Problem Chosen	2022 MCM/ICM Summary Sheet	Team Control Number
D		**123456**

Teamwork Evaluation System Based on Network and PCA-CART Methods

Summary

As football a very popular sport across the world, A coach of a famous team tends to use objective analysis of the football team's performance as the main way to set training targets, and improve the team's level. This paper adopts the perspective of players' passing network, constructs the analysis framework of "identifying cooperation network–expanding cooperation model-exploring influence effect–putting forward suggestions", quantitatively analyzes the cooperation relationship of Huskies football team and its influence law on competition performance. We have mainly solved the following four problems:

Firstly, we should build a paasing network and recognize and inspect the patterns and properties in the network, which may be useful in solving next problem. We preprocess the large dataset and give a general description of it. We exam the spatial distribution of all kinds of events (passing events especially) using KDE techniques. We consider dyadic and triadic configurations and other 4 kinds of indexes (MaxEigenvalue, Connectivity, Clustering Coefficient and Shortest-path Length), and show the variation of them during a match or during a season.

Secondly, we are supposed to identify performance indicators that reflect successful teamwork and consider other team level processes to create a model that captures some aspects of teamwork. We define a index (GD) to represent the outcome of a match, and construct the comprehensive evaluation system from two dimensions (Individual and Team), with 6 different indexes (Flow Centrality, Team PlayeRank, MaxEigenvalue, Connectivity, Clustering Coefficient, Shortest-path Length). Then, by using CART combining PCA, we successfully build a model to forecast the outcome of a match according to the given 6 indexes. Furthermore, we exam the correlation coefficients between GD and other 6 indexes to find the most important index, and generate strategies according to it.

Thirdly, we should use the insights gained from our model to inform the coach of effective structural strategies. By analysizing the outcome of our model, we can find the most important factors that lead to a team's success, and consider how to increse these factors. By further demonstrating, we concluded three important suggestions for structural strategies.

Fouthly, providing our existing knowledge, we are supposed to shed some lights on how to design more effective teams in other fields. We have extended the model established in previous parts. We get the following conclusion: in order to improve team effectiveness, we should not only focus on personal improvement, but also focus on team building. No matter what index is chosen, it is impossible to correctly reflect the effectiveness of the team a hundred per cent. We need to think about it dynamically and make it more practical when it comes to evaluating a team and everyone in that.

To sum up, we build a passing network based on network science. We extract 6 important indexes, and build a forecasting model based on PCA and CART.

Keywords: Passing Network; PCA; CART; Spearman Correlation Coefficient

Contents

1 Introduction

The application of Network Science to social systems has introduced new methodologies to analyze classical problems such as the emergence of epidemics, the arousal of cooperation between individuals or the propagation of information along social networks. More recently, the organization of football teams and their performance have been unveiled using metrics coming from Network Science, where a team is considered as a complex network whose nodes (i.e., players) interact with the aim of overcoming the opponent network. In this paper, based on complex network methods, we're supposed to give a solution to the Huskie coach's requests for teamwork evaluation, which would like to be heuristic for identifying specific strategies that can improve teamwork and instructive for training in the future.

In this section, we'll provide a brief overview of the problems to be solved, basic methods we'd like to use in our solutions, the nomenclature and assumptions for the paper.

1.1 Problem Restatement

The coach of the Huskies Football Team wants to analyse the potential strategies and teamwork pattern with mathematical and complex network methods, in order to help understand the team's dynamics and improve their performance in the future. The Huskies have provided data detailing information from last season, including all 38 games they played against their 19 opponents. Overall, the data covers 23,429 passes between 366 players (30 Huskies players, and 336 players from opposing teams), and 59,271 game events.

We are required to explore how the complex interactions among the players on the field impacts their success, including the interactions that lead directly to a score, team dynamics throughout the game or over the entire season and other specific strategies that can improve teamwork next season. In particular, Quantification and formalization of the structural and dynamical features that have been successful (and unsuccessful) for the team are needed.

To respond to the Huskie coach's requests, solutions for the following problems are required,

- ***PRO1:*** Create a network for the ball passing and identify network patterns, such as dyadic and triadic configurations and team formations.
- ***PRO2:*** Identify performance indicators that reflect successful teamwork and create a model to capture structural, configurational, and dynamical aspects of teamwork.
- ***PRO3:*** Give suggestions on what changes the network analysis indicates based on PRO1 and PRO2 to help the team improve.
- ***PRO4:*** Extent the model and Generalize the findings to discuss how to design more effective teams and What other aspects of teamwork would need to be captured to develop generalized models of team performance?

1.2 Problem Analysis

The core of the solution lies in establishing a model which is convenient to explore its internal mechanism and mode with complex network tools and using the model to verify and improve the traditional team evaluation indicators, at the same time exploring new valuable indicators,such as team level processes, to present reasonable suggestions.

For PRO1, considering the value and complexity of temporal information, we can use temporal-value networks and mean-value networks to model the performance of a single game and a whole season, respectively, in order to simplify the model while preserving as much data integrity as possible. We can use the color partition algorithm to mine the potential patterns of core position, core player, core duo and trio, etc.

For PRO2, firstly, we can identify several characteristics to characterize team performance and team level processes and demonstrate their relevance to match outcomes. Based on SVM, we can establish the relationship model between characteristic indicators and competition results with historical data and the accuracy of the model can be proved by cross validation. In addition, all matches are classified based on the opponent's different competition strategies, and the difference of results can be tested in different categories to verify whether the strategy is universal.

For PRO3, analysis will be carried out from two aspects: team network patterns and teamwork characteristics. Based on the conclusions in PRO1 and PRO2, we can understand what kind of performance or modes owns more probability to win, and Therefore, reasonable suggestions on strategy can be prevented.

For PRO4, we can generalize the characteristics of the model and, in the meantime, improve the evaluation scheme of the model according to requirements. The goal is to obtain a more universal model, to be applied to more common scenarios and provide guidance in how to design more effective teams.

We will show the detailed solutions and results for the 4 problems in Section 2 to 5.

1.3 Nomenclature

We'd like to begin by defining a list of the nomenclature used in the paper as Table 1.

Table 1: Nomenclature

Abbreviation	Description
$V^(t)$	Time-varying description of the nodes in network.
$E^(t)$	Time-varying description of the edges in network.
N	Numbers of nodes in the passing network.
$G^(t)$	Time-varying network of passing.
S_ijk	Quantity of the pass among groups of players i, j, k.
λ_1	Largest eigenvalue of adjacency matrix for passing nework.
$\widetilde{\lambda_2}$	Algebraic connectivity for passing network.
$C_w^{(t)}$	Clustering coefficient for player i.
C	Overall clustering coefficient for passing network.
d	Shortest-path length for passing network.
$r(u, m)$	PlayeRank for player u in match m.
$Gini_{split}(T, i)$	GINI information gain for attribut i in dataset T.
a_*	Attribute which minimizes gini after partitioning.

1.4 Assumptions and Justifications

To simplify the problem and make itconvenient for us to simulate real-life conditions, we make the following basic assumptions, each of which is properly justified.

- **All data used is real and reliable.** This is very important for the validity and accuracy of our model. The data we used is provided by the MCM/ICM, and we can presume that it is real and reliable.
- **Using (geometric) mean of all players in some index to represent the team's index is valid.** The Flow Centrality and PlayeRank are all indexes about a player. But a team consists of many players, and obviously, only if all the players perform well will a team have a good outcome.
- **Omitting the substitution of players won't cause a big influence on our model.** Though a team consists of 11 members at a specific time, there are more than 11 nodes in our passing network, because of substitution during the match. However, we don't reckon this as a big deal, because all the nodes and edges did exist in some periods, which shows the structure and connection of a team.

2 Passing Model Based on Complex Network Method

2.1 Data Preprocessing

There are altogether 11 types of event data given, as shown in Figure 1. As the most parts of all data, pass and duel will be the focus of modeling and analysis.

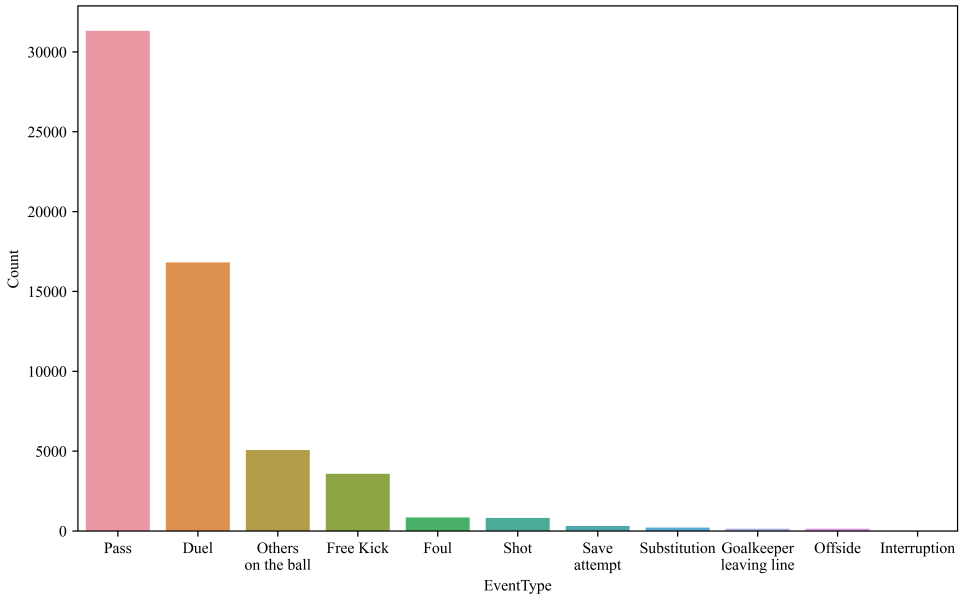


Figure 1: Types of Data

After data cleaning, we drew the thermal diagram of the frequency of each behavior at different positions, as shown in Figure 2. It describes the spatial distribution characteristics of team behavior. After data preprocessing, we find it convenient to construct the time-varying passing network based on graph theory.

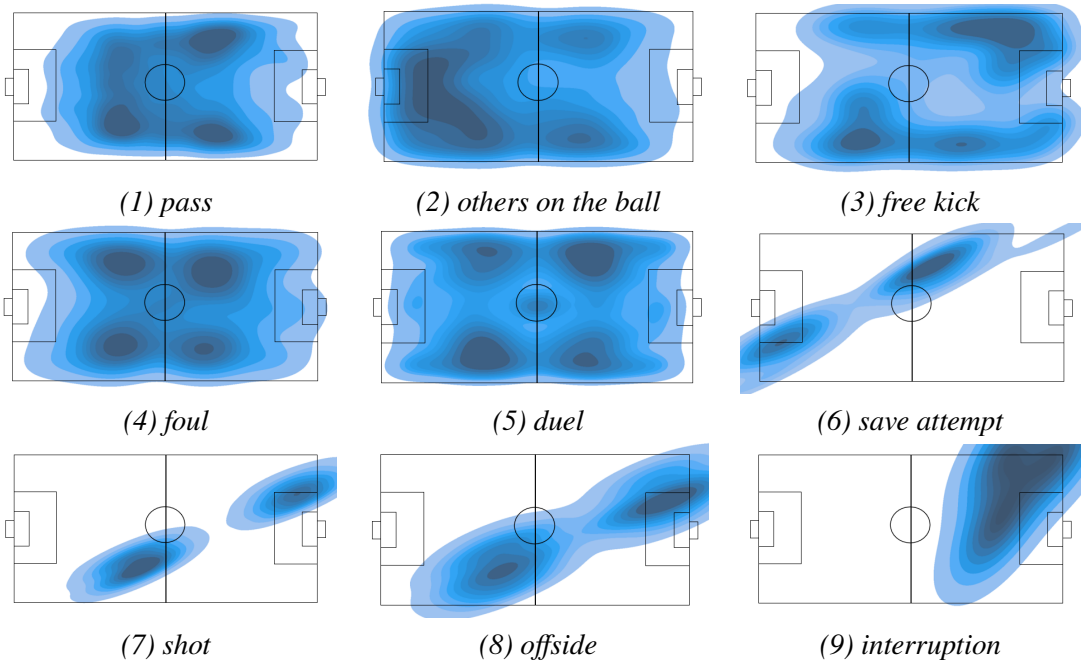


Figure 2: Thermal Diagram of Behaviors

2.2 Passing Network Construction

2.2.1 Structure

Based on graph theory, we described the characteristics of nodes and edges in a network, with time-varying state matrices of nodes and adjacency matrices of edges, to construct the passing network. For the long-time passing network, considering the low importance of timing information, we adopt the mean value method to simplify the model.

Node. State vectors at each moment are used to describe the node. According to the data sets, the state of node at moment t could be determined by parameters pid(player ID, used to discriminate against different players), cid(contest ID, used to describe which contest the node belongs to), r(role in the team, including goalkeeper, defender, midfield, forward), x, y(position at t moment, used to locate the node), b(behavior at moment t, including pass, duel, free kick, foul, offside, save attempt, shot, goalkeeper leave the line, interruption, none and others). Define

$$v_i^{(t)} = [pid_i^{(t)}, cid_i^{(t)}, r_i^{(t)}, x_i^{(t)}, y_i^{(t)}, b_i^{(t)}]^T \tag{1}$$

to describe the state change of node v_i over time, and then we use the matrix composed of node state vectors to represent the set of node $V^{(t)}$, where

$$\begin{aligned} V^{(t)} &= [v_1^{(t)}, v_2^{(t)}, \ldots, v_n^{(t)}] \\ &= [PID^{(t)}, CID^{(t)}, R^{(t)}, X^{(t)}, Y^{(t)}, B^{(t)}] \end{aligned} \tag{2}$$

It should be emphasized that the state vector of nodes changes with time due to different player states at different moments. In this paper, for the long-time passing network, timing information is relatively less effective and difficult to extract. We use average value method to simplify the model. Within a certain time span T, establish the average state matrix of nodes

$$\begin{aligned} \overline{V} &= [\overline{v_i}, \overline{v_i}, \ldots, \overline{v_i}] \\ &= [\overline{PID}, \overline{CID}, \overline{R}, \overline{X}, \overline{Y}, \overline{B}] \end{aligned} \tag{3}$$

to describe the overall state of a node within the time span T.

Edge. We use the adjacency matrix to define the edge set E of the passing network and describe the interaction between nodes

$$E^{(t)} = \begin{bmatrix} e_{11}^{(t)} & e_{12}^{(t)} & \dots & e_{1n}^{(t)} \\ e_{21}^{(t)} & e_{22}^{(t)} & \dots & e_{2n}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1}^{(t)} & e_{n2}^{(t)} & \dots & e_{nn}^{(t)} \end{bmatrix} \tag{4}$$

Actually, edge set is also a time-varying matrix $E^{(t)}$ with timing information. In formula (4), e_{ij} represents the interaction between node v_i and node v_j at moment t. Define

$$e_{ij}^{(t)} = \begin{cases} 1, & \text{if pass between } v_i \text{ and } v_j \text{ at moment } t; \\ 0, & \text{else.} \end{cases} \tag{5}$$

For the long time passing network, we also use the mean value method to simplify the model. Within a certain time span T, establish the average adjacency matrix

$$\overline{E} = \begin{bmatrix} \overline{e_{11}} & \overline{e_{12}} & \dots & \overline{e_{1n}} \\ \overline{e_{21}} & \overline{e_{22}} & \dots & \overline{e_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{e_{n1}} & \overline{e_{n2}} & \dots & \overline{e_{nn}} \end{bmatrix} \tag{6}$$

where $e_{ij} = k \geq 0$, indicating that, within T time, the total number of passes which v_i and v_j participate in is k (namely, mean edge weight is k).

Network. Combine the model of nodes and edges, and we finally built the passing network based on graph theory

$$G^{(t)} = (V^{(t)}, E^{(t)}) \tag{7}$$

The information contained in the model includes $V^{(t)}$, which represents the individual performance information of each team member, and $E^{(t)}$, which represents the interaction information of each team member, basically covering all important information of the data set. For the long time passing network, we use the mean value method to simplify.

2.2.2 Category

According to the demand of the research, we built three kinds of passing network based on the method in section 2.1.1. Model is visualized, using the average position of nodes to determine node coordinates, different colors of nodes to represent different roles, and edge thickness to represent passing frequency.

According to the season. First, to describe the situation in the whole season, we constructed the mean value network as shown in Figure 3.

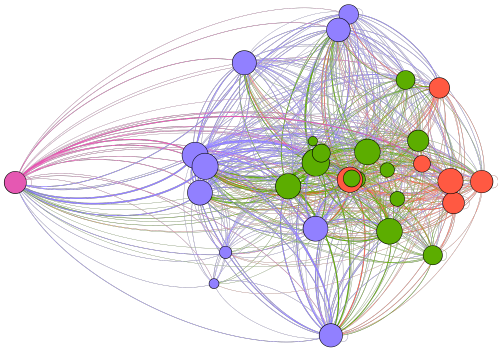
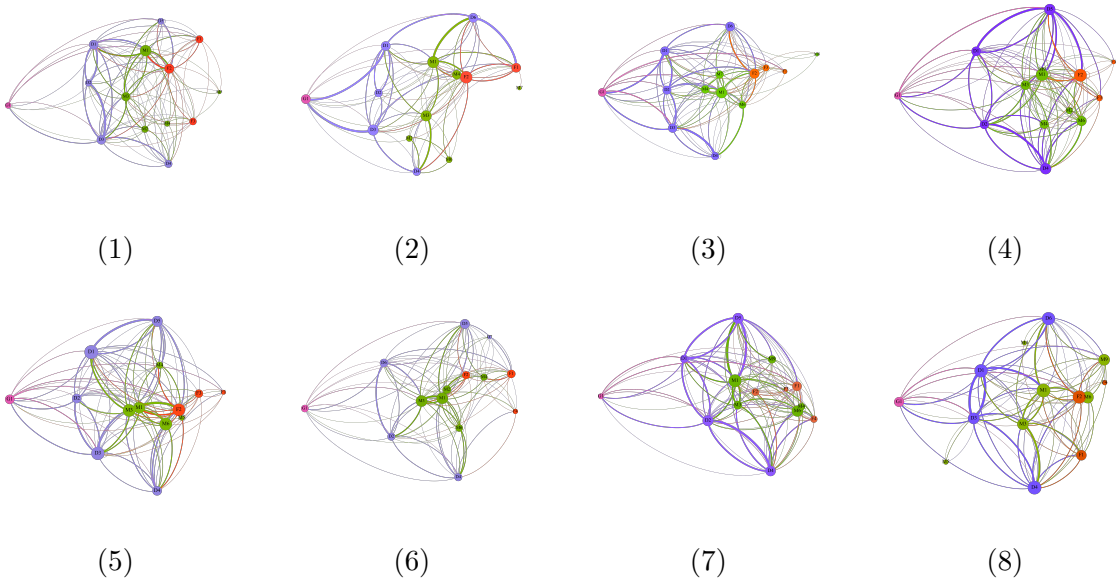


Figure 3: Network for the whole season

According to matches. We constructed the model for each match also, to describe detailed features. Here, we displayed 8 matches as examples in Figure 4.



(1) (2) (3) (4)

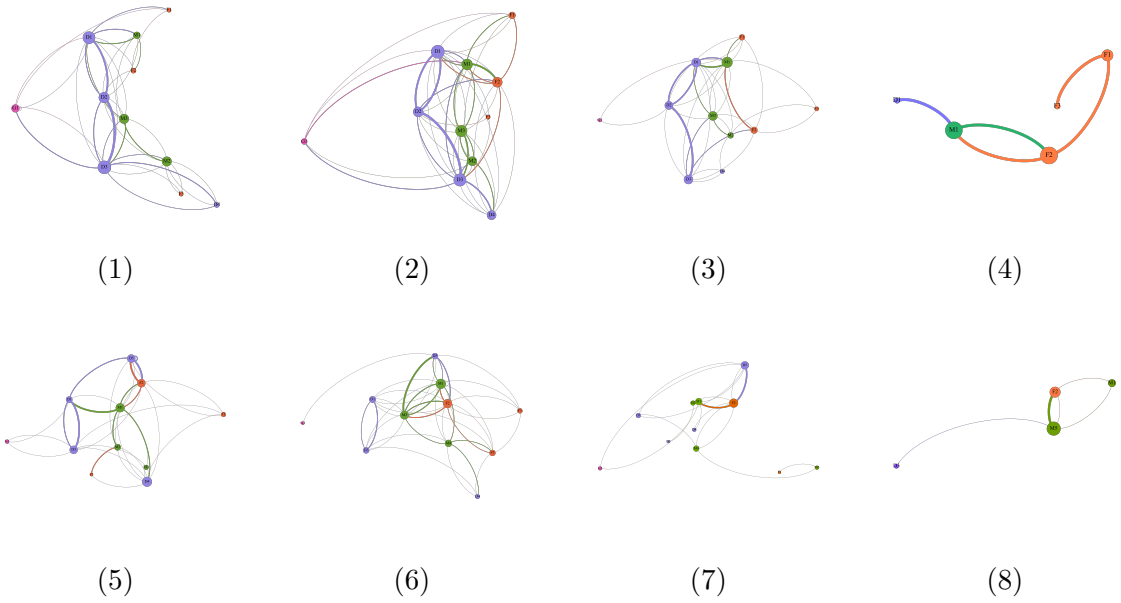
(5) (6) (7) (8)

Figure 4: Network for a match

With a span of 15 mins. To describe periodic changes, we We'll split the game every 15 minutes, and model for every periods. Here, in figure 5, we showed 8 graphs in a match, let's say, match 1. After modeling and visualization, we could indentify the network patterns with graph theory methods.

2.3 Network Patterns Identification

To identify the network patterns, we firstly found the structural indicators and network properties according to the requirements and literature in section 2.3.1. We explore their potential information and the changesover time and the visualized results will be listed in section 2.3.2.



(1) (2) (3) (4)

(5) (6) (7) (8)

Figure 5: Network for a match with 15 mins span

2.3.1 Structural Indicators and Network Properties

Dyadic and triadic configuration. The dyadic and triadic configuration represent the relationship involving pairs of players and group of three players. In a general sense, the exploration of dyadic and triadic configuration could help find the main cooperation groups and cooperation modes in the team. It's represented by the sum of the total passes in the group. Define S_{ij} to be the sum of the total passes between pairs of players, which represents dyadic configuration, and S_{ijk} to be the sum among group of three players, which represents triadic configuration.

Largest Eigenvalue of Adjacency Matrix. The largest eigenvalue λ_1 of the weighted adjacency matrix A of a network is a measure of network strength. The weighted adjacency matrix A is a $N \times N$ matrix whose elements a_{ij} contain the number of passes going from player i to player j. The largest eigenvalue of $A$is bounded by the average number between players $\langle S \rangle$, as $\lambda \geq \langle S \rangle$, and also by $s_{max} \geq \lambda_1 \geq max(\langle S \rangle, \sqrt{s_{max}})$, where s_{max} is the maximum numnber of passes that a player has made to any other player of his team. As a rule of thumb, networks with higher number of links (passes) will have a higher λ_1 and networks with the important nodes connected between them will have higher λ_1 than networks where the hubs (i.e.,importent players) are not directly connected between them.

Algebraic Connectivity. The algebraic connectivity $\widetilde{\lambda}_2$ corresponds to the second smallest eigenvalue of the Laplacian matrix $\widetilde{L}$, which is defined as $\widetilde{L} = S - A$. The algebraic connectivity is an indicator of the modular structure of a network. The lower the $\widetilde{\lambda}_2$, the clearer the existence of independent groups inside the network. In the framework of multilayer networks, one can interpret the value of $\widetilde{\lambda}_2$ as a way to quantify structural integration and segregation of different network layers. On the other hand, $1/\widetilde{\lambda}_2$ is proportional to the time required to reach equilibrium in a linear diffusion process. Additionally, the time t_{sync} to reach synchronization of an ensemble of phase oscillators that are linearly and diffusively coupled also proportional to $1/\widetilde{\lambda}_2$.

Clustering Coefficient. In general, the local clustering coefficient of a node i is obtained as the percentage of the nodes directly connected to it that, in turn, are connected between them.

This measure can be averaged along the N nodes of network to obtain the average clustering coefficient. However, when the network is weighted, we can not simply account for the number of nodes 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 in not constant. In this way, we use the weighted clustering coefficient $C_w^{(i)}$ to measure the likelihood that neighbours of a given player i will also be connected between them.

$$C_w^{(i)} = \frac{\sum_{j,k} w_{ij} w_{jk} w_{ik}}{\sum_{j,k} w_{ij} w_{ik}} \tag{8}$$

where j and k are any two players of the team and w_{ij} and w_{ik} the number of passes between a third player i and both them. Finallly, the clustering coefficient of the whole network is obtained by averaging $C_w^{(i)}$ over all players, i.e., $C = \frac{1}{N} \sum_{i=1}^{N} C_w^{(i)}$. 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.

Shortest-path Length. In a passing network, the shortest path length is the minimum number of players that must be traversed by the ball to go from one player to any other. Since passing networks are weighted (i.e.,the number of passes between players is different), we have to take into account the different weights of the links, considering that, the higher the weight, the shorter the topological distance between two nodes. The topological length l_{ij} of the link etween two players i and j is defined as the inverse of the link weight, $l_{ij} = 1/w_{ij}$. However, when computing d for weighted networks, the shortest-path length between a pair of players may not be a direct link, since there could exist a shorter path by combining two (or more) alternativ links. Therefore, we compute the minimal shortest-path p_{ij} between all pairs of players using the Dijkstra's algorithm. Next, we define the average shortest path d of the whole team as

$$d = \frac{1}{N(N-1)} \sum_{i \neq j} p_{ij} \tag{9}$$

where $N = 11$ is the total number of players of the team.

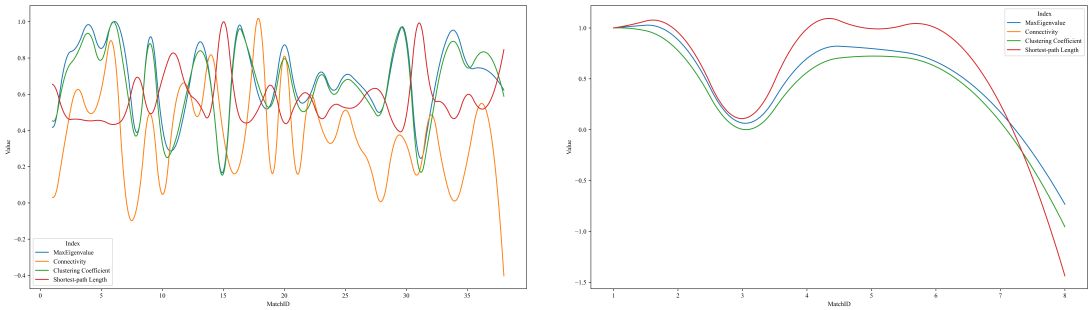
2.3.2 Results

Dyadic and triadic configuration. According to S_{ij} and S_{ijk}, we sorted all the dyadic and triadic configuration, as shown in Table 2. In the passing network, the rank represents the relationship involving pairs of players and group of three players. Nodes with higher rank represent more important part in the network, meanwhile, pairs or groups of players who have higher rank may play more significant roles in the team.

Table 2: Dyadic and Triadic Configuration

Dyadic Configuration			**Triadic Configuration**		
Pairs of Nodes	**Pass Count**	**Rank**	**Groups of Nodes**	**Pass count**	**Rank**
M3&M1	311	#1	M3&M1&F2	731	#1
M1&F2	299	#2	M1&D5&F2	634	#2
D3&D1	203	#3	M3&M1&D1	613	#3
D1&G1	183	#4	M3&D4&M1	603	#4
D5&F2	181	#5	M3&M1&D5	601	#5
M1&D1	177	#6	M1&F2&D1	580	#6
...	...	...	...	...	...

Other indicators. We mainly explored how other indicators change with the flow of time. We calculated the change of largest eigenvalue of adjacency matrix, algebraic connectivity,clustering coefficient, shortest-path length over the season and over a match, just as shown in Figure 6. Note that in order to compare the trend of changes, all indicators are normalized to interval $[0, 1]$.



(1) Over the season *(2) Over a match*

Figure 6: Changes of Other Indicators

2.4 Conclusion

We can see from Figure 6 that there is a strong correlation between the largest eigenvalue of adjacency matrix and clustering coefficient. The trend of both is basically the same throughout the season. They both show how closely the team members cooperated with each other during a match. It's obvious that the values fluctuated within the normal range indicating the team was in a good state during the 38 games. From the angle of a single game, we can see that all the indicators varied obviously and consistently. In the first half, the players needed to take time to warm up their bodies and they preferred to slow down their tempo in order to save energy for the violent competition in the next game stage. In the second half, all the indicators kept high values which were related to the seesaw battle between the two teams and the game turned white-hot.

3 Teamwork Evaluation Model based on PCA-CART

3.1 Performance Indicators & Team Level Processes

Accoding to the literature, the teamwork evaluation model was built based on the player level and team level including 6 indicators. We will list them representively in section 3.1.1.

3.1.1 Indicators for Player

Flow centrality. Combining the flow network with the passing and shooting accuracy of the players, the probability that each path definable on the network finishes with a shot can be obtained. This procedure suggests a natural measure of performance of a player — the betweenness centrality of the player with regard to the opponent's goal, which is denoted as flow centrality. To make some simplification, we choose the ratio of pass event of one player as flow centrality, and the average of all players is defined as the team flow centrality.

PlayeRank. The key task addressed by PlayeRank is the evaluation of the performance quality of a player u in a football match m. This consists of computing a numerical rating $r(u, m)$, called performance rating, that aims at capturing the quality of performance of u in m

given only yhe set of events related to that player in that match. This is a complex task because of the many events ovserved in a match, the interactions among players within the same team or against players of the opponent team, and the fact that players performance is inextricbly bound to the performance of their team and possibly of the opponent team. PlayeRank addresses such complexity by means of a procedure which hinges onto a massive database of football-logs and consists of three phases: rating phase, a ranking phase and a learning phase.

To make some simplification, we assess the performance of a player by the ratio of his actions during the match, namely, the PlayeRank. And the geometric mean of all players is defined as the team PlayeRank. It can be seen that, only if all the players perform well, will the team PlayeRank be high.

3.1.2 Indicators for Team

The same as section 2, we still uesed the largest eigenvalue of the adjacency matrix, algebraic connectivity, clustering coefficient and shortest-path length as the indicators of a team in the teamwork evaluation model.

Therefore, a evaluation system was finally built as shown in Figure 7.

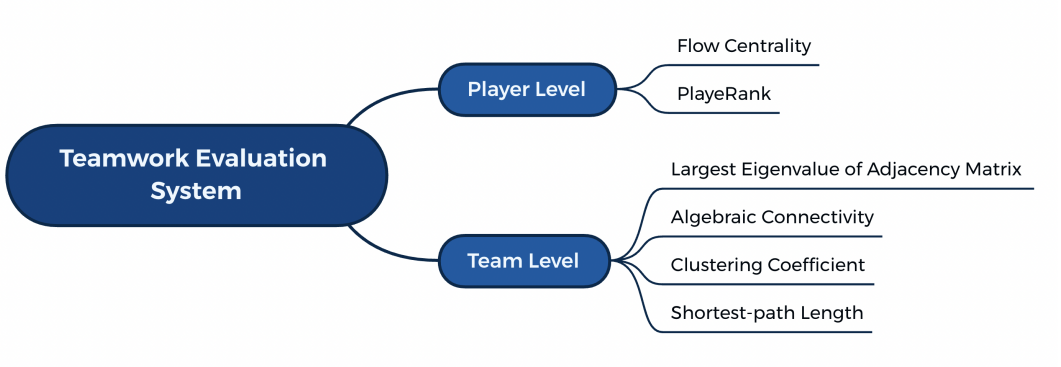


Figure 7: Teamwork Evaluation System

3.2 PCA-CART-based Teamwork Evaluation Model

3.2.1 Classification and Regression Tree

Classification and regression tree, that is, CART algorithm is used to construct the classification tree. The CART algorithm is based on two-minute recursive segmentation, so that each non-leaf node has two branches, so the decision tree generated by CART algorithm is a simple structure of the binary tree, which in each step decision can only be divided into 'yes' or 'no', even if a feature has multiple values, the data is divided into two parts. CART algorithm flow has the following two steps: recursively classify the data samples to construct the binary tree; pruning with the validation data to avoid overfitting.

The node division of CART is based on the GINI index, which is an inequality measure that can be used to measure any uneven distribution, between 0 and 1, which 0 is equal, 1 is completely different. When measuring the GINI index of all the values of a feature which belongs to the data set, we can get the GiniGain of the feature. When the pruning is not considered, the process of classification decision tree recursively creates is selecting GiniGain of the smallest node is the bifurcation point until the sub-datasets belong to the same class or

all the features are used. For a data set T, the GINI is calculated as

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2 \tag{10}$$

where n is the number of classes and p_j is the probability of different classes for the dataset samples. Gini Split Info, which measures the GINI index for all the values of a feature, is associated with the letter in the ID3 algorithm.

$$Gini_{split}(T, i) = \sum \frac{N_i}{N} gini(T_i) \tag{11}$$

where i represents the i-th value of the feature. The gain is similar, which can be called GINI information gain, that is, GiniGain. For CART, $i = (1, 2)$, get GINI information gain in Binary Split cases

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2) \tag{12}$$

Therefore, in the candidate attribute set A, we choose the attribute that minimizes gini after partitioning as the optimal partitioning attribute, that is

$$a_* = \arg_{a \in A} \min Gini_{split}(T, i) \tag{13}$$

3.2.2 PCA-based Improvement

Principal component analysis (PCA), is to map the n-dimensional feature to the k dimension ($k < n$), which is a new orthogonality. This k-dimensional feature is called the principal component, the reconstructed k-dimensional features, rather than simply removing the remaining nk-dimensional features from the n-dimensional features.

PCA eliminates the correlation between variables, which is known as one of the biggest obstacles for decision trees, and improves the accuracy and applicability of the model.

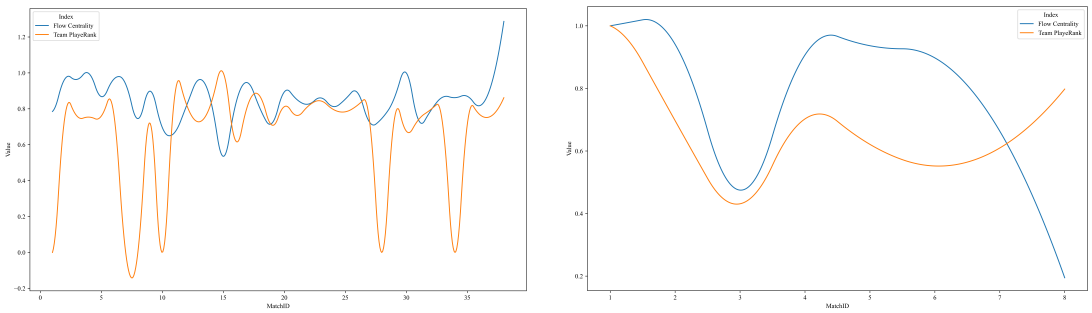
3.2.3 PCA-CART-based Model

Based on the CART and PCA methods, we finnally built the teamwork evaluation system. Here, we discribe the model with Algorithm 1 in appendix A.

3.3 Model Implementation and Result

3.3.1 Indicator Analysis

We first analysed the change of player-level indicator over the season and over a match. Figures for the change of flow centrality and PlayeRank were made as shown in Figure 8. Note that in order to compare the trend of changes, all indicators are normalized to interval $[0, 1]$.



(1) Over the season *(2) Over a match*

Figure 8: Changes of Player-level Indicators

3.3.2 Establishment of Decision Tree

According to PCA-CART algorithm, we successfully implied our model into football passing dateset, and built the decision tree for the teamwork evaluation system as shown in appendix B. Based on the decision tree, we could easily find how the change of indicators influences the result of a match.

In order to measure the performance of model, we made 10- fold cross validation, and define

$$ace(f; D) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i) = y_i) \tag{14}$$

to be the overall accuracy rate of the model f in dataset D.

After simulation and calculation based on the given data, the acc of the model was proved to be 83%, which is excellent, in relative speaking.

3.4 Conclusion

Based on PCA-CART methods, we use player-level and team-level indicators, including flow centrality, playerank, largest elgenvalue of adjacency matrix, algebraic connectivity, clustering coefficient and shortest-path length, we establish the teamwork evaluation system, which could generate the decision tree to help the determination. By the mean of 10-fold cross validation and judged by overall accuracy rate, the model is proved to be accurate and reliable.

4 Analysis and Suggestions for Structural Strategies

4.1 Correlation Analysis for Structural Sterategies

4.1.1 Spearmann Correlation Analysis

Spearman correlation coefficient is defined as Pearson correlation coefficient between hierarchical variables. For samples with a sample size of n, n raw data are converted to hierarchical data. The correlation coefficient ρ is

$$\rho = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \tag{15}$$

The original data is assigned a corresponding rank based on its average descending position in the total data. In practical applications, the links between variables are of no great importance.

So, a simple step can be used to calculate the grade difference of the observed two variables. $\rho = r_s$ is

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{16}$$

where $d_i = (x_i - y_i)$, x_i and y_i represent the rank of two variables in descending order, respectively.

When testing for significance, we set 'H_0: X and Y are independent;' and 'H_1: X is positively correlated with Y.' Then, the rejection fields of hypothesis testing are $W = \{\rho_s >= c_\alpha\}$ and $W = \{\rho_s <= d_\alpha\}$ respectively, where c_α and d_α are critical values. If the null hypothesis is true,

$$t = r_s\sqrt{\frac{n-2}{1-r^2}} \tag{17}$$

obeys a T distribution with $v = n - 2$ degrees of freedom.

At significance level α, if the value of the statistic falls in the negation field $\{|t| > \frac{t_\alpha}{2(n-2)}\}$, we reject the null hypothesis, and we can say the Spearman correlation coefficient is significant; or, we accept the null hypothesis, and the Spearman correlation coefficient is not significant.

4.2 Results

Here we define a new index, Goal Difference(GD), as the assessment of the team's performance in a match, which is calculated by the number of goals scored by Huskies minus that of the opponent. As shown in Table 3, the Goal Difference(GD) is most relevant to indexes MaxEigenvalue and Clustering Coefficient (actually, it can be seen from the Table that these two indexes have the most relevance). And the relation between GD and Flow Centrality is also positive. So, the key to increasing GD is to increase MaxEigenvalue and Clustering Coefficient, and Flow Centrality. Also, we have assessed both dyadic and triadic configurations previously. They are important to our strategies too.

Table 3: Result of Correlation Analysis

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1.000	0.043	-0,062	0.194	-0.118	0.209	-0.164
x_2	0.043	1.000	-0.040	0.840	0.290	0.803	-0.768
x_3	-0.062	-0.040	1.000	-0.166	0.756	-0.118	-0.115
x_4	0.194	0.840	-0.166	1.000	0.269	0.967	-0.837
x_5	-0.118	0.290	0.756	0.269	1.000	0.298	-0.437
x_6	0.209	0.803	-0.118	0.967	0.298	1.000	-0.833
x_7	-0,164	-0.768	-0.115	-0.837	-0.437	-0.833	1.000

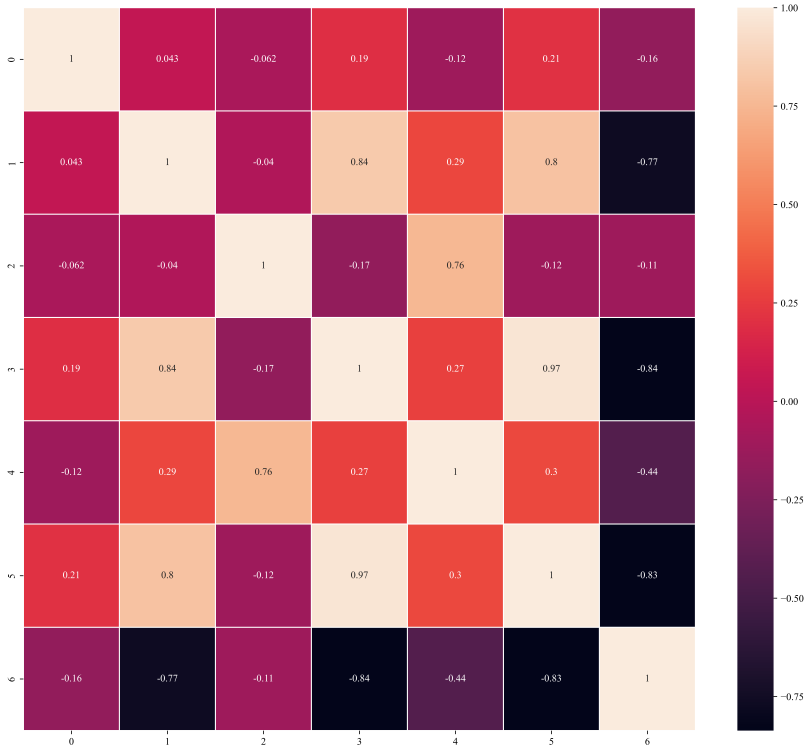


Figure 9: Result of Correlation Analysis

4.3 Suggestions

- Looking back on the definition and meaning of MaxEigenvalue and Clustering Coefficient, it can be seen that, to increase the index, the key is to make players more united, and to maintain and expand advantages in passing flexibility. The Huskies team's minimum distance is lower than the average, indicating that the passing flexibility is higher than the average, which is beneficial. It is necessary to exert its advantages.

- It's very important to pay attention to the level balance of players, which makes Flow Centrality index higher. The number of passes received by different players varies greatly, and the positions of players vary greatly. Coaches should pay attention to the balance between players. Different players should be treated roughly equally to get higher scores.

- From Table 3, it can be seen that some kinds of dyadic and triadic configurations are very frequent, which influence the outcome of a match to some degree. Generally speaking, these players are the core of a team, so it is very important to train the members of these dyadic or triadic configuration and make them cooperate with each other better.

5 Extension of the Model

5.1 From Football to Life

To extent the model, we explore the inner workings of a team, and how to achieve an excellent team. By the means of extension of every indicators, we make it frome football to life.

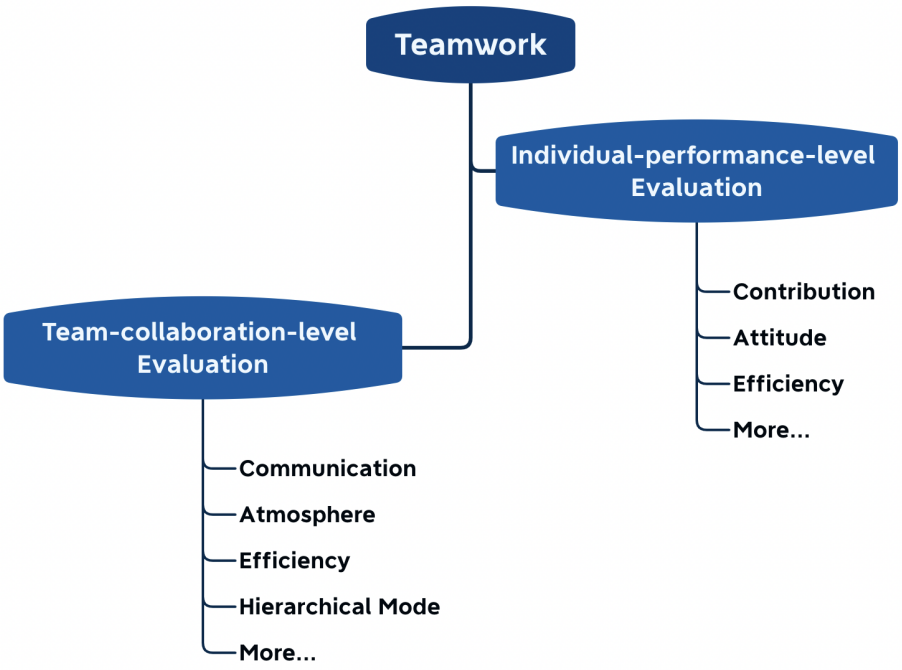


Figure 10: From Football to Life

5.2 Individual Level

Flow centrality focuses on the performance of one member, which can be extended to different treatment of performance appraisal. Different criteria are adopted to judge people with different division of labor in the team, which is conducive to the formation of a more effective team.

The PlayeRank can be extended to the difference of individual contributions. The contribution of everyone in the team tends to be heterogeneous, which is conducive to the formation of a more effective team.

5.3 Team-colloboration Level

The MaxEigenvalue can be extended to the coverage of communication in a team. The team should not have a number of key figures to shift the center of the network. As far as possible, the transmission of information can cover every member evenly, which is conducive to the formation of a more effective team.

Connectivity can be extended to the effectiveness of communication. Effective communication is conducive to the formation of more effective teams.

Clustering coefficient can be extended to the frequency of communication among team members. Members of the team should not communicate as often as possible. They should pay attention to the efficiency of communication.

Shortest-path Length can be extended to the hierarchical mode. The more complex the hierarchical structure of the team organization, the more the shortest path will be added, while the flattening mode will have the smaller shortest path. The size of the shortest path has no

significant effect on team performance, therefore, for the hierarchical model, the appropriate one is the best.

5.4 Conclusion

So far, we have extended the model established above. We get the following conclusion: In order to improve team effectiveness, we should not only focus on personal improvement, but also focus on team building. No matter what index is chosen, it is impossible to correctly reflect the effectiveness of the team a hundred per cent. We need to think about it dynamically and make it more practical when it comes to evaluating a team and everyone in that.

6 Final Remark

In this section, we'll make a general summary of our work, including the model we built, the strategies we found, the pros & cons of our methods, and the future development of our model.

6.1 Strengths and Weaknesses

6.1.1 Strengths

- **Accurate and intuitive.** The result of the solution has been proved to be accurate in Section 2 and 3. We also made pairs of figures to help the result comprehensible.
- **Contractor-friendly.** Methods could be used just with some sets of contest data, which is not difficult for a team or club to provide.
- **Extendable.** The parameter of the model shares high interpretability, which, therefore, makes the solution easy to extend just as Section 5 says.

6.1.2 Weaknesses

- **Data-based only.** The football contest is an extremely complex system which is impossible for any model to do mechanism analysis completely. Our work is supposed to be carried out based on historical data, as detailed as possible.
- **Simplified temporal networks.** Considering the complexity of model, we simplify the long-time network by average methods. Some kinds of temporal information could be ignored.
- **Hard to be exhaustive.** For comprehensive evaluation methods, it's hard to reach every aspect of a matter and impossible to be totally objective, our methods are no exception.

6.2 Future Model Development

- **Further exploration of mechanism analysis.** More work could be done to find the kernel of the football passing network, which could give more explicable ideas for strategies in the future.
- **Wider exploration of performance characteristics.** Limited by the variety of the data sets and the complexity of the model, characteristics used in our solution remain restricted, which could be developed to make the model have stronger robustness in the future.

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] 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.

[3] Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D., & Malvaldi, M. (2015). The harsh rule of the goals: Data-driven performance indicators for football teams. *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 1-10, 7344823.

[4] Duch J., Waitzman J.S., Amaral L.A.N. (2010). Quantifying the performance of individual players in a team activity. *PLoS ONE*, 5: e10937.

[5] Boccaletti S, Latora V, Moreno Y, *et al*. Complex Networks: Structure and Dynamics[J]. *Complex Systems and Complexity Science*, 2006, 424(4-5):175–308.

[6] Lewis Ted G.. Network Science:Theory and Practice[M]. Hoboken, NJ, USA : John Wiley & Sons, Inc., 2008

[7] Herrera-Diestra J L, Echegoyen I, JH Martínez, *et al*. Pitch networks reveal organizational and spatial patterns of Guardiola's F.C. Barcelona[J]. *Chaos Solitons & Fractals, 2020*, 138:109934.

[8] Anthony, C, Constantinou, *et al*. Pi-football: A Bayesian network model for forecasting Association Football match outcomes[J]. *Knowledge Based Systems*, 2012.

[9] Narizuka T, Yamamoto K, Yamazaki Y. Degree Distribution of Position-Dependent Ball-Passing Networks in Football Games[J]. *Journal of the Physical Society of Japan*, 2015, 84(8).

[10] Arriaza-Ardiles E, JM Martín-González, Zuniga, M.D, *et al*. Applying graphs and complex networks to football metric interpretation[J]. *Hum Mov*, 2017:S0167945717306280.

[11] Felix L, Barbosa C M, Vieira V, *et al*. A Social Network Analysis of Football with Complex Networks[C]. *XXV Simpósio Brasileiro de Sistemas Multimídia e Web*. 2019.

[12] Torres D, Rocco C. Reliability assessment of complex networks using rules extracted from trained ANN and SVM models[C]. *International Conference on Hybrid Intelligent Systems. IEEE Computer Society*, 2005.

Appendices

Appendix A: Algorithm for PCA-CART

Algorithm 1: PCA-CART

Input: training dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
attribute dataset $A = \{a_1, a_2, \ldots, a_d\}$.

Progress:

process variables with PCA;
generate $node$;
if The samples in D all belong to the same category C **then**
  mark $node$ as a leaf node of class C; **return**
end if
if $A = \emptyset$ **or** samples of D have the same values in A **then**
  mark $node$ as a leaf node, mark its category as the most-sample class in D; **return**
end if
select the optimal partition attribute a_* from A
for every value a_*^v of a_* **do**
  generate a branch for $node$; let D_v represent the subset of D in which the value of a_* is a_*^v;
  if $D_v = \emptyset$ **then**
    mark branch nodes as leaf, mark its category as most-sample class in D; **return**
  else
    let $TreeGenerate(D_v, A \setminus \{a_*\})$be the branch node
  end if
end for

Output: a decision tree with $node$ as the root.

Appendix B: Decision Tree for Teamwork Evaluation