral to understanding a team’s ball control strategy. As the strongest structures in a soccer team, triplets can indicate among which players the strength of a team lies during a match. By identifying the most popular triplets in a passing network, we are able to learn both about where the ball spends the most time during a team’s possession as well as which groups of players are able to keep the ball for longer during a match.

A “triplet” in a match is found by using the event data to identify sequences of consecutive, completed passes that involve exactly three players. The “popularity” of a triplet is determined by summing up the number of passes that are completed among these three players, only counting passes if they are part of a consecutive sequence of passes involving only and all of these three players. A consecutive sequence of passes in considered “complete” if a player loses the ball, a pass is intercepted, or a shot is taken.

In the average network, we select the “top triplets” by taking the top 5 triplets from each contributing match and finding which triplets appear most often in the
Figure 5.5: Average Network With Highest Betweenness Edges. The edges highlighted in green are those with the top 10 edge betweenness scores in this network. The brightest edge corresponds to the edge with the highest score while the darkest color corresponds to the edge with the 10th highest score. The top 10 edges are directionally aware and thus if both edge $(p_1, p_2)$ and edge $(p_1, p_2)$ are in the top 10, it will appear as a single edge in this network diagram, highlighted with the brighter of the two colors.

Table 5.6 lists the top triplets on average for the Sharks. Notice that even the most popular triplets (3, 5, 6) and (2, 5, 6) appear among the top 5 triplets in little over half of the matches played by the Sharks over the course of their season. Given the lack of a set of triplets that are popular in a large majority matches, we may find triplets to be a particularly useful feature in identifying differences between networks, as we will seek to do in Chapter 6.

The passing network on its own loses information about individual events and their order, so it is valuable to extract triplets and place them in the context of the passing network by displaying them on top of the visualized team formation, as in Figure 5.6. Considering that the basis of the clustering coefficient is groups of three players, it comes as no surprise that the players involved in the most triplets are those with the highest clustering coefficients (largest nodes). We can see in Figure 5.6 that the top triplets are all completed amongst players who are positioned closest to each other on the pitch. Thus, we conclude that the strongest triplets are local, as players are intuitively able to complete longer series of passes with neighboring players, since longer passes cover more distance and are more likely to be intercepted by opposing players.
<table>
<thead>
<tr>
<th>#</th>
<th>Average</th>
<th>Season Appearances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(3, 5, 6)</td>
<td>15/25</td>
</tr>
<tr>
<td>2</td>
<td>(2, 5, 6)</td>
<td>15/25</td>
</tr>
<tr>
<td>3</td>
<td>(2, 4, 5)</td>
<td>11/25</td>
</tr>
<tr>
<td>4</td>
<td>(5, 6, 8)</td>
<td>9/25</td>
</tr>
<tr>
<td>5</td>
<td>(2, 5, 8)</td>
<td>9/25</td>
</tr>
</tbody>
</table>

Table 5.6: Top 5 Triplets on Average. Of the top 5 triplets in each match played by the Sharks this season, the triplets listed here appeared in the top 5 most often. The appearances column indicates in how many matches out of the 25 matches played in the season did each of these triplets appeared in the top 5.

5.8 “Time-Series” Metrics

Before proceeding to the comparison of two passing networks, it is a good exercise to look at the metrics for individual matches over the course of the season. Plotting these metrics out as a “time-series” over the course of the season will help in identifying outlying features of specific matches. We place “time-series” in quotes because we are not actually looking at these metrics as a function of time and are not searching for match trends over time. That being said, the matches in the plots in Figure 5.7 are presented in chronological order. Figure 5.7 presents average betweenness and average closeness - calculated by averaging betweenness and closeness, respectively, of all players in the network - alongside network strength and algebraic connectivity over the course of the Sharks’ season.

First, we note that average closeness and network strength in Figure 5.7 mirror each other very closely. This is expected, as both metrics measure similar network qualities. Network strength measures a team’s ability to distribute the ball around the network and delivers value to matches with stronger connections. Closeness of a player reflects how well the player is able to distribute the ball to others around the network, so averaging the closeness of all players will also measure a team’s overall ball flow.

Figure 5.7 also exposes that the average team network has a very low average betweenness compared to individual matches, as shown by the light blue line in the Average Betweenness time series plot. This is a function of the average team network having more connections than a typical match network due to all existent connections from every match being averaged into one network, though these connections may not necessarily be stronger than those found in any given match. Recall that low average betweenness in a network is a sign of a team having good ball flow [14]. When there are more connections overall to utilize, there is less reliance on individual players to receive and distribute the ball throughout the network, and therefore players have lower betweenness on average.

This makes Match 22 particularly interesting because it is the only match in Figure 5.7 that has a lower average betweenness than the average match. As stated
Figure 5.6: Average Full Network Most Popular Triplets. The Sharks’ top 5 triplets from the season are marked on top of the average team formation. The most popular triplet is highlighted in the brightest shade of green and the fifth most popular triplet is shaded in the darkest shade of green.

by Gonçalves et al. [14], low average betweenness and high average closeness are indicators of good ball flow and thus it is unsurprising that Match 22 also has higher-than-average average closeness and one of the highest network strengths in the season. Algebraic connectivity is also higher than average, although it is not among the highest recorded in the season, suggesting that while these metrics may be related, they are not necessarily correlated. This match ended up in a 1-1 tie, although this information is not necessarily relevant for appropriate match analysis.

Match 24 is also of note as a match with among the lowest average closeness, network strength, and algebraic connectivity and among the highest average betweenness. High closeness, network strength, and algebraic connectivity and low betweenness are desirable, so a match such as this one would likely generate a passing network representative of poor ball movement and heavy reliance on very few players on the team. Interestingly enough, this match also ended up in a tie (2-2), thus exemplifying the necessity for metrics outside of final score to characterize the performance of a team.
Figure 5.7: “Time Series” of Network Metrics. These time series plots show average network closeness, average network betweenness, network strength, and algebraic connectivity over the course of the Sharks’ season. The light blue line in each plot represents the corresponding value for the average network referenced earlier in this chapter. Matches are presented in chronological order on the x-axis. Matches 1 and 17 are circled on the x-axis because they will be analyzed and compared in Chapter 6, and these average network values may be of interest for back-reference.
Chapter 6

Match Comparison

We proceed by comparing the passing networks for the same team from two different matches in order to identify and explain any performance differences. Given the emphasis on both numerical metrics and graphical visualization techniques, there are two viable methods of interpreting similarity between passing networks. Comparisons can be made by relating numbers (as in metrics) to one another, interpreting characteristics based on whether values are less than, greater than, or equal to the others. Visual comparisons and interpretations can also be made if we think about the networks in the context of a soccer match and its events.

<table>
<thead>
<tr>
<th>Match #</th>
<th>17</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>6-1</td>
<td>1-2</td>
</tr>
<tr>
<td>Home/Away</td>
<td>Away</td>
<td>Home</td>
</tr>
<tr>
<td>Opponent</td>
<td>Peanuts</td>
<td>Peanuts</td>
</tr>
<tr>
<td>Opposing Formation</td>
<td>4-4-2</td>
<td>5-4-1</td>
</tr>
</tbody>
</table>

Table 6.1: Details About Match 17 and Match 1. Formations are determined via methods and classifications put forward in [20].

In this chapter, we have selected two matches in which some features are controlled. The Sharks play both Match 17 and Match 1 against the same opposing team, the Peanuts. Some details about these two matches are given in Table 6.1. Particularly of note is the difference in opposing formation: in Match 17, the Peanuts use a 4-4-2 formation (4 defenders, 4 midfielders, and 2 strikers) while in Match 1, the Peanuts use a 5-4-1 formation (5 defenders, 4 midfielders, and 1 striker). The Peanuts use each of these formations for the duration of the match, and don’t make any changes at halftime. Impacts of opponent changes made at halftime will be explored in Chapter 7.


### 6.1 Visual Network Comparison

The most straightforward method of comparing two networks is to generate role-based representations and observe the visual differences.

By looking at the two diagrams for Match 17 and Match 1 in Figure 7.8 side-by-side, we observe that the nodes, which are sized by their respective clustering coefficients (see Section 3.2.4), in Match 17 are on average larger than the nodes in Match 1, meaning that the players in Match 17 cluster more (i.e., participate in more passing triplets) than the players in Match 1. In addition, generally thicker edges appear in Match 17 to indicate overall stronger passing relationships between more pairs of players. From this, we may predict that further metrics would indicate that the Sharks are able to distribute the ball around the entire network better in Match 17 than in Match 1. The thickest edges, and thus the most frequently used passing relationships, appear evenly distributed along the full center-back (defense and center midfield) of the network in Match 17 while the strongest edges in Match 1 are more skewed towards the left of the network, perhaps suggesting that the Sharks were disrupted by the opposing team in some way on the right side of their network in Match 1.

### 6.2 Differencing Networks

While side-by-side comparison of networks is useful, we can also condense this information into one difference network via the differencing method described in Section 5.2. For the full-match passing networks of the Sharks in matches 17 and 1, the difference network 17 - 1 in Figure 6.2 gives information about the differences in player and edge strengths between the two matches. Recall that blue nodes indicate larger clustering coefficients in Match 17 while orange nodes indicate larger clustering coefficients in Match 1. Similarly, blue edges indicate higher pass frequency between the two involved players in Match 17 while orange edges indicate higher pass frequency in Match 1. Finally, the sizes of nodes and thickness of edges represent the absolute value of the difference in clustering coefficients and pass frequencies, respectively, between the two networks.

We can infer a number of details from this difference network. The majority of the nodes are blue and large, meaning that the players in Match 17 have much higher clustering coefficients than the players in Match 1. The two orange nodes - players 1 and 9 - are also small, indicating that although these two roles may have higher clustering coefficients in Match 1, they are not larger by much. From these details about the clustering coefficients, we find that the players in Match 17 were more prone to forming passing relationships with other pairs of players, building cluster relationships that lead to better ball distribution during the match.

Perhaps more interesting is the selection of edges that are blue and orange in the difference network. It appears that all edges involving player 8 (a center midfielder) are among the passing relationships that are significantly more utilized in Match
Figure 6.1: A side-by-side comparison of the passing networks for matches 17 and 1 allows us to see that the clustering coefficients (node sizes) of the members of the Sharks are on average larger in Match 17 than in Match 1 and that the connections in Match 17 are generally stronger than those in Match 1. Node sizes and edge thicknesses are not normalized within the network and thus can be compared across networks.
Figure 6.2: Difference Network Between Matches 17 and 1. Blue nodes and edges correspond to larger clustering coefficients and pass frequencies, respectively, in Match 17, while orange nodes and edges indicate larger clustering coefficients and pass frequencies, respectively, in Match 1. All relationships among the central defenders and midfielders (players 5, 6, 4, and 8) are also more popular in Match 17 than in Match 1. The relationship between players 4 and 3 is the edge with the most noticeably higher strength in Match 1 than in Match 17, and it crosses over player 8 in order to be completed. The fact that player 8 has the largest clustering difference from Match 17 to Match 1 and is also passed over in a significant connection in Match 1 indicates a major difference in the role that player 8 plays in both matches. Perhaps this is attributable to the Peanuts’ formation difference between matches, as they have an additional defender in Match 1 that could contribute towards disrupting more connections in the center of the network, maybe by defending player 8 more closely.

6.3 Full-Network Connectedness

The connectedness of two passing networks can be compared numerically by comparing the network strengths and algebraic connectivities (see Section 3.2.1). Since every soccer team has 11 players on the field at once, we know that every single network that we are comparing will have the same number of nodes. This means that if network A has a larger network strength value than network B, network A has more and stronger connections than network B rather than more nodes.

Similarly, if network A has a larger algebraic connectivity than network B, net-
work A is more difficult to break into disconnected subnetworks than network B. If network A were to have more nodes than network B, then network A would require more strong edges to yield the same algebraic connectivity as network B. Since we are dealing with two networks with the same number of nodes, both networks have the same number of possible edges, and a higher algebraic connectivity in this case is synonymous with more edges connecting all of the nodes together.

<table>
<thead>
<tr>
<th>Match #</th>
<th>17</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Strength ($\lambda_1$)</td>
<td>74.4560</td>
<td>49.4352</td>
</tr>
<tr>
<td>Algebraic Connectivity ($\tilde{\lambda}_2$)</td>
<td>27.8472</td>
<td>23.3545</td>
</tr>
</tbody>
</table>

**Table 6.2: Network Connectedness Measures for Matches 17 and 1**

We find from the network connectedness values in Table 6.2 that the match that we had visually inspected in Section 6.1 to have stronger passing connections (Match 17) also has higher network strength and algebraic connectivity. The implication of this finding is that the behavior and setup of the Sharks in Match 17 yields a network that is both denser and harder to break up. The higher network strength means that more pairs of players have stronger connections (i.e., pass to each other more often) in Match 17 than in Match 1. The higher algebraic connectivity is interesting because it indicates that there are less opportunities for the opposing team to disrupt the connectedness of the entire network simply by interrupting individual passing connections. A lower algebraic connectivity translates to it being easier for the opposing team to block individual connections between players in order to make it highly unlikely for one subnetwork of the team to reach the other players in the team via any passing paths.

### 6.4 Centrality Measures

In Section 5.5, we plotted the betweenness and centrality measures (see Section 3.2.2) of each player against each other to identify role clustering patterns and noticed the approximately linear relationship between betweenness and closeness. When comparing two matches against each other, we instead plot closenesses of one match against closenesses of the other match and plot betweennesses of one one match against betweennesses of the other match in order to see how role importances relate to each other across networks.

The plot of Match 17 betweenness vs Match 1 betweenness in Figure 6.3 reveals that a majority of players have higher betweenness in Match 1 than in Match 17. Recall from [14] that higher betweenness values for players in a network are an indicator of poor ball movement while higher closeness values indicate the opposite. Curiously, the Match 17 closeness vs Match 1 closeness in Figure 6.4 shows that many of the players have very similar, yet still higher closeness in Match 1 than in Match 17 - a similar finding to the betweenness comparison. This detail is an
Figure 6.3: Match 17 vs Match 1 Betweenness. Green points indicate offensive players, yellow points midfielders, and red points defenders; the blue point is the goalkeeper. The red dashed line demarcates equivalence in betweenness between both matches.

indication that, contrary to assumptions made about network distributivity using the clustering coefficients from Section 6.1, most of the players in Match 1 were able to distribute the ball throughout the network better than the players in Match 17.

Of note are player 5 (a center defender) and player 8 (a center midfielder), who have significantly higher betweenness and closeness in Match 17. The central positioning of these players in the network suggests that the Sharks’ ability to distribute the ball around the network is more center-balanced in Match 17 than in Match 1. The main difference between the two matches was the formation of the opposing team, the Peanuts, which may be able to explain the change in centrality balance within the network. In Match 17, the Peanuts use a 4-4-2 formation, employing 4 defenders to attempt to intercept and block connections within the Sharks’ network. Meanwhile, in Match 1, the Peanuts use a 5-4-1 formation, enlisting an additional defender to disrupt the Sharks’ network. By inserting an additional player into the center of their network, the Peanuts appear to have been able to shift the balance of the Sharks’ ball distribution from the center of the network in Match 17 to the outside of the network in Match 1.
Certain player-to-player relationships may define a team’s performance during a match, and edge betweenness centrality provides a metric for indicating which relationships are the most “important” in a match by highlighting which edges are the most likely to be traversed in order to get the ball between any two points in the network. We look at the top 5 edges in order of highest betweenness.

In order to understand the role that the edges with the top 10 highest betweennesses from Table 6.3 may play in the network, it helps to visualize high-betweenness edges on the network itself. The top ten edges from each match are highlighted in Figure 6.5, with the edge with the highest betweenness highlighted in the lightest shade of green and the edge with the lowest betweenness of the top ten highlighted in the darkest shade of green. Both networks have a clear reliance on passing relationships between defensive players in the networks, as the edges connecting players 3, 6, 5, and 2 are present in the top 10 list for both Match 17 and Match 1. While Match 17 appears to rely on passing relationships in the center of the network (i.e., between players 5, 6, and 8), there are no edges in the top ten list for Match 1 that connect the central players 4, 5, 6, and 8 to each other outside of edge (5, 6). Match 17 clearly has players distributing the ball through the center of the network while Match 1 reveals players focusing ball distribution on the outsides of the network, a prediction alluded to through a discussion of centrality measures in Section 6.4.
Figure 6.5: The most important edges in Match 17 tend to go through and connect central players while the most important edges in Match 1 connect players along the outer edges of the network. Each of these diagrams highlight the 10 edges with the highest edge betweenness measures for their respective networks, where the lightest shade of green indicates the highest betweenness value.
is possible that the Peanuts in Match 1 were able to defend central connections in the Sharks’ network, forcing the Sharks to use outer relationships to distribute the ball around the network.

### 6.6 Top Triplets

Triplets are the strongest structures involved in a team maintaining possession, so we can take a look at the top 5 triplets in each match to compare the areas of possession in the team’s passing network. As discussed in Section 5.7, the strength of a triplet is determined as the number of consecutive passes that are completed between a group of three players within a match. The ball must consecutively (event-wise) be passed from one player to another until all three players have touched the ball (either receiving or delivering a pass) at least once for the pass number to count towards the triplet strength. The triplets are then ranked in order of total popularity.

<table>
<thead>
<tr>
<th>#</th>
<th>Match 17</th>
<th>Match 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(5, 6, 8)</td>
<td>(3, 5, 6)</td>
</tr>
<tr>
<td>2</td>
<td>(4, 5, 6)</td>
<td>(2, 5, 6)</td>
</tr>
<tr>
<td>3</td>
<td>(3, 5, 6)</td>
<td>(3, 4, 11)</td>
</tr>
<tr>
<td>4</td>
<td>(3, 8, 11)</td>
<td>(3, 4, 6)</td>
</tr>
<tr>
<td>5</td>
<td>(2, 4, 5)</td>
<td>(3, 6, 8)</td>
</tr>
</tbody>
</table>

**Table 6.4:** Top 5 Triplets for Matches 17 and 1. Triplets are ranked in descending order of popularity within each match.

By looking at Table 6.4, we see that many of the same players (e.g. players 2,
5, 6, and 8) are involved in the top triplets in both Match 17 and Match 1. Given this information, it is also important to consider where these triplets occur on the network and with what others each players is interacting with the most. We can derive more comparative information by looking at a visual representation of these triplets placed atop a formation.

In Figure 6.6, we notice that Match 17 finds that possession power is centralized in the center of the network with players 5, 6, and 8. Additionally, strong triplets are composed of players that are near each other physically in the formation, but Match 1 finds that even the strongest triplets are composed of players farther apart in the network; most of the triplets in Match 1 connect players who are not closest to each other positionally. For example, triplet (3, 4, 11) crosses over player 8 without involving him, which, unless player 4 was in a different position when passing sequences occurred, undermines the structural usefulness of a triplet as a tight possession-preserving tool.

We can also look at the color fill of the network with highlighted triplets and notice that the top triplets in Match 17 tightly connect players in both the midfield and defense throughout the center and outsides of the network while the top triplets in Match 1 are more heavily skewed towards the left side. The strong central triplets of (5, 6, 8) and (4, 5, 6) that we find in Match 17 are not present in the network for Match 1, perhaps indicating that the Peanuts have placed more pressure in the center of the field in this match, disrupting some of the Sharks’ connections that normally would have provided strength to central triplets.

### 6.7 Summary

By looking at a number of different metrics and match comparison methods, we have found that the Sharks were able to maintain better possession and ball distribution in Match 17 than in Match 1. Ball movement and clustering is focused in the defense and center of the network in Match 17 while it is pushed to the fringes of the network in Match 1. In addition, player 8 in particular experiences significantly less action in Match 1 than in Match 17, as he is not part of several high-betweenness edges, is passed over in a popular triplet, and has lower betweenness and closeness scores. Considering that many of the features of these two matches were held constant, it is a good assumption to make that the Peanuts’ formation difference may have contributed to these differences between Match 17 and Match 1. The Peanuts have a 4-4-2 formation in Match 17 while they remove a striker and add a defender to create a 5-4-1 formation in Match 1, likely putting more pressure on the Sharks in the center of the network and generating many of these central relationship disruptions observed in this chapter.
Figure 6.6: Top triplets in Match 17 are distributed throughout the midfield and defense of the network while top triplets in Match 1 are heavily skewed towards the left side of the network. Each of these diagrams highlights the top 5 most popular triplets in each of the matches. The brighter the shade of green, the more popular the triplet is in the match.
Chapter 7

In-Match Tactical Changes

In a soccer match, a team can make tactical changes and/or changes in personnel at halftime that drive passing network transformations either within its own team or in the opposing team. In order to visualize and quantify the impacts of these changes, we generate a passing network for each half of a single match and compare the two using the methods implemented in Chapter 6.

<table>
<thead>
<tr>
<th>Match Period</th>
<th>Half 1</th>
<th>Half 2</th>
<th>Full Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>1-0</td>
<td>1-1</td>
<td>2-1</td>
</tr>
<tr>
<td>Opposing Formation</td>
<td>5-2-3</td>
<td>4-3-3</td>
<td></td>
</tr>
</tbody>
</table>

Table 7.1: The Sharks played Match 9, a home game, against Dynamite, a team that implemented a formation change at halftime. Formations are determined via methods and classifications put forward in [20].

As an implementation example, we will use Match 9 to assess impacts of an opposing team’s in-match tactical changes. In this match, the Sharks played the opposing team, Dynamite, at home. In the first half, the Sharks score one goal against Dynamite in the 9th minute of the game. At halftime, Dynamite changes its formation from 5-2-3 (5 defenders, 2 midfielders, and 3 forwards) to 4-3-3 (4 defenders, 3 midfielders, and 3 forwards), moving one of its defenders up into the midfield. In the second half, Dynamite scores a goal in the 49th minute, but then the Sharks score another goal in the 58th minute. This example is ideal because Dynamite implemented a tactical change at halftime, deciding to shift its formation. Comparing the two halves will allow us to observe the impacts that this activity had on the performance of Sharks. The specifications for this match are given in Table 7.1.

We also select a “control match” to try to determine what changes may be attributable to Dynamite’s formation change at halftime in Match 9 as opposed to normally expected changes from half to half. We select Match 10 as the control match, where the opposing team, Tsunami, has a 5-3-2 formation for the duration of the match, meaning that there are 5 defenders, 3 midfielders, and 2 strikers. In the
first half, the Sharks score a goal in the 24th minute, but then Tsunami returns in the second half with two goals - one in the 74th minute and one in the 78th minute. Specifications for this match are given in Table 7.2.

### 7.1 Visual Network Comparison

We begin by once again comparing the two networks visually in Figure 7.1. It appears that the passing network for the Sharks in the first half has overall higher clustering (see Section 3.2.4) across roles as well as stronger passing connections in the defense and midfield, as the edges connecting the defensive players 3, 6, 5, and 2 are all noticeably stronger in the first half than in the second. Additionally, edges connecting players 2 and 3 to the center midfielders 4 and 8 are somewhat stronger in the first half than in the second. In the second half, clustering has shrunk overall, except for the left center midfielder (player 4). Almost all pairwise passing connections have become weaker, except for some of those involving player 4. From a visual standpoint, it appears that Dynamite’s formation change has disrupted the strength of the Sharks’ passing network, making it more difficult for players on the Sharks to cluster and complete passes. This visual designation is actually confirmed by event data: the Sharks complete 323 passes in the first half compared to just 263 in the second half. This is an understandable outcome in response to the shift from the original 5-2-3 formation to a more offensive formation of 4-3-3 on the opponent’s part, as an additional midfielder gives Dynamite the opportunity to disrupt more passing connections in the Sharks’ midfield.

When looking at the passing networks in Figure 7.2 for each half in the control Match 10 (where the opponent, Tsunami, did not change formation at halftime) as comparison, there seems to be a less concrete change in node sizes from the first half to the second - some nodes have higher clustering coefficients while others’ become lower. There is also no concrete pattern to which edges have become stronger or weaker from the first half to the second - there is actually an increase in the number of completed passes from 293 in the first half to 330 in the second half.

### 7.2 Differencing Networks

By differencing the passing networks from both halves of Match 9, we get a more concrete look at how the dynamics of the Sharks changed from the first half to the
Figure 7.1: A side-by-side comparison of the passing networks for the halves of Match 9 shows a general shrinkage in clustering coefficient (given by node size) from the first half to the second as well as an overall weakening of defensive edges.
Figure 7.2: A side-by-side comparison of the passing networks for the halves of Match 10 demonstrates change from the first half to the second half that is harder to pinpoint in terms of changing clustering of nodes and strength of edges.
second, as seen in Figure 7.3. Here, blue nodes indicate higher clustering coefficients in the first half than in the second while orange nodes indicate higher clustering coefficients in the second half than in the first. Blue edges indicate that the passing connection is stronger in the first half than in the second while orange edges indicate that the passing connection is stronger in the second half than in the first. Finally, the size of the node reflects the difference in clustering coefficient while the edge thickness reflects the difference in pass frequency between the two halves.

We find from Figure 7.3 that every player clusters more in the first half than in the second half, except for the left center midfielder (player 4). The frequency of tight passes (passes between neighboring players) in the team appears to be much higher in the first half, as indicated by the quantity and thickness of the blue lines linking closely positioned players in the difference network. Players in the second half appear to be favoring long passes - punts from the goalie and cross-field passes, for example - with the exception of the connections involving role 4. From this difference network, we may even be able to infer that the additional midfielder that Dynamite introduced in the second half began blocking and marking player 8 more, causing player 4 - perhaps a weaker player - to become more central to the network in the second half. This theory is further supported by the fact that the players in roles 4 and 8 were not substituted between halves, indicating that the change in clustering is unlikely to have come from a personnel change and more likely to have come from the formation change implemented by the opposing team.
Looking at the half difference network for Match 10, our control match, in Figure ??, we find that some of the players have higher clustering in the first half (indicated by the blue nodes) while others have higher clustering in the second half (indicated by the orange nodes). Additionally, it seems like a large variety of pass types - including long punts and crosses - are more prevalent in the second half than in the first. It is true, though, that there exist a number of offensive passing connections in Match 10 that are stronger in the first half than in the second. These findings for the control match contrast with those for Match 9, as only one player in the Match 9 difference network appears to have a higher clustering coefficient in the second half. Also, only one major type of pass (cross-field) seems to be more popular in the second half than in the first half of Match 9. These differences between Match 9 and Match 10 indicate a probable impact of Dynamite’s halftime formation change on the Sharks’ passing network dynamics in Match 9.

### 7.3 Full-Network Connectedness

<table>
<thead>
<tr>
<th>Match Period</th>
<th>Full Match</th>
<th>Half 1</th>
<th>Half 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Strength ((\lambda_1))</td>
<td>65.3</td>
<td>38.2</td>
<td>28.2</td>
</tr>
<tr>
<td>Algebraic Connectivity ((\tilde{\lambda}_2))</td>
<td>20.0</td>
<td>12.9</td>
<td>10.0</td>
</tr>
</tbody>
</table>

*Table 7.3: Network Connectedness Measures for Match 9*

The calculated network strength and algebraic connectivity (see Section 3.2.1) values for the halves of Match 9 are given in Table 7.3. We find that both the network
strength and the algebraic connectivity for the first half of Match 9 are larger than those for the second half. The trend in network strength indicates a higher volume of passes in the first half than in the second, allowing the network to contain more and stronger passing relationships. Algebraic connectivity decreasing from the first half to the second means that the ball was less distributed around the entire team in the second half than in the first half. The lower algebraic connectivity is a result of there being fewer and weaker connections between subnetworks of players, so the second half’s lower algebraic connectivity could indicate a concentration of the ball in one section of the passing network, while the formation change by Dynamite in the second half disrupted the Sharks’ ability to distribute the ball around the team, perhaps due to key players being blocked from receiving or initiating passes.

<table>
<thead>
<tr>
<th>Match Period</th>
<th>Full Match</th>
<th>Half 1</th>
<th>Half 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Strength ($\lambda_1$)</td>
<td>69.0</td>
<td>32.9</td>
<td>37.9</td>
</tr>
<tr>
<td>Algebraic Connectivity ($\tilde{\lambda}_2$)</td>
<td>21.8</td>
<td>14.6</td>
<td>7.8</td>
</tr>
</tbody>
</table>

Table 7.4: Network Connectedness Measures for Match 10

We then also look at the values for network strength and algebraic connectivity in both halves of control Match 10, given in Table 7.4, in which there is no opposing formation change at halftime. Notice that the network strength in Match 10 actually increases from the first half to the second due to the Sharks completing more passes in the second half than in the first and thus creating stronger connections throughout the network. Algebraic connectivity in Match 10 follows a similar downward trend from the from the first half to the second half that we observed in Match 9, suggesting that it will generally be more difficult for the Sharks to maintain strong connection of all parts of the network in the second half than in the first.

Inferences about changes in network strength and algebraic connectivity from the first half to the second are useful to see what kinds of full-network changes are occurring in a match. It is difficult to say whether the changes in these connectedness values observed for Match 9 are a result of Dynamite’s formation change on its own, but in the context of the rest of the metrics, we could posit that Dynamite’s formation change was at least a partially influential factor in reducing network strength and algebraic connectivity from the first half to the second.

7.4 Centrality Measures

By looking at the centrality measures across both halves, we can infer what effects changes made at halftime had on the roles that each player performs in the team. We choose to focus on betweenness as a measure for comparative centrality, as there are two main insights we can draw from looking at betweenness measures of players compared against each other across halves. The first is about the network as a whole: higher overall betweenness scores indicate poorer ball movement and heavier reliance
Figure 7.5: Match 9 First Half vs Second Half Betweenness. Green points correspond to forwards, yellow points to midfielders, and red points to defenders; the blue point is the goalkeeper.

on individual players to distribute the ball around the network. The second is about individual players: if certain players increase in betweenness from one half to the other while other players decrease, we have some indication about the role change that these players have undergone from one half to the next.

The betweenness-vs-betweenness points for many players in Match 9 seem to fall fairly close to the center line, as shown in Figure 7.5, indicating that about half of the team maintains approximately the same role in ball distribution from the first half to the second. Of the other half of the team, some players have increased in betweenness from the first half to the second while other players have decreased in betweenness. Of note are players 2, 3, 5, and 8, who all experience an increase in betweenness from the first half to the second, and players 4 and 6, both of whom experience a decrease in betweenness from the first half to the second. In the control Match 10, Figure 7.6 seems to show that players 2, 3, 6, 7, and 8 all experience an increase in betweenness between halves while players 4 and 11 experience a decrease. Outside of players 2, 3, and 6 who all increase in betweenness from the first half to the second inf both Match 9 and Match 10, Match 9 finds the center defender (player 5) increasing in betweenness while Match 10 finds a center defender and a wide midfielder (players 6 and 7) increasing in betweenness. While it is possible that Dynamite’s formation change at halftime in Match 9 caused a defensive player to
Figure 7.6: Match 10 First Half vs Second Half Betweenness

become more involved in the network passing distribution, there appears to be no particular pattern to which players fall on either side of the line. Betweenness, as an indicator of how likely each player is to be in the middle of a passing path between any pair of players in the network, does not seem to follow a particular trend from the first half to the second half given both the match with a halftime opposition formation change - Match 9 - and the match with no such change - Match 10.

7.5 Highest-Betweenness Edges

By taking a look at which edges change in importance between halves, we may be able to get insight into what impact that changes made by the opposing team have on the team in question’s ball distribution channels. We compare diagrams of the edges with the top 10 highest edge betweennesses in both halves of Match 9 in Figure 7.7. From these edge betweenness network diagrams, we can see that the “most important” edges in the first half are focused in the defensive center of the team, with betweenness indicating that the ball is distributed from one side of the network to the other. In the second half, however, the “most important” edges are on the outer margins of the network and a commonly integral connection between defensive players 3 and 6 no longer has high betweenness. This is an indication that when the opposing team, Dynamite, made a formation change at halftime towards a more
(a) Match 9 First Half Highest Betweenness Edges

(b) Match 9 Second Half Highest Betweenness Edges

Figure 7.7: Match 9 Highest Betweenness Edge Diagrams Across Halves. Each of these diagrams highlight the edges with the top 10 highest edge betweenness values for each half. The lighter the shade of green highlighting the edge, the higher the edge betweenness value.
offensive structure, it disrupted the Sharks’ defensive connections. Of course, there are many other factors at play, so we also compare the highest-betweenness edges of the halves of the control Match 10.

Match 10’s network diagrams with the top 10 betweenness edges highlighted in Figure 7.8 indicate that ball distribution and possession remains strong in the central defense of the network between halves, which is different from the half-to-half comparison from Match 9. One detail worth noting in this control match is the relative increase in the number of high-betweenness edges present in the offensive end of the network in the second half, indicating that the team makes more of an offensive effort as time runs out in the match. Although Figure 7.7b also demonstrates that there are more well-traversed connections to the offensive end of the network in the second half (e.g., the passing connection between players 3 and 9), the general consistency of the top edges in Match 10 contrast with the change in the top edges of Match 9, suggesting that Dynamite’s formation change at halftime in Match 9 likely disrupted the Sharks’ dynamics in the central region of the network.

7.6 Top Triplets

Similar to edge betweenness, identifying the top triplets in each half can provide insights to centers of possession in a team and how they change from half to half. As we can see in Figure 7.9, the Sharks are focusing their possession in the center and defensive regions of the network, controlling the ball in the middle of the pitch and maintaining network-wide ball control. In the second half, though, the top triplets shift to the outer edges of the network, with only one triangle (3, 4, 6) even partially covering the defensive region of the network. Before concluding that this is an effect of the opposing team, Dynamite, adding a midfielder and adding pressure to the Sharks’ defense after halftime, we also take a look at the control Match 10.

Unlike in Match 9, Match 10 sees a general maintenance of important triplets in Figure 7.10 from the first half to the second half, at least in terms of network region and involved players. In particular, the Sharks are able to maintain possession strength via triplets in the defensive and central regions of the network in both the first and second halves. This contrast with Match 9 provides a confirmation that the formation change of the opponent at halftime for Match 9 disrupted possession patterns. Dynamite’s choice to shift to a more offensive formation structure at halftime forced the Sharks to shift their possession centers from central players and defenders to different players - particularly players situated on the outer edges - in the second half.

7.7 Summary

Teams generally implement tactical changes in order to solicit some change in behavior from either their own team or the opposing team. By examining the halves of
(a) Match 10 First Half Highest Betweenness Edges

(b) Match 10 Second Half Highest Betweenness Edges

Figure 7.8: Match 10 Highest Betweenness Edge Diagrams Across Halves
Figure 7.9: Match 9 Most Popular Triplets Across Halves. Each of these diagrams highlights the top 5 most popular triplets in each half, with brighter shades of green indicating more popular triplets. Note the shift from central triplet coverage in the first half to top triplets only covering the outside edges of the network in the second half.
Figure 7.10: Match 10 Most Popular Triplets Across Halves. Note in this match that the triplets, although changing slightly, still cover the same approximate central defensive region of the network.
Match 9 as an example of a match where the opposing team changed formation at halftime, we observe how match comparison strategies can reveal response patterns in a team. The Sharks sought to maintain possession in the center and full defense of their network, which they were able to do in the first half, as evidenced by the most important edges and most popular triplets being quite central to the passing network. When Dynamite changed its formation from 5-2-3 to 4-3-3 at halftime, becoming more offensive by adding an additional midfielder, top triplets and important edges were pushed to the outsides of the network and both network strength and algebraic connectivity decreased. We did notice, though, that algebraic connectivity went down from the first half to the second anyway in the control Match 10 played against Tsunami, but these other observations still hold. Similar to the effect of a different opposing formation in Chapter [6], the Sharks are unable to maintain control in the center of the network in response to a tactical change at halftime by their opponents, Dynamite, and it is reflected in the weaker ball distribution and possession metrics observed in the second half as opposed to the first.
Chapter 8
Conclusion

In the world of sports analytics, soccer is relatively under-explored, especially considering its sheer volume of worldwide fans. The main difficulty in generating useful analysis of soccer teams is that there exists no clear metric for quantifying the quality of a team or player, making it difficult to identify trends and make comparisons between players and teams and across matches. Existing literature utilize graph theory and metrics in order to compare teams against each other in the context of individual matches, but generally fail to provide a way to characterize changes in performance of a single team over the course of a match or a season. Quantifiable team characterization is a key step towards explanatory analytics and, ultimately, predictive analytics.

Passes are fundamental building blocks in the game of soccer and thus represent much of the activity and player-to-player relationships in a match. By reducing the events of a soccer game down to a concise passing network, we have opened up the ability to use graph theory fundamentals in conjunction with derived metrics representative of a team’s match activity in order to produce comparable metrics placed in a context that allows us to explain certain match results and team behavior. Since graphs can only contain a limited amount of information in the form of vertices, edges, and weights, it was key to first produce an appropriate visualization method that encompassed as much relevant information about a team’s match activity.

Following this, we then were able to characterize a single team’s average performance patterns over the course of an entire season through passing network visualization, half-by-half network differencing, calculation of network strength and algebraic connectivity measures, calculation and explanation of centrality measures of individual players on the team, and identification of high-importance relationships within the network, both by calculating edge betweenness and by finding popular triplets within the network. We also calculated average centrality and network connectedness measures for every match over the duration of the team’s season, generating a brief summary of the team’s activity in each match in the context of ease of ball movement and ability to move the ball around the entirety of the network.

Using the same metrics, we demonstrated the methods through which one can
compare a team’s performance between matches or between halves of the same match. Comparisons such as these help explain the impacts that match factors such as opposing team formation or in-match tactical changes have on the behavior of a single team. This thesis has shown that quantitative methods are useful in explaining changes in the roles that different players have in a team network as well as in explaining differences in overall team cohesion. When combined with visual methods, the relevance of these quantities becomes clear when considering the soccer match context in which they must be placed.

8.1 Future Work

This thesis has set forth a framework through which soccer matches can be analyzed and compared through graphical and quantifiable means. While we have been able to explain match behavior and strategy of an entire team in individual game situations, there are a number of applications of these methods that would prove useful in learning more about teams as entities in and of themselves rather than just as actors in individual matches. In particular, we may be able to learn more about the roles of individual players and their contributions to team dynamics, determine the effects that a team’s changes has on its own network, and use some of these metrics as classification features in predictive contexts.

A team’s roster shifts from match to match and even undergoes personnel changes in the middle of matches, and our analyses in this paper simply combined the effects of players based on what positions they fall into on the pitch. In order to determine what the effects of individual players are on a team, we could compare a team’s passing networks during stretches of game time where the player is on the field and where the player is not on the field. Following similar network comparison methods, we may be able to narrow down the effects that individual players have on team strength. Identifying player effects is useful not only from a recruiting perspective, but also from a team development perspective, as a coach will want to put resources towards identifying which players need work (perhaps they appear to weaken the team when they play) or which players are the biggest contributors to team strength.

Chapters 6 and 7 particularly looked at what impacts an opposing team’s choices may have had on the target team’s performance in whole matches or match halves. While it is expected for a team to react to an opposing team’s tactical choices, we would also want to identify what effects these choices have on the team implementing these changes. For example, in a match between Team A and Team B, say that Team B changes formation from 5-4-1 to 5-3-2 at halftime. In this paper, we have discussed how to analyze and explain the effects that Team B’s tactical change would have had on Team A’s performance, but it may prove useful to also assess what effects Team B’s tactical change had on Team B’s performance as well. It is possible that a team performs with certain formations better than others, for example, and comparing a team’s passing networks across time periods where the team itself institutes change
would allow us to take a step towards identifying a team’s optimal setup.

The most exciting potential application of the methods and findings in this thesis is for predictive purposes. From a pattern-finding standpoint, we want to be able to generalize passing network characterizations for matches with different features. We can then actually use the similarity heuristics discussed in order to determine what features bring the networks in each group together. By finding thresholds of similarity between passing networks of certain types, we can try to take an unidentified passing network and determine which groups it falls into, or use predefined characteristics of a team and the match (e.g., home or away or opposing team formation) it is playing in as a way to project what the passing network will look like. Coaches and trainers can use such patterns as the basis for models that may help them to make decisions such as which player to enter into the game when the opposing team has made a tactical change, what formation to implement based on the typical expected formation of an opposing team, or which passing relationships in the opposing team should be defended well in order to restrict ball movement and possession. There are an infinite number of choices that a coach, player, or opponent can make, and delivering predictive power to the question around the impact of these choices opens up the opportunity for a team to improve its overall training efficiency and game-play effectiveness.
Bibliography


