Problem Chosen	2020	Team Control Number
D	MCM/ICM Summary Sheet	**2011703**

Analysis of Teaming Strategies Based on Adjusted Passing Network and MLRM

As social problems and challenges become more complex, the importance of teamwork becomes increasingly prominent. In order to quantify and formalize the structural and dynamical features that have been successful for the team, based on data of Huskies from last season, we establish several models to solve the questions.

For task 1, we first construct a passing network of directed graph, whose node denotes the player and weight is processed weighted passing times. In order to determine the weight of the edges, apart from the number of passes, we comprehensively consider factors such as the distance of the pass, the type of pass and the player's playing time. Then we define the connection strength C_{ij} as the weight of edge E_{ij} and depict the passing network based on the average coordinate of each player. To identify network patterns, we apply **Warshall-Floyd algorithm** to compute the node intermediate $P(i)$ so that we can measure the influence of each node on complex networks. According to the product of the node intermediates of two nodes and connection strength of them, we classify these nodes into two categories using the **K-means clustering model**. Dyadic and triadic configurations and team formations are then determined by the thresholds. Additionally, we also consider the impact of changes in multiple scales and time on network properties.

For task 2, to identify performance indicators that reflect successful network, we introduce four types of indicators ($ADDI$, PDI, ACC, $GINI$) to reflect types of the play, coordination among players and distribution of contributions respectively. As for team level processes, we also select four kinds of metrics to describe the adaptability, flexibility, tempo and flow of the team. Afterwards, we normalize the data and use them as the independent variables while we choose the goal difference(GD) in every match as the dependent variable to establish a **multiple linear regression model($MLRM$)**. We use the weight of each indicators to measure the contributions of each factors towards successful teamwork performance. Meanwhile, to clarify the universality of our team strategies, we introduce the **Nash equilibrium model** in game theory, and consider the pure and mixed strategy Nash equilibrium to analyze the basis of strategy selection.

For task 3, We mainly instruct Huskies' coaches to determine effective structural strategies based on the most important index ACC in the multiple linear regression results, as well as the secondary TA, $ADDI$, $GINI$. Besides, we combine the indexes with network patterns, such as dyadic and triadic configurations. Then we provide the coach with feasible suggestions to increas the win rate in the next season.

For task 4, based on our findings in a controlled setting of a team sport, we conclude five factors which are beneficial to design more effective teams. Furthermore, we also take other three aspects of teamwork that may help us develop generalized models of team performance into account.

Key words: K-means clustering algorithm; multiple linear regression; Nash equilibrium; network pattern; team strategies

Contents

1 Introduction

1.1 Background

In light of the rapid growth of information technology, people are prone to get in touch with each other. Meanwhile, the society is becoming increasingly interconnected and sophisticated. Many studies indicate the network science can be a crucial tool to analyze the propagation of information among people. Team, as it consists of people with interdisciplinary backgrounds, is considered to be more creative and competitive to tackle complex problems. Team can be seen as a network model. If we apply network science to optimize the cooperation efficiency of the team, chances are that we could obtain more inception of the teamwork and higher odds of ultimate team success. Soccer is a typical team sport, the result of which is concerned with multiple factors. Team members shall stick to specific rules and employ their team skills to gain victory. We are seeking ways to find out the potential metrics, analyze their influence during a game and give scientific instructions for future soccer training.

1.2 Literature review

Generally speaking, the analysis of sports, especially soccer, has been a heated research topic for a long time. *Charles Reep* suggested that the key to scoring goals and winning games was to transfer the ball as quickly as possible from back to front already in the early 1950s.[1].In 1970s, *Gould* and *Gatrell* proposed the concepts of passing networks related to the football game through a structural analysis of a particular game.[2]But these researches did not arouse full attention until the release of *Duch's* work, where network science is used to unveil the process passing network affects the performance of a football team.[3]Looking at the recent studies, some focus on the individual performance quantification and its influence on the whole teamwork[4], some propose a data-driven approach, extract a set of pass-based performance indicators and summarize them as H indicator to simulate the real football games[5], while others take the temporal nature of the football network into account as well as emphasize the evolution of the network properties instead of merely the average.[6]

1.3 Our work

In our paper, we mainly focus on the following tasks:

- We will create the football passing network based on the data in Match 1 and 14, and apply **K-means clustering model** and **Warshall-Floyd algorithm** to identify network patterns. We will also discuss some strategy transformations over time qualitatively.

- We will first find the performance indicators and team level processes. Then we apply **multiple linear regression analysis** to construct the evaluating model. And we also intend to use **Nash Equilibrium** to discuss the universality of our team strategies.
- We will give several suggestions to the coach based on the indicators and models in the previous section.
- We will generalize the characteristics of an effective team and then come up with other aspects which affect the team performance.

1.4 Assumptions

Our models are based on the following assumptions:

- The distance of passing is the Euclidean distance between the passer and the receiver at two points in the two-dimensional space, regardless of the special case where the trajectory is an arc, and the speed of passing is also based on the simplified linear distance divided by the time difference.
- In the network diagram of passing, the player's position on the field is based on the arithmetic mean value of the set of points where the player is in the coordinate position when participating in passing events in a match, while the player's position in other situations such as duel events is not included in the point set.
- The formation of all teams is based on the type of player that starts with 11 players. If there are four defenders, four midfielders and two strikers in the starting line-up, the formation is classified as 4-4-2, regardless of the deeper divisions such as 4-2-2-2 that are common in official matches.
- The performance of the team only considers the team's own factors, ignoring the impact of different opponents on the team's performance.
- We ignore the impact of changing coaches on the team's adaptation to the new tactics.
- Other assumptions for one particular model will be introduced when needed.

2 Task 1: Passing network model

Task One requires our team to create a network showing the balls passed between players, and then identify the network patterns including dyadic as well as triadic configurations according to the model designed. Additionally, we are supposed to find out the tactical choice transformations of the players on the field. Firstly, we construct the passing network on the basis of match 1 and 14. Our chief job is to calculate the weights of each edge in the network, and derive the formula of the connection strength. Secondly, we use **K-means clustering model** to identify the directed edges, which helps us classify them into two patterns: dyadic and triadic configurations. Last but not least, we will qualitatively analyze the strategies changes over time. By the way, we only consider Huskies in this section.

Before the model construction, we first claim our notations used in this section.

Table 1: Notations of Task 1

Symbols	Definition
α	The weight of different passing styles
D_{ijk}	The k^{th} passing distance from node i to j
H_{ijk}	Adjusted Pass Difficulty Factor
C_{ij}	The connection strength from node i to j
S_i	Normalization coefficient due to substitution
E_{ij}	Edges from node i to j
$P(i)$	The node intermediate of node i
f_i	The frequency of player i showing on the field
$(\bar{x_i}, \bar{y_i})$	Average coordinate for player i
d_{mn}	Numbers of the shortest path linking node m and n
$d_{mn}(i)$	Numbers of the shortest path through node i

2.1 Establish the passing network

In this section, we first emphasize the weight of each edge. To determine the value of these weights, apart from **the frequency of passes**, we also take **passing styles** and **passing distance** into consideration. We add up the different line-ups of Huskies in each match, where we discover 4-3-3 (i.e. the number of defense, midfield and forward players is 4, 3 and 3 respectively, same as below) and 4-4-2, are the most typical formations used by Huskies (shown in Figure 1). Thus, we primarily extract the data from **Match 1 and 14** played by Huskies whose line-ups are exactly 4-3-3 and 4-4-2 respectively. Afterwards, we will compute the connection strength between two nodes and depict the network graph with weights.

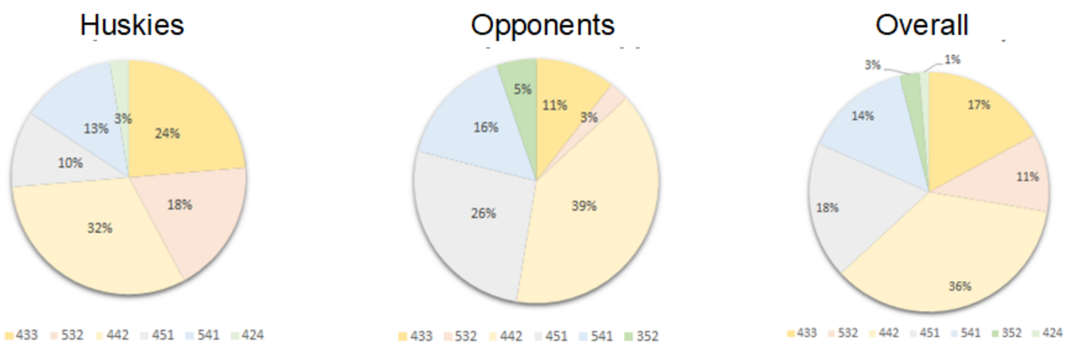

Figure 1: The Line-ups

2.1.1 Construction process

- **Calculate the average coordinate of the eleven starters' nodes**

  We first figure out the original players in each match. As assumed, we only take the first eleven different players shown on the field into account. By programming, we locate the starters in Match 1 and 14 and list them in Table 2.

Table 2: The starters in Match 1 and 14

Match 1	M1	M2	D1	G1	D2	D3
	D4	M3	F2	F3	F1	
Match 14	F2	F1	M6	D5	D2	G1
	D6	M8	M1	M4	D7	

  As for each player, they are located by their coordinate x_i and y_i on the field, where x_i denotes the vertical position from the perspective of Huskies and y_i denotes the horizontal position. Particularly, when x_i equals to zero, it means Huskies' goal. And y_i indicates the left-hand side when it reaches zero. We define $(\bar{x}_i, \bar{y}_i)$ as the average coordinate for each player i and f_i indicates the times players i shows on the field. Hence, we can work out the formula of the average coordinate.

$$\begin{aligned} \bar{x}_i &= \frac{\sum_{i=1}^{n} x_i}{f_i} \\ \bar{y}_i &= \frac{\sum_{i=1}^{n} y_i}{f_i} \end{aligned} \tag{1}$$

- **Compute the passing distance each time**

  In this step, we have to count the passing distance of the Huskies every time. Here we add up all the passes, including the failed ones. Supposing

D_{ijk} denotes the k^{th} passing distance from node i to j, we can get it from the coordinate (x_{ki}, y_{ki}) of the players.

$$D_{ijk} = \sqrt{(x_{kj} - x_{ki})^2 + (y_{kj} - y_{ki})^2} \tag{2}$$

- **Reckon the difficulty coefficient of the pass**

  The difficulty of one pass is concerned with the passing distance and the passing styles. Since different passing styles vary in difficulty and diversity, we distribute them with different weight α. Here we suppose the weight of a simple pass is 1, the defined weight of other passing styles are listed in Table 3.

Table 3: Weight of Different Passing Styles

Passing Styles	Weight α	Passing Styles	Weight α
Head Pass	2	High Pass	2
Launch	3	Cross	3
Hand Pass	4	Smart Pass	4

Based on the weight, we are accessible to define the passing difficulty coefficient H_{ijk} as follows:

$$H_{ijk} = \alpha \times D_{ijk} \tag{3}$$

- **Derive the formula of the connection strength**

  Considering the existence of substitution in the match, we define a variable S_i to eradicate the deviation. In a 90-minute match, the standardized coefficient S_{ij} can be defined as

$$S_i = \frac{90}{t_s} \tag{4}$$

Here the variable t_s denotes the substitution time. Then we can define the **connection strength** C_{ij}. It means how strong the two nodes i and j is linked, which is related to the passing styles, passing distance and the normalization coefficient due to substitution. Therefore, we can derive the formula of the connection strength C_{ij}.

$$C_{ij} = S_i \times \sum_{k=1}^{n} H_{ijk} \tag{5}$$

The coefficient C_{ij} describes the coordination of two players. As defined, the higher C_{ij} is, the stronger node i and j is connected, and the better player i and j cooperates.

- **Construct connected directed graphs**

  We utilize C_{ij} as the weight of the edge E_{ij} and construct two connected directed graphs. Each node is depicted as a circle, the size of which represents the total weights of the node. And we use the width of connection line to denote the connection strength of each node. Besides, color bar is employed to picture the weights of each connection line. The weights increase as the color gets darker.

2.1.2 Model solution

With the help of Matlab, we visualize the two pass-through network graphs as follows:

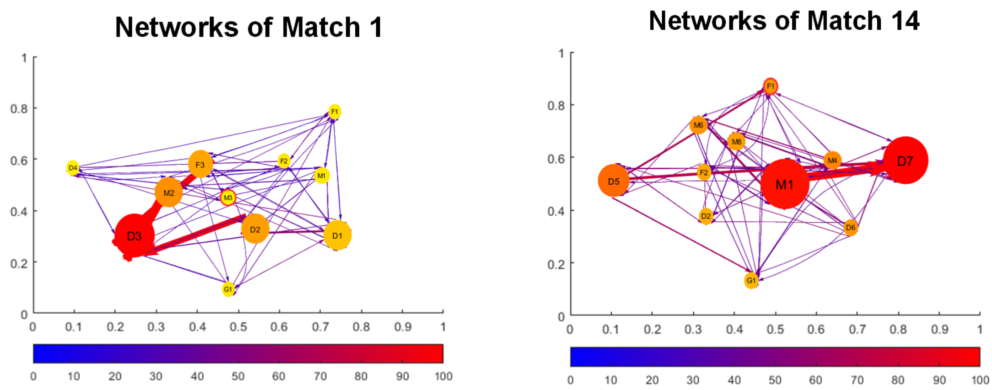

Figure 2: The network of Match 1 and 14

2.2 Identify network pattern

In this part, we choose **K-means clustering model** to identify the network pattern. To begin with, we define the **node intermediate** $P(i)$, which is the proportion of the number of nodes that pass through the shortest path at any two nodes in the network. It measures the influence of nodes in the entire complex network and reflects the influence of nodes on the information of the network. We conclude the formula below according to the definition

$$P(i) = \sum_{m \neq n \neq i} \frac{d_{mn}(i)}{d_{mn}} \tag{6}$$

where d_{mn} denotes numbers of the shortest path linking node v_i and v_j. To compute the node intermediate, we have to calculate the shortest path. Here we use the **Warshall-Floyd algorithm**. We define the **reciprocal** of the connection strength C_{ij} as the distance connecting node i and j. With the help of Matlab, we can get $P(i)$ of the 11 nodes.

Table 4: The the node intermediate of each player in Match 1 and 14

Match 1	PlayerID	M1	M2	D1	G1	D2	D3
	$P(i)$	23	11	21	18	16	26
	PlayerID	D4	M3	F2	F3	F1	
	$P(i)$	12	10	21	10	14	
Match 14	PlayerID	F2	F1	M6	D5	D2	G1
	$P(i)$	20	13	18	16	15	21
	PlayerID	D6	M8	M1	M4	D7	
	$P(i)$	16	10	10	11	17	

Afterwards, with the help of SPSS, we apply K-means algorithm based on connection strength C_{ij} and product of the node intermediate $P(i) \times P(j)$. The results of the clustering are shown below:

Table 5: The results of the clustering

	1	2
C_{ij}	11.540037	26.476698
$P(i) \times P(j)$	100	598

According to the results, we categorize the edges into strong connected directed edges and weak connected directed edges. And we select the strong connected directed edges as the basis of the network patterns' classification. So we connect strong connected directed edges in a scatter plot showing the players' position in Match 1 and 14 and picture it in Figure 3.

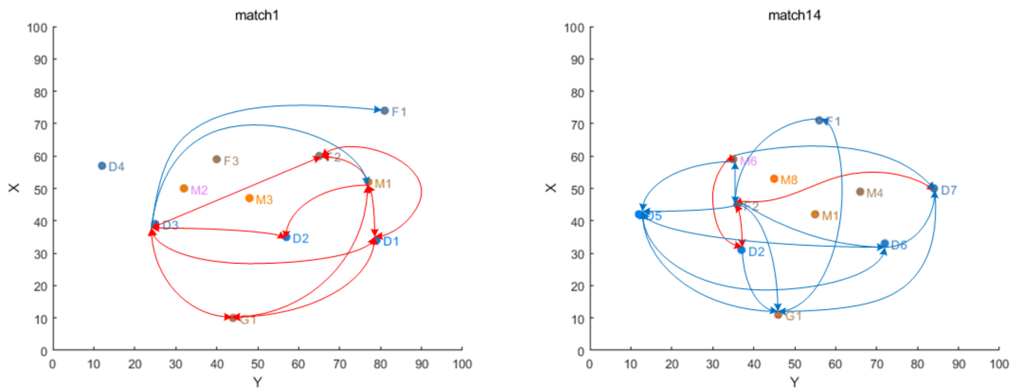

Figure 3: The strong dircted graph in Match 1 and 14

From the graph, we can conclude that if the number of strong edges between three points reaches six, then the structure is triadic. And if the number of strong edges between two points reaches two, the network is a dyadic configuration.

2.3 Other network properties

In this section, we focus on the transformations of the players' tactical choice. First, we add up the number of duels, fouls, shots and passes in every five minutes, which is depicted in Figure 4 .

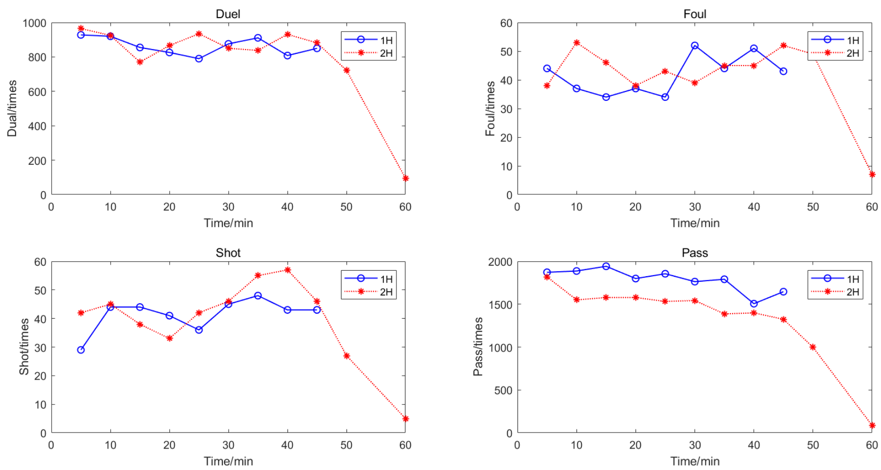

Figure 4: The number of duels, fouls, shots and passes in every five minutes

From the line graph we can see the passes differ greatly between the first and second half. The number of passes during the first half is consistently larger than that during the second pass, which indicates the physical exertion of the players is growing considerably as the game continues. Moreover, it is obvious that the number of shots reaches its peak at approximately the end of the game, proving the tension and fierce when the match is about to end. It can also reveal that players are becoming increasingly adaptable to their opponents. Although the number of duels and fouls is fluctuating all the time, we can still conclude that players are more likely to foul and duel during the second half.

Afterwards, we count the average coordinate $(\bar{x}_i, \bar{y}_i)$ for player i of Match 1 and 14 during the first and second half respectively. The results are pictured in Figure 5.

As shown, in Match 1, the Huskies uses the 4-3-3 formation. From the change of the coordinates, especially the transformation of $\bar{x}_i$, we can easily reach the conclusion that the players are more radical because most of the $\bar{x}_i$ during the second half is larger than that during the first half, which means they are getting closer to the opponents' goal. In contrast, in Match 14, the line-ups of Huskies is 4-4-2, where they choose a relatively conservative game style.

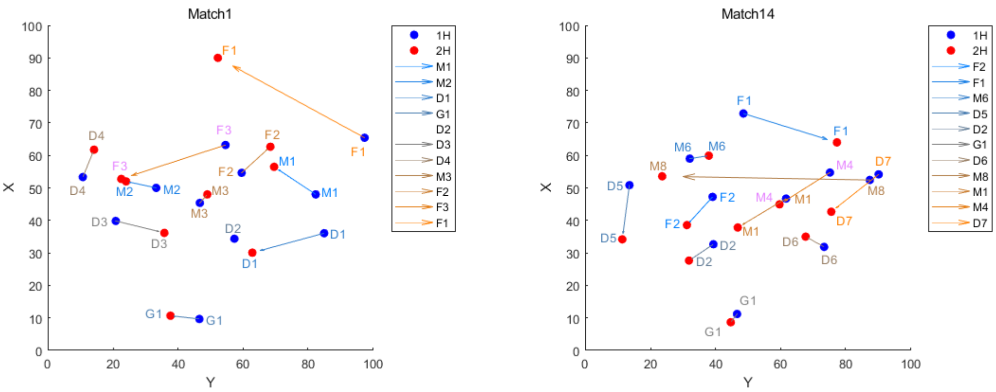

Figure 5: The coordinate change of Match 1 and 14

3 Task 2: Evaluation index

In this section, we are working on the performance metrics that reflect successful teamwork. We first determine the evaluating indicators based on the assessment of the performance and the team level processes. Then we apply **multiple linear regression analysis** to establish the evaluating model, which captures structural, configurational, and dynamical aspects of teamwork. Finally, we will use **Nash Equilibrium of Complete Information Static Game** to discuss whether the tactics concluded from our assessing systems are universally effective. As assumed, unless otherwise specified, our research project is Huskies.

3.1 Performance indicator

As for the performance of the team, we primarily consider three aspects: diversity in the types of the plays, coordination among players and distribution of the contributions. The specific metrics of them are depicted in Figure 6.

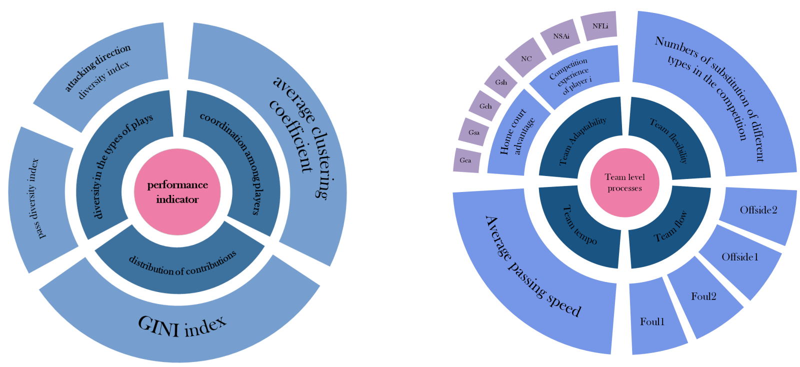

Figure 6: The performance indicator and the team level processes

3.1.1 Diversity in the types of plays

- **Attacking Direction Diversity Index (ADDI)**
  ADDI is used to evaluate the diversity of the attacking directions. We assume the falling point of the goalkick as the starting area of the attack. Then we divide the field into left, middle and right area at a ratio of 3:4:3 in width. Here we define the standardized attack direction is 0.3,0.4 and 0.3 respectively. After that, we are able to calculate the deviation of the actual ratio falling in three areas from the standard.

$$ADDI = \sqrt{(LAP - 0.3)^2 + (MAP - 0.4)^2 + (RAP - 0.3)^2} \tag{7}$$

  Here LAP denotes left field percentage, MAP denotes middle field percentage and RAP denotes right field percentage.

- **Pass Diversity Index (PDI)**
  PDI focuses on the different types of passes. We define PDI as the ratio of the non-simple pass.

$$PDI = 1 - \frac{SimplePass}{Pass} = \frac{NonsimplePass}{Pass} \tag{8}$$

  The diversity of the passes increases as the PDI grows larger, which means the cooperation approaches between teammates are more diverse.

3.1.2 Coordination among players

- **Average Clustering Coefficient (ACC)**
  We defined C_{ij} as the connection strength from node i to j, which reflects the cooperation quality of player i and j. Here we introduce a new variable $C_w(i)$ to weigh the agglomeration of a node in a particular area:

$$C_w(i) = \frac{\sum_{j,k} C_{ij}C_{jk}C_{ik}}{\sum_{j,k} C_{ij}C_{jk}} \tag{9}$$

  where j and k denotes the two players of the team and i represents the third player. The weighted clustering coefficient $C_w(i)$ measures the likelihood that neighbors of a given player i will also be connected with player j and k.[7] Therefore, we use the arithmetic mean of clustering coefficient to describe the starters' degree of coordination.

3.1.3 Distribution of the contributions

- **Gini coefficient (GINI)**
  We use the clustering coefficient to define the contributions of players. To measure the distribution of these contributions, we introduce Gini coefficient. It is a measure of statistical dispersion intended to represent the

income or wealth distribution of a nation's residents, and is the most commonly used measurement of inequality.[9] The Gini index ranges from 0 to 1, and the degree of unevenness the clustering coefficient distributes increases monotonically with it. There are many ways to compute the Gini coefficient. Here we use the following formula:

$$GINI = \frac{1}{2n^2u} \sum_{j=1}^{n} \sum_{i=1}^{n} |C_w(j) - C_w(i)| \tag{10}$$

where n denotes the number of the nodes and u denotes the average value of $C_w(i)$ for $i = 1, 2, \cdots, n$.

3.2 Team level processes

At the team level, we mainly focus on four aspects: team adaptability, team flexibility, average passing speed and the relative fluency. The indicators related to each factor are pictured in Figure 6.

- **Team adaptability(TA)**

  To specifically analyze the adaptability of the team, we select competition experience of player i and home court advantage as our starting point.

  Competition Experience of player i
  The team consists of multiple players, but there are only 11 players who can appear as first lineup which is crucial in the match. Players shown as the first lineup is said to have more experience and skills than the substitute. To weight the experience of each player, we add up the number of first lineup, substitute appearance and competitions in the past year for each member of the team. For narrative convenience, we list the related notations in Table 6.

Table 6: Notations in $CE(i)$

Symbols	Definition
$NFL(i)$	Number of first lineup for player i
$NSA(i)$	Number of substitute appearance for player i
NC	Number of competitions
$CE(i)$	Competition experience of player i

We tend to believe the $CE(i)$ is concerned with $NFL(i)$, $NSA(i)$ and NC. We conclude the relation with the following function:

$$CE(i) = \frac{NFL(i) + w_c \times NSA(i)}{NC} \tag{11}$$

where w_c denotes a weight coefficient, ranging from 0 to 1.

Home court advantage
Since the team have to compete both at home and away, we need to count the advantage when they play at home to describe their adaptability. In this part, we first define the factors we need and list them in Table 7.

Table 7: Notations in HCD

Symbols	Definition
GSH	Goals scored at home
GCH	Goals conceded at home
GSA	Goals scored away
GCA	Goals conceded away
HCD	Home court advantage

Then we use these factors to define HCD:

$$HCD = \frac{GSH/GCH}{GSA/GCA} \tag{12}$$

As defined, the indicator HCD describes the superiority of players played at home. If HCD is high, it means players score more at home than away, indicating their adaptability is higher than those with lower HCD.

Team adaptability
As discussed above, we can derive the formula of team adaptability using the indicators defined:

$$TA = \frac{\sum_i^{11} CE(i)}{11} \times HCD \tag{13}$$

where we define TA as the adaptability of the team. The formula illustrates that the higher HCD is, the better their adaptability is. Also, the average of $CE(i)$ describes the average experience of the team, whose growth will lead to the increase of team adaptability.

- **Team flexibility(TF)**

  Numbers of substitution of different types in the competition(NDUSB)
  We count the substitution frequency of different types to evaluate the flexibility of the team. The substitution of different types always means the change of formations, which reflect how flexible a team is when they are confronted with different opponents. Here we define the team flexibility(TF) as follows:

$$TF = e^{\frac{NDUSB}{3}} \tag{14}$$

- **Average passing speed(TP)**

  Passing speed reflects the coordination and tempo between the teammates. Generally speaking, if the passing speed of a team is high, the team have better control of their pace. Thus, their performance is more likely to be better and their team is more united. Note that, the passing speed we discuss here refers to the speed of successful passes. We define a successful pass that the original sending team of the ball are the same as the receiver's team. Suppose the original and destined coordinate of one pass is (x_{i0}, y_{i0}) and (x_{i1}, y_{i1}) and the passing event happens at $t(i)$, and we can compute the average passing speed TP.

$$TP = \frac{\sqrt{(x_{i1} - x_{i0})^2 + (y_{i1} - y_{i0})^2}}{t(i+1) - t(i)} \tag{15}$$

  The equation holds only when $TP = TR$ (i.e. TP denotes the team of passing and TR denotes the team of receiving).

- **The relative fluency(RF)**

  The relative fluency(RF) indicates the flow of the team. We calculate the reflective fluency using four indicators listed below:

Table 8: Notations in RF

Symbols	Definition
$Foul1$	Numbers of fouls by our side
$Foul2$	Numbers of fouls by the opponent
$Offside1$	Numbers of offside by our side
$Offiside2$	Numbers of offside by the opponent

So we compute the relative fluency by the following formula:

$$RF = \frac{Foul2 + Offside2}{Foul1 + Offside1} \tag{16}$$

3.3 Establish the evaluating model

Based on the indicators above, we get the $ADDI$, PDI, ACC, $GINI$, TA, TF, TP and RF from the 38 matches. In order to facilitate subsequent data processing, we normalization is performed on each data[8]. The formula of the normalization is

$$x' = \frac{x - minx}{max(x) - min(x)} \tag{17}$$

Therefore, we can acquire the processed data $ADDI'$, PDI', ACC', $GINI'$, TA', TF', TP' and RF'. All of them range from 0 to 1, the maximum and minimum number of which are 1 and 0 respectively within the 38 matches. Next, we view these numbers as independent variables and set the goal difference(GD) of Huskies in 38 matches as the dependent variable. Then we can get the following multiple linear regression model.

$$GD = b_0 + b_1 ADDI' + b_2 PDI' + b_3 ACC' + b_4 GINI' + b_5 TA' + b_6 TF' + b_7 TP' + b_8 RF' \tag{18}$$

With the help of Matlab, we get the following results listed in Table 9.

Table 9: The outcome of multiple linear regression

Parameter	Parameter estimates	Parameter confidence interval
a_0	-3.4412	[-7.2152, 0.3328]
a_1	-1.3564	[-3.8588, 1.1459]
a_2	-0.5596	[-4.2216, 3.1023]
a_3	6.1793	[2.2067, 10.1519]
a_4	-1.2864	[-4.2743, 1.7015]
a_5	1.4394	[-0.1083, 2.9871]
a_6	0.0198	[-1.7803, 1.8199]
a_7	0.8258	[-1.7786, 3.4301]
a_8	-0.7821	[-3.1875, 1.6232]
R^2=0.5337	F=3.4311	p=0.0090 s^2=1.8842

Theoretically speaking, $ADDI'$ represents the the deviation degree from diversity and $GINI'$ indicates the degree of uneven contribution distributions. So the lower they are, the more diverse the team plays, and the more even the contributions distribute, that is, the more successful the teamwork is. And the teamwork is better when other indicators are higher. According to the calculation results, we can find that the two weights b_1 and b_4, which are the two with the largest negative values, meet the requirements of lower $ADDI'$ and $GINI'$ mentioned earlier. Meanwhile, it is obvious that b_3 reaches over 6, which far exceeds b_5, the second largest one. It indicates that ACC is a effective indicator to reflect the successful teamwork. b_5 equals to 1.4394, indicating that TA is also of significance. Other values have no considerable effect on the prediction of team performance.

Strangely, b_2 and b_8 are negative, indicating that the team's variety of passing types and the number of fouls cannot explain the strengths and weaknesses of teamwork. A reasonable explanation is that penetration of a large number of short passes and the reasonable use of tactical fouls can help the ball The team achieve better performance.

3.4 Universality of the strategies

In this section, we will use Nash Equilibrium of Complete Information Static Game[10] to clarify whether strategies are universally effective or dependent on opponents counter-strategies. We first clarify the definition of certain conception.

- RWR_{ij}:relative win rate between i and j. We can acquire it from the history data.
- Using frequency φ: We define using frequency φ as the the ratio of tactics i used by coach $a(N_{ai})$ and coach a's total coaching sessions(N_{total}).

$$\varphi = \frac{N_{ai}}{N_{total}}(N_{ai} \geq 5) \tag{19}$$

- EXP_{ai}: the proficiency of coach a. It increases exponentially with the using frequency φ, the formula is listed below.

$$EXP_{ai} = e^{\varphi} \tag{20}$$

In this game, the player is the coach of the two teams that are about to compete in the football match. Let us be the player 1, and the opponent is the player 2. Since the game is static, it is assumed that there is no sequence in the choice of actions of the two parties, but that decisions are made simultaneously. The behavior space of each player is the choice of tactics and strategies.

To simplify the problem, we assume there are only two strategies for each player. The strategy for player 1 S_{1i} $(i \in \{1,2\})$ and that for player 2 S_{2j} $(j \in \{1,2\})$ are mutually restrained. That is, based on the historical win-loss relationship of the tactics, if there is an advantage of S_{11} over S_{21}, there must be a disadvantage of S_{11} over S_{22}. Otherwise, if there is a tactic of one player S_{ij} , which is relatively restrained against all tactics of the other, we can deduce that the strategy is the best, which is universally effective. We have defined the relative win rate RWR_{ij} to describe the win rate of strategy S_{1i} in all S_{1i} and S_{2j} confrontations in a match above a certain level in a certain period of time. If $RWR_{ij} > 1$, then we can assert S_{1i} relatively restrain S_{2j}. We also defined the coach a's proficiency EXP_{ai}. The higher EXP_{ai} indicates that coach a is more adept in strategy S_{ai}.

Therefore, we can get our theoretical gain U_{ij} when when we use tactics S_{1i} and our opponents use tactics S_{2j}.

$$U_{ij} = RWR_{ij} \times e^{F_{1i} - F_{2j}} \tag{21}$$

Since the outcome of a football match is a zero-sum game, the theoretical opponent's theoretical gain is $-U_{ij}$. The benefit matrix is shown below.

When there exists a strategy combination (S_{1i}, S_{2j}) , for player 1,

$$U_{ij} \geq U_{kj}, (\forall k \in \{1,2\})$$

Table 10: The benefit matrix of the football match

Our side \ Opponent	Y_1	Y_2
X_1	$(U_{11}, -U_{11})$	$(U_{12}, -U_{12})$
X_2	$(U_{21}, -U_{21})$	$(U_{22}, -U_{22})$

and for player 2,

$$-U_{ij} \geq -U_{ik} (\forall k \in \{1, 2\})$$

then we call (S_{1i}, S_{2j}) a Nash equilibrium of G. At this time, according to the consistent predictability of the Nash equilibrium, on the premise that both coaches are economically rational people, neither part of the game will deviate from this prediction result. So the prediction result will become the final outcome of the game.

When there is no pure strategy Nash equilibrium as described above, and the possibility of optimal strategy equilibrium is ruled out, the player 1 will randomly choose tactics S_{a1} and S_{a2} among the two optional strategies with probability distribution

$$P_a = (P_{a1}, P_{a2})$$

this situation is called a "mixed strategy", where

$$0 \leq P_{ai} \leq 1$$

is true for $i = \{1, 2\}$, and

$$P_{a1} + P_{a2} = 1$$

According to the idea of mixed strategy selection, in order to prevent player 2 from anticipating their own preferences, the player 1 will choose the idea of equalizing the expected benefits of the other two strategies to determine the probability distribution of his own strategy selection. For the counterpart, the probability distribution is determined by satisfying

$$U_{11} \times P_{21} + U_{12} \times P_{22} = U_{21} \times P_{21} + U_{22} \times P_{22} \tag{22}$$

For our side, it is satisfied

$$-U_{11} \times P_{11} - U_{21} \times P_{12} = -U_{12} \times P_{11} - U_{22} \times P_{12} \tag{23}$$

to determine the probability distribution.

All in all, to discuss the universality of our strategies, we are supposed to consider the relative win rate RWR_{ij} between tactics i and j. Additionally, the EXP_{ai} of each coach to their potential use of tactics have to be taken into account to determine strategies of our side.

4 Task 3: Suggestions for coach of Huskies

In task 3, we are required to inform the coach the effective team strategies for Huskies using insights acquired from our model constructed. Then we should give the coach some suggestions to help improve their team success in the next season.

According to the results of the multiple linear regression computed above, we find that the value of average clustering coefficient (ACC) is extremely crucial for the success of the successful teamwork. Hence, the coach first have to attach great importance to improving the ACC value of the team in each match. To improve the value of ACC, we give two feasible ideas:

- **The first lineups should be chosen from players with high ACC as many as possible.** Here we calculate the ACC value for each player and list them in the appendix.

  Also, we calculate the average ACC for each type of players shown in Table 11.

  Table 11: Average ACC for each type of players

Types	Average ACC
Defense	35.22
Midfield	35.46
Forward	36.56

  We view these three average values as thresholds of ACC for each type, and players whose ACC are above the threshold should be chosen as starters more in the next season. For defense, we suggest coach should promoteD_1, D_2, D_3 and D_5 more in the next season. For midfield, players M_2, M_8, M_9 and M_{12} are better options for the first lineups. Similarly, coach need to pay more attention to the forward players including F_2 and F_3. Moreover, we find the Huskies is weak in the forward players because the eligible forward players appear fewer than the other two types of players. Therefore, coach are supposed to **strengthen the training for forward players.**

- **The coach should use first lineups with higher ACC as many as possible.** It can be seen from the historical data of 38 matches that Huskies' GD and ACC are basically positively correlated. From the data, we find that Huskies' priority for the formation next season should be 4-3-3, 5-4-1, 4-4-2, 4-5-1, 5-3-2, 4-2-4, basically accord with the order of ACC value

However, at the same time, the coach must **take Nash equilibrium-based behavioral choices into consideration**, because the excessive abuse of the formation of high ACC values would lead to counter-tactics targeted by opponents.

In addition, at the level of team level processes, TA's contribution to teamwork also plays an extremely important role. Therefore, TA **must also be expanded** through two aspects. One is to discover young players with high ACC values during normal training or competition. Provide him with more playing opportunities to appear as starters to enrich his competition experience. The second is to attract more fans to cheer at the game site by recruiting stars or reducing ticket prices and other commercial operations so as to **strengthen their home court advantage**.

The results of multiple linear regression analysis also revealed the negative effects of a single offensive organization direction and uneven distribution of ACC contributions. According to Huskies s left-center-right three-way offensive organization ratio last season, which is $16 : 43 : 41$, we can easily get that the team needs to **reinforce the relatively weak lineup on the left in the transfer window**, and at the same time **enrich the tactics of the left to organize attacks**. On the other hand, we must **avoid rely too heavily on one or two players** to prevent the surge of the team's $GINI$ coefficient.

Finally, before the arrival of the new season, we must also **pay more attention to the dyadic and triadic configurations** between the small groups of players, just like the top midfielder of the Barcelona / Spanish national team Xavi plus Iniesta and Busquets. Pay close attention to the "chemical reaction" between some players and avoid breaking up their combination in the main match.

5 Task 4: Build more effective teams

In this section, we will first generalize our findings to give a brief view of designing more effective teams. Then we will discuss other aspects of teamwork to develop generalized models of team performance.

We think an effective team should have the following characteristics:

- **Clear, measurable team goals**
  A team must first understand the meaning of its existence, which is a common goal recognized by all members. Football embodies many elements perfectly and centrally, but its final measure is extremely simple and clear (this is why we focus on the goal difference of the multiple linear regression variable). The team goals indicate the work of its members The final direction.

- **Fair internal competition and reasonable distribution of contributions**
  Although an efficient team will not annihilate the demands of personal interests, it does not mean that personal interests can take precedence over collective interests. The best team is definitely not a single team. Only when the contribution is reasonably distributed to each team Only when you have a member (such as preventing the $GINI$ coefficient from being

too high) can you really strengthen the team's cohesion.

- **Flexible organization with small teams**
  Just as there are dyadic and triadic configurations on the team, in a small team, the collaboration is effective when everyone fully understands the other team members.

- **Reasonable division of duty and sharing awareness**
  An efficient team can timely transfer resources, knowledge and information among team members to help achieve the team's common goals. In a complex network, nodes with a large number of nodes play a vital role in the information flow of the entire network. Meanwhile, just like the roles of forwards, midfielders, defenders and goalkeepers on the court, in an efficient team, each member knows his own division of functions and performs efficiently.

- **Strengthen a sense of belonging and encourage innovation**
  In order to enhance the adaptability of members, on the one hand, it is necessary to pay attention to individual needs. A player who cannot play for a long time is more likely to feel marginalized; on the other hand, it is necessary to provide an effective working environment, such as building a home court advantage for the team. At the same time, innovation is the primary driving force for development. Football's tactical style requires diversification of play and timely change, and a team that adheres to rules will eventually be eliminated.

Afterwards, we are thinking about other aspects which may influence the team performance.

- **Leadership and management level**
  A leader is a person who plays a leading role in the team. An excellent leader is not only good at influencing the thinking of subordinates, promoting decision-making and ensuring the execution of plans, but also possessing enough personality charm to play an exemplary role.

- **Team stress resistance**
  Just like the alternation of backward and leading in a football match, a team will inevitably encounter various ups and downs in its own development. Maintaining stability in good times and facing difficulties in adversity is an ideal form of team performance.

- **Impact of team culture**
  Team culture has a strengthening effect on the collective awareness of team members, which can promote the effective operation and development of the team and improve the overall effectiveness of the organization.

6 Strengths and weaknesses

6.1 Strengths

- **Sufficient consideration is given to the size of the connection strength of each edge in the pass-through network diagram.** In addition to the number of passes, the difficulty of the different types of passes and the impact of the distance of the passes are also taken into account, while the impact of the substitution of the starter is excluded.
- **A reliable formation identification standard is established.** The clustering model is used to classify the directed edges and dyadic and triadic configurations are obtained. The internode product and the connection strength C_{ij} of the directed edge are taken as the reflection of individual influence and node local aggregation degree respectively, and the role of individual and the connection between nodes are comprehensively considered.
- **Set up the parameters diversity of performance indicators and team level processes.** It covers eight elements and matches the parameters well.
- **Using the commonly used Gini coefficient in economics to measure the equilibrium degree of contribution distribution** is a great innovation.
- **The introduction of game theory**, through the prediction of each other's behavior to optimize the strategy, is in line with the rational choice.

6.2 Weaknesses

- The data of multiple linear regression are all from the single-season matches of huskies' team, and the weight coefficient of evaluation index may lack some universality.
- Due to the lack of time and the lack of availability of some data, most of the indicators are based on the status of players with the ball, and fail to consider the impact of players' movement without the ball and the position of opponents.
- There is no discrimination in the evaluation of individual players. And our models lack different criteria for different types of players.

References

[1] C. Reep, B. Benjamin, "Skill and chance in association football", Journal of the Royal Statistical Society, vol. 131, pp. 581-585, 1968

[2] Gould, P. and Gatrell, A. A structural analysis of a game: The Liverpool vs manchester united cup final of 1977. Soc. Netw. 2, 253273, 1979

[3] Buldú, J. M. et al. Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. Front. Psychol. 9, 1900, 2018

[4] Duch, J., Waitzman, J. S. and Nunes Amaral, L. A. Quantifying the Performance of Individual Players in a Team Activity. PLoS ONE, 5(6), 17, 2010

[5] P. Cintia, F. Giannotti, L. Pappalardo, D. Pedreschi and M. Malvaldi, "The harsh rule of the goals: Data-driven performance indicators for football teams," 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, pp. 1-10, 2015

[6] Buldú, J.M., Busquets, J., Echegoyen, I. et al. Defining a historic football team: Using Network Science to analyze Guardiolas F.C. Barcelona. Sci Rep 9, 13602, 2019

[7] Ahnert, S. E., Garlaschelli, D., Fink, T. M. A. and Caldarelli, G. Ensemble approach to the analysis of weighted networks. Phys. Rev. E 76, 016101, 2007

[8] Fenghong Xie, Dawei Zhang, Dan Huang, Fuding Xie.Text keyword extraction based on weighted complex network[J].Systems Science and Mathematics, 30(11):1592-1596, 2010

[9] http://en.wikipedia.org/wiki/Gini_coefficient

[10] http://en.wikipedia.org/wiki/Nash_equilibrium

Appendices

- **Main Matlab codes**
  Here are simulation programmes we used in our model as follow.

```
%calculate Clustering Coef?cient
clear;clc;
N = 11;

Sij_1 = 46; %substitution time
Sij_2 = 61;
Sij_3 = 83;
N1 = 3;
N2 = 10;
N3 = 8;

Cij = sumHijk('38');
Cij(N1,:) = Cij(N1,:)*90/Sij_1;
Cij(N2,:) = Cij(N2,:)*90/Sij_2;
Cij(N3,:) = Cij(N3,:)*90/Sij_3;

t = Cij;
for i = 1:N
    for j = 1:N
            Cij(i,j) = t(i,j) + t(j,i);
    end
end

for i = 1:N
    numerator = 0;
    denominator = 0;
    for j = 1:N
        for k = 1:N
            numerator = numerator + Cij(i,j) * Cij(j,k) * Cij(i,k);
            denominator = denominator + Cij(i,j) * Cij(i,k);
        end
    end
    cw(i) = numerator/denominator;
end
cw
```

```
%% Initialization environment
close all
clear
clc

%% read and process data
coord = xlsread('coordinate.xlsx');

for i = 1 :11
    x(i) = coord(i,1);
    y(i) = coord(i,2);
end
```

```
%substitution time
Sij_F2 = 85;
Sij_M8 = 89;
Sij_F1 = 90;

Cij = sumHijk('14');
Cij(6,:) = Cij(6,:)*90/Sij_F2;
Cij(11,:) = Cij(11,:)*90/Sij_M8;
Cij(5,:) = Cij(5,:)*90/Sij_F1;
cm = Cij;

%% establish a gragh with weight
IDS={'M1','D7','F1','M4','D2','D5','D6','G1','F2','M8','M6'};
bg=biograph(cm,IDS);
n = length(bg.edges);

for i = 1:n %the size and color of the edges are determined by Cij
    c = sqrt(sqrt(bg.edges(i,1).Weight));
    bg.edges(i,1).LineWidth = bg.edges(i,1).Weight/50;
    mycolormap = [0:0.002:1;zeros(1,501);1:-0.002:0]';
    count(i) = c;
    mycoloridx(i) = round(500 * count(i)/ max(max(sqrt(sqrt(Cij)))) );
    mycoloridx(mycoloridx<1) = 1;
    bg.edges(i,1).LineColor = mycolormap(mycoloridx(i),:);
end

for i = 1:11 %the size and the color of the nodes are determined by flu
    flu(i) = sum(Cij(i,:))+sum(Cij(:,i));
end
flu = flu/300;

k=500;
mycolormap = autumn(k);
max_data = max(flu);
for i = 1:11
    count(i) = flu(i);
    mycoloridx(i) = 500 - round( k * count(i)/ max_data );
    mycoloridx(mycoloridx<1) = 1;
    r(i) = round(flu(i));
    bg.nodes(i,1).Color = mycolormap(mycoloridx(i),:);
    bg.nodes(i,1).LineColor = mycolormap(mycoloridx(i),:);
    bg.nodes(i,1).FontSize = 5*r(i);
    axis square;
end

%% view the graph
set(bg.nodes,'shape','circle')
set(bg,'layoutType','radial');
bg.showWeights='off';
set(bg,'arrowSize',3,'edgeFontSize',1);
view(bg);
```

```
%calculate the shortest parth between two nodes
```

```
function[P u]=n2shorf(W,k1,k2)
%P is the shortest path between k1 and k2. Nodes are sorted in sorted
%order.
%u is the length of the path

n = length(W);
U = W;
m = 1;

%% use Warshall-Floyd algorithm to identify network patterns
while m <= n
    for i = 1:n
        for j = 1:n
            if U(i,j) > U(i,m) + U(m,j)
                U(i,j) = U(i,m) + U(m,j);
            end
        end
    end
    m = m + 1;
end
u = U(k1,k2);

%% calculate P
P1 = zeros(1,n);
k = 1;
P1(k) = k2;
V = ones(1,n) * inf;
kk = k2;
while kk ~= k1
    for i = 1:n
        V(1,i) = U(k1,kk) - W(i,kk);
        if V(1,i) == U(k1,i)
            P1(k+1) = i;
            kk = i;
            k = k + 1;
        end
    end
end
k = 1;
wrow = find(P1~=0);
for j = length(wrow):(-1):1
    P(k) = P1(wrow(j));
    k = k + 1;
end
P;
```

```
%% Initialization environment
close all
clear
clc
N = 11;

%% import and process data
Sij_1 = 46; %substitution time
Sij_2 = 60;
```

```
Sij_3 = 77;
N1 = 5; %player number
N2 = 2;
N3 = 10;

%% calculate Cij
Cij = sumHijk('1');
Cij(N1,:) = Cij(N1,:)*90/Sij_1;
Cij(N2,:) = Cij(N2,:)*90/Sij_2;
Cij(N3,:) = Cij(N3,:)*90/Sij_3;
Cij = ones(N)./Cij;

for i = 1:N
    for j = 1:N
        if i == j
            Cij(i,j) = 0;
        end
    end
end

Cij = round(Cij*1000);


%% calculate node intermediate
for i = 1:N
    for j = 1:N
        P = n2shorf(Cij,i,j);
    end
end

P = zeros(N,1);
k = 1;
for i = 1:N
    for j = 1:N
        if i ~=j
            p = n2shorf(Cij,i,j);
            x = size(p);
            d = zeros(N,1);
            for n = 1:x(1)
                for m = 1:x(2)
                    if p(m) == j
                        break
                    elseif p(m) == 1
                        d(1) = d(1) + 1;
                    elseif p(m) == 2
                        d(2) = d(2) + 1;
                    elseif p(m) == 3
                        d(3) = d(3) + 1;
                    elseif p(m) == 4
                        d(4) = d(4) + 1;
                    elseif p(m) == 5
                        d(5) = d(5) + 1;
                    elseif p(m) == 6
                        d(6) = d(6) + 1;
                    elseif p(m) == 7
```

```
                    d(7) = d(7) + 1;
                elseif p(m) == 8
                    d(8) = d(8) + 1;
                elseif p(m) == 9
                    d(9) = d(9) + 1;
                elseif p(m) == 10
                    d(10) = d(10) + 1;
                elseif p(m) == 11
                    d(11) = d(11) + 1;
                end
            end
        end
        for a = 1:N
            P(a) = P(a) + d(a)/x(1);
        end
    end
end
end
P;
```

```
function y = sumHijk(MatchID)
N = 11;

%% read and process data
data0 = importdata('passingevents.xlsx');
data_p = importdata('originlineup.xlsx');
id = str2num(MatchID);
data_0 = data0.data;
textdata_0 = data0.textdata;

%% identify origin players
for i = 1:38
    if id == i
        for j = 1:N
            Player(j) = data_p.textdata(id,j);
        end
    end
        continue
end

%% import data of huskies
j = 1;
for i = 1:length(textdata_0)
    if data0.data(i,1) == id && strcmp(data0.textdata(i,1),'Huskies') == 1
        data1(j,:) = data_0(i,:);
        textdata1(j,:) = textdata_0(i,:);
        j = j + 1;
    end
end

x1 = data1(:,8);
y1 = data1(:,9);
x2 = data1(:,10);
y2 = data1(:,11);
origin = textdata1(:,2);
```

```
destination = textdata1(:,3);
score = data1(:,12);

% calculate s
n = length(origin);
k = ones(11,11);
D = zeros(11);
s = zeros(11);
for i = 1:n
    for j = 1:N
        if strcmp(origin(i),Player(k)) == 1
            for k = 1:N
                if strcmp(destination(i),Player(k)) == 1
                    d = Distance(x1(i),y1(i),x2(i),y2(i));
                    a = char(origin(i));
                    b = char(destination(i));
                    D(j,k,k(j,k)) = d;
                    k(j,k) = k(j,k) + 1;
                    s(j,k) = s(j,k) + d * score(i);
                end
            end
        end
    end
end

%calculate sumHijk
k0 = 1;
a = size(D);
for i = 1:11
    for j = 1:11
        for m = 1:a(3)
            c(k0,m) = D(i,j,m);
            m = m + 1;
        end
        k0 = k0 + 1;
    end
end

for i = 1:11
    for j = 1:11
            s(i,j) = s(i,j)/k(i,j);
    end
end

y = s;
```

- **The ACC value of each player**

- **The first lineups statistics of Huskies**

- **ACC&DG of Different Line-ups**

Table 12: ACC value of each player

ID	ACC	ID	ACC	ID	ACC
D_1	37.06	F_1	31.08	M_2	39.10
D_2	37.91	F_2	36.79	M_3	35.41
D_3	35.36	F_3	43.17	M_4	34.89
D_4	32.55	F_4	32.12	M_6	34.21
D_5	34.80	F_5	34.68	M_8	42.41
D_6	35.73	F_6	34.92	M_9	38.59
D_7	33.86	G_1	34.16	M_{12}	38.07
D_8	34.52	M_1	35.36	M_{13}	34.21

Table 13: The first lineups statistics of Huskies

Huskies	Type	433	532	442	451	541	424
	Number	9	7	12	4	5	1
	Rate	23.58%	18.42%	31.58%	10.53%	13.16%	2.63%

Table 14: ACC&DG of Different Line-ups

Line-ups	ACC	GD
424	20.86	-2
433	38.14	0.78
442	35.98	-0.5
451	32.55	-0.75
532	29.35	-1.14
541	40.68	-0.4