

## Analysis of Team Performance Through the Network

To win a football game, it is not enough to rely on the player's ability. We also need high team performance. Therefore, for team competitions, it is important to develop a cooperative strategy.

For question 1, using players as nodes and passing to form links between players, we constructed a matching average passing network and a "50-pass" network. From the two dimensions of time (short-term and long-term) to space (macro and micro), we calculate the structural indicators and network attributes that can reflect player interaction information in the cooperative network. In order to reflect the overall network structure, we establish a matching average passing network and draw a weighting map of the passes. We solve the problem of time evolution by establishing a "50-pass network". Among them, we counted the number of rings in a complex network on the macro scale of space. In the micro space, we calculated the network centroid coordinates, dispersion and other indicators that can reflect the player's group ability. In terms of time, we selected 3 indicators to discuss changes in network indicators over time.

For question 2, we built a factor analysis model to study the structure, configuration, and internal dynamics of the team. First, we followed some of the indicators from the first question and introduced the indicators of action diversity and team contribution. According to the KMO test, some variables were screened out, and the performance team cooperation configuration factor ( $F_1$ ), dynamic factor ( $F_2$ ), and structural factor ( $F_3$ ) were obtained. The factor regression equation was obtained according to the factor score function  $F = 0.6568 - 0.0626Y$ . Finally, we did a correlation analysis for the season results and proved the reliability of the model.

For question 3, we provide advice to the coach from the perspective of team ability and tactics. For team abilities, we compared the configuration, dynamics, and structure of the Huskies team with the top three teams this season. In the end, we recommended that coaches improve their personal skills. For strategy, we provide suggestions for both formation and staffing. we used the ullmann algorithm to obtain sub-graphs of 3 and 4 nodes that are different from each other, and counted the number of different sub-graphs of 3 and 4 nodes of Huskies and opponent teams this season. According to the frequency of different sub-graphs, we performed cluster analysis. The games that were divided into one category showed that the same formation or strategy was used. Based on the strength of each formation, the Huskies team's winning matrix is obtained. We use decision theory to give the optimal formation or strategy that each coach should adopt. We also found the best staffing for each coach. Finally, we analyzed the formations that could cause the team to fail and made some suggestions to the coach.

For question 4, in order to study the characteristics of excellent teams and build a more comprehensive team performance model. First, we explain how to better design the team based on the models established in the first three questions. We divide the performance of the excellent team into two aspects: ability and strategy. Capability focuses on the individual aspect of the team, and strategy focuses on how to build an excellent collaborative network. Then, based on this model, we add other factors to make the model more For universal. Finally, we believe that those factors that are difficult to digitize are the keys to improving the model, such as execution and team morale.

Finally, we analyze the advantages and disadvantages of the model, and propose the generalization and application of the model.

**Key words:** Dijkstra, factor analysis model, cluster analysis, ullmann algorithm, game theory

# Content

1	Introduction.....	1
1.1	Background of the Problem.....	1
1.2	Previous Works.....	1
1.3	Our Work.....	1
2	Symbol Descriptions.....	2
3	Model Hypothesis.....	2
4	Task 1: Passing network between players.....	3
4.1	Analysis of the problem.....	3
4.2	Data preprocessing.....	3
4.3	Index selection and definition.....	3
4.4	Construction of Delivery Network.....	5
4.4.1	Matching the average delivery network.....	5
4.4.2	50-pass network.....	5
4.5	Calculation of indicators.....	6
4.6	Analysis of results.....	7
4.6.1	Definition of terms.....	7
4.6.2	Analysis of spatial indicators.....	7
4.6.3	Time index analysis.....	8
5	Task 2: Teamwork Model.....	9
5.1	Analysis of the problem.....	9
5.2	Model establishment.....	9
5.2.1	Defining variables.....	10
5.2.2	KMO and Bartlett's Test.....	10
5.2.3	Calculation steps.....	10
5.3	Analysis of results.....	11
5.4	Factor score function and factor interpretation.....	12
5.5	Hypothesis testing.....	13
6	Task 3: Suggestions for Outcome Strategies.....	13
6.1	Problem analysis.....	13
6.1	Suggestions for structural strategies.....	14
6.2	Tactical strategy recommendations.....	14
6.2.1	Definition of terms.....	14
6.2.2	ullmann algorithm.....	15
	Algorithm Related Theorem.....	15
6.2.3	System Clustering.....	16
6.3	Analysis of results.....	16
7	Task 4: Summarize our findings.....	18
7.1	Analysis of the problem.....	18
7.2	Analysis of results.....	18
8	Model Extensions.....	19
9	Conclusion.....	20
9.1	Strengths.....	20
9.2	Weaknesses.....	20
10	References.....	1
11	Appendix.....	2

# 1 Introduction

## 1.1 Background of the Problem

The success of the exploration team has always been a complex issue. Social systems have become one of the many fields that have benefited greatly from the network science framework, and the team is considered a complex network [3]. Competitive team sports are one of the useful methods for studying team processes. The success of a team is not just the sum of the capabilities of individual members, but a combination of factors. For example, the teamwork ability balance between teams. The barrel effect reflects a collective power. It refers to the amount of water in a wooden barrel, which depends on the overall effect of the shortest wooden board.

The nature of team sports like football lies somewhere between game abstraction and complex social systems, combining the unique size and composition of this data set, providing an ideal basis for solving a variety of data science problems, including measuring and evaluating individuals And collective performance, as well as the factors that determine success or failure [1]. In order to understand team dynamics and team dynamics throughout the season, network analysis of football is important. If you want to study the success of a football game and develop a teamwork strategy, it is essential to check the interaction of team scores and explore team motivation throughout the season.

## 1.2 Previous Works

With the improvement of the level of sensing technology, this technology provides a high-fidelity data stream for each game. Therefore, the analysis of football has attracted more and more interest in academia and industry [1]. In the 1970s, Gould and Gatrell published a groundbreaking article about the concept of a delivery network related to the football match, but it didn't get much attention. In 2010, after ten years of hard work, Duch and collaborators witnessed how to reveal key information about the organization of football, the performance of teams and players through network science [3].

In 2010, Jordi Duch developed a network approach that can perform on individual participants and teams as a whole. In 2014, Guisheng Y et al. Proposed the SA algorithm and verified the effectiveness of the algorithm in real social networks. Also in 2014, Gupm N and Singh A thought that network topology information could be used to predict links in the network and proposed Link prediction method combining topological information of non-common neighbor nodes. In order to obtain higher prediction accuracy, some prediction algorithms based on global and quasi-local structure information are proposed [9]. In 2019, Buldú created a delivery network model to analyze the reasons for the team's failure through various indicators[3]. With the increase of football data, More and more research. The analysis of teamwork through the establishment of models has an important influence on the actual cooperation strategy.

## 1.3 Our Work

In order to understand the team's dynamics and develop strategies to improve teamwork, we used the game data provided to do the following:

In question 1, with each player being a node, the pass constitutes a link between the players. Based on this we created two passing networks. From the four dimensions of time (a game and

the entire season) to space (macro and micro), we have selected the corresponding network indicators and network attributes. Based on the data, each network indicator is calculated, and the network mode and interaction of Huskies and the opponent team are analyzed.

In question 2, we kept some of the indicators of the first question, added the variety of actions and player level variables, and established a factor analysis model. The scoring factor function was used to analyze the structure, configuration and dynamic aspects of team work. Finally, a hypothesis test was performed for the season results.

In question 3, we made our recommendations to the coach from both competence and strategy. In terms of capabilities, we compared the configuration, dynamics, and structural indicators of the Huskies team with other teams based on the model of problem two and made our recommendations. In terms of strategy, we used the ullmann algorithm to count the number of mutually different subgraphs of 3 and 4 nodes. We performed cluster analysis on the subgraphs to summarize the formation or strategy. Finally, we use the knowledge of decision theory to point out the best strategies for each coach and suggestions for team success.

In question 4, the model established based on the first three questions of us explained how to design a more effective team; based on the model, we made changes and added other factors to make the model more general.

## 2 Symbol Descriptions

Symbol	Definition
$C$	Overall clustering coefficient
$w_{ij}$	The number of passes from player $i$ to player $j$
$l_{ij}$	Topological distance of the link between two players $i$ and $j$
$p_{ij}$	Shortest path for all players
$D$	Team's average shortest path
$\lambda_1$	Maximum eigenvalues of network weighted adjacency matrix
$\lambda_2$	Laplacian matrix's second smallest eigenvalue
Md	Match type diversity variable
$A_2$	Load matri
$d_{ij}$	Distance between pairs of sample points

Note: For the description of other symbols, see the text.

## 3 Model Hypothesis

- **The level of players on the court will not change significantly.**

It is not possible to study the network of a team over the span of the entire season, because the player level has changed significantly, and the same two games will generate a large difference in weighting maps, resulting in large strategic classification errors.

- **The player's personal ability will not be changed by the coach, and the home and away changes will be greatly affected.**

In order to accurately analyze the strategy for improving team performance, even if the player executes the coach command more familiar, it is required that the player's personal ability does not change significantly due to the change of coach. If the level of the home and away

teams is uncertain in each game, the true level of the team cannot be accurately estimated, which will affect the analysis of team performance evaluation.

- **The team has no internal conflicts.**

The team needs cooperation between players. If there are contradictions within the team, the possibility of losing is increased. This may lead to inaccuracies in the improvement strategies derived from data analysis. Therefore, there should be no internal conflicts between the teams.

- **A game has only one formation or strategy.**

If a match has multiple formations or strategies, the formation strategy after the cluster analysis will be inaccurate.

## 4 Task 1: Passing network between players

### 4.1 Analysis of the problem

In the first question, we use players as nodes, and each pass constitutes a link between players, forming the empowerment map for the entire game. The number of passes between two nodes is the weight of the link. In order to reflect the overall network structure, we establish a matching average passing network. The network changes over time. We solve the problem of time evolution by establishing a "50-pass network". In order to dig out the effective information in the network, we calculate the structural indicators and network attributes that can reflect the player's interactive information in the cooperative network from four aspects: time (a game and the entire season) to space (macro and micro). Among them, we counted the number of rings in the complex network on the spatial micro level. In the macro space, that is, on the entire network, we calculated the network centroid coordinates, dispersion, and other indicators that can reflect the ability of the player group. In terms of time, we have selected 3 indicators to discuss changes in network indicators over time. Pave the way for discussion of team performance.

### 4.2 Data preprocessing

The data covers all 38 games played by Huskies and 19 opponents, including 23,429 passes and 59,271 matches between 366 players. We screen the given data to reduce redundancy. For example, we only study one football situation and screen out other games. We also corrected logical problems encountered in the data. For example, there is the problem of players playing without leaving the field.

### 4.3 Index selection and definition

When looking at interactions between teams, we analyze from both spatial and temporal scales. Space is divided into micro and macro (the entire network); time is divided into one game and the entire season.

Passing is related to the spatial characteristics of the player's position. We select the macro space index as the number of rings, and the ternary network is reflected by the number of rings. The micro spatial indicators are as follows:

(1) **X-coordinate and Y-coordinate of the network centroid** indicates the position of the passer or receiver. The average coordinates of all passing x and y during the game define the network centroid.

(2) In order to analyze the team's passing direction, we specified **the advance ratio**  $(\Delta y) / (\Delta x)$ . The  $\Delta y = y_2 - y_1$  of the pass is the difference between the receiver's coordinate y and the

pitcher's coordinate  $y$ , and the corresponding  $x$  coordinate defines  $\Delta x$ . This indicator has nothing to do with the number of passes.

(3) **The clustering coefficient** is a parameter related to the topology of the average transfer network. In order to ensure a triangular relationship between the three players, we chose a clustering coefficient. The clustering coefficient (local index) is related to the number of triangles created between any triad players. And when a triangle with three nodes exists, and the link between the two nodes cannot pass, there is another way to reach another node that passes through the other two edges of the triangle [3].

(4) **Shortest-path length ( $D$ )** indicates the degree of connection between players in the team. It is a measurement of the "topological distance" that any two players in the team must pass.

(5) **Largest eigenvalue of the connectivity matrix** has been used as a measure of network strength because it increases with the number of nodes and links.

(6) **Algebraic connectivity of the Laplacian matrix.** The second minimum eigenvalue is related to several network attributes and can be used as an index to quantify team division. In the context of a football pass network, the second smallest eigenvalue can be interpreted as an indicator to quantify the division of teams. The reason is that the lower second minimum eigenvalue indicates that the network is easily divided into two groups, and eventually the second minimum eigenvalue is broken when it is zero. In this way, the higher the second minimum eigenvalue, the closer the team is, which is a measure of team cohesion.

(7) **Dispersion of the players' centrality** is the importance of players in the passing network. This is an index calculated from the eigenvectors related to the maximum eigenvalues of the connection matrix.

The difference in space may be caused by the number of passes, so study the evolution of the network over time and therefore the evolution of the parameters. We divide **the time indicator into one game and the entire season**. We study the position of the  $x$ -coordinate of the network centroid passing or receiving team members, the advance ratio  $(\Delta y) / (\Delta x)$  and the centrality. dispersion. The specific index classification is shown in the following figure:

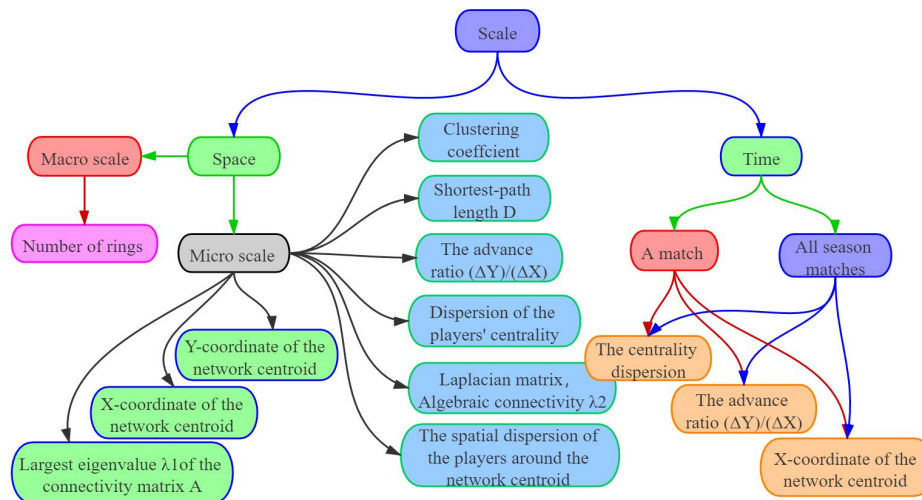


Figure 1: Classification of network indicators

## 4.4 Construction of Delivery Network

The interaction of football matches is reflected in the scale of space and time, so we use two ways to build the network. The first focuses on the overall spatial network structure, rather than individual teams. The second type mainly considers the evolution of time.

### 4.4.1 Matching the average delivery network

The first is to establish a matching average passing network, with players as nodes and links representing the number of passes between players. The passing direction between players is one-way and weighted according to the number of passes. Replacement rules: Each player is assigned a node at the start of the game. If a player is replaced, a new player is added to ensure that the football member of the game is eleven. A football team has a total of eleven players, including forward (F), midfielder (M), defensive (D) and goalkeeper (G).

Note that both the  $x$  and  $y$  coordinates of this field are limited to  $[0,100]$  and are measured in “field units”, since not all fields have the same dimensions. Where players are represented by circular nodes, the size of the circular nodes and their characteristics Vector centrality is directly proportional. The position of each player is given by the average of all passing positions of that player along the game. We mark the weight of each link, the number of passes between players. From this we drew Huskies' football pass network (weighted directed graph) in the first game, the effect is as follows:

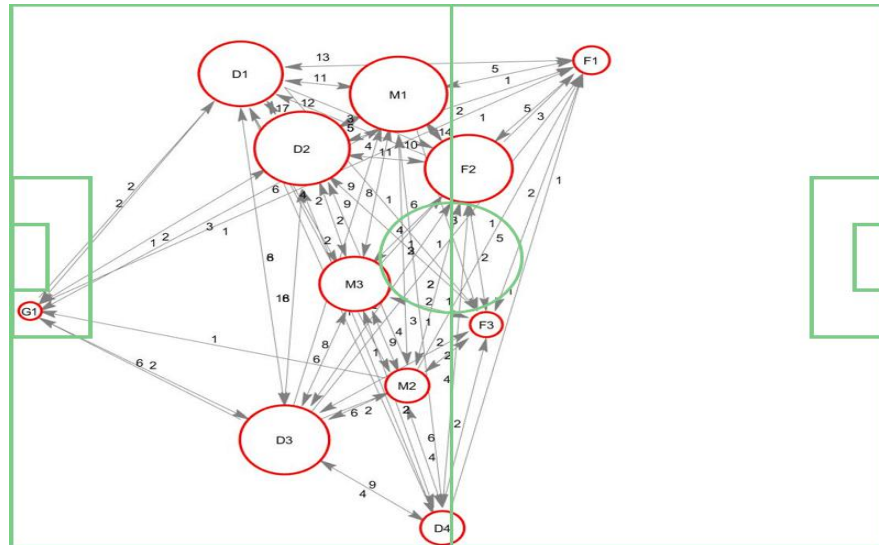


Figure 2: Empowerment diagram

As can be seen from the above picture, the striker accepts the most balls for offense and has more defensive passes. The picture is consistent with the different priorities for different types of players, such as striker first with the ball, midfielder with the ball first, and defense with the ball first. We can intuitively see the passing situation between various players.

### 4.4.2 50-pass network

Second, we build a "50-pass network" to solve the problem of time evolution. Similarly each player is a node, and each pass constitutes a link between the players. The 50-pass network contains only 50 consecutive passes, and the time of the last pass is allocated. The purpose of this allocation is that when  $1 = 50$  occurrences, we record the time at this time and construct and draw a 50-pass network. Next, every time a new delivery is made, we can ignore the old delivery

and allocate the time of the last delivery to the new network. Looking at the effect of different numbers of passes, 50-pass is the best, so we build a "50-pass".

## 4.5 Calculation of indicators

### • Average clustering coefficient

The aggregation coefficient of the entire network is defined as the average of the local aggregation coefficients of all nodes  $N$ . If the network is weighted, we must not only calculate the number of nodes connected between them, but also calculate the distribution of the weight of each link. Describes the team forming a balanced triangle among the players. This is an indicator of the local level. We use weighted clustering coefficients to measure the likelihood that a given player's neighbors will also be established between them.

The weighted clustering coefficient is:

$$C_i = \frac{\sum_{j,k} w_{ij} w_{jk} w_{ik}}{k(k-1)/2} = \frac{\sum_{j,k} w_{ij} w_{jk} w_{ik}}{\sum_{j,k} w_{ij} w_{ik}} \quad (4-1)$$

Where  $j$  and  $k$  are any two players on the team,  $w_{ij}$  is the weight of the link,  $w_{ij}$  and  $w_{ik}$  are the third player  $i$  and the number of passes between them.

The overall clustering coefficient is the average of the weighted clustering coefficients of all players, that is:

$$C = \frac{1}{N} \sum_{i=1}^N C_i \quad (4-2)$$

### • Shortest-path length $D$

The shortest path length  $D$  is the minimum number of players to pass from one player to another during the pass. Since the links of the network are weighted by the number of passes, we define the topological distance as the inverse of the number of passes. The more passes between two players, the shorter the topological distance between the two points. The topological distance of the link between two players  $i$  and  $j$  is defined as the inverse of the link weight, that is:

$$l_{ij} = \frac{1}{w_{ij}} \quad (4-3)$$

When calculating the shortest path length, there may be the shortest path through two players or more instead of the link. Therefore, we use Dijkstra's algorithm to calculate the shortest path  $p_{ij}$  for all players.

The basic idea of Dijkstra is to find the shortest path and distance between the vertices of  $u_0$  and  $G$  in order from near to far according to  $u_0$ , until all vertices of  $G$  are reached, and the algorithm ends. Dijkstra's specific algorithm is:

*Step1*: Let  $l(u_0) = 0$ , for  $v \neq u_0$ , let  $l(v) = \infty$ ,  $S_0 = \{u_0\}$ ,  $i = 0$ .

*Step2*: For each  $v \in \bar{S}_i (\bar{S}_i = V / S_i)$ , use  $\min_{u \in S_i} \{l(v), l(u) + w(uv)\}$  instead of  $l(v)$ . Calculate  $\min_{v \in \bar{S}_i} \{l(v)\}$ , record a vertex that reaches this minimum as  $u_{i+1}$ , and let  $S_{i+1} = S_i \cup \{u_{i+1}\}$ .

*Step3*: If  $i = |V| - 1$ , Stop; if  $i < |V| - 1$ , use  $i + 1$  instead  $i$ , turn *Step2*.

Furthermore, the average shortest path of the team is the average of the shortest paths among all players, that is:



$$D = \frac{1}{N(N-1)} \sum_{i,j} p_{ij} \quad (i \neq j) \quad (4-4)$$

Among them,  $N = 11$  is the total number of players on the team.

#### • Largest eigenvalue of the connectivity matrix

The maximum eigenvalue  $\lambda_1$  of the network weighted adjacency matrix  $A$  is a measure of the strength of the network. For a weighted directed graph, the weighted matrix  $A$  is an  $N \times N$  matrix, that is, an  $11 \times 11$  matrix, where the element  $w_{ij}$  represents the number of passes from player  $i$  to player  $j$ , which is the weight of the link. The matrix coefficients are:

If  $i \neq j$ , and  $E(G) < i, j > E(G)$ , it shows that vertex  $i$  and vertex  $j$  are adjacent points to each other, then  $A[i][j] = w_{ij}$ ;

If  $i = j$ , then  $A[i][j] = 0$ ; Other situations, then  $A[i][j] = \infty$ .

Between players  $S$ , the maximum characteristic value of  $A$  is the average number of passes  $\lambda_1 > S$  and  $S_{\max} \geq \lambda_1 \geq \max(S, \sqrt{S_{\max}})$ , where  $S_{\max}$  is the maximum number of times a player passes to any other player in his team. According to experience, a network with a higher link will have a higher  $\lambda_1$ , and a network that is not directly connected between important participants will also have a higher  $\lambda_1$  than a network connected between important nodes.

#### • Algebraic connectivity of the Laplacian matrix

The algebraic connectivity  $\lambda_2$  corresponds to the second smallest eigenvalue of the Laplacian matrix  $L$ , which is defined as:

$$L = S - A \quad (4-5)$$

Among them,  $A$  is a weighted adjacency matrix;  $S$  is a diagonal matrix, where each element is the sum of each player's passes. Algebraic connectivity is closely related to the structure and dynamics of the network. Among them, algebraic connectivity is an indicator of the modular structure of the network. The lower the second minimum eigenvalue, the clearer the existence of independent groups within the network.

#### • Center dispersion

The coordinates of the center of gravity of  $x$  and  $y$  correspond to the average position of all poses in the network, that is, only 50 passes of all matches in the network are in 50-pass passes. The centroid dispersion corresponds to the standard deviation of the distance of the player from the coordinate position of the network centroid.

## 4.6 Analysis of results

### 4.6.1 Definition of terms

- (1) Definition of robustness: Robustness refers to the control system maintains certain other performance characteristics under a certain (structure, size) parameter perturbation.
- (2) Dyadic Configurations: relationships involving pairs of players.
- (3) Triadic Configurations: relationships involving groups of three players

### 4.6.2 Analysis of spatial indicators

Parameters related to the characteristics of the network space: network centroid coordinates, lead ratio, dispersion. In order to better analyze the structure of the average transfer network, we

analyze and compare other parameters. The comparison of the other five network parameters is shown in Appendix 1. Huskies and Opponent analysis is as follows:

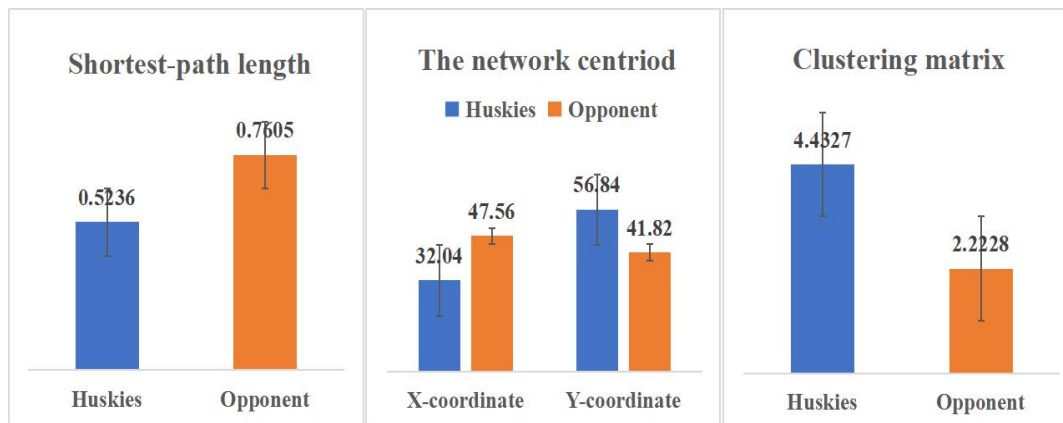


Figure 3: Comparison of 4 network parameters

The standard deviation of each indicator is illustrated by the error bars. We can see from the shortest chart that the value of Huskies is lower than Opponent, indicating that the players in the opponent are more closely linked. We can see from the network centroid coordinate graph that Huskies is more inclined to the network centroid y, while Opponent is more inclined to the network centroid x, indicating that the team may have a preference for one side.

In the graph of clustering coefficients, Huskies has a much higher value than Opponent clustering coefficients, indicating that the relationship between the three members of Huskies is richer.

From the quantitative diagram of the three rings, we observe that Huskies has more than the opponent's three-ring network, indicating that Huskies has passed more times and is more biased towards Triadic Configurations. From the maximum eigenvalue of the weighted adjacency matrix, it can be seen that Huskies passes There are many times, Huskies has a higher value than Opponent. This indicator explains the higher robustness of the Huskies pass network.

As can be seen from the algebraic connectivity diagram, Huskies is much higher than Opponent. This eigenvalue is a measure of team division, indicating that Huskies' attack and defense lines are more complicated; from the advance ratio we observe that Huskies' forward ratio is much higher than Opponent, which indicates that the pass is more parallel to the opponent's goal than the rest of the opponent's goals; from the divergence diagram Seeing little difference between the two, Huskies is slightly higher than Opponent.

#### 4.6.3 Time index analysis

Based on the analysis of the space, it can be seen that the average network indicates the difference between Huskies and Opponent. According to the "50-pass" network and the corresponding formula, we use MATLAB programming to get the network centroid coordinate parameters Huskies and Opponent competition. The abscissa is the number of passes per pass, and the ordinate is the value of the centroid of the network, as shown below. For the comparison of the other two parameters, see Appendix 2.

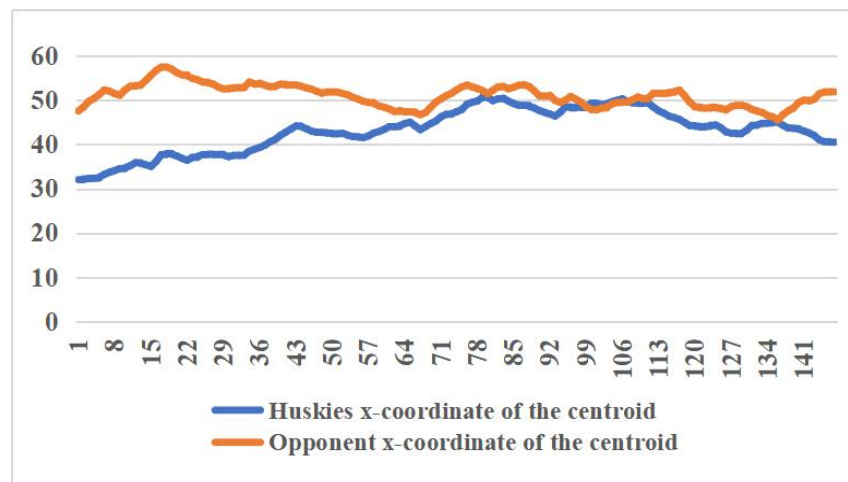


Figure 4: x-coordinate of the network centroid

As can be seen from the figure above, Huskies's x-network centroid coordinates fluctuate more than Opponent's fluctuations, and the opponent's x-network coordinate values are higher than Huskies on each pass.

The advance ratio of Huskies is lower in the first few seconds, and then it increases over time and is higher than the opponent's value. From the Central Dispersion, the central dispersion fluctuations of Huskies and the opponent's players are relatively small and fluctuate within a range. Changes may be related to the style of the game.

## 5 Task 2: Teamwork Model

### 5.1 Analysis of the problem

Based on the indicators requested in the first question, we introduced a variable  $Md$  that reflects the diversity of game types and a variable  $lamde2$  that reflects the importance of players in the team, and deleted some indicators such as the total number of passes. For indicators, we use SPSS to test its suitability for factor analysis and dimensionality reduction. In order to study the dynamics of team structure, configuration, and cooperation networks, we established a factor analysis model. Factor analysis is used to summarize the factor categories, and the factor regression function is obtained using the factor score function. According to the degree of influence of each variable, we named it teamwork and configuration factor  $F_1$  to reflect the team configuration, teamwork motivation factor  $F_2$  to reflect team motivation, and teamwork structure factor  $F_3$  to reflect team structure. And finally get a score  $F$  that reflects the overall performance of the team.

### 5.2 Model establishment

Factor analysis is the synthesis of variables with intricate relationships into a small number of factors. "Factors" are called abstract variables and can reflect the main information of many original variables. It is also commonly used in mathematical modeling for number-weighted analysis.

### 5.2.1 Defining variables

In order to better reflect the performance indicators of teamwork, based on question 1, we introduced a variable Md that reflects the diversity of game types and a variable lamda2 that reflects the importance of players in the team. The total number of passes and the network were deleted. Some indicators such as the center of mass position coordinates.

Competition type diversity variables:

$$Md = \frac{x_i}{\sum_{k=1}^M x_i} \sqrt{\sum_{j=1}^{M_i} (x_{ij} - \bar{x}_i)^2 / M_i} \quad (5-1)$$

$x_i$  : The total number of occurrences of the i-th event ;  $x_{ij}$  : the j-th occurrence of the i-th event;  $\bar{x}_i$  : the average number of occurrences of each event in the i-th event;  $M_i$  : The sub-event type of the i-th event; M: Event type.

Reflect player importance variables in the team:.

$$lamda2 = \frac{\max_i \Delta \lambda_{2i}}{\lambda_2} \quad (5-2)$$

$$\Delta \lambda_{2i} = \lambda_2 - \tilde{\lambda}_{2i} \quad (5-3)$$

Among them,  $\tilde{\lambda}_{2i}$  is the algebraic connectivity after removing the i-th player.

In the end, we selected seven index variables, namely Shortest path (D); Pass successful (PS); Pass Success Rate (PSR); Maximum Eigenvalue of Connectivity Matrix (Lamda1); Shots (S); Movement dispersion (Md); The maximum change rate of the second smallest eigenvalue of the Laplacian matrix after removing a player (Lamda2). Represented by  $x_1, x_2, \dots, x_7$ . Twenty teams serve as the twenty evaluation targets. Among them, the value of the j-th index of the i-th evaluation object is  $a_{ij}$ .

### 5.2.2 KMO and Bartlett's Test

We use SPSS to test whether the above indicators are suitable for factor analysis. The test results of KMO and Bartlett's are as follows:

Table 1: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy	0.717
Bartlett's Test of Sphericity Approx. Chi-Square	58.078
df	21
Sig.	0.000

It can be seen from the table that the KMO statistic is 0.717, which is greater than the minimum standard, indicating that it is suitable for factor analysis, Bartlett spherical test,  $P < 0.001$ , which is suitable for factor analysis.

The related formulas of the factor analysis model we have established and the model establishment process are shown in Appendix 4.

### 5.2.3 Calculation steps

(1)The basic equations are established from the data calculation  $\bar{x}_k, s_k$ ;

- (2) Obtain the eigenvalue  $j$  and the variance contribution, contribution rate and cumulative contribution rate of each common factor from the correlation coefficient matrix  $R$ ,  $j=1,2,\dots,m$  and determine the number  $p$  retained by the common factor according to the cumulative contribution rate.
- (3) Determine the factor load matrix  $A$  by principal component analysis.
- (4) The variance is maximally orthogonalized and rotated, and the variable coefficients are mechanized (as close as possible to 0 or 1).
- (5) Get the factor score function and calculate the factor score.

### 5.3 Analysis of results

The SPSS software was used to obtain the list of principal components (see Appendix 4). From the table, it can be seen that the characteristic values of the first three principal components are relatively large, and their cumulative contribution rate has reached 81.201%.

#### 5.3.1 Common factor variance ratio result graph

Table 2: communalities

Zscore	D	PS	PSR	Lamda1	S	Md	lamda2
Initial	1	1	1	1	1	1	1
Extraction	0.802	0.849	0.846	0.763	0.792	0.815	0.817

Note: extraction method is principal component analysis.

The results show that the common variance of each index variable is above 0.76, indicating that the three common factors can reflect most of the original index variables.

#### 5.3.2 Load scatter plot

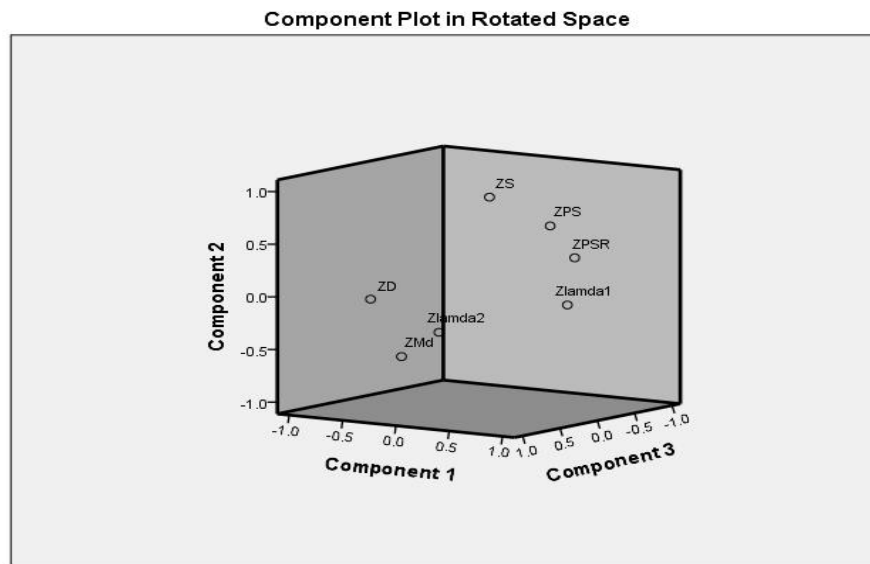


Figure 6: Load scatter plot

From the load scatter plot, it can be seen that the first common factor can explain the variable number of shots, the number of successful passes, and the pass success rate; the second common factor can explain the eigenvalues and algebraic connectivity of the connectivity matrix

The third common factor can explain the shortest path of the variable and the variety of actions. It shows that the reduction of the index dimension has not caused a lot of information loss.

### 5.3.3 Factor load after rotation

Table 3: Component Score Coefficient Matrix

Zscore	D	PS	PSR	Lamda1	S	Md	lamda2
Component1	-0.240	0.129	0.323	0.475	-0.267	-0.124	0.208
Component2	0.313	0.326	0.035	-0.265	0.644	-0.284	-0.050
Component3	0.619	0.117	0.026	0.065	0.076	-0.018	0.662

Note: Extraction method is principal component analysis. Rotation method is varimax with kaiser normalization.

After rotation, the maximum eigenvalue of connectivity matrix (Lamda1) has a larger load on factor one; shots ( $S$ ) has a larger load on factor two; shortest path ( $D$ ) and the maximum change rate of the second smallest eigenvalue of The Laplacian matrix after removing a player (Lamda2) has a larger load on factor three.

## 5.4 Factor score function and factor interpretation

The established factor analysis model was solved using MATLAB. We keep the coefficients of the function to three decimal places.  $x_1, x_2, \dots, x_7$  represents the shortest path, the number of shots, the success of the pass, the success rate of the pass, the eigenvalue of the connectivity matrix, the diversity of actions, and the algebraic connectivity. The three common factor score functions are as follows:

$$F_1 = -0.208x_1 + 0.224x_2 + 0.492x_3 + 0.236x_4 + 0.381x_5 + 0.282x_6 + 0.021x_7$$

Among them,  $F_1$  is mainly related to the number of shots, the number of successful passes, the pass success rate, and the eigenvalue of the connectivity matrix. The first three of them indicate that the player's technical ability and connectivity characteristic values reflect the robustness of the passing network, that is, the impact of a passing failure on the player, reflecting the player's psychological ability. It can be seen that the main indicators in  $F_2$  are related to the player's own ability, so we believe that  $F_1$  represents the configuration of teamwork, that is, the ability of personnel.

$$F_2 = -0.036x_1 - 0.194x_2 - 0.009x_3 + 0.007x_4 + 0.222x_5 - 0.085x_6 + 0.954x_7$$

$F_2$  is mainly related to algebraic connectivity. Algebraic connectivity reflects the synchronization time required for a team to reach the passing network and the diffusion time of the passing network. It can measure the degree of connection between teams. The cooperation between teams can change at any time, leading to changes in algebraic connectivity. So we think  $F_2$  represents the dynamic aspect of teamwork.

$$F_3 = -0.565x_1 - 0.024x_2 + 0.316x_3 - 0.043x_4 + 0.214x_5 - 0.651x_6 + 0.126x_7$$

$F_3$  is mainly related to the shortest path and action diversity. The shortest path reflects the passing match of each person between the teams. The variety of actions reflects the tactics and spectacles used by the team. We believe that  $F_3$  represents the structure of teamwork.

The score of each factor is as follows:



Figure 7: Factor analysis scores for each team

Finally, we make a simple ranking of the data in match.xlsx based on the number of scores, and use the comprehensive score, that is, the teamwork ability  $F$  and the ranking  $Y$  to perform a correlation analysis. among them

$$F = 0.4824F_1 + 0.1653F_2 + 0.3523F_3$$

Using the comprehensive factor score formula, the factor regression equation is:

$$F = 0.6568 - 0.0626Y$$

## 5.5 Hypothesis testing

In order to measure the reliability of the model, we performed a correlation analysis between the overall performance ranking and the team's simple ranking this season. When the significance level of the regression equation is 0.05,  $p = 0.0054$ , which passes the hypothesis test. The model is accurate.

## 6 Task 3: Suggestions for Outcome Strategies

### 6.1 Problem analysis

In order to improve the performance of the Huskies team, we consider from the perspective of team configuration and strategy. Among them, according to the results of the second question factor analysis, we compared the Huskies team with the top three teams in this season's configuration dynamic structural indicators and proposed corresponding improvements.

In terms of strategy, we are divided into two aspects: formation and staffing. In terms of formation, for different coaches, we used the ullmann algorithm to count the subgraphs of 3 and 4 nodes that are different from each other. We performed a cluster analysis on the frequency of the subgraphs and obtained Huskies according to the strength of each formation. The team's winning matrix. Using decision theory, we got the optimal formation that every coach should take. In terms of staffing, we counted a total of more than 10 staffing methods in 38 games, and obtained the Huskies team's winning matrix according to the scoring situation. According to the decision theory, the optimal personnel matching plan for different coaches is obtained. Finally, we analyzed the formations that could cause the team to fail and made recommendations to the coach.

We have drawn the following mind map for the third question's solution:

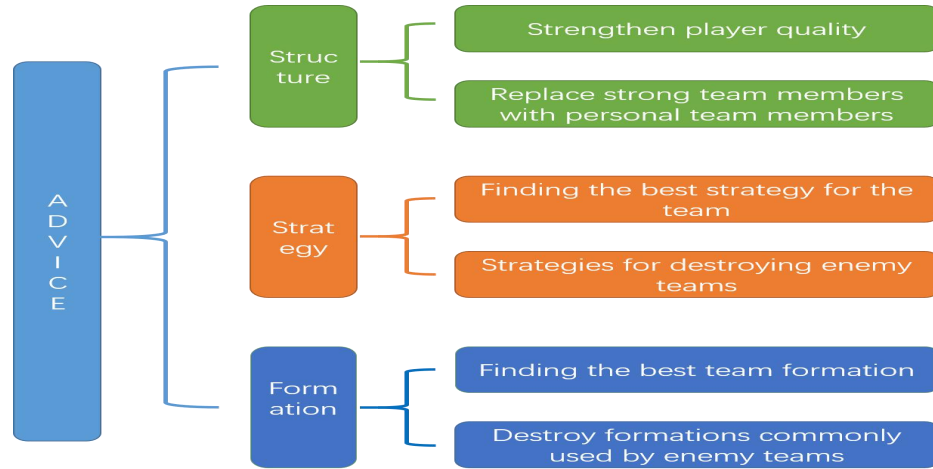


Figure 8: Question three mind map

## 6.1 Suggestions for structural strategies

The third question first let us adjust the team's strategy based on the insights obtained from the team cooperation model. In the conclusion of the second question, we obtained a model based on the structure, configuration and dynamic aspects of team cooperation, and then based on match.xlsx The data obtained by our team ranked 13th. In order to explore how to improve the winning rate of the Huskies team, we obtained 20 teams' scores in three aspects from the factor analysis and compared them with the data of the top three.

Table 4: Team work scores in three areas

Team	Score ranking	Configuration score	Dynamic score	Structure score
Huskies	The 13th place	-2.791	-0.227	3.163
Opponent9	The first place	0.841	1.771	0.637
Opponent4	The second place	0.738	0.034	0.211
Opponent5	The third place	0.991	-0.723	0.448

According to the table, we can see that the top three teams are far higher than the Huskies team in terms of configuration. Although the Huskies team is stronger in structure than the other teams, it cannot play a decisive role. Specifically, although the Huskies team's cohesion is dominant, the player's hit rate and technical ability are not strong enough. Therefore, based on the insights obtained from the team cooperation model, we recommend that **the coach can conduct separate training for the players to strengthen the players' own qualities Or replacing some players with strong team cooperation ability with players with strong personal ability will effectively improve the team's winning rate.**

## 6.2 Tactical strategy recommendations

### 6.2.1 Definition of terms

- Subgraph isomorphism: Suppose there are two graphs  $A=(V_H, E_H)$  and  $B=(V, E)$ . The subgraph isomorphism is a function from  $A$  to  $B$ , and  $f: V_H \rightarrow V$ , makes  $(f(u), f(v)) \in E$  the same holds.  $f$  is called a map of a subgraph isomorphism.
- Three-node sub-graphs with different structures: each game has a corresponding weighting graph, and there are node sub-graphs in the complete network structure. A three-node subgraph is



composed of three nodes, and different directions form different node subgraphs with different structures. There are at most 13 kinds of node subgraphs in each match (Appendix 6). The four-node subgraph is the same, with a maximum of 199 node subgraphs.

- If it depends on only one participant, we call this kind of decision model as decision theory; if the result depends on the decision of more than one participant, we call this kind of decision model as game theory [14].

### 6.2.2 ullmann algorithm

In order to study the tactics or formation characteristics of each team, we analyze it through three- and four-node subgraphs of different configurations. Among them, the ullmann algorithm is the simplest of the subgraph isomorphic algorithms.

#### Algorithm Related Theorem

If graph  $A$  is isomorphic to graph  $B$  with respect to mapping  $f$ , let

$$MC = M'(M'MB)^T \quad (6-1)$$

Where  $T$  is the transpose of the matrix. Then there are:

$$\forall i \forall j : (MA[i][j] = 1) \Rightarrow (MC[i][j] = 1) \quad (6-2)$$

Positions that are one in the  $A$  matrix are also one in the  $C$  matrix. The detailed algorithm design is shown in Appendix 5. We use the ullmann algorithm to construct mutually different subgraphs and record the number of statistics in the weighted graph as M3, M4.

Taking the fourth game as an example, we used ullmann and the matching algorithm to obtain the four-node type of the fourth game Opponent4.

Table 5: Types of the fourth game

Type number	1	2	3	5	12	17	28	40	41	99
The number of occurrences	64	10	2	3	5	1	3	4	1	1

Draw a corresponding four-node sub-graph according to the category number, and sort them according to the number of occurrences. The passing method is as follows:

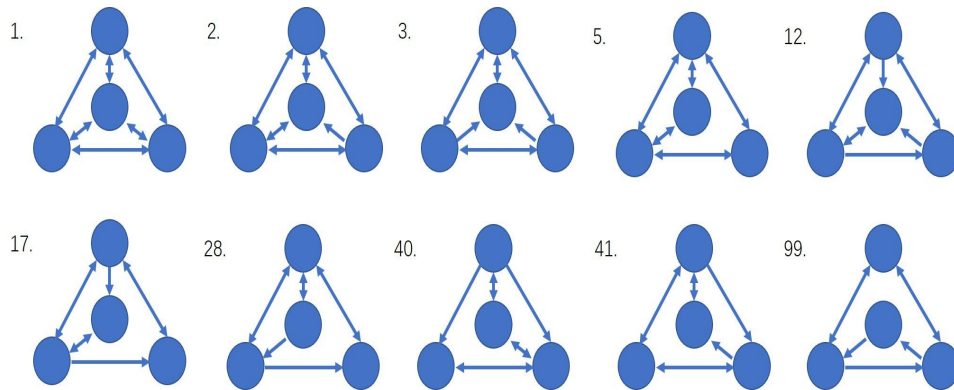


Figure 9: Four-node subgraph

As can be seen from the figure above, the four players pass each other the highest number of times. The top four-node subgraphs all focus on passing each other, and the opponent 4 teams score higher in the team work structure, configuration and dynamics.

In order to analyze the team's tactics, we classified the Huskies and opponents' sub-maps in 38 games. Among them, systematic clustering is widely used for index classification and data

classification. Therefore, we use cluster analysis to treat each category as a tactical or formation characteristic.

### 6.2.3 System Clustering

The method of calculating the distance between two sample points is based on the shortest distance method and the class average method. The theorem is as follows:

If there are two sample classes, the shortest distance is

$$D(G_1, G_2) = \min_{\substack{x_i \in G_1 \\ y_j \in G_2}} \{d(x_i, y_j)\} \quad (6-3)$$

Among them, the distance between two sample points is intuitively the distance between the nearest two points in the two classes.

#### Class averaging

$$D(G_1, G_2) = \frac{1}{n_1 n_2} \sum_{x_i \in G_1} \sum_{y_j \in G_2} d(x_i, y_j) \quad (6-3)$$

It is equal to the average of the distance between two sample points in  $G_1, G_2$ , where  $n_1, n_2$  is the number of sample points in  $G_1, G_2$ .

Write a program through MATLAB, the classification is as follows:

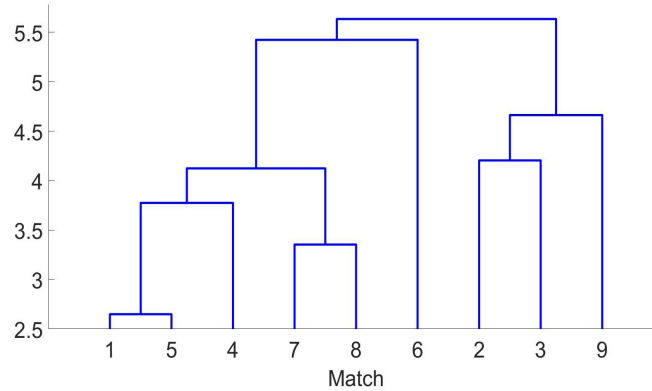


Figure 10: Cluster tree diagram

The **clustering results** are:

For staffing:

Coach 1: 5 types; Coach 2: 4 types; Coach 3: 8 types; 9 types of each other.

For formation:

There are 3-node coach 1: 3 types; coach 2: 3 types; coach 3: 3 types; opponent: 5 types.

There are 4-node coach 1; 5 types; coach 2: 2 types; coach 3: 2 types; opponent: 3 types.

## 6.3 Analysis of results

Then, based on the scoring rate for each game,

$$\text{Scoring rate per game} = \frac{\text{Huskies formation scoring rate}}{\text{Huskies formation scoring rate} + \text{opponent formation scoring rate}}$$

Calculate the winning ratio of the Huskies i-th formation to the enemy's j-th formation, match the winning ratio of the i-th personnel match with the i-th personnel match. Finally get the winning matrix of each coach's 3, 4 node subgraph. According to the game theory, the optimal strategy

and the optimal winning percentage are obtained. The winning percentage will fluctuate between the winning percentages according to the level of the teams encountered.

### 6.3.1 Decision Theory Finds Results

After obtaining the winning matrix for each coach's different nodes, we found that each winning matrix has saddle points, that is,  $\{ \text{Let } f(x,y) \text{ be a time-valued function defined on } x \in A \text{ and } y \in B. \text{ If } x^* \in A, y^* \in B \text{ exists and } f \text{ exists, then } f(x,y^*) \leq f(x^*,y^*) \leq f(x^*,y), (x^*,y^*) \text{ is called saddle point of function } f. \}$ . As long as the opponent does not change the strategy at this time, any player in any game It is impossible to increase the winning or reduce the loss by changing the strategy, that is, we have found the optimal strategy.

Coach 1: For the three-node formation selection, selection strategy 1 will win the game with a 40% probability. For the four-node formation selection, selection strategy 1 will win the game with a 60% probability.

Coach 2: For a three-node formation, choose strategy 1, with a 38% probability of winning the game, and for a four-node formation, choose strategy 1, with a 60% probability of winning the game.

Coach 3: For a three-node formation, choose Strategy 2, with a 35% probability of winning the game, and for a four-node formation, choose Strategy 2, with a 44% probability of winning the game.

Based on the formation, we add the situation of staffing. The types of coaching staff are as follows:




<b>Coach 1</b>		Choosing this strategy has more than 33% expectation of victory.
<b>Coach 2</b>		Choosing this strategy has more than 33% expectation of victory.
<b>Coach 3</b>		Choosing this strategy has more than 50% expectation of victory.

Figure 11: Three strategies

Can be seen from the above figure:

Coach 1: Choose Strategy 3, with 33% expecting to win 541; Coach 2: Choose strategy 1, with 33% expecting to win 343; Coach 3: Choose strategy 8 with a 50% expectation of winning 640.

### 6.3.2 Network analysis part

One of the three or four nodes in the opponent's first and fourth games is as follows:

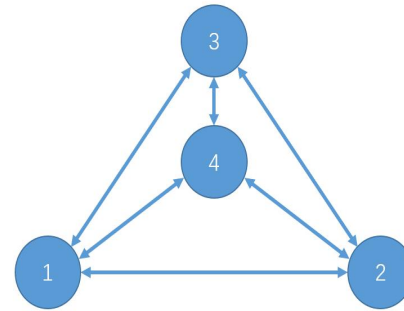
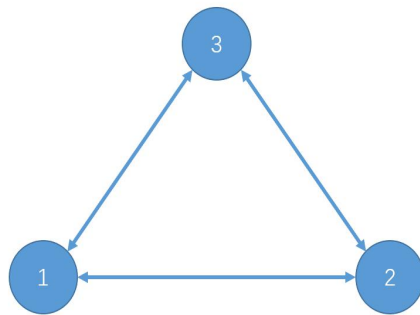


Figure 12: First three-node subgraph      Figure 13: First node four-node graph

If we want to disrupt the way our opponents play, we must first know the passing patterns that teams often use. Determining how to break these patterns will make a significant contribution to the success of the team. System clustering divides games with similar frequencies into a strategy. Game theory is to find the best tactics of their own against other teams, but it is still possible to lose the game due to team abilities and enemy tactics.

Therefore, we find the enemy strategy that is most likely to let us lose the game, and analyze the frequency of various networks of this strategy and destroy the enemy's more frequent passing network in future games, such as game 4. Their most commonly used ternary and quaternary passing network is (Figure 12, Figure 13). So we need to deliberately prevent opponents from forming the cooperative network in the game, but the most important flaw of the analysis method of passing network is that the analysis The player's position on the court is not considered in any way.

## 7 Task 4: Summarize our findings

### 7.1 Analysis of the problem

We divided this question into two questions. One is to build a model based on our first three questions to explain how to better design the team. The other is to make changes on the basis of this model and add other factors to make the model more Universal. First of all, for the first question, we divided the team into two aspects: ability and strategy. Ability focuses on the individual aspects of the team. The strategy focuses on the entire team. The key is execution, team morale.

### 7.2 Analysis of results

We divided this question into two questions. One is to build a model based on our first three questions to explain how to better design the team. The other is to make changes on the basis of this model and add other factors to make the model more Universal. First of all, for the first question, we divided the team into two aspects: ability and strategy. Ability focuses on the individual aspect of the team, and strategy focuses on the entire team.

In terms of team strategy, we can find the optimal network from all the cooperative networks, and we can analyze and obtain the subnets most commonly used by the opponent, thereby changing the team's network structure to confront each other. According to the capabilities of the team members, we use The strategy is also different. In terms of team capabilities, we are divided into three aspects: structure, dynamics, and configuration. The

cooperation structure depends on the professionalism and coordination of members; the dynamics of cooperation depends on the emergency capabilities of team members; the configuration of cooperation depends on the technical capabilities and stress resistance of members .

For a more intuitive description, we have drawn the following mind map:

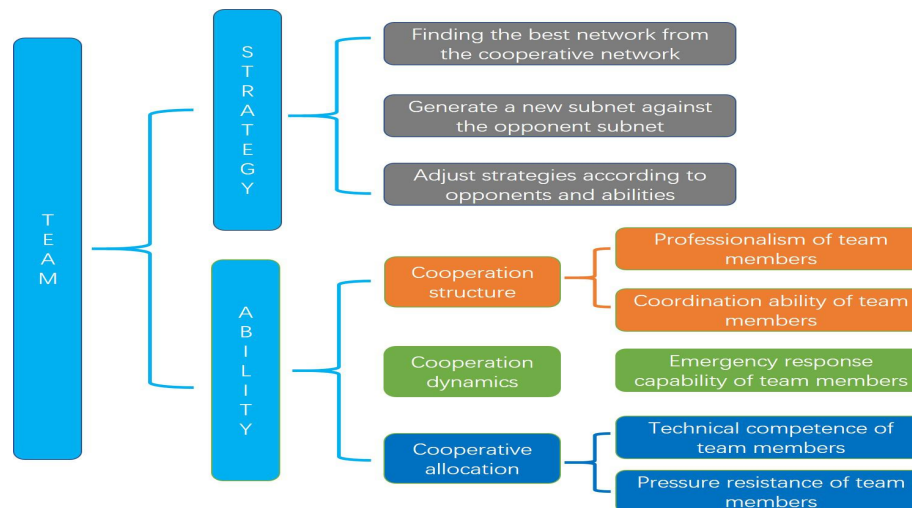


Figure 16: Mind map of team strategy

For the second question, we believe that those factors that are difficult to digitize are the key to improving the model. Some of the factors in the figure below:

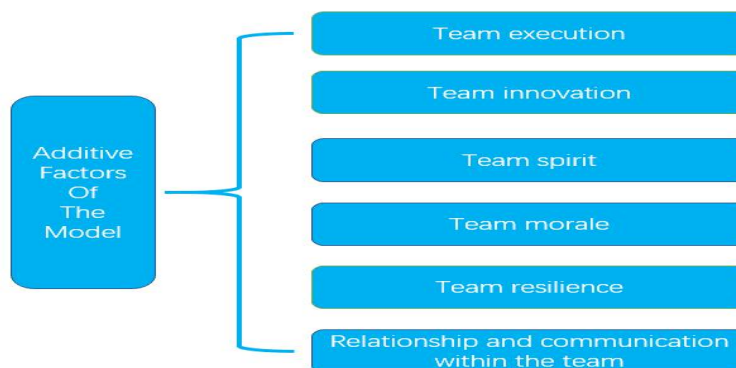


Figure 17: Other information

Among them, team morale and spirit are aspects we need to quantify. The flexibility and execution of the team in frustration can be obtained through more detailed court information. Relations and communication within the team can be obtained by putting a questionnaire in the team. For the team's innovation, we need a longer time span and players' positions per game to obtain. That is, if we can provide more accurate information, we can make a general model of team performance more comprehensive.

## 8 Model Extensions

(1) If we can know the time of the goal, we can analyze it according to the indicators before and after the goal so that we can provide the coach with more detailed advice for each game.

- (2) By studying the various indicators before and after the substitution time, you can see the cooperation between the player and the team after playing.
- (3) We can quantify more indicators of the team by putting questionnaires in the team, and understand the team's such as execution, innovation, and internal relationships in various ways, so as to specify the optimal formation dedicated to each team.
- (4) We can extend the performance of football teams to various areas of teamwork or industries in which both teams compete; We can use the ullmann algorithm to extract more networks with four nodes or more in the network, but the running time will also increase.
- (6) If we can get more records of Huskies' past games, we will give the best formation for different teams through statistics.

## 9 Conclusion

### 9.1 Strengths

- (1) Factor analysis is not a choice of the original variables, but recombination based on the information of the original variables  
Find common factors that affect variables, simplify the data, and make the factor variables interpretable according to the rotation, making the naming clear.
- (2) Use cluster analysis to study different formations or strategies
- (3) Index selection We look for indicators from the micro, macro, short-term, and long-term aspects of time and space.
- (4) The third question is that our suggestions for coaches are divided into team ability and formation or tactics. In addition, we also analyzed the more powerful formations that may occur in the place.

### 9.2 Weaknesses

- (1) Factor analysis In calculating the factor score, the least square method is used, and this method may sometimes fail.
- (2) Disadvantages It is difficult to obtain clustering conclusions when the sample size is large. Because the similarity coefficient is based on the reflection of the participants to establish an index that reflects the internal connection between the participants, and in practice, sometimes although they have found a close relationship between them from the data obtained by the participants, there is nothing between them. Intrinsic connection. At this time, it is obviously inappropriate to obtain the results of cluster analysis based on distance or similarity coefficient. However, the cluster analysis model itself cannot identify such errors.
- (3) Because there is no specific time and space data for each player, network analysis has limitations.

## 10 References

- [1] Pappalardo, L., Cintia, P., Rossi, A. et al. A public data set of spatio-temporal match events in soccer competitions. *Sci Data* 6, 236 (2019).
- [2] Ahnert, S. E., Garlaschelli, D., Fink, T. M. A. & Caldarelli, G. Ensemble approach to the analysis of weighted networks. *Phys. Rev. E* 76, 016101 (2007).
- [3] Buldú, J.M., Busquets, J., Echegoyen, I. et al. (2019). Defining a historic football team: Using Network Science to analyze Guardiola's F.C. Barcelona. *Sci Rep*, 9, 13602.
- [4] Almendral, J. A. & Díaz-Guilera, A. Dynamical and spectral properties of complex networks. *New Journal of Physics* 9, 187 (2007).
- [5] Cintia, P., Rinzivillo, S. & Pappalardo, L. A network-based approach to evaluate the performance of football teams. In *Machine Learning and Data Mining for Sports Analytics Workshop*, Porto, Portugal (2015).
- [6] Cotta, C., Mora, A. M., Merelo, J. J. & Merelo-Molina, C. A network analysis of the 2010 FIFA world cup champion team play. *J. Syst. Sci. Complex.* 26, 21 (2013).
- [7] Mendes Bruno, Clemente Filipe Manuel, Maurício Nuno. Variance in Prominence Levels and in Patterns of Passing Sequences in Elite and Youth Soccer Players: A Network Approach.[J]. *Journal of human kinetics*, 2018, 61.
- [8] Thomas U. Grund. Network structure and team performance: The case of English Premier League soccer teams[J]. *Social Networks*, 2012, 34(4).
- [9] Gao Yang, Zhang Yanping, Qian Fulan, Zhao Yan. Link prediction algorithm combining node degree and node clustering coefficient [J]. *Small Micro Computer System*, 2017, 38 (07): 1436-1441.
- [10] Hughes M, Franks I (2005) Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Science* 23: 504–514.
- [11] Scott J (2000) *Social Network Analysis: A Handbook*. London, UK: SAGE Publications Ltd., 2 edition.
- [12] Guimera` R, Uzzi B, Spiro J, Amaral L (2005) Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308: 697–702.
- [13] <https://blog.csdn.net/wz11997/article/details/79059034>
- [14] <https://blog.csdn.net/DOUBLE121PIG/article/details/100881373>

# 11 Appendix

## Appendix 1

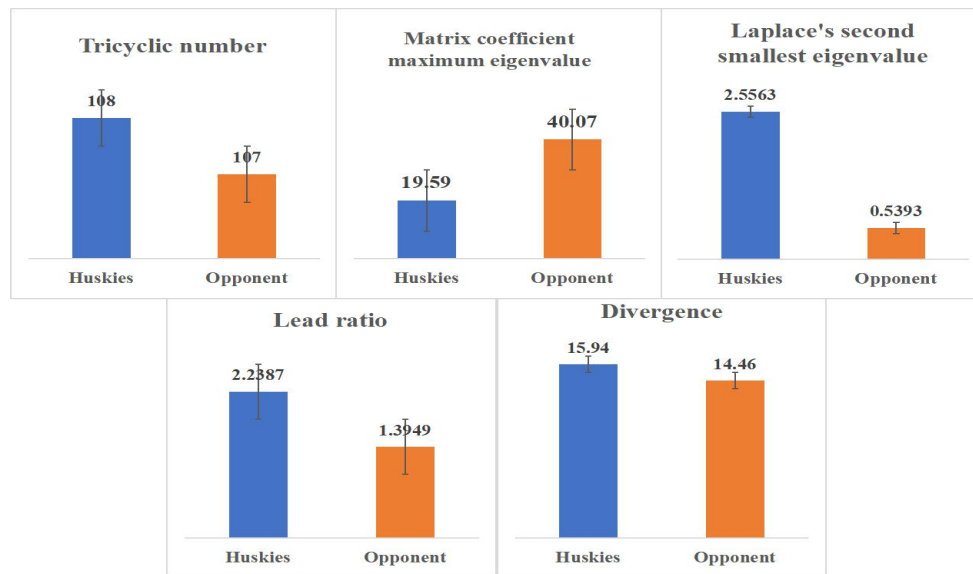


Figure 1: Comparison of 5 network parameters

## Appendix 2

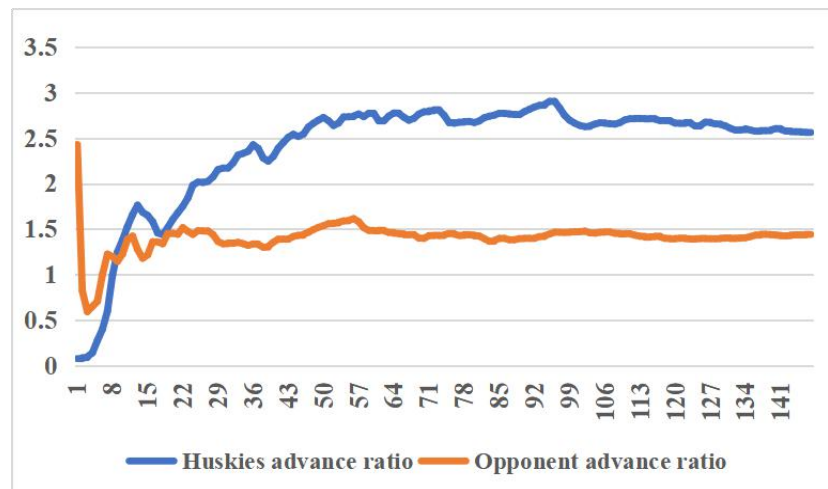


Figure 2: Advance ratio



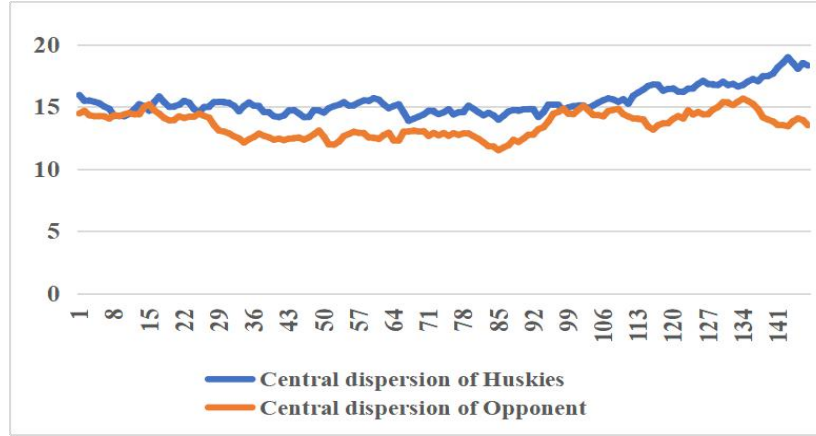


Figure 3: Central dispersion

## Appendix 4

### Variable normalization

Converting each indicator  $a_{ij}$  into a standardized indicator  $\bar{a}_{ij}$ , there are:

$$\bar{a}_{ij} = \frac{a_{ij} - \bar{\mu}_j}{s_j}, i = 1, 2, \dots, 20; j = 1, 2, \dots, 7 \quad (1)$$

Among them,  $\bar{\mu}_j = \frac{1}{20} \sum_{i=1}^{20} a_{ij}$ ,  $s_j = \sqrt{\frac{1}{20-1} \sum_{i=1}^{20} (a_{ij} - \bar{\mu}_j)^2}$ . This is,  $\bar{\mu}_j, s_j$  is the mean and standard deviation of the j-th index. The standardized indicator variables are:

$$\bar{x}_j = \frac{x_j - \bar{\mu}_j}{s_j}, j = 1, 2, \dots, 7 \quad (2)$$

### Calculate correlation coefficient matrix R

Correlation coefficient matrix is  $R = (r_{ij})_{7 \times 7}$ , where  $r_{ij} = 1, r_{ij} = r_{ji}$ ,  $r_{ij}$  is the correlation coefficient between the i-th index and the j-th index, there are:

$$r_{ij} = \frac{\sum_{k=1}^{20} \tilde{a}_{ki} \cdot \tilde{a}_{kj}}{20-1}, (i, j = 1, \dots, 7) \quad (3)$$

### Calculate Elementary Load Matrix

Calculate the eigenvalues  $\lambda_1 \geq \dots \geq \lambda_7 \geq 0$  of the correlation coefficient matrix  $R$  and the corresponding eigenvectors  $\mu_1, \dots, \mu_7$ , where  $\mu_j = [\mu_{1j}, \dots, \mu_{7j}]^T$ , the elementary load matrix is

$$A = [\sqrt{\lambda_1} \mu_1, \sqrt{\lambda_2} \mu_2, \dots, \sqrt{\lambda_7} \mu_7] \quad (4)$$

### Select $m(m \leq 4)$ main factors

According to the elementary load matrix, the contribution rate of each common factor is calculated, and m main factors are selected. The extracted factor load matrix is rotated to obtain a matrix (where the first m columns are,  $T$  is an orthogonal matrix), and the factor model is constructed as follows:

$$\begin{cases} x_1 = \alpha_{11}F_1 + \dots + \alpha_{1m}F_m \\ x_2 = \alpha_{21}F_1 + \dots + \alpha_{2m}F_m \\ \vdots \\ x_7 = \alpha_{71}F_1 + \dots + \alpha_{7m}F_m \end{cases} \quad (5)$$

Use regression to find the score function of each factor:

$$\hat{F}_j = \beta_{j1}x_1 + \beta_{j2}x_2 + \dots + \beta_{j4}x_4 \quad (6)$$

Where  $j$  is from 1 to  $m$ . Then

$$\begin{bmatrix} \beta_{11} & \dots & \beta_{m1} \\ \vdots & & \vdots \\ \beta_{17} & \dots & \beta_{m7} \end{bmatrix} = R^{-1}A_2 \quad (7)$$

$$\hat{F} = (\hat{F}_{ij})_{20 \times m} = X_0 R^{-1}A_2 \quad (8)$$

Where  $X_0 = (\tilde{a}_{ij})_{20 \times 7}$  is the normalized data matrix of the original data;  $R$  is the correlation coefficient matrix;  $A_2$  is the load matrix obtained in the previous step.

Table 1: SPSS Derived Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.456	49.369	49.369	3.456	49.369	49.369	3.032	43.312	43.312
2	1.330	18.999	68.368	1.330	18.999	68.368	1.754	25.056	68.368
3	0.898	12.834	81.201						
4	0.529	7.561	88.763						
5	0.431	6.157	94.920						
6	0.245	3.500	98.420						
7	0.111	1.580	100.000						

Note: Extraction method is principal component analysis.

## Appendix 5

算法如下:

```

Step 1   $M = M^0$ ,  $d := 1$ ;  $H_1 = 0$ ;
        for all  $i = 1, \dots, p_\alpha$ , set  $F_i := 0$ ;
Step 2  If there is no value of  $j$  such that  $m_{dj} = 1$  and  $F_j = 0$  then go to step 7;
         $M_d = M$ ,
        if  $d = 1$  then  $k := H_1$  else  $k := 0$ ,
Step 3   $k := k + 1$ ,
        if  $m_{dk} = 0$  or  $F_k = 1$  then go to step 3;
        for all  $j \neq k$  set  $m_{dj} := 0$ ,
Step 4. If  $d < p_\alpha$  then go to step 6 else use condition (1) and give output if an isomorphism is found;
Step 5  If there is no  $j > k$  such that  $m_{dj} = 1$  and  $F_j = 0$  then go to step 7;
         $M := M_d$ ,
        go to step 3;
Step 6   $H_d = k$ ,  $F_k = 1$ ;  $d = d + 1$ ;
        go to step 2,
Step 7  If  $d = 1$  then terminate algorithm,
         $F_k = 0$ ;  $d := d - 1$ ,  $M = M_d$ ,  $k := H_d$ ,
        go to step 5,

```

Figure 4: ullmann algorithm design

## Appendix 6

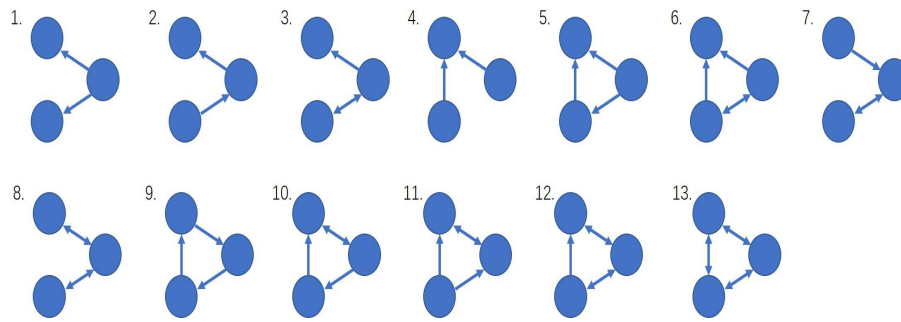


Figure 4: Thirteen differently constructed three-node subgraphs

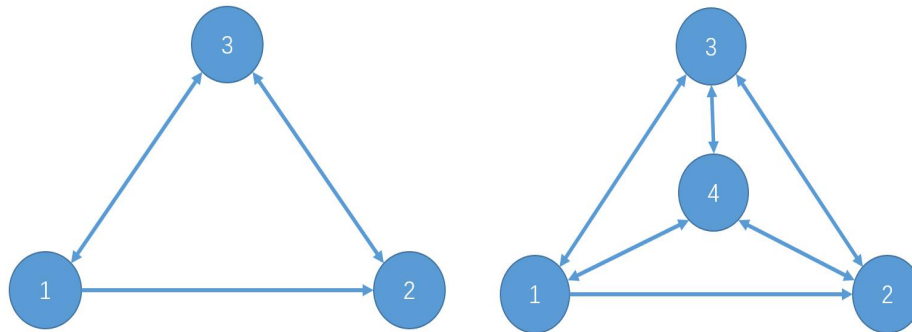


Figure 5: Fourth node three-node graph    Figure 6: Fourth node three-node graph

## Appendix 7

### Program 1.1 Calculating the shortest path

```
function [distance,path]=Dijkstra
(A,s,e)
% returns the distance
and path between the start
node and the end node.
% A: adjacent matrix
% s: start node
% e: end node
% initialize
n=size(A,1);% node number
D=A(s,:);% distance vector
path=[];% path vector
visit=ones(1,n); % node visibility
visit(s)=0;% source node is invisible
parent=zeros(1,n);
% parent node

% the shortest distance
for i=1:n-1 % BlueSet has n-1 nodes
    temp=zeros(1,n);
    count=0;
    [value,index]=min(temp);
    j=index; visit(j)=0;
    for k=1:n
        if D(k)>D(j)+A(j,k)
            D(k)=D(j)+A(j,k);
            parent(k)=j;
        end
    end
    end
    distance=D(e);

% the shortest distance path
if parent(e)==0, return; end
path=zeros(1,2*n);% path preallocation
t=e; path(1)=t; count=1;
while t~=s && t>0
    p=parent(t);
    path=[p path(1:count)];
    t=p; count=count+1;
end
if count>=2*n, error(['The path
```

<pre> for j=1:n     if visit(j)         temp=[temp (1:count) D(j)];     else         temp=[temp (1:count) inf];     end     count=count+1; end </pre>	<pre> preallocation length is too short.',...     'Please redefine path preallocation parameter.']); end path(1)=s; path=path(1:count); </pre>
---	--

### Program 1.2 Calculate the Laplacian matrix

<pre> function lamda2_A = laplace(A,t) l=11; if t~=0 A(t,,:)=[]; A(:,t,:)=[]; l=10; end Laplace_a=A; for i=1:20     for j=1:l         Laplace_a(j,j,i)=-sum(A(j,,:i));     end end </pre>	<pre> for i=1:20     laplace_e=eig(A(:,i,i));     [~,id1]=min(laplace_e);     laplace_e(id1)=[];     lamda2_A(i)=min(laplace_e); end end </pre>
---	---

### Program 1.3 Draw the passing network

<pre> clear clc close all %% [~,~,Data] = xlsread('data1.xlsx'); data = Data(2:end,3:4); num = unique(data); data = char(data); num = char(num); result = zeros(11,11); for i = 1:11     for k = 1:length(data)/2         if strcmp(num(i,:),data(k,1:2)) == 1             for j = 1:11                 if                     strcmp(num(j,:),data(369+k,1:2)) == 1 </pre>	<pre>                         if strcmp(num(i,:),data(k+369,1:2))                             == 1                                 Num(i) = Num(i)+1;                         end                     end                 end             end             Num =             Num/sum(Num)*80;Num=[Num',Num'];             %% Drawing             IDS={'D1','D2','D3','D4','F1','F2','F3','G1'             ,'M1','M2','M3'};             bg=biograph(result,IDS);             set(bg.nodes,'shape','circle','color',[1,1,1],'             lineColor',[1,0,0]);             set(bg,'layoutType','radial');             set(bg,'edgeType','straight'); </pre>
---	---

<pre>                 result(i,j) = result(i,j)+1;             end         end     end end %%% Find a location xy = cell2mat(Data(2:end,8:9)); Site = zeros(11,2); for i = 1:11     t = 0; site = [];     for k = 1:length(data)/2         if strcmp(num(i,:),data(k,1:2)) == 1             t = t+1;             site(t,:) = xy(k,:);         end     end     Site(i,:) = mean(site); end %%% Find the size Num = zeros(1,11); for i = 1:11     t = 0;     for k = 1:length(data)/2 </pre>	<pre> bg.showWeights='on'; bg.nodeAutoSize = 'off'; set(bg.nodes,'textColor',[0,0,0],'lineWidth',2,'fontSize',10); set(bg,'arrowSize',2,'edgeFontSize',10); dolayout(bg); for i = 1:11     bg.nodes(i).Position = Site(i,:); end dolayout(bg, 'Pathonly', true); for i = 1:11     set(bg.nodes(i),'size',Num(i,:)) end dolayout(bg, 'Pathonly', true); h = view(bg); for i = 1:11     set(h.nodes(i),'size',Num(i,:)) end </pre>
--	--

### Program 2.1 Computing adjacent matrices

<pre> function [distance_H,distance_O] = mind(H,O) name_H = unique(H(:,3)); name_O = unique(O(:,3)); data_O = O(:,3:4); data_H = H(:,3:4); result_H = zeros(length(name_H),length(name_H)); result_O = zeros(length(name_O),length(name_O)); for i = 1:length(name_H)     for k = 1:length(H)         if             strcmp(char(name_H(i,:)),char(data_H(k,1             ))) == 1                 for j = 1:length(name_H)                     if                         strcmp(char(name_H(j,:)),char(data_H(k,2                         ))) == 1                             result_H(i,j) = </pre>	<pre> result_O(i,j)+1;                     end                 end             end         end     end end  d_H = 1./result_H;d_O = 1./result_O;distance_H = zeros(length(d_H),length(d_H));distance_ O = zeros(length(d_O),length(d_O)); for i = 1:length(d_H)     for j = 1:length(d_H)         distance_H(i,j) =             dijkstra(d_H,i,j)+distance_H(i,j);     end end distance_H(distance_H == inf) = 0; for i = 1:length(d_O)     for j = 1:length(d_O) </pre>
---	--

<pre> result_H(i,j)+1;         end     end end end for i = 1:length(name_O)     for k = 1:length(O)         if strcmp(char(name_O(i,:)),char(data_O(k,1 ))) == 1             for j = 1:length(name_O)                 if strcmp(char(name_O(j,:)),char(data_O(k,2 ))) == 1                     result_O(i,j) = </pre>	<pre>         distance_O(i,j) = dijkstra(d_O,i,j)+distance_O(i,j);         end     end distance_O(distance_O == inf) = 0; distance_H = sum(sum(distance_H)); distance_O = sum(sum(distance_O));  distance_H = distance_H/(length(d_H)*(length(d_H)- 1)); distance_O = distance_O/(length(d_O)*(length(d_O)- 1)); </pre>
---	---

### Program 2.2 The number of times all tricks occurred in the match

<pre> clear clc %%% [~,~,Name] = xlsread('kind.xlsx'); [~,~,Data] = xlsread('fullevents.xlsx'); Name = Name(2,2:end); data_kind = zeros(39,37); %%% Points match for k = 1:38     t = 0;     for i = 2:length(Data)         if cell2mat(Data(i,1)) == k             t = t+1;             data(t,:) = Data(i,:);         elseif cell2mat(Data(i,1)) &gt; k             break         end     end end %%% tt = 0; ttt = 0; for i = 1:length(data)     t = char(data(i,2));     if strcmp(t(1),'H') == 1         tt = tt+1;         H(tt,:) = data(i,:);     else         ttt = ttt+1;         O(ttt,:) = data(i,:); </pre>	<pre> end %%% Small class data clearvars -except data_kind [~,~,Class] = xlsread('matches.xlsx'); class = char(Class(2:end,2)); class = class(:,end-1:end); class = str2num(class); data_kind(2:end,end+1) = class; t = zeros(19,38); for i = 1:19     for j = 2:39         if data_kind(j,end) == i             t(i,:) = t(i,:) + data_kind(j,:);         end     end end kind = [data_kind(1,:);t]; kind(:,end) = []; %%% Big class data clearvars -except kind data_Kind = {'Duel','Foul','Free Kick','Goalkeeper leaving line','Interruption','Offside','Others on the ball','Pass','Save attempt','Shot','Substitution'}; num = [4 8 7 1 2 1 3 7 2 1 1]; Kind = zeros(20,length(data_Kind)); for i = 1:20     t = 0; </pre>
--	--

<pre> end end %% Count for i = 1:length(H)     if isnan(H{i,8}) == 1         data_kind(1,23) = data_kind(1,23)+1;     else         for j = 1:length(Name)             if j == 23                 continue             end             if strcmp(char(H(i,8)),char(Name(j))) == 1                 data_kind(1,j) = data_kind(1,j)+1;             end         end     end end for i = 1:length(O)     if isnan(O{i,8}) == 1         data_kind(k+1,23) = data_kind(k+1,23)+1;     else         for j = 1:length(Name)             if j == 23                 continue             end             if strcmp(char(O(i,8)),char(Name(j))) == 1                 data_kind(k+1,j) = data_kind(k+1,j)+1;             end         end     end end end </pre>	<pre> for j = 1:length(data_Kind)     Kind(i,j) = sum(kind(i,t+1:t+num(j)));     t = num(1); end end %% clearvars -except kind Kind num ave_kind = kind./2; ave_kind(1,:) = ave_kind(1,:)/19; ave_Kind = Kind./2; ave_Kind(1,:) = ave_Kind(1,:)/19; zhi = zeros(1,20); for i = 1:20     t = 0;t3 = 0;     for j = 1:11         t1 = 0;         for k = 1:num(j)             t1 = (ave_kind(i,t+i)- ave_Kind(i,j))^2 + t1;         end         t = num(j);         t2 = sqrt(t1/num(j));         t3 = ave_Kind(i,j)/(sum(ave_Kind(i,:))- 1)*t2+t3;     end     zhi(i) = t3; end zhi = zhi'; </pre>
---	---

**Program 3.1 Ullmann algorithm**

```

function r=Ullmann(a,b)
% Ullmann algorithm: the simplest of
the subgraph isomorphic algorithms
% Determine if there is a subgraph
isomorphic in a
[p1,~]=size(a);

```

```

m=matrixlist(:,d);
end
while k<=p2
    if m(d,k)==1&&F(k)==0
        break;
    end

```

<pre> [p2,~]=size(b); da=sum(a); db=sum(b); m=zeros(p1,p2); for j=1:p2     for i=1:p1         if db(j)&gt;=da(i)             m(i,j)=1;         end     end end for i=1:p1     if max(abs(m(i,:)))==0         r=0;return     end end F=zeros(p2,1);H=zeros(p1,1); matrixlist=zeros([size(m),p1]); d=1; k=1; while 1     if H(d)==0         k=1;         matrixlist(:,d)=m;     else         k=H(d)+1; </pre>	<pre>         k=k+1;     end     if k==p2+1         H(d)=0;d=d-1;     else         m(d,:)=zeros(1,p2);m(d,k)=1;         H(d)=k;F(k)=1;d=d+1;     end     if d==0         r=0;return;     end     if d==p1+1         c=m*b*m';         tmp=1;         for i=1:p1*p1             if a(i)==1 &amp;&amp; c(i)~=1                 tmp=0;break;             end         end         if tmp             r=1;return;         else             d=p1;         end     end end </pre>
---	--

<b>Program 3.2 Calculate the type and number of four nodes for each team</b>	
<pre> function result = huan_4(linjie) %Input adjacency matrix and matching set load pipei n = 11; num = linjie; num(num == 0) = inf; linjie = linjie./linjie; linjie(isnan(linjie))=0; result = zeros(1,length(pipei)); for i = 1:8     for j = i+1:9         for k = j+1:10 </pre>	<pre>             t = linjie([i,j,k,y],[i,j,k,y])             t2 = num([i,j,k,y],[i,j,k,y]);             t1 = 0;             for k1 = 1:length(pipei)                 t1 = t1+1;                 if t == pipei(:,k1)                     result(t1) = result(t1)+min(min(t2))*1;                 end             end         end     end end </pre>