Problem Chosen	2020	Team Control Number
D	MCM/ICM	2005195
	Summary Sheet	

There are many factors that go into the performance and cohesiveness for a team. This makes the job as a coach difficult as they need to take in an extensive amount of information combined with all possible variables for their team. We developed a ranking method that can help generate the best possible starting lineup for a soccer team. Using performance indicators determined by match statistics and centrality between players of the team on the field, we can determine whether the lineup is expected to perform better than others.

We analyzed the starting line up of all of games provided. We then created a passing network for each game. Each starting line up configuration was analyzed separately so that the efficiency of each system could be determined. FIFA world cup analysis was used to determine which performance indicators are most helpful in determining the result of a game. We then weighted valuable performance indicators utilizing betweenness centrality.

As with every model, we were required to make some assumptions. We assumed that the game took place on a rectangular field. Although a player is not bound to a particular region of the field, we assumed that each team member played in the position which they are assigned. As previously mentioned, we are assuming that performance indicators for FIFA analysis will continue to be valuable in our model. Soccer analytics is typically accompanied by video footage of games, Since we did not have such footage, we analyzed events and their time stamps in the game to determine shots on target. The data set provided did not indicate if a shot was successful or not. Therefore, we assume that a shot on target occurred if a shot was immediate followed by a save attempt from the opposing foal keeper.

How to EXCEL in Soccer

Team Control Number #2005195

Thursday 20th February, 2020

Contents

1 Introduction

Diverse societies have become increasingly more connected. Such a network of relationships, participants, and stake holders has prompted further research into how these teams produce successful outcomes. As it turns out, a mathematical analysis of how the members of a team interact has become an indisposable tool in making sense of these vast networks. Perhaps one of the most primed networks for a mathematical approach to analysis is that of a sports team.

We, at Intrepid Champion Modeling (ICM) will utilize data collected from our home team's, the Huskies', previous season to develop a model representing on field connections which are impactful to the result of the game. Last season, the team posted 13 wins, 15 losses, and 10 ties in total. During this time, the team competed twice against 19 different opponents, one match at the Huskies' home field, and one at their opponent's.

Since the soccer ball makes its way around the field as a result of passing events, we will analyze the passing network between starting players in a game. We will construct our passing network "...from the observation of the ball exchange between players, where network nodes are [soccer] players and links (or edges) account for the number of passes between any two players of a team" [4]. We will consider the success of the whole team, as well as pairs and triads of players. These interactions will enable us to make suggestions to the coach for increased wins through the up coming season.

1.1 Restatement of Problem

Our efforts will be to understand how individual players on the team interact with each other and how these interactions impact the outcome of a game. This will include how a player impacts the whole team and how they impact another member of the team. A player's effect on the team will be quantified using betweenness centrality. A visualization of the network will be used to leverage understanding of how these relationships play out on the field. We will utilize data provided to us by the Huskies' coach in our analysis. Typically, analysts watch videos of past matches as well as considering raw statistical data from the game. As we do not have video footage, we will make reasonable assumptions on our data to aid in analysis. We will provide suggestions to the coach to improve next season's outcomes.

1.2 Assumptions and Parameters

1.2.1 Coordinate System

The provided data includes x and y coordinates of the origin and destination of all game event. Utilizing this data, we will create a Cartesian coordinate system to represent the field. Both x and y axes have a range from 0 to 100. It is worth noting that not every type of event produces a destination location, for example a player substitution or foul. However,

physical soccer fields are not perfect squares, but rather are rectangles. One must keep this in mind when creating visualizations else the network will be compressed horizontally. The standard dimensions for a soccer field run from 110 - 120 yards long, and 70 - 80 yards wide, which makes the field rectangular [3]. Due to the key provided with the data, we know the points are presented from the attacking team's view, making 0 on the x axis in line with the attacking team's own goal. They are attacking the opponent's goal located at 100 on the x axis from their point of view. Looking at the data for corner kicks, a free kick taken by the attacking team from a corner of the field on their opponent's goal line, we see that all are taken at or next to the coordinates $(0, 100)$ and $(100, 100)$. This shows that by looking at the field from a side view, we have our plane with a team attacking from left to right. Thus, one unit on the x axis is not equal to one unit on the y axis. We will assume a 120 yard long, 80 yard wide field.

1.2.2 Team Formations

Although soccer is a dynamic game, we will assume that each team member plays exclusively in the in the position indicated in their Player ID. For example, players with an "F" in their player ID will be assumed to exclusively play as a forward. Based on the FIFA mandated rules of the game each team must play with one goal keeper [3]. From there we will assume that each team is playing with any combination of defenders, midfielders, and forwards where each team has 11 players on the field. We assume the number of defenders ranges from 3 - 5 defenders, midfielders ranges from 3 - 5 midfielders, and forwards from 1 - 3. These assumptions are stemmed from what player formations are most commonly used throughout the sport. [5] [1]

1.2.3 FIFA Research Applies

To build a model with performance indicators for each player, we need to determine which match statistics are calculable for to a given player. Furthermore, which statistics are significant indicators of the game outcome. To do this we chose to look at those that have been shown to be significant in professional soccer matches. [7]. We assume that the Huskies these indicators, shots on target, crosses, pass success rate, tackles, and ball possession, will also be significant for the Huskies. We are unable to look at aerial advantage (%) as our data only provides the event of an aerial duel, but not who won it.

Since the provided data does not reflect how a goal was scored or which player scored it, we are unable to factor in a player's impact in the team network based off of how many goals they were directly or indirectly involved in. Instead, the shots on target performance indicator will weight higher for the forwards, who tend to shoot more, as well as the plays with the best connections to them.

1.2.4 Data Erros

The data indicates that a player occasionally passes the ball to themselves. Since a pass must occur from one player to another, we will exclude these data points from our analysis.

Locations of some goalie events are listed at or around the points (0, 0) and (100, 100). As noted earlier, these should be corners of the field. Happily, the data within `passingevents.xlsx`, lists passes from the goalkeeper (launch, pass, etc.), at a point of origin around 50 on the y axis. This allows us to use the data in `passingevents.xlsx` for creating our network for the team.

With the data provided, we are not given whether a shot was on target or not. The `DestinationOrigin` for a shot has the same error as the goal keeper location for goal kicks. All shots are directed at (0, 100) or (100, 100). Instead of using a shot location, we will assume that a shot is on target, if it forced the goal keeper to make a save attempt. Therefore, if a shot was immediately ($\pm$ 5 seconds), then we can count it as a shot on target.

2 Methods

Our aim is to create a model that can be adjusted to look at the structural indicators, network properties, and performance indicators of players on a macro and micro level. This would mean that we can use the model to look at how the team dynamics have changed across the season, as well as at specific games (or time intervals within games). To do this we wish to build a visual representation of the passing network between players in the game, as well as a formula to calculate the impact each player has on the team's performance, as well as on other players he interacts with.

2.1 Developing a Network of Passes

We wish to study the relationship between different players within formations most used by the Huskies as well as those used by their opponents. With the data provided, we are able to divide the connections between players based on the team, match or matches, and times within game.

First we need to understand the interaction between players on the field and the importance of them; namely, the passing events between such players. A natural way to relate the players on the field is through a passing network: according to [4], "passing networks are constructed from the observation of the ball exchange between players, where network nodes are [soccer] players and links (or edges) account for the number of passes between any two players of a team".

2.1.1 Directed Graphs as Networks

For each pass exchange between players, we would like to take into account whether the ball was passed or received by the player. Since there is a direction on each pass, we utilize a directed graph.

Definition 2.1.1 (Directed Graph, [10])**.** *A directed graph or digraph G consists of a vertex set $V(G)$ and an edge set $E(G)$, where each edge is an ordered pair of vertices. Given the edge (u, v), we say u is the tail and v is the head. We write $u \to v$ if there is an edge from u to v in G. Note that the choice of head gives the direction of the edge.*

Each `PlayerID` is given as "`Team Name_Positional Notation`", where positional notation includes a letter denoting the position and a number ID for the unique player at that position on the team. Such as Huskies_G1 is goalkeeper 1 for the Huskies, and Opponent19_F2 is forward 2 on team Opponent19. We also have "`D`" for defender, "`M`" for midfielder, and "`G`" for goalie.

Let $\mathbf{P}$ denote the set of all players for the Huskies and let $p_i \in \mathbf{P}$ be the distinct team player with $i \in \{\text{D1}, \ldots, \text{D10}, \text{F1}, \ldots, \text{F6}, \text{M1}, \ldots, \text{M13}, \text{G1}\}$. As there are 30 players on the Huskie's team, $\#\mathbf{P} = 30$, with 1 goalie, 10 defenders, 13 midfielders, and 6 forwards.

We'd like our model to best represent what is happening during the games, so instead of creating a network of all 30 players, we will look to create different networks for the 11 players on the field over a time of interest. By placing each node at the average position for the respective player, we can look for trends within their game play. For the sake of simplicity we will just consider the starting 11 players as they played the majority of the game. However, we note that it is possible to adjust the model to account for the players substituted throughout the match.

2.1.2 Network of Passes Between Players

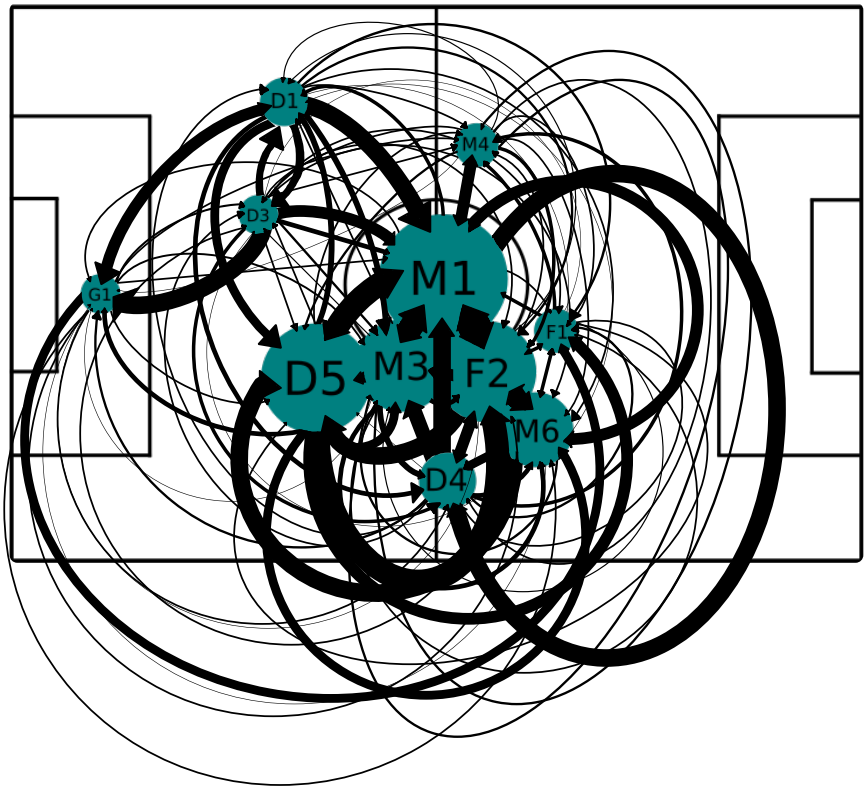

Figure 1: Passing Network for Huskie's

As requested by the coach, our networks will consist of only passing combinations. Hence, we will utilize the `passingevents` spreadsheet provided. Each passing event has an `OriginPlayerID`. If the pass was successful, the passing event includes a `DestinationPlayerID` if not, the `DestinationPlayerID` is left blank. Using the data provided, we can filter by team and match or matches as desired. Note that we converted all `.csv` files into an *Excel Workbook* (`.xlsx`), to create pivot tables for filtering the data, adjacency matrices to build networks, as well as simple calculations.

We will consider one type of team formation at a time in our analysis. Considering multiple formations in one network will not produce valuable results, as the strengths and weaknesses of any formation or tactic will be covered by the noise of the other formations. To build a passing network for best representing the Huskies in their season, we need to know at what formation they utilized. We will analyze game results with different formations to make suggestions for next season.

By cross referencing players substituted during each match, for each team, with the total passing events for each player, we can conclude which players each team started with for any of the games. From there we can deduce the formations based on the types of players for each team. For example, creating a pivot table from `passingevents.xlsx`, shown in Excel Table 1 below, we see counts for all passing events registered by Huskies in matches 14 and 38 vs Opponent14. Then, using filters for Huskies, Substitutions, and MatchID's 14 and 38 in `fullevents.csv`, we see who did not start on the field for those games (substituted on as `DestinationPlayerID`).

By Excel Table 2 we know the Huskies substituted players M11, M10, and M9 into the game during the second half. Comparing that to Excel Table 1, we see the Huskies began the game with one goalie, four defenders, four midfielders, and two forwards, putting them in a 1-4-4-2 formation.

Let $\mathbf{P}^{(11)}$ be the finite set of starting 11 players for which each $p_i \in \mathbf{P}$, or line up of players, is distinct. If p_1 and p_2 denote two distinct players, called player 1 and player 2, respectively, then each element $s_i \in S$ is of the following form: (1) if there exists a pass from p_1 to p_2, then we denote this pass with $p_1 \to p_2$; (2) if there exists a pass to p_1 from p_2, we denote this pass with $p_1 \leftarrow p_2$.

To find the total number of passes to and from a fixed arbitrary player, $p_1 \in \mathbf{P}^{(11)}$, we have the following equations:

$$\#(p_1 \to p) = \sum_{\substack{s_i \in S, \\ i=1}}^{n} \begin{cases} 1 & \text{if } p_1 \to p \in S, \\ 0 & \text{otherwise} \end{cases} \tag{2.1.1}$$

and,

$$\#(p_1 \leftarrow p) = \sum_{\substack{s_i \in S, \\ i=1}}^{n} \begin{cases} 1 & \text{if } p_1 \leftarrow p \in S, \\ 0 & \text{otherwise} \end{cases} \tag{2.1.2}$$

where $\#(p_1 \to p)$ is the total passes from p_1 to every other player p in the starting 11 formation, and $\#(p_1 \leftarrow p)$ is the total passes to p_1 across the season. In other words, we find the total number of passes from player 1 each of the other 10 players in the starting line up over the entire season which is denoted by $\#(p_1 \to p)$.

Creating a pivot table with counts of all passing connections between players that started the game (`OriginPlayerID` x `DestinationPlayerID`) we then have an adjacency matrix for all connections between players within the game. This can be altered or filtered to consider the whole season, specific games, halves or a game, or times in a game.

Count of DestinationPlayerID	Column Labels		
Row Labels	14	38	Grand Total
Huskies_D1		30	30
Huskies_D10		6	6
Huskies_D2	38		38
Huskies_D3		18	18
Huskies_D4		25	25
Huskies_D5	31		31
Huskies_D6	36		36
Huskies_D7	30		30
Huskies_D8		19	19
Huskies_F1	12		12
Huskies_F2	54		54
Huskies_F4		22	22
Huskies_F5		8	8
Huskies_F6		10	10
Huskies_G1	13	8	21
Huskies_M1	46	50	96
Huskies_M10	3		3
Huskies_M11	6		6
Huskies_M12		2	2
Huskies_M2		14	14
Huskies_M3		35	35
Huskies_M4	26	28	54
Huskies_M6	28		28
Huskies_M8	21		21
Huskies_M9	3		3
Grand Total	347	275	622

Figure 2: Excel Table 1

MatchID	TeamID	OriginPlayerID	DestinationPlayerID	MatchPeriod	EventTime	EventType	EventSubType
14	Huskies	Huskies_F2	Huskies_M11	2H	2400	Substitution	Substitution
14	Huskies	Huskies_M8	Huskies_M10	2H	2640	Substitution	Substitution
14	Huskies	Huskies_F1	Huskies_M9	2H	2760	Substitution	Substitution
38	Huskies	Huskies_D10	Huskies_M2	2H	60	Substitution	Substitution
38	Huskies	Huskies_F5	Huskies_F6	2H	960	Substitution	Substitution
38	Huskies	Huskies_M4	Huskies_M12	2H	2280	Substitution	Substitution

Figure 3: Excel Table 2

Definition 2.1.2. *Let A be a $n \times n$ matrix and let p_i and p_j be two distinct players in the set $\mathbf{P}^{(11)}$. Then, we say that p_i and p_j are adjacent if there exists a pass between the two players. So, for each total number of successful passes between players $i, j \in \{1, \ldots, |\mathbf{P}^{(11)}|\}$, the a_{ij} entry of A is given by,*

$$a_{ij} = \begin{cases} \#(p_i \to p) & \text{if there exist a pass from } p_i \text{ to any other } p \in \mathbf{P}^{(11)}, \\ \#(p_j \leftarrow p) & \text{if there exist a pass to } p_j \text{ from any other } p \in \mathbf{P}^{(11)}, \\ 0 & \text{otherwise} \end{cases}$$

then, we call A the adjacency matrix *of successful passes between players.*

By finding the connections from an `OriginPlayerID` $\to$ `DestinationPlayerID`, we get all successful passes between those players. Recall, if a player has a pass event with no `DestinationPlayerID`, then it is not a completed pass, failing to connect two players. Let S denote the set of all successful passes between distinct players in $\mathbf{P}$ over an entire season. We note that given the data from `passingEvents.csv`, there are a total of $10,435$ successful passes between Huskie players. Hence, $\#S = 10,435$.

2.2 Closeness Centrality

According to [8], a geodesic path is a self-avoiding path between two nodes in a network such that the length is the number of edges traversed between nodes. Closeness centrality is a measure that takes into account the shortest path between nodes as a metric for centrality. Let d_{ij} be the *shortest* distance between nodes $i, j \in G$ where G is the network and let N be the number of nodes within the network. We wish to take the average distance from node i to any other node j in the network. From [8], we have

$$\ell_i = \frac{1}{N} \sum_{j \in G}^{N} d_{ij}. \tag{2.2.1}$$

In essence, nodes with lower averages will have more influence throughout the network. By the definition of a geodesic path, we will we only consider passes from the player of focus to any other player without using a player that has already received the ball. Therefore, the player that has the lowest average for their geodesic paths will be the most connected within the team.

Now, we define *closeness centrality*, denoted by $C_C(i)$, to be the inverse of Equation 2.2.1, given by

$$C(i) = \frac{1}{\ell_i} = \frac{N}{\sum_{j \in G}^{N} d_{ij}} \tag{2.2.2}$$

(as defined in [8]).

Closeness centrality is an effective measure of centrality when considering the shortest distances between nodes and how important each node becomes once paths are considered. Furthermore, closeness centrality conveys how information is passed through the network using the shortest paths [9]. Given that we are tasked to model configurations of players in dyadic and triadic configurations, we know that the passes between players should reflect the shortest distance that will result in a successful play. A player that receives a lower closeness centrality measure is said to be more central to the team because the player requires less passes to get the ball to the rest of the team.

2.3 Betweenness Centrality

Now, we wish to consider the influence the player has in the network in relation to their placement within the network and how many passes are played and received by the player. We can use the concept of betweenness centrality–closeness centrality's counterpart–to determine how influential a node is throughout the network with the relation between paths. For example, "nodes with higher betweenness centrality may have considerable influence within a network by virtue of their control over information passing between others" [8]. In other words, since we are considering dyadic configurations of players within a network because it takes into account how efficient the pathways are that involve the player in question. Betweenness proves to be essential to evaluating which players cannot be removed from the starting 11.

In order to define betweenness centrality, we first consider a geodesic path: a self-avoiding path throughout a network [8]. Let A be the adjacency matrix as defined in Definition 2.1.2. The geodesic path from a fixed node is the number of edges traversed for which the fixed node can reach every other node. Simply, if there is a path connection between two nodes we count this connection as 1, otherwise, we do not account for the path. Therefore, we can sum over the entire network for each fixed node and obtain geodesic path for each node within the network.

Now, let d_{ij} be the geodesic path between p_i and p_j in the network of nodes given by $\mathbf{P}^{(11)}$. If p_k is a node that lies on a path between p_i and p_j, then d_{ij}^k is 1, otherwise 0 if p_k is not on the geodesic path d_{ij}. Recall that d_{ij} is the total number of paths between two nodes. We obtain the following formula for betweenness centrality:

$$C_B(i) = \sum_{\substack{p_k \in \mathbf{P}^{(11)}, \\ k=1}}^{N} \frac{d_{ij}^k}{d_{ij}}, \tag{2.3.1}$$

where N is the number of players in the particular formation we are considering. Thus,

$N = \#\mathbf{P}^{(11)} = 11$. By [8], if d_{ij}^k and d_{ij} are both zero, then $\frac{d_{ij}^k}{d_{ij}} = 0$.

Therefore, when a player is more efficiently connected within the dyadic pairings, there will be more passes that connects an influential player to the rest of the team. Efficiency is important to consider as the team needs to move the ball to an attacking position before the defending team has a chance to adjust and potentially prevent a goal.

2.4 Finding Performance Indicators

With all of the match events provided, there are some that are more important than others. Using team statistics that were found to be significant towards predicting the results of matches in the group stage at the FIFA World Cup [7], we can deduce which player statistics could be strong indicators for their importance to the team.

Looking at those that were significant in the 2014 tournament for all games, as well as those that we could find for the Huskies with the data provided, we came up with Shots on Target, Crosses, Tackles, and Pass Succession (%). With these performance indicators per player, we can do a logarithmic regression test to make sure they hold significance in our data. If they are significant, we can then use the log-odds coefficient for its respective performance indicator within our model. We use the log-odds in order to have the match result be the dependent variable while taking into account 1 as a win, 0 for a loss, and 0.5 for a tie.

Performance Indicators	Log Coefficients	p-value
SOT	0.450912	$p = 0.0002$
Pass Accuracy (%)	4.24027	$*p = 0.1001$
Crosses	-0.0956477	$p = 0.0112$
Tackles	0.00247793	$*p = 0.8258$
Ball Possession (%)	4.24369	$p = 0.0186$

Note, each match statistic chosen to analyze is adjusted for ball possession as done in [7]. This will take into account if a team possesses the ball for the majority of the game, then their defending match statistics will be lower, and vice versa. To adjust our parameters, we will use the equations given from [7]. Performance indicators that are already given in a percentage (pass success rate and ball possession) do not need to be adjusted.

Let V be the match statistic in questions where V_{ajstd} is our target adjusted value and $V_{original}$ is the original value of the statistic for the team within that game. We then have BP_{team} and $BP_{opposition}$, such that BP_{team} is the Ball Possession (%) for the them within the game, and $BP_{opposition}$ is the ball possession for their opponents within the game. This

gives us the following two equations. 2.4.1 is used for adjusting match statistics related for attacking, while 2.4.2 is used for the adjustment of defending statistics.

$$V_{ajstd} = (V_{original}/BP_{team}) \times 50\% \tag{2.4.1}$$

$$V_{ajstd} = (V_{original}/BP_{opposition}) \times 50\% \tag{2.4.2}$$

For example, in match 1, the Huskies had a BP proportion of 0.6435247 and 3 shots on target. The opposing team, Opponent1, possessed the ball for roughly 35% of the game, or 0.3564753. This would give the Huskies an adjusted shots on target of 2.330913, $(2.330913 = (3/0.6435247) \times 0.5)$

By running the findings for adjusted variables and results from matches into *gretl*, we are able to calculate the log coefficient for each variable as well as check the p-value to make sure it is a good indicator towards the end result for a game within our data set.

To find the total Performance Indicator value for each player, we can multiply the coefficient for the respective match statistic by the value of the statistic by the player in question over the extended time frame being analyzed. Summing these values for each performance indicator for the player gives their impact to the team based on their match statistics. For example, over the course of the season, `Huskies_F1` had a total of 16 shots of target, 23 crosses, and a pass success rate of 0.657458564. Using the log coefficients found above, we found their total Performance Indicator Value as 7.802496723.

3 Building the Model

Our aim for this model is to take into account the performance indicators chosen to analyze for each player, with their betweenness centrality to help the coach decide which lineup is best suited for the team. This will quantify how much of a factor each player is towards the success of the team, as well as how strong specific pairings are. A method with this aim has already been attempted before by Ryan Beal and Professor Gopal Ramchurn who presented their findings at the *Statsbomb Innovation in Football Conference* of 2019 [2].

3.1 The Parameters

By creating a table of all 76 team observations, one observation per team in each game, of their total shots on target, ball possession (%), pass success rate (%), crosses, talks, and result for the game, we can run it through logarithmic regression analysis. By loading the *Excel* table into *gretl*, we find that all except tackles are significant. Where the p-value and log coefficient for each performance indicator is shown in Figure 3.

Performance Indicators	Log Coefficients	
SOT	0.450912	p = 0.0002
Pass Accuracy (%)	4.24027	* p = 0.1001
Crosses	−0.0956477	p = 0.0112
Tackles	0.00247793	* p = 0.8258
Ball Possession (%)	4.24369	p = 0.0186

Figure 4: p-value and log coefficient for performance indicator

Looking back at the data collected over the course of the season for the Huskies, we see that what may be considered as a tackle, `Ground defending duel`, most likely is an attempted tackle instead of a successful tackle. Hence why it is most likely is not significant with our data. Due to ball possession as a team having such a low p-value, we can justify using individual pass success rate with a borderline p-value (0.1) as the more successful a player is at completing their passes, the more they will contribute to the team keeping the ball.

3.2 The Model

We first find the performance indicators for each Huskie player as determined by the following equation. So, for each $p_i \in \mathbf{P}$ we have the following formula for PI_i:

$$PI_i = SOT_i + C_i + PS_i,$$

where SOT_i is the shots on target, C_i is the crosses, and PS_i is the pass success percentage for player p_i, respectively. Let x_i be a binary variable for which $x_i = 1$ if PI_i suggests that the player should be chosen, and $x_i = 0$ otherwise.

Now, we consider the different combinations of dyadic players.

3.2.1 Selection of Dyadic Configurations

We first want to consider the most played formations for the Huskie's, which are the 1-4-4-2 and 1-4-3-3. The Huskies started with a 1-4-4-2 for 14 matches, and the 1-4-3-3 for 13. With the 1-4-4-2, they posted a record of 6-5-3 (Wins-Losses-Ties), while having a record of 7-3-3 with the 1-4-3-3 formation. To start we will look at the configurations of players who played in games with the 1-4-4-2 team formation.

Now, we assume that the formation chosen for the starting 11 will not change throughout the time frame being analyzed. As mentioned in the problem, we are allowed substitutions in

the game–however, since it needs more data to be more effective, theoretically we will have the best representation with those that have the most connections already. Therefore we will limit the players we look at to those with the most passing connections when modeling the team over more than one game. As noted before, we can filter the successful passes between players on a given team within excel to pick our 11 players of focus.

Recall $\mathbf{P}$ denotes the set of all players in the starting 11. We now, consider all dyadic configurations from the starting 11 formations, denoted by the set $\mathbf{P}_{dyadic}$. Then, the total number of dyadic configurations is the permutation:

$$\#\mathbf{P}_{dyadic} = \frac{11!}{(11-2)!} = 110.$$

3.2.2 Centrality Measures with Performance Indicators

Let c_{ij} be some combination of dyadic player configuration in $\mathbf{P}_{dyadic}$ where i denotes player $p_i \in \mathbf{P}$ and j denotes the player $p_j \in \mathbf{P}$. Then, let R_{close} denote the ranking of each players configuration within the team using the closeness centrality metric.

$$R_{close} = 11 \cdot \sum_{i=1}^{110} \left(\sum_{j=1}^{110} (PI_i + PI_j) \cdot \left(\frac{x_{c_{ij}}}{\sum_{j \in G}^{11} d_{ij}} + \frac{x_{c_{ji}}}{\sum_{i \in G}^{11} d_{ji}} \right) \right) \tag{3.2.1}$$

Then, we also want to analyze the connections between players through betweenness centrality measures where the ranking is given by $R_{between}$:

$$R_{between} = \sum_{i=1}^{110} \left(\sum_{j=1}^{110} (PI_i + PI_j) \cdot \left(x_{c_{ij}} \cdot \sum_{\substack{p_k \in \mathbf{P}^{(11)}, \\ k=1}}^{11} \frac{d_{ij}^k}{d_{ij}} + x_{c_{ji}} \cdot \sum_{\substack{p_k \in \mathbf{P}^{(11)}, \\ k=1}}^{11} \frac{d_{ji}^k}{d_{ji}} \right) \right) \tag{3.2.2}$$

3.3 Limitations of the Model

Like any tool, there are some weaknesses to our model. As noted in [6], there are some limitations to a visual display of passing networks. With the position of each node being relative to the average position of the player, if such player often plays on different sides of the field, or in different positions, then the positioning of their node will be a misrepresentation of where they usually are located during the game(s). In the case of a player consistently switching from a wide right to a wide left position over the time frame in question, their node will show up more central on the field.

The network of passes visual only gives a snapshot of what happened in the game(s). Not every pass from player $p_i \to p_j$ went along the same distance as shown by the figure. This

would make the model best paired with video analysis and more data to see in the game how or why the issues found by the model are happening on the field.

In regards to selecting the performance indicators to best represent the impact of players, we are limited to the data provided. Unfortunately we are unable to calculate Aerial Duel Success Rate (%) as well as successful tackles which would better represent the impact of defenders than shots on target. This leaves pass success rate as the main match statistic that will show the most positive impact from defenders. Forwards and midfielders on the other hand, have pass success rate and shots on target which they are able to have more often from playing closer to the opponent's goal.

Lastly, our model is limited to the data we are given. If there are errors in some of the data provided to build it, then there will be errors in the calculations. Also, if the time frame of game play in question is less than others, then the model will not be as accurate.

4 Results

Using the data provided with our developed model, we can now quantify the connections between players while taking into account their performance indicator values to generate a ranking for a specific lineup by formation and players. This will help solve the question as to which was the best lineup for the Huskies over the course of their season.

We looked at both the 1-4-4-2 and 1-4-3-3 as those are the two formations most used and most successful for the Huskies. Other formations used by them were 1-4-5-1, 1-5-4-1, and 1-5-3-2. By cross referencing the games played with each formation to the results, we can find the biggest wins and losses for each formation. Note that the Huskies failed to win any games when starting with any formation other than 1-4-4-2 or 1-4-3-3. Also all of the games lost with the 1-4-3-3 can be considered a close game (goal differential between teams $\leqslant 2$) [7], while with the games started with the 1-4-4-2 had two games (matches 4 and 23) that resulted in losses with a > 2 goal differential. Match 4 resulted in a 0-3 loss at home, and a 0-4 loss away in match 23.

Running our model in respect to each type of centrality (closeness and betweenness) provides us with different recommendation's for a lineup. First we look to generate a recommended lineup with the 1-4-4-2 formation using closeness centrality. As seen in figure 5, we have the lineup of Huskies players [G1, D1, D2, D3, D9, M1, M5, M6, M7, F1, F4]. Using our two models we can see the team lineup ranking for both betweenness and closeness of the given lineup. As noted in the methods, a higher ranking is better.

In order to know which type of centrality is a better indicator for recommending a team

playerID	PIV	$C(i)$	R_{close}	$C_B(i)$	$R_{between}$
Huskies_D1	3.18755	0.966667	79.2305	20.8191	1706.38
Huskies_D2	5.13329	0.935484	93.0565	14.1162	1404.2
Huskies_D3	3.7241	0.966667	83.8985	19.1428	1661.43
Huskies_D9	3.71839	0.644444	55.8992	0.238034	20.6471
Huskies_F1	7.8025	0.90625	111.919	13.9905	1727.78
Huskies_F4	9.34583	0.763158	104.848	7.34366	1008.92
Huskies_G1	2.58127	0.878788	67.2326	10.6992	818.554
Huskies_M1	5.65983	1	104.213	24.1047	2512.03
Huskies_M5	3.60379	0.659091	56.4898	0.224034	19.2017
Huskies_M6	4.4622	0.935484	87.4064	14.4108	1346.47
Huskies_M7	4.05591	0.591837	53.1338	0.0555556	4.98766
		Total =	897.328	Total =	12,230.6

Figure 5: Round 1 with 1-4-4-2 formation

lineup, we will now adjust the lineup by switching players out based on their betweenness centrality. This allows us to switch in Huskie's players D6, M3, and M4 for D9, M5, and M7. As shown in figure 6, both the totals values for R_{close} and $R_{between}$ went up. We can then conclude that betweenness centrality is a stronger indicator in our model.

Now we can compare the two formations we aim to analyze. By selecting a starting 11 based on the formation 1-4-3-3 as well as their Performance Indicator Values and Betweenness Centrality, get the players [G1, D1, D2, D3, D6, M1, M3, M6, F1, F2, F4]. As shown in figure 7, the 1-4-3-3 has the rankings $R_{close} = 953.863$ and $R_{between} = 15,341.6$. Comparing these to those calculated for the strongest 1-4-4-2 lineup, we see that the 1-4-3-3 has a higher ranking.

Our model only takes into account the attacking abilities of the team as noted before. To see how the switch to a 1-4-3-3 could help defensively, we can look for a trend to where the Huskies give up the majority of shots from their opponents. Looking at Figure 7 we see that most shots by the opponents against the Huskies are centrally located in front of the goal.

There are a plethora of variables that go into a team's success on the field, but we can conclude that based on our model the 1-4-3-3 is a stronger formation for the Huskies. This may depend on the other factors such as what formation the opponent is using, weather, players are out for injuries, and more. In high level sports, every little advantage helps. Therefore the better team cohesion by using a 1-4-3-3 could be the difference in producing better results throughout the whole season.

playerID	PIV	$C(i)$	R_{close}	$C_B(i)$	$R_{between}$
Huskies_D1	3.18755	0.966667	76.518	20.8191	1647.96
Huskies_D2	5.13329	0.935484	90.4315	14.1162	1364.59
Huskies_D3	3.7241	0.966667	81.1859	19.1428	1607.71
Huskies_D6	2.15027	0.90625	63.2753	10.9254	762.823
Huskies_F1	7.8025	0.90625	109.376	13.9905	1688.53
Huskies_F4	9.34583	0.763158	102.707	7.34366	988.318
Huskies_G1	2.58127	0.878788	64.7667	10.6992	788.531
Huskies_M1	5.65983	1	101.407	24.1047	2444.39
Huskies_M3	3.11938	0.966667	75.9249	16.5115	1296.86
Huskies_M4	3.30237	0.935484	75.0163	15.6685	1256.46
Huskies_M6	4.4622	0.935484	84.7814	14.4108	1306.03
		Total =	925.39	Total =	15,152.2

Figure 6: Round 2 with 1-4-4-2 formation

playerID	PIV	$C(i)$	R_{close}	$C_B(i)$	$R_{between}$
Huskies_D1	3.18755	0.966667	78.1469	20.8191	1683.05
Huskies_D2	5.13329	0.935484	92.0079	14.1162	1388.38
Huskies_D3	3.7241	0.966667	82.8149	19.1428	1639.97
Huskies_D6	2.15027	0.90625	64.8024	10.9254	781.233
Huskies_F1	7.8025	0.90625	110.903	13.9905	1712.1
Huskies_F2	4.98747	0.90625	87.9433	12.2585	1189.58
Huskies_F4	9.34583	0.763158	103.993	7.34366	1000.69
Huskies_G1	2.58127	0.878788	66.2475	10.6992	806.56
Huskies_M1	5.65983	1	103.092	24.1047	2485.01
Huskies_M3	3.11938	0.966667	77.5539	16.5115	1324.69
Huskies_M6	4.4622	0.935484	86.3578	14.4108	1330.31
		Total =	953.863	Total =	15,341.6

Figure 7: Round 3 with recommended 1-4-3-3 formation

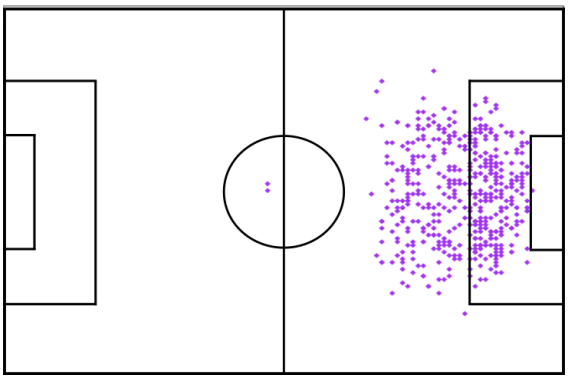

Figure 8: Scatterplot of Opponents

5 Going Forward

Going forward we wish to analyze the connections between triadic configurations of players. The following shows how we can update our current model to consider all triadic pairs. The next two equations denote the binary decision variable x with the betweenness centrality measure for each particular paring of players. As there are 6 combinations of two player pairings, we have 6 equations.

$$q_1 = x_{c_{ij}} \cdot \sum_{n=1}^{11} \frac{d_{ij}^n}{d_{ij}}, \quad q_2 = x_{c_{jk}} \cdot \sum_{n=1}^{11} \frac{d_{jk}^n}{d_{jk}}, \quad q_3 = x_{c_{ki}} \cdot \sum_{n=1}^{11} \frac{d_{ki}^n}{d_{ki}}, \tag{5.0.1}$$

$$q_4 = x_{c_{ji}} \cdot \sum_{n=1}^{11} \frac{d_{ji}^n}{d_{ji}}, \quad q_5 = x_{c_{kj}} \cdot \sum_{n=1}^{11} \frac{d_{kj}^n}{d_{kj}}, \quad q_6 = x_{c_{ik}} \cdot \sum_{n=1}^{11} \frac{d_{ik}^n}{d_{ik}}. \tag{5.0.2}$$

Therefore, the ranking of each Huskie player within a triadic configuration (denoted $R_{between}^{(3)}$), measured by the betweenness centrality is given by,

$$R_{between}^{(3)} = \sum_{i=1}^{990} \left(\sum_{j=1}^{990} \left(\sum_{k=1}^{990} PI_{ijk} \left(q_1 + \cdots + q_6 \right) \right) \right) \tag{5.0.3}$$

where $PI_{ijk} = PI_i + PI_j + PI_k$ and each q_i is defined by Equations 5.0.1 and 5.0.2.

The triadic configurations, coupled with more data for specific match statistics, such as tackles and aerial duels win percentage, will help strengthen our model. We can also adjust it based on specific criteria. For example, we can use the model to predict a best lineup during a time frame within a match or matches, such as how the team performs during the 1st half of games. Another consideration could be the best lineup for home or away matches.

6 Conclusion

The dynamic nature of soccer allows us to consider multiple types of relationships between players. Using performance indicators to quantify specific player actions of importance towards the match result, as well as their ability to connect with other players, we can then rank players and lineups to best suit the team. From logistic regression analysis to centrality in graph theory, we can provide the coach with a predicted best lineup for a given game. The model can be adjusted to take various factors in to consideration, such as the best performing team in a specific time frame for the game (1st half vs 2nd half), location of the game (home vs away), as well as who they're playing against. There are some limitations as the visuals created are an average, which only provides a snapshot of the game(s) being analyzed and would be best paired with game film. Our model only uses data that accounts for the attacking abilities of players, but with the proper data can be adjusted for defensive capabilities as well.

7 Recommendations for Coach

One of the toughest decisions a coach needs to make is which formation and lineup is the best decision come game day. Our goal has been to develop a model to make that decision easier for you. The model takes into account the successful passing connections between configurations of two or three players. We also consider match statistics of players for specific performance indicators desired.

For our calculations, combined with the data provided to us, we only used the shots on target, crosses, and pass success rate as those have been found in a recent study to be significant indicators towards a team's chances of winning. Other statistics for players we would add with a better data set would be the Aerial Duel Win Rate (%) and successful tackles for each player.

In the study conducted on the 2014 FIFA World Cup, crosses were found to have a significantly, negative impact towards the match result for a team. This is often a popular approach to create a goal scoring opportunity, but often times can result in losing possession and opening up the team to a counter attack by the opponents. Looking at your past season, your team attempted 500 crosses, with 130 of them being successful, giving the team a 26% success rate on crosses. This means that the team should have more success by trying to play through the central players into the attacking area.

Of course this all depends on many factors for in the game, such as the opponent's strategies, but we believe we found the best formation and lineup for doing such. Looking over your matches from last season, we noticed that you set the team up in a 1-4-4-2 in

14 matches, and a 1-4-3-3 in 13. For the other 11 games you alternated between a 1-4-5-1, 1-5-4-1, and 1-5-3-2, but the 1-4-4-2 and 1-4-3-3 were the most successful.

Based on our calculations, we found that the 1-4-3-3 with players [G1, D1, D2, D3, D6, M1, M3, M6, F1, F2, F4] is the strongest lineup of the two formations. This only takes into account the attacking capabilities of the team, which means the defensive configurations will need to be looked at. If given the proper defensive statistics for players (aerial duel win % and completed tackles), then that would allow us to take that into account. Looking at the shot map for all of the team's opponents last season, we see that most shots were taken from central areas. By using the 1-4-3-3 formation, the midfielders will have to play in more central positions on the field, adding an extra player to defend in the center.

8 Appendix

PlayerID	Performance Indicator Values
Huskies_D1	3.18755
Huskies_D10	3.09866
Huskies_D2	5.13329
Huskies_D3	3.7241
Huskies_D4	-0.919253
Huskies_D5	-1.32594
Huskies_D6	2.15027
Huskies_D7	-1.26159
Huskies_D8	0.500829
Huskies_D9	3.71839
Huskies_F1	7.8025
Huskies_F2	4.98747
Huskies_F3	2.55827
Huskies_F4	9.34583
Huskies_F5	5.79349
Huskies_F6	2.45445
Huskies_G1	2.58127
Huskies_M1	5.65983
Huskies_M10	3.05657
Huskies_M11	3.28116
Huskies_M12	0.271005
Huskies_M13	3.01108
Huskies_M2	3.39059
Huskies_M3	3.11938
Huskies_M4	3.30237
Huskies_M5	3.60379
Huskies_M6	4.4622
Huskies_M7	4.05591
Huskies_M8	2.02748
Huskies_M9	2.12767

Figure 9: Performance Indicator Values for Huskie's Players

9 Glossary

Shot: an attempt to score a goal, made with any (legal) part of the body, either on or off target [7].

Shot on Target: an attempt to goal which required intervention to stop it from going in or resulted in a goal/shot which would go in without being diverted [7].

Pass: an intentional played ball from one player to another [7].

Long Pass: an attempted pass of 25 yards or more [7].

Short Pass: an attempted pass of less than 25 yards [7].

Corner: ball goes out of play for a corner kick [7].

References

[1] Q. Amarikwa. The 5 most popular soccer formations. *Perfect Soccer*, 2016.

[2] R. Beal and P. G. Ramchurn. Valuing player influence within teams. from statsbomb innovation in football, oct 2019. *Statsbomb*, 2019.

[3] T. I. F. A. Board. Law's of the game. *FIFA*, page 34, 2018.

[4] J. M. Buldu, J. Busquets, I. Echegoyen, et al. Defining a historic football team: Using network science to analyze guardiola's fc barcelona. *Scientific reports*, 9(1):1–14, 2019.

[5] M. Hasic. Learn popular formations with our illustrated guide. *Soccer Training Guide*, 2020.

[6] T. Knutson. Explaining xgchain passing networks. *Statsbomb*, 2018.

[7] H. Liu, M. Ángel Gomez, C. Lago-Peñas, and J. Sampaio. Match statistics related to winning in the group stage of 2014 brazil fifa world cup. *Journal of Sports Sciences*, 33(12):1205–1213, 2015. PMID: 25793661.

[8] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[9] F. A. Rodrigues. Network centrality: An introduction. *Nonlinear Systems and Complexity*, page 177–196, Jun 2018.

[10] D. B. West. *Introduction to Graph Theory*. Prentice Hall, 1996.