Problem Chosen	Fall 2021 Mathematical Modelling and Practice	Author
D		Zeyu Zheng

Improving Team Performance During a Football Match

Summary

"If you want to go fast, go alone. If you want to go far, go together."

In order to quantify the impact of team and player performance on team goals in soccer, we were commissioned by the Huskies coach to quantify the data from 38 games in a season for the Huskies and to formalize the structure and dynamic metrics in the passing network.

First, we adopted a social network analysis-based treatment for Problem 1 and constructed a weighted directed graph about the passes and its weighted adjacency matrix. The team performance was analyzed separately from the team scale and the time scale. The team scale included the overall network, local network and individual network. The overall network was analyzed for the characteristics of the weighted adjacency matrix and the pass center dispersion, maximum radius and average clustering coefficient. The local network analysis includes core analysis and vertex configuration, where core analysis is used to distinguish core players from edge players, and vertex configuration specifically analyzes dyadic and triadic combination forms and ratios. And the individual network analysis analyzes the degree centrality and eigenvector centrality of the vertices. Meanwhile, the variation of goals in the time series was analyzed on the time scale.

Second, to address problem two, we analyzed in detail the many factors that influence team performance through performance metrics and team-level processes, and quantified these factors in relation to known passing data. After correlation analysis and multivariate linear analysis of these factors as independent variables with the number of goals against, eligible control variables were screened. These variables were utilized to provide a comprehensive portrayal of the Huskies' season performance at three levels: structural, configurational, and dynamic.

In addition we tested the proposed hypothesis model against the proposed hypothesis model and made targeted recommendations to the coach on the basis of the optimized model for the weak areas of the Huskies. We came up with three strategies to strengthen the tacit cooperation within the group, improve the team's ball control skills and control the fast pace of the opening game.

Finally, we extended from the model metrics to a broad social network and summarized recommendations for improving group cooperation from a team perspective. The model associated with this paper makes full use of data related to passing and the resulting structure has been validated and can be used to assess the level of passing cooperation of teams and also to further describe the effectiveness of cooperation in social groups.

Keywords: social network , regression and analysis, network patterns and properties.

Contents

1 Introduction

1.1 Problem Background

"Alone we can do so little; together we can do so much."

Great things are accomplished through cooperation, and the charm of teamwork is much more than 1+1=2, it can create many new possibilities. To take advantage of the strengths of a team, it requires not only the outstanding ability of individual team members, but also the use of cooperative strategies and the characteristics of organizational coordination to achieve the effect of 1+1>2. To fully understand the approach of teamwork, it is necessary to use sociological theories and the methods of big data.

With the progress of the times and the popularity of information technology, people are using big data to analyze problems more and more often. In the face of increasingly complex problems, conventional means of solving them may be out of reach, while big data can often find alternative ways to find solutions to complex problems. For example, soccer, as a long history of team sports, the technical and tactical aspects of soccer seem to have been mined thoroughly enough by traditional methods, but improving team rankings by analyzing players' teamwork data is what all team coaches want. Can new discoveries be made by analyzing soccer season performance through big data, which in turn can provide strategic guidance for team coaches?

1.2 Problem Restatement

The Huskies coach wanted ICM to help them analyze their team's data, and we were given quantitative data from the team's previous season. Overall, the data covers 23,429 passes between 366 players (30 Huskies players, and 336 players from opposing teams), and 59,271 game events.

At the request of the Huskies coach, our team was to analyze the form of player interaction on the field and the dynamic characteristics of team success or failure.

- Create a team passing network composed of players and explore different scales of passing network patterns.
- Develop appropriate performance metrics so that these metric models can be used to evaluate whether the strategy is effective.
- Provide formation strategies to coaches through network analysis.
- Summarizing the lessons learned from the above analysis and outlining the important factors for improving team effectiveness.

2 Basic Assumption

- There is no difference between players, and all players on the field are numbered in order in the passing network.

- Home and away games and coaching changes have no bearing on team performance.
- Further analysis of the weighted adjacency matrix ignores the effect of the self-loop, i.e., the player's own pass to himself.
- The respective variables in the pass network model are not related to each other and are linearly related to the dependent variable.

3 Notations

The primary notations used in this paper are listed in Table 1.

Table 1: Notations

Symbol	Definition
Adj	The weighted adjacency matrix
(x_0, y_0)	The x, y coordinates of the network centroid
(x_i, y_i)	The x, y coordinates of each pass
C_{disp}	Center dispersion of pass coordinates
N	Total number of passes
n	Number of vertices in the graph
ω_{ij}	Number of passes from i to j
R_max	Maximum network radius
$C_w(i)$	Clustering coefficient of vertex i
$C_w(i)$	Clustering coefficient of vertex i
C	The average clustering coefficient of the graph
C_D	Degree centrality of the vertex
C_E	Eigenvector centrality of the vertex
d_i	Degree of the vertex i
x	Eigenvector of the adjacency matrix
χ	Independent variable
ψ	Dependent variable

4 Model

4.1 Data Pre-processing

- To be more intuitive and realistic, we converted the coordinates of the starting and ending points of the passes in the given data to the x and y coordinates of the range [0,100] to the dimensions of a standard World Cup soccer field, 105 meters long and 68 meters wide.
- Treat the Huskies' and opponents' passing data separately and ignore the players' passing data to themselves.

4.2 The Football Passing Network

It is generally believed that a network is a collection of vertices and certain relationships between them, and a social network is a collection of social actors as vertices and the relationships between them. Soccer is a same-field confrontation sport, and because of the large playing field, passing becomes the main means for the attacking team to control the ball and create scoring opportunities, so passing is the most used and most important technique in soccer. For the attacking team, the route of the ball in the process of passing is like a network that organically connects all players and lines. The more passes players make to each other, the closer the relationship becomes. We can abstractly consider the passing relationship between players and players as a kind of "social network", in which players are the "network vertices", and the passing between players can be seen as a social network between them. Passes between players can be seen as "edges" between them. The players (vertices) and the passes (edges) constitute the "social network" in the soccer game, which we call "passing network". The isomorphism of "passing network" and "social network" indicates the feasibility of quantitative study of "passing network" in soccer game by using social network analysis method.

4.2.1 The whole network

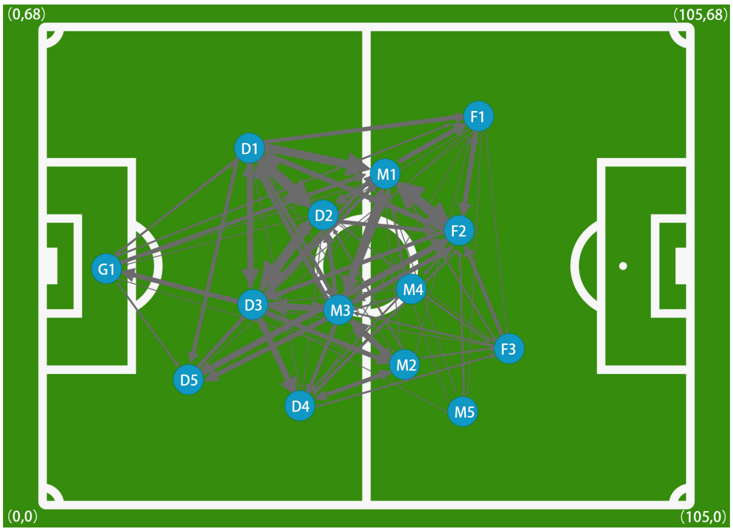


Figure 1: Huskies' passing network of match 1

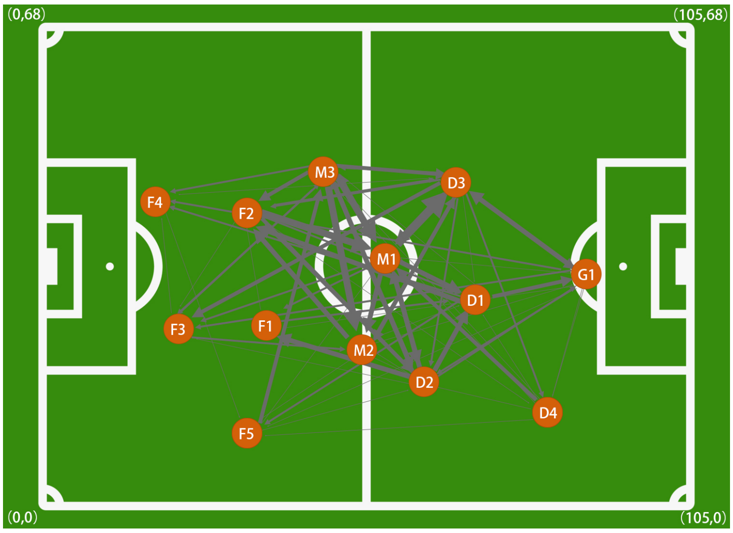


Figure 2: The opponent1's passing network of match 1

As shown in Figures 1 and 2, the passing network in a game consists of vertices and directed edges. The vertices in the network represent the individual players on the court, and to simplify the numbering, they are subsequently numbered in the order 1, 2, 3... in order instead. The directed edges in the network represent the direction of the pass, and the width of the line in the diagram represents the number of passes, the wider the line the greater the number of passes.

The matrix of the passing network contains the numbers of the serving and receiving players and the positions of the serving and receiving players (x and y coordinates).

The players (vertices) and the directed passing routes (edges) of the passing network form a weighted directed graph.

4.2.2 Network patterns and properties

This paper analyzes the team's passing network from macro (whole team), local (passing group) to micro (individual player). In addition, the identification of different network patterns within the passing group and their configurations will be analyzed.
(1) The whole network

For the macro passing network of the whole team, we analyze the weighted adjacency matrix and aggregation coefficients composed of the passing data and calculate the team structure of the whole game such as the centroid and dispersion of the network and the maximum radius of the network.

- The weighted adjacency matrix

A directed graph composed of a passing network can generate an adjacency matrix that represents the relationship between vertices. The adjacency matrix is an $n \times n$ square matrix with matrix dimension n being the number of vertices, and in this matrix implementation, each row and column represents a vertex in the graph. The value stored in the cell at the intersection of row v and column w indicates the presence or absence of an edge from vertex v to vertex w. The value stored in the cell at the intersection of row v and column w indicates the presence or absence of an edge from vertex v to vertex w. When two vertices are connected by an edge, we say that they are adjacent. The values in the cells indicate the weights of the edges from vertex v to vertex w.

For example, the weighted directed graph and weighted adjacency matrix of the Huskies in Match 1 are shown in Figures 3 and 4.

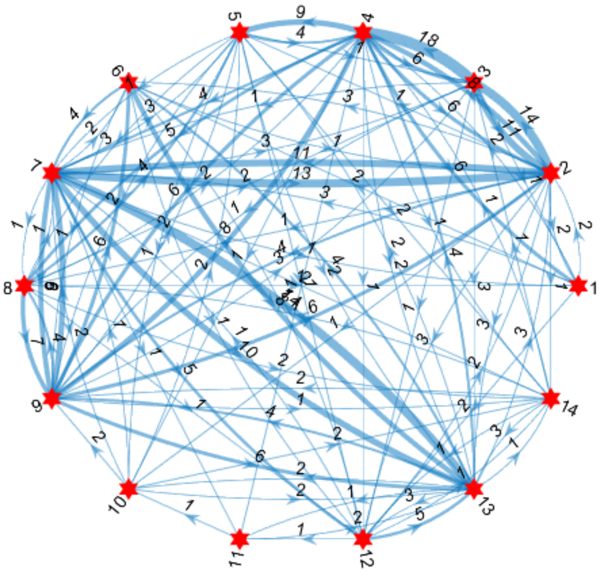
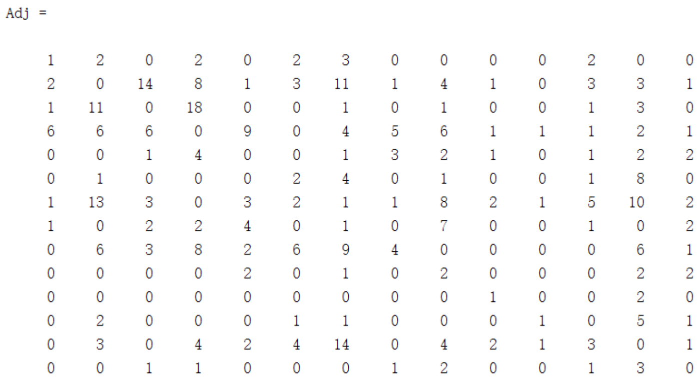


Figure 3: Huskies' weighted directed graph in match 1



```
Adj =

     1     2     0     2     0     2     3     0     0     0     0     2     0     0
     2     0    14     8     1     3    11     1     4     1     0     3     3     1
     1    11     0    18     0     0     1     0     1     0     0     1     3     0
     6     6     6     0     9     0     4     5     6     1     1     1     2     1
     0     0     1     4     0     0     1     3     2     1     0     1     2     2
     0     1     0     0     0     2     4     0     1     0     0     1     8     0
     1    13     3     0     3     2     1     1     8     2     1     5    10     2
     1     0     2     2     4     0     1     0     7     0     0     1     0     2
     0     6     3     8     2     6     9     4     0     0     0     0     6     1
     0     0     0     0     2     0     1     0     2     0     0     0     2     2
     0     0     0     0     0     0     0     0     0     1     0     0     2     0
     0     2     0     0     0     1     1     0     0     0     1     0     5     1
     0     3     0     4     2     4    14     0     4     2     1     3     0     1
     0     0     1     1     0     0     0     1     2     0     0     1     3     0
```

Figure 4: Huskies' weighted adjacency matrix in match 1

- Centroid coordinates and dispersion

A game consists of many passes, both offensive and defensive. In general, passes in the backfield are mainly used for defense; while passes in the frontfield are mainly on offense. Therefore, we separate the front court and back court passes and consider the central position of the passing network separately. Specifically, we only consider the coordinates of the serve.

$$\begin{cases} (x_{OFF}, y_{OFF}) = (\frac{\sum_i x_i}{N}, \frac{\sum_i y_i}{N}) \ , \quad x_i \geq 52.5 \\ (x_{DEF}, y_{DEF}) = (\frac{\sum_i x_i}{N}, \frac{\sum_i y_i}{N}) \ , \quad x_i < 52.5 \end{cases} \tag{1}$$

In addition, the center dispersion corresponds to the variance of the vector distance between the mean position of each pass and the center position of the network. We also consider the offense and defense separately.

$$C_{disp\{OFF,DEF\}} = \frac{\sum_i [(x_i - x_0)^2 + (y_i - y_0)^2]}{N}, \\ x_0 = x_{OFF}, x_{DEF} \;\; ; \;\; y_0 = y_{OFF}, y_{DEF} \tag{2}$$

- The maximum radius of the network

The maximum radius of the network is the farthest distance from the average passing point of the network players to the center of the passing network. The network maximum radius can describe the distribution range of the passing network.

$$R_{\max} = \max\{\sqrt{(x_i - x_o)^2 + (y_i - y_o)^2}\}, i = 1, 2, ..., N \tag{3}$$

- Network average clustering coefficient

In graph theory, the clustering coefficient is a measure of the degree to which vertices in a graph tend to cluster together. There is evidence that in most real-world networks, particularly social networks, vertices tend to create tightly connected groups characterized by a relatively high density of relationships, a likelihood that tends to be greater than the average probability of a randomly established tie between two nodes. There are two versions of this metric:global and local. The global version is designed to give the overall degree of clustering in the network, while the local version gives the degree of aggregation of individual vertices.

The average agglomeration coefficient of the network as a whole is the arithmetic average of the local agglomeration coefficients of all vertices. The average agglomeration coefficient measures the degree of agglomeration of a graph on the whole.

In general, the local clustering coefficient of a vertice i is obtained as the percentage of the vertices directly connected to it that, in turn, are connected between them. This measure can be averaged along the N vertices of the network to obtain the average clustering coefficient. However, when the network is weighted, we can not simply account for the number of vertices connected between them but, also, how the link weights are distributed. This is the case of passing networks, where the number of passes between pairs of players is not constant. In this way, we use the weighted clustering coefficient $C_w(i)$ to measure the likelihood that neighbours of a giver player i will also be connected between them.

$$C_w(i) = \frac{\sum_{j,k} w_{ij} w_{jk} w_{ik}}{\sum_{j,k} w_{ij} w_{ik}} \tag{4}$$

Note that, the weighted version of the clustering coefficient characterizes the tendency of the team to form balanced triangles between players and it is a measure of local robustness.

Finally, the clustering coefficient of the whole network is obtained by averaging $C_w(i)$ over all players, i.e.,

$$C = \frac{1}{n} \sum_{i=1}^{n} C_w(i) \tag{5}$$

(2) Local networks—identification of patterns

- Core-peripheral importance analysis

Eigenvector centrality is a way to measure the influence of vertices on the network. For vertices with the same number of connections, nodes with higher scores on adjacent vertices will have higher scores than nodes with lower scores on adjacent vertices, and all vertices are assigned corresponding scores based on this principle. A higher eigenvector centrality score means that the vertex is connected to many vertices with higher scores of their own. The following eigenvector equation is available for the adjacency matrix.

$$Mx = \lambda x \tag{6}$$

The eigenvector corresponding to the largest eigenvalue of the adjacency matrix is the eigenvector centrality of each vertex.The core players and edge players can be analyzed by ordering the eigenvector centrality.

Typically, a team has a number of core and peripheral players, with the peripheral players usually being relatively unimportant positions (e.g. goalkeeper) and substitutes. Core players are defined as those who have a higher number of passes or who are closely associated with players who have a higher number of passes. The average centrality can be derived from the eigenvector centrality, and those above the average can be considered as core players.

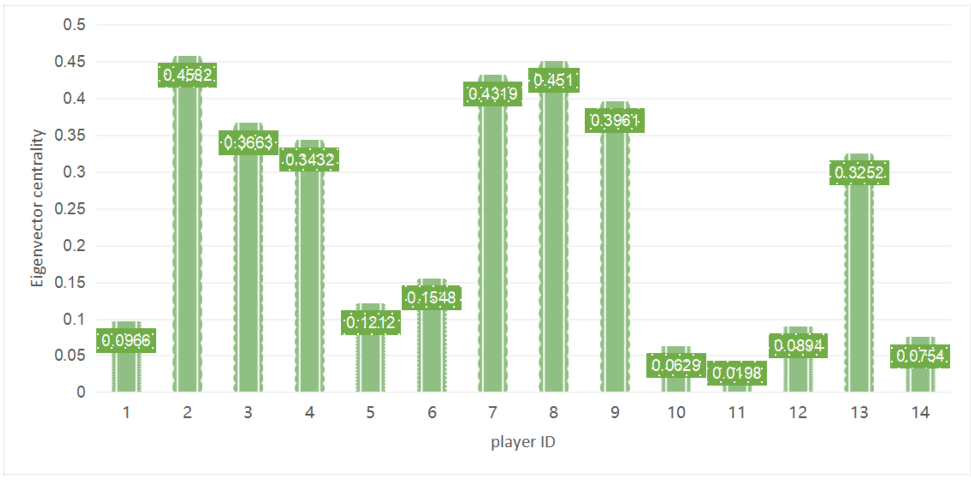


Figure 5: Huskies' the result of the core-periphery analysis in match 1

As shown in Figure 5, the core players of the Huskies in match 1 are numbered 2, 3, 4, 7, 8, 9, 13, corresponding to field players D1,D2,D3,M1,M2,M3,F2, while the peripheral players of the Huskies are numbered 1, 5, 6,10,11,12,14, corresponding to field players G1,D4,D5,M4,M5,F1,F3.

- The configurations of a team

Considering the difference in opponent and strategy, the ball passing network we build varies from match to match. Therefore, in order to identify the network pattern, we need to analyze a specific game in detail. In this chapter, we take the match 1 as an example.

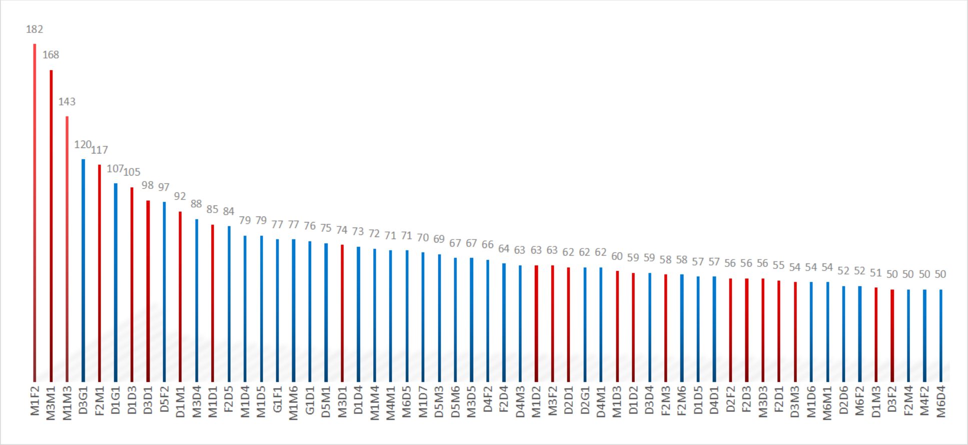


Figure 6: Routes with 50 or more interchangeable passes

In *Core-peripheral importance analysis*, we discussed the core player configuration for a game. Of the 10,454 passes the Huskies made throughout the season, passes related to the core players accounted for more than 66%, and typically the team's flexible inter-passes occurred between the core players. Figure 6 shows a histogram of routes with 50 or more interchangeable passes, with passes between core players in red.

In a partial passing network, dyadic configurations refer to passing interactions between two players and triadic configurations refer to passing interactions between three players. Obviously, the interaction of three players contains the passing of dyadic configurations. In order to analyze the specific passing situations between players, we compute all combinations of induced subgraphs between core players in the pass directed graph.

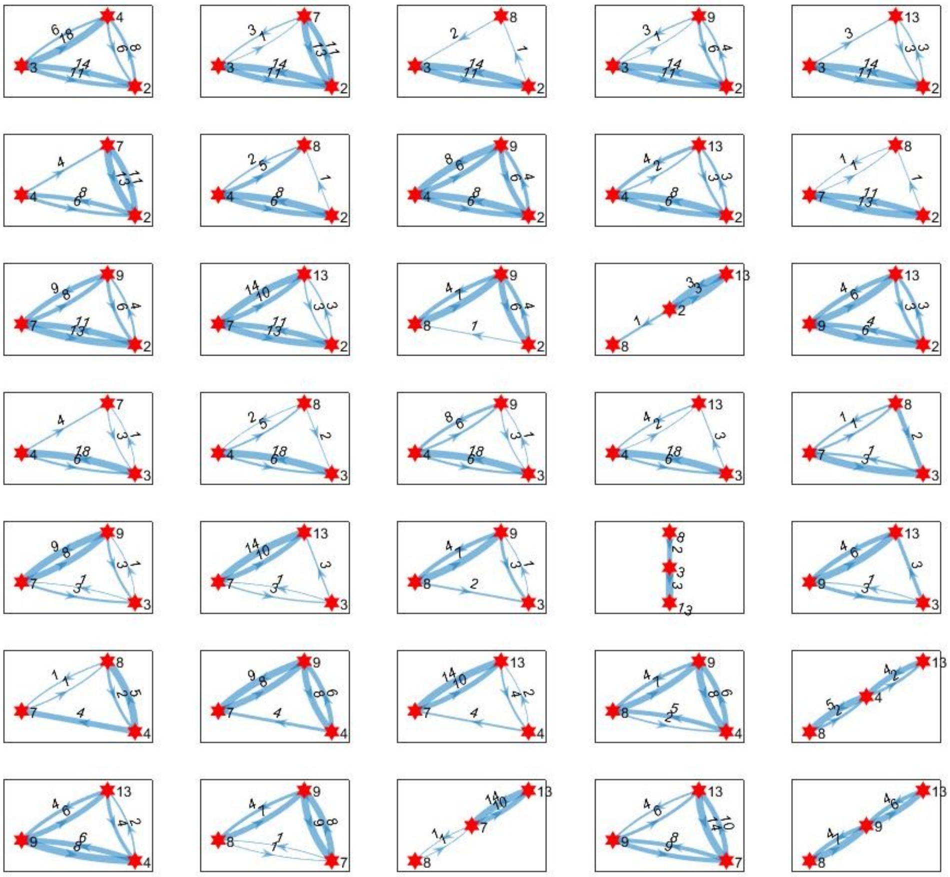


Figure 7: Induced subgraphs of the core player's passing routes in Match 1

Figure 7 shows that all induced subgraphs of the core player's passing routes in Match 1.

By counting all the matches, the most typical passing forms and the triads with the highest number of occurrences can be obtained. Figure 8 shows that the most typical pass forms are two-way passes between vertices 1,2 and between vertices 1,3, while one-way passes are between vertices 2,3. This form of passing is the most technical and tactical in style, and is also useful for goal scoring attacks. As shown in Figure 9, the most frequent triadic configurations are M1,M3 and F2, with a high number of 731 interactive passes between them.

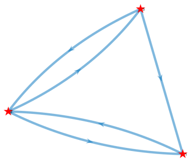


Figure 8: The most typical pass forms

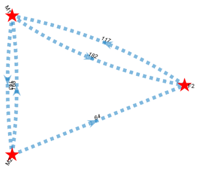


Figure 9: The most frequent triadic configurations

(3) Individual networks

- Degree centrality

In social network analysis, "centrality" is often used to determine the importance or influence of vertexs in a network. The most direct measure is degree centrality, and the greater the degree centrality of a vertex, the more important the vertex is. The degree centrality of a vertex can be defined as the following equation.

$$C_D\left(v_i\right)=d_i=\sum_j A_{ij} \tag{7}$$

When it is necessary to compare the importance of vertices in different networks, a normalization of the degree centrality values can be implemented. The normalized degree centrality value of a vertex of degree d_i is defined as:

$$C_D\left(v_i\right)=\frac{d_i}{n-1} \tag{8}$$

- Eigenvector centrality

In *Core-peripheral importance analysis*, we introduced eigenvector centrality. Eigenvector centrality is different from degree centrality in that a vertex with high degree centrality, i.e., having many connected vertices, does not necessarily have high eigenvector centrality, because all the connectors may have low eigenvector centrality. Similarly, high eigenvector centrality does not mean that it has high vertex degree centrality; it can have high eigenvector centrality even if it has few but important connectors. Thus, eigenvector centrality centrality can also evaluate the structural characteristics of single player networks from another perspective.

$$C_E{=}x_j=c\sum_{j=1}^{n}\omega_{ij}x_j \tag{9}$$

4.2.3 The timeline of passes in a match

In addition to the spatial dimension of the passing network, the analysis of the temporal dimension is equally important. For this purpose we analyzed the passing characteristics at the minute level, as well as the passing timeline from the whole game to the whole season.

A figure of the Huskies' total number of passes over time for the first three games is shown below. The overall trend in the number of passes is that there is a period of intense passing at the beginning of the first and second halves, and another round of intense passing at the quarter-half time. At the same time, we can find that the overall number of passes tends to decrease for the players towards the end of the first and second half, which may be caused by the decrease in physical strength.

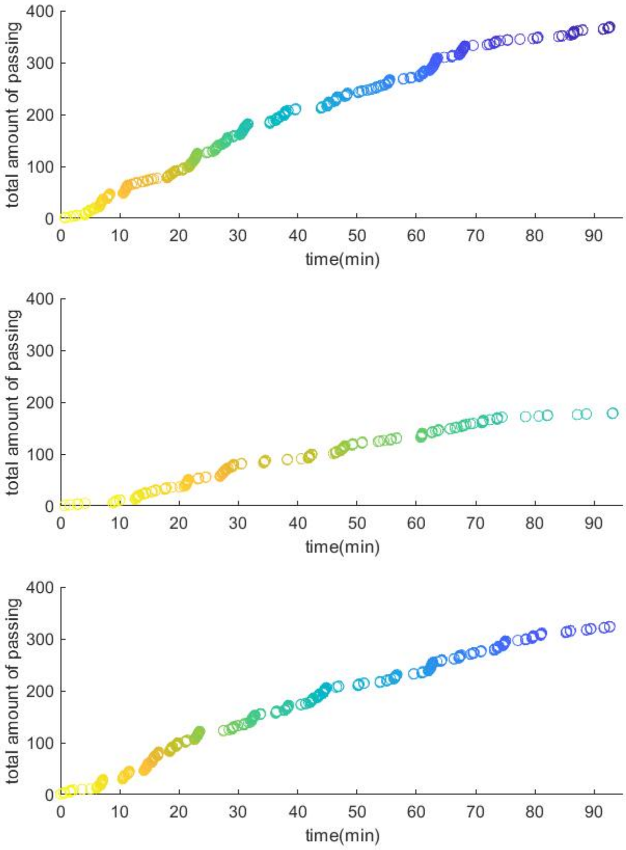


Figure 10: Total number of passes over time for the first three games

4.3 The Performance Indicators and Team Level Processes

4.3.1 Performance indicators

1. For a team, more passes means more interaction and running between players, which means better teamwork. At the same time, more total passes increase the team's chances of possession, thus ensuring the initiative of the attack and an increase in the number of successful shots on goal. Therefore, the total number of passes N in a single match is a direct reflection of the team's cooperation and is denoted as χ_1.

$$\chi_1 = N \tag{10}$$

2. Considering the group perspective, the analysis of the first problem yields a local aggregation coefficient and an average aggregation coefficient for each team We define the overall aggregation as the average aggregation coefficient of the team and use it as an indicator of the degree of player contribution. If a team implements a single core, then this metric will be large; conversely, if a team implements no core, then this metric will be small. This metric is denoted as χ_2.

$$\chi_2 = C = \frac{1}{N}\sum_{i=1}^{N} C_w(i) \tag{11}$$

3. From an individual perspective, we chose the CV (coefficient of variation) of the number of passes made by each member of a team as a measure of coordination among players. This metric captures whether a team's major passes are distributed among a few core players or among each player, while eliminating the effect of variation in total passes on the results. It also reflects the configuration of individual passing ability within a team and the structural characteristics of the team. The indicator is denoted as χ_3.

$$\chi_3 = c_v = \frac{\sigma}{\mu} \tag{12}$$

4.3.2 Other Team level processes

1. In the model of *4.2*, we introduce the definition of the central coordinates of the passing network and its dispersion, and we consider the offensive zone and the defensive zone separately. Obviously, the dispersion of the passing network can measure the formation of a team, and the larger the dispersion, the more open the team's formation is, which is conducive to long-distance passing. However, the greater the dispersion, the less cohesive the team is, which is not conducive to stealing and defensive counterattacks. Therefore, the greater the dispersion in offense and the smaller the dispersion in defense, the better the team's overall ball control ability. We denote this indicator as χ_4.

$$\chi_4 = \frac{C_{dispOFF}}{C_{dispDEF}} \tag{13}$$

2. The same maximum radius of the network presented in the first question can also be used to characterize the degree of dispersion of a team's formation. Team configurations are arranged with forwards and defenders standing to ensure vertical depth, when the radius of the network on the x-axis needs to be large enough. Some formations also require the use of the wings to attack, which requires a large enough y-axis radius to ensure cross-running from both wings.This indicator is denoted as χ_5.

$$\chi_5 = R_{\max} = \max\{\sqrt{(x_i - x_o)^2 + (y_i - y_o)^2}\} \tag{14}$$

3. In addition, to measure the team's initiative throughout the game, we introduced the classic metric of possession into the model. Possession is the percentage of time that a team controls the soccer ball during the course of a match. The literature shows that, in general, team performance is positively correlated with a team's ability to control the ball, and the higher the possession rate, the better the chances of winning the game.This indicator is denoted as χ_6.

$$\chi_6 = \frac{T_{team}}{T_{team} + T_{opp}} \tag{15}$$

4. Finally, we analyze the rhythmical nature of the team from the point of view of time. We chose to characterize the time required for the first 50 passes of each ballgame. The shorter the time required for the first 50 passes, the faster the team is consistently passing the ball quickly at the beginning of the game. A team's fast start facilitates mastering the rhythm of the game, and the rhythm will be effectively controlled afterwards. Controlling the rhythm of the passing game can disrupt the

opponent's tactical rhythm and thus attack effectively. This indicator can be expressed as χ_7.

$$\chi_7 = T_{50} \tag{16}$$

4.4 Screening and Testing Indicators

4.4.1 Descriptive analysis of indicators

To clarify whether the strategy is generally effective, the team's goal difference can be used as the dependent variable to measure the team's goal scoring success and thus confirm whether our choice of performance indicators and chosen team-level processes can improve success rates.

$$\psi = Score_{team} - Score_{opp} \tag{17}$$

In addition, to eliminate the influence of the scale range of each independent variable on further data processing, all variables were normalized to have a mean of 0 and a standard deviation of 1.

The following figure shows the scatter plot of the 7 independent variables χ_1 -χ_7 and the dependent variable ψ after normalization. The scatter points are the adjusted data. The solid line is the fitted regression line, and the larger the slope indicates the larger the correlation coefficient and the more correlated the relationship between the independent variable and the dependent variable. The dashed line is the 95% confidence boundary, the more points within the boundary, the more the data meet the 95% confidence interval, the more reliable the data are. The p-value at the top of the graph is the significance, the smaller the p-value, the more significant the result.

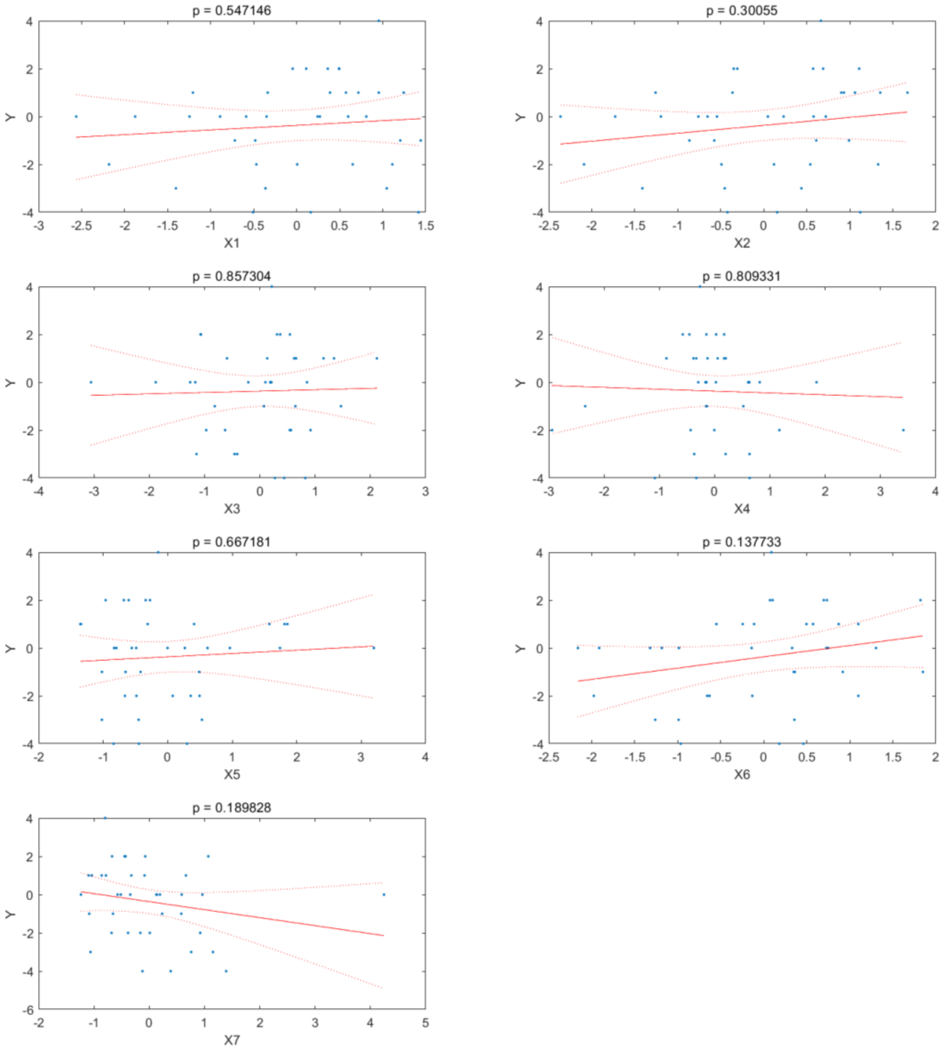
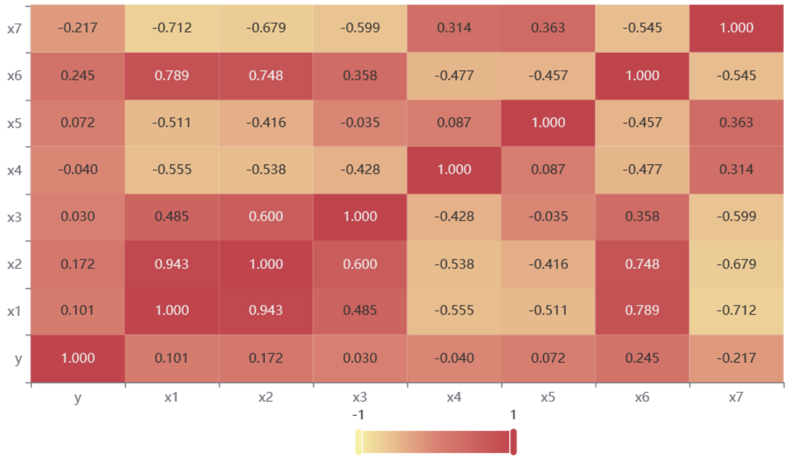


Figure 11: Scatter diagram of independent and dependent variables



	y	x1	x2	x3	x4	x5	x6	x7
x7	-0.217	-0.712	-0.679	-0.599	0.314	0.363	-0.545	1.000
x6	0.245	0.789	0.748	0.358	-0.477	-0.457	1.000	-0.545
x5	0.072	-0.511	-0.416	-0.035	0.087	1.000	-0.457	0.363
x4	-0.040	-0.555	-0.538	-0.428	1.000	0.087	-0.477	0.314
x3	0.030	0.485	0.600	1.000	-0.428	-0.035	0.358	-0.599
x2	0.172	0.943	1.000	0.600	-0.538	-0.416	0.748	-0.679
x1	0.101	1.000	0.943	0.485	-0.555	-0.511	0.789	-0.712
y	1.000	0.101	0.172	0.030	-0.040	0.072	0.245	-0.217

Figure 12: Correlation coefficient of independent and dependent variables

Figure 12 shows the correlation coefficients between the independent and dependent variables. It can be seen that the correlation coefficients between χ_3 and χ_4 and the dependent variable are very low, so we chose to exclude them from the regression analysis afterwards. It can also be seen that χ_1, χ_2 and χ_5,χ_6 are positively correlated with the dependent variable, where χ_5 is weakly correlated, and the correlation between χ_1 and χ_2 is large and there may be multicollinearity. And χ_7 shows negative correlation with the dependent variable.

4.4.2 Regression and analysis

Assuming first that there is no non-linear relationship between several independent and dependent variables, we can perform a multiple linear regression analysis on the data by treating all the data factors of a game as linear relationships. The regression equation is shown below.

$$\psi = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \beta_5\chi_5 + \beta_6\chi_6 + \beta_7\chi_7 + \varepsilon \tag{18}$$

Due to the large fluctuation of data points, we choose the robust regression method to perform robust regression to analyze the multiple linear regression model. This method can avoid the effect of heteroskedasticity caused by cross-sectional data, and can eliminate the effect of outliers.

Table 2: Linear regression results

variables	β	se	t	p
χ_0	-0.4166	0.3098	-1.3448	0.1881
χ_1	2.0480	1.1233	-1.8231	0.0776
χ_2	1.2430	0.9706	1.2806	0.2095
χ_5	0.2725	0.3774	0.7219	0.4756
χ_6	0.9644	0.5150	1.8727	0.0703
χ_7	-0.5963	0.4480	-1.3308	0.1927

The results of the multiple linear regression analysis show that the p-value of the t-test for the independent variable χ_5 is large and cannot pass the t-test, so the independent variable χ_5 is discarded.

Table 3: Further Linear regression results

variables	β	se	t	p
χ_0	-0.4134	0.3073	-1.3451	0.1878
χ_1	2.3440	1.0565	-2.2187	0.0335
χ_2	1.4322	0.9367	1.5291	0.1358
χ_6	0.9208	0.5078	1.8132	0.0789
χ_7	-0.5921	0.4444	-1.3322	0.1919

After removing one independent variable and re-running the multiple regression analysis, it can be seen that χ_1 meets a p-value less than 0.05 and the significance level is statistically significant. While the other independent variables are greater than 0.05, the p-values are not too large and the p-values are still acceptable for such a complex model as soccer, so I chose to keep these variables.

It is worth mentioning that the p-value of the χ_1 indicator is highly significant, but the regression coefficient is negative, which may be due to the effect of cointegration of χ_2. This means that the number of passes and the number of net goals do not have the most direct effect, but rather to reduce the number of passes with the same team clustering coefficient. This means that the number of passes is quantity and the agglomeration coefficient is quality, and quantity affects the number of goals by affecting quality and thus goals.

The above regression analysis shows that the regression model we developed works well overall and has a small root mean square error of 1.8946.In summary, it shows that the model has a significant linear effect on the team's goal difference prediction. Among the four indicators finally selected, χ_6 has the largest correlation coefficient with the dependent variable, indicating that the possession rate helps a lot in the team's goal scoring role. The next most important is χ_7. The correlation coefficient of χ_7 is also large and the regression coefficient is negative. This indicates that they are negatively correlated, i.e. the smaller the time spent on the first 50 passes, the more effective the team starts and scores goals.

Therefore, further analysis of the meaning of each index shows that the number of passes represents the team interaction and the agglomeration coefficient represents the core team system, which have a significant joint effect. Possession rate represents team possession strength with a significant positive effect, and the first 50 pass times represent the team's opening rhythm with a significant negative effect. Overall these indicators can summarize the court strategy, so there is no need to consider the negative impact of the opponent's strategy on our team. The reason for losing the game can be explained by the insufficient amount of certain indicators.

4.5 Performance Model of Passing Network and Description of Team Cooperation

After successfully building a suitable multiple linear regression model, we can use this performance model to describe the strengths and weaknesses of the team's passing network. Using game data from all teams throughout the season, we can perform a side-by-side comparison of the Huskies' teamwork levels.

4.5.1 Structure

The metric related to team structure is χ_1, the number of passes. The Huskies' average number of passes per game is 549, ranking 12 out of 20 teams. This indicates that the overall level of players of the Huskies is weak compared to the level of players of other teams in the league, so the amount of passes cannot compete with them. It can also be seen that the Huskies' teamwork level is average and they are not able to make better passes to improve the team's goal scoring strength.

4.5.2 Configuration

Team configuration can be interpreted as the team's choice of tactical system and fit. At the team tactical level, the ternary configuration is the most frequently occurring tactical system. A ternary configuration implies a high level of interaction between any two and three vertices and a high level of connectivity throughout the team. The Huskies are at the bottom of this average ratio of 0.3658, ranking 15th out of 20 teams. It shows that the team is not enough in the number of three-player team configurations and the passing consistency needs to be strengthened, including the passing accuracy and offensive ability.

Specifically in terms of the quality of team configuration, we can assess the direct team configuration difference between the Huskies and other teams by the agglomeration factor. The Huskies' agglomeration coefficient for all games averaged 3.95, ranking eighth among all teams. This indicates that the Huskies' play system revolves around a few key core players, and everyone's passing is focused on the core players.

4.5.3 Dynamic

Dynamics refers to the time elicited between periods of each game and between multiple games to observe the team's performance. At the micro level, the team averaged 50 passes in 803.87 seconds after the start of the game, which ranked seventh out of 20 teams. On the macro level, the team averaged 287.24 passes per game over 38 games, which ranked ninth and was mediocre overall.

As for possession the Huskies are averaging less than 45%, well below their season average. Not reaching half of the possession rate indicates that the Huskies are weak at stealing and controlling the ball and cannot hold the ball against for long periods of time. This is partly due to the team's fast-paced start, and partly due to the team's low number of passes that were snapped and lost.

4.6 Strategies

4.6.1 Cooperation training of three-member groups

From the previous analysis, we can see that team configuration helps a lot with the number of goals scored, and among the team configurations, dyadic and triadic configurations are the most common. At the micro level, the tactics and skills among players can be differently boosted by different configurations, and excellent team configurations can bring out the maximum technical ability. Obviously, among all the top link combinations, more sides represent more frequent cooperation between players.

The passing configuration that occurs most frequently on the Huskies is not optimal, thus resulting in a lower number of passes and a lower set coefficient for the team. Compared to the best teams, the Huskies' core players were similar in their coordination with each other, but there was a big gap in their coordination with non-core players, and conceded goals often occurred in the process.

Therefore, we suggest that coaches should pay attention to the tacit understanding of group cooperation, simulate the environment and interference on the real court through tactical cooperation in daily training, and maximize the cooperation between core players and non-core players, so that a more complete and flexible cooperation can be realized in actual combat, better organize the synchronization rate of the whole team and make the team become a whole.

4.6.2 The importance of physical training

Possession of the ball is a very critical winning factor, and if you can't control the ball efficiently, you will be in a passive situation. Our performance model shows that a team's performance relies heavily on the team's possession rate, which, from a macro perspective of length of play, relies heavily on passing between players. This is because players have limited physical strength, and in order to control the ball effectively, they need excellent passing to consume the opponent's physical strength, and to do so by trading space for time.

Therefore, the daily training communication needs to pay attention to the team's basic ability development, and improve the players' ball control skills and anti-stealing ability in the training.

4.6.3 Control of opening attack routine and training of personal skills

The effect of the independent variable χ_7 makes it clear that the team's opening effect is crucial. A good start is half the battle. The average time spent by the team on the 50 passes at the beginning of the game is 13.4 minutes, which is a relatively short duration. In the model, this indicator has a negative effect on the final goal difference, which just means that the shorter the time spent on the first 50 passes, the better the final result.

Therefore, our strategic suggestion to the coach is to master the opening offensive routine by using fast attacking tactics, and while destroying the opponent's rhythm, seize the favorable moment to quickly break through the defense into the opponent's penalty area. At the same time, if the opponent catches the moment, also through the excellent passing set to reverse according to the reality, turning passivity into initiative and effectively apply this tactical strategy.

4.7 Design of a Generalized Model

According to the soccer model of team performance obtained from the study, there are four main aspects to reflect the effectiveness of teamwork, namely χ_1 , χ_2 , χ_6 and χ_7 .

- Looking at χ_1, the higher the pass total tends to be the higher the number of goals scored. In a broader sense the number of passes corresponds to the amount of team interaction, so it is important to have frequent interaction in the team.
- Looking at χ_2, the team's agglomeration factor can help a lot with the team's formation structure. Therefore, it can be noted that the team needs a strong core and strong cohesion in order to maximize the team effect.
- Looking at χ_6, team possession is critical, and in the broad team concept, team execution can be viewed as the equivalent concept. The higher the team's execution, the better the team's control over its work and the better it can solve complex problems.
- Looking at χ_7, the team's opening tempo helped a lot in scoring goals, and the fast pace helped the team gain the opening initiative. Therefore the team must have its own rhythm if it wants to control the tempo without being disturbed by the outside world. It is only in the rhythm that you are familiar with that you can eliminate all difficulties.

5 Evaluation and Promotion of Model

5.1 Strength and Weakness

5.1.1 Strength

- For the relationship between soccer networks and graph theory, variable extraction and regression analysis of team passing networks were conducted based on social network analysis theory.
- After testing, the selected evaluation metrics were able to describe well the attributes of the passing network and the cooperative performance of the team.
- The established regression model was subjected to correlation analysis and t-test. A sensitivity analysis was completed with high reliability.

5.1.2 Weakness

- In the model assumption phase, changes in time series, such as changes in tactical fits over time intervals and differences in the top and bottom halves, are ignored.
- For the given data, no deep analysis was performed using the difference in the types of passes, such as some special forms of passes.
- There is no consideration of home and away factors and coaching.

5.2 Promotion

The weighted directed graph can be transformed into an undirected graph for some analyses. Each player can be targeted and differentiated for independent analysis.

6 Conclusions

By modeling the passing between players as a network, the metrics related to player passing were meticulously analyzed and used to create an assessment model that can evaluate team performance. The factors influencing the team's goal difference in terms of the structure, configuration and dynamics of teamwork were analyzed, including the number of passes, team cohesion, possession rate and opening rhythm. These lessons can also be generalized to a broader social group.

References

[1] Feasibility Analysis of Passing Performance in Football Match by Social Network Analysis, Li Bo, Wang Lei, Journal Of Beijing Sport University.

[2] 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.

[3] GÜRSAKAL, N., YILMAZ, F., ÇOBANOĞLU, H., ÇAĞLIYOR, S. (2018). Network Motifs in Football. Turkish Journal of Sport and Exercise.

[4] Visualization Analysis on Passing Technique of Spanish Tiki-Taka Tactics Based on Social Network Analysis Method,Cao Weihua,Journal of Chengdu Sport University.

[5] D. J. Watts , Steven Strogatz (1998). "Collective dynamics of 'small-world' networks". Nature.

[6] P. W. Holland , S. Leinhardt (1971). "Transitivity in structural models of small groups". Comparative Group Studies.

```
clear;
data=xlsread('passnetwork_match␣(38).xlsx','A2:C112');
player=data(:,1:2);
N=length(unique(player));

%% Calculate the number of passing balls between each player
Adj=zeros(N,N);
len=length(data);
A=zeros(30,30);
for i=1:len
    w=data(i,1);
    v=data(i,2);
    weight(i)=data(i,3);
    A(w,v)=data(i,3);
end
j=0;
row=all(A==0,2);
column=all(A==0,1);
for i=1:30
    if row(i)==1 && column(i)==1
        A(i-j,:)=[];
        A(:,i-j)=[];
        j=j+1;
    end
end
%  A(all(A==0,2),:)=[];
%  A(:,all(A==0,1))=[];
% Row=zeros(30);
% Column=zeros(30);
% i=0;
% while 1
%     i=i+1;
%     if i > length(A)
%         break
%     end
%     if find(A(i,:)==0)
%         Row(i)=1;
%     end
%     if find(A(:,i)==0)
%         Column(i)=1;
%     end
%     if Row(i)==1 && Column(i)==1
%       A(i,:)=[];
%       A(:,i)=[];
%     end
%
% end
Adj=A;
```

```
 %% plot directed graph
  G=digraph(Adj,'omitselfloops');
   LWidths = 5*G.Edges.Weight/max(G.Edges.Weight);
  plot(G,'Layout','circle','EdgeLabel'
,G.Edges.Weight,'LineWidth',LWidths,'Marker',
'h','MarkerSize',8,'NodeColor','r');

 %% Calculate eigenvector centrality

 [V,D]=eig(Adj);

 % Calculate the cluster coeffient of each player
 Cw=zeros(N,1);
 for i=1:1:N
     s=0;
     for j=1:1:N
         for k=1:1:N
             if i~=j && j~=k && i~=k
                 Cw(i)=Cw(i)+(Adj(j,i)+Adj(i,j)
 * (Adj(j,k)+Adj(k,j)) * (Adj(k,i)+Adj(i,k));
             end
             if i~=j && i~=k
                 s=s+(Adj(j,i)+Adj(i,j)) *(Adj(k,i)+Adj(i,k));
             end
         end
     end
     Cw(i)=Cw(i)/s;
 end
 Cw;

 %% Calculate the cluster coeffient
 C=0;
 for i=1:1:N
     C=C+Cw(i);
 end
 C=C/N

 % X2=zscore(Adj);
 % d=zeros(14,14)+100;
 % for i=1:1:14
 %     for j=1:1:14
 %         if X2(i,j)~=0
 %             d(i,j)=1./X2(i,j);
 %         end
 %     end
 % end
 % for i=1:1:14
 %     d(i,i)=100;
 % end
 %
```

```
% Z2=linkage(d);
%
% C2=cophenet(Z2,d);          %//0.94698
%
% T=cluster(Z2,6);
% H=dendrogram(Z2);

%% Calculate the induced subgraph
cp=[2,3,4,7,8,9,13];
nch = nchoosek(cp,3);
insub=zeros(3,3);
nlen=length(nch);
for i=1:nlen
    insub=zeros(3,3);
    for j=1:3
        for k=1:3
            insub(j,k)=Adj(nch(i,j),nch(i,k));
        end
    end
    nodename={ num2str(nch(i,1)), num2str(nch(i,2)), num2str(nch(i,3))}
    if rank(insub)~= 0
        Gsub=digraph(insub,nodename,'omitselfloops');
        subplot(7,5,i);
        LWidths = 4*Gsub.Edges.Weight/max(Gsub.Edges.Weight);
        plot(Gsub,'EdgeLabel',Gsub.Edges.Weight,
'LineWidth',LWidths,'Marker',
'h','MarkerSize',8,'NodeColor','r');
    end
end

%% plot the most frequent triadic configurations
asub=[0,117,0
    182,0,143
    64,168,0];
nodename1={'F2','M1','M2'};
agsub=digraph(asub,nodename1);
plot(agsub,'Layout','circle','LineStyle',':',
'EdgeLabel',agsub.Edges.Weight,'LineWidth',5,'Marker','p',
'MarkerSize',18,'ArrowSize',15,'NodeColor','r');

%% plot time line
datime=xlsread('passingevents2.xlsx','Sheet1','A2:J370');
y1=datime(:,1);
x1=datime(:,8);
x2=datime(:,9);
x3=datime(:,10);
color = linspace(10,1,length(x1));
subplot(3,1,1);
scatter(x1,y1,[],color);
axis([0,95,0,400]);
```

```
xlabel('time(min)');
ylabel('total␣amount␣of␣passing');
subplot(3,1,2);
scatter(x2,y1,[],color);
axis([0,95,0,400]);
xlabel('time(min)');
ylabel('total␣amount␣of␣passing');
subplot(3,1,3);
scatter(x3,y1,[],color);
axis([0,95,0,400]);
xlabel('time(min)');
ylabel('total␣amount␣of␣passing');

%% Calculate the x3
for i=1:N
    Rsum(i)=sum(Adj(i,:));
end
deta=std(Rsum)/mean(Rsum)

% Calculate the x4

datacoord=xlsread('Centroid␣coordinates.xlsx','A2:C10436');
% xsumoff=zeros(38,1);
% ysumoff=zeros(38,1);
% xsumdef=zeros(38,1);
% ysumdef=zeros(38,1);
% coordnumoff=zeros(38,1);
% coordnumdef=zeros(38,1);
% for i=1:10435
%     for j=1:38
%         if datacoord(i,1)==j
%
%             if datacoord(i,2)>=52.5
%                 coordnumoff(j)=coordnumoff(j)+1;
%                 xsumoff(j)=xsumoff(j)+datacoord(i,2);
%                 ysumoff(j)=ysumoff(j)+datacoord(i,3);
%             else
%                 coordnumdef(j)=coordnumdef(j)+1;
%                 xsumdef(j)=xsumdef(j)+datacoord(i,2);
%                 ysumdef(j)=ysumdef(j)+datacoord(i,3);
%             end
%
%         end
%     end
% end



for j=1:38
    k=1;
```

```
        for i=1:10435
            if datacoord(i,1)==j
                matchcoord(j,k,:)=[datacoord(i,2),datacoord(i,3)];
                k=k+1;
            end
        end
    end

    for i=1:38
        k=1;
        m=1;
        for j=1:length(matchcoord)
            if matchcoord(i,j,1)>=52.5
                matchoff(i,k,1)=matchcoord(i,j,1);
                matchoff(i,k,2)=matchcoord(i,j,2);
                k=k+1;
            else
                matchdef(i,m,1)=matchcoord(i,j,1);
                matchdef(i,m,2)=matchcoord(i,j,2);
                m=m+1;
            end
        end
    end

    for i=1:38
        meanoffx(i)=mean(matchoff(i,:,1));
        meanoffy(i)=mean(matchoff(i,:,2));
        meandefx(i)=mean(matchdef(i,:,1));
        meandefy(i)=mean(matchdef(i,:,2));
    end
    Coffsum=zeros(38,1);
    Cdefsum=zeros(38,1);
    for i=1:38
        for k=1:length(matchoff)
            roff(i,k)=(matchoff(i,k,1)-meanoffx(i))^2
+(matchoff(i,k,2)-meanoffy(i))^2;
            Coffsum(i)=Coffsum(i)+roff(i,k);
        end
        Coff(i)=Coffsum(i)*(1/length(matchoff));
        for k=1:length(matchdef)
            rdef(i,k)=(matchdef(i,k,1)-meandefx(i))^2
+(matchdef(i,k,2)-meandefy(i))^2;
            Cdefsum(i)=Cdefsum(i)+rdef(i,k);
        end
        Cdef(i)=Cdefsum(i)*(1/length(matchdef));
        x4(i)=Coff(i)/Cdef(i);
    end

    %% Calculate the x5
    for i=1:38
```

```
      x5(i)=max(max(roff(i)),max(rdef(i)));
  end

  % Calculate the x6
  datatime=xlsread('time.xlsx','A2:C23430');

  for j=1:38
      k=1;
      for i=1:23429
          if datatime(i,1)==j
              matchtime(j,k,:)=[datatime(i,2),datatime(i,3)];
              k=k+1;
          end
      end
  end
  Ti=zeros(38,2);
  for i=1:38
      ball=matchtime(i,1,1);
      for j=2:length(matchtime)
          if matchtime(i,j,1)==ball &&ball>0
              if matchtime(i,j,2)-matchtime(i,j-1,2)>0
                  Ti(i,ball)=Ti(i,ball)
+matchtime(i,j,2)-matchtime(i,j-1,2);
              end
          else
              ball=matchtime(i,j,1);
          end
      end
  end
  for i=1:38
      x7(i)=Ti(i,1)/(Ti(i,2)+Ti(i,1));
  end


  dataind=xlsread('indicators.xlsx','A2:H39');
  Y=dataind(:,1);
  X1=zscore(dataind(:,2));
  X2=zscore(dataind(:,3));
  X3=zscore(dataind(:,4));
  X4=zscore(dataind(:,5));
  X5=zscore(dataind(:,6));
  X6=zscore(dataind(:,7));
  X7=zscore(dataind(:,8));
  [coef1, pval1] = corr(X1, Y);
  [coef2, pval2] = corr(X2, Y);
  [coef3, pval3] = corr(X3, Y);
  [coef4, pval4] = corr(X4, Y);
  [coef5, pval5] = corr(X5, Y);
  [coef6, pval6] = corr(X6, Y);
  [coef7, pval7] = corr(X7, Y);
```

```
subplot(4,2,1);
set(gcf,'InvertHardCopy','off','color','white');
mdl = fitlm(X1, Y);
h = plotAdded(mdl);
set(h(1),'Marker','.');
xlabel('X1');
ylabel('Y');
title(sprintf('p␣=␣%g', pval1));
subplot(4,2,2);
set(gcf,'InvertHardCopy','off','color','white');
mdl = fitlm(X2, Y);
h = plotAdded(mdl);
set(h(1),'Marker','.');
xlabel('X2');
ylabel('Y');
title(sprintf('p␣=␣%g', pval2));
subplot(4,2,3);
set(gcf,'InvertHardCopy','off','color','white');
mdl = fitlm(X3, Y);
h = plotAdded(mdl);
set(h(1),'Marker','.');
xlabel('X3');
ylabel('Y');
title(sprintf('p␣=␣%g', pval3));
subplot(4,2,4);
set(gcf,'InvertHardCopy','off','color','white');
mdl = fitlm(X4, Y);
h = plotAdded(mdl);
set(h(1),'Marker','.');
xlabel('X4');
ylabel('Y');
title(sprintf('p␣=␣%g', pval4));
subplot(4,2,5);
set(gcf,'InvertHardCopy','off','color','white');
mdl = fitlm(X5, Y);
h = plotAdded(mdl);
set(h(1),'Marker','.');
xlabel('X5');
ylabel('Y');
title(sprintf('p␣=␣%g', pval5));
subplot(4,2,6);
set(gcf,'InvertHardCopy','off','color','white');
mdl = fitlm(X6, Y);
h = plotAdded(mdl);
set(h(1),'Marker','.');
xlabel('X6');
ylabel('Y');
title(sprintf('p␣=␣%g', pval6));
subplot(4,2,7);
set(gcf,'InvertHardCopy','off','color','white');
```

```
mdl = fitlm(X7, Y);
h = plotAdded(mdl);
set(h(1),'Marker','.');
xlabel('X7');
ylabel('Y');
title(sprintf('p = %g', pval7));
X=[X1,X2,X6,X7];
[beta,stats]=robustfit(X,Y)
```