Problem Chosen	2020	Team Control Number
D	MCM/ICM Summary Sheet	**2008803**

Summary

Team sports is one of the most informative settings to conduct our cooperation related research. Recently, invited by the coach of the Huskies, we exploit our modeling skills to analyze this soccer team from multiple aspects. Through developing a series of models, namely the Ball Passing Graph Model (BPGM), the Competitive-Oriented Index Estimation System (COIES), and the Optimal Passing Route Model (OPRM), we conduct a complete analysis for the Huskies' team cooperation properties. Also, we put forward the long-term and on-field strategy advice for the coach.

Firstly, based on graph theory, we construct a well-defined network (BPGM) as our basic model. Through the static analysis, we gain a deep insight into the cooperation patterns. In general, their defenders have strong passing ball abilities while the forwarders seem to drag the team down. In most of the matches, the Huskies takes on the triadic configurations between D1,D2 and D3. Sometimes, they will also try the attacking dyadic cooperation between F1 and M1. This reflects its relatively weak coodinate abilities. Refering to the dynamical analysis, we abstract their overall time dependent strategy 'Defend-Attack-Preserve' (DAP). The whole season analysis also reveals their defects in attacking abilities and cooperaton.

Secondly, in order to quantify the parameters in the model, we create a sub-model COIES and explore the universal long-term improvement strategies for the Huskies. We start with extracting six quantitative performance Indicator (QPI) indices which serve as our basis for modified logistic regression (MLR). After ranking the six QPI indices according to their weight, we conclude that the contribution difference and coordinate coefficient are among the highest. The model overall estimation score is 76.6 along with $81.6\%$the prediction accuracy. Furthermore, we introduce a hyperplane method as a possible extension. This allows as to point out the Huskies should enhance their coordinate coefficient by 39.2% and the contritbution difference by 21.1% to achieve the average level among all teams. They need to get trained on cooperation capabilities and cultivate their own superstars.

Thirdly, based on the established one single team, we combine the opponent team to form a full-connected network (OPRM) which could be used to provide on-field strategies for the Huskies to fight with the specific opponent. Refering to Markov decision process (MDP), we construct our own tuple elements and policies for this soccer competition background. After using value iteration method to maximize the value function, we give out a concrete ball passing route against one opponent where F2 plays the dominant role. The visualization effect of this process can help the players make plans even as the competition progresses.

In addition, we consider the influence of two interesting parameters in model sensitivity. Finally, in response to the conclusions of this paper, we write a letter to the coach of the Huskies about the team's current situation and their future developments.

Keywords: cooperation; network; regression; Markov

Content

1 Introduction

1.1 Restatement of the Problem

With the increasing mutual bonds between different social sectors, the set of challenges we face are becoming more and more complex. It is common now to rely on people from all walks of life with diverse skills to tackle many of these tough issues. Over the past 50 years, researchers have been struggling to explore the best strategies for team management, optimal relations among teammates, and effective leadership styles. In this project, we focus on team processes in one of the most informative settings, the soccer competition, where team success is much more than the sum of the abilities of individual players. Diversity of skills, team balances and the abilities to effectively coordinate over time all contribute to the final triumph.

Now following the requests from the Huskies' coach, we have four major tasks:

- Create a network for the ball passing between players, which means **only consider the passing events**, and then identify various network patterns and structural properties from static and time-dependent viewpoint.
- Identify performance indicators like coordination, flexibility and diversity etc. to reflect **how possible the team will win**. Then analyze the strategy dependence due to the changing of opponents. Finally, create a model that captures structural and dynamical charateristics of teamwork.
- Inform the coach about the **long-term improvements and certain countermeasures to opponents** then give some advice.
- Generalize our findings on how to design more effective teams. Capture other key cooperation features for further model development.

1.2 Review of Previous Work

Because of the usefulness of network analysis (NA), many scientists have been concerned about its potential to gain insight into many aspects in our daily life, such as vehicle analysis [1, 2, 3], optimal route choices [4, 5], communication detections [6, 7] and sports competitions [8, 9]. The review of the networks developments in soccer team investigation has been presented in [10] which also serves as the basis of our current project. Although big data analytics (BDA) has many other technical algorithms for data processing, NA still preserve its popularity for the simple but clear mathematical forms.

1.3 Our Work

Here, we briefly introduce our work. In Section 3, we use graph theory method to make analysis of team patterns, including individual performance, whole team estimation and sub-pattern searching. In Section 4, based on logistic regression method and a possible hyperplane extension, we provide concrete suggestions for the Huskies with the help of overall indicator and weight coefficient. In Section 5, we exploit the Markov decision process to find the best ball passing route for the Huskies to fight against certain opponents.

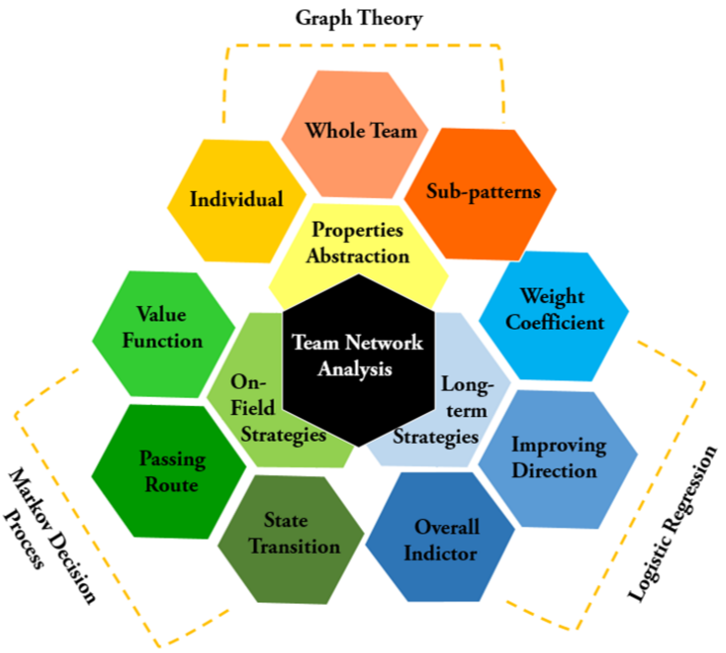


Figure 1: Our work

2 Preliminaries

2.1 Basic Assumptions

It is difficult to study the actual team cooperation with many uncertain conditions in reality, such as the multiple shape of the soccer fields, various soccer shoes the players wearing and the unpredictability of the weather when the competitions start. In order to simplify the problem and to hold the correctness of analysis at the same time, we make several assumptions below and explain our reasons.

- We assume that the soccer field only has one shape like Fig.2. The position ranges in the given data sets are x, y both from 0 to 100 code unit. However, according to the international soccer regulation, the competition field is not a full square. We then scale x ranges to $[0\ \mathrm{m},\ 100\ \mathrm{m}]$ and y to $[0\ \mathrm{m},\ 64\ \mathrm{m}]$. All the competitions are held in this field no matter what sides are considered.

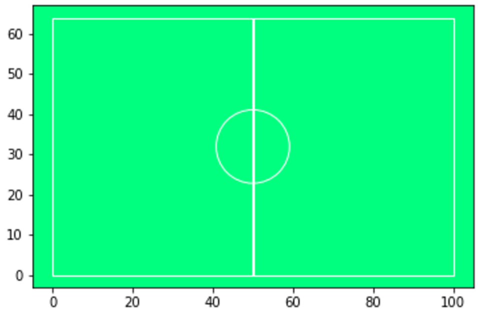


Figure 2: The soccer field with length $100\ \mathrm{m}$ and width $64\ \mathrm{m}$.

- When analyzing the ball passing network properties in Section 3, for simplicity, we assume that there are only **three types** of pass scenarios, namely long pass, medium pass and short pass, without considering any other complicated actions. When we identify the performance indicators, we begin to consider all **different actions** even including fouls.

- We assume that the players' physical conditions are always good enough. They can just feel tired, but without being hurt by other teamates and opponents.
- We also assume that the **time** for each soccer match is always 100 min i.e. 6000 s, each half section lasting 50 min i.e. 3000 s. We do not consider the injury time.

2.2 Basic Notations

In this section, we briefly summarize our variable notations in Table.1 for future use.

Symbol	Definition/Explanation
$G(V, A)$	Graph G with vertices V and edges A
w_{uv}	Weight of arcs between nodes u and v
c_{pass}	Value for different kinds of passes
$Odeg(u)$	Outdegree for vertex u
$Ideg(u)$	Intdegree for vertex u
H_{hod}	Number of passing ball events from teammates
L_{los}	Number of losing ball events
I_{int}	Number of intercepting ball events
$d(u, v)$	The shortest-path distance between nodes u and v
$\sigma_{u,v}$	The number of shortest (u, v)-paths
$\sigma_{u,v\|t}$	Number of paths through node t other than u, v
$\hat{w}_{uv}$	Weight normalized by the maximum weight in the network

Table 1: Basic Notations

3 The Ball Passing Graph Model (BPGM)

3.1 Mathematical Structure of BPGM

We create the network for the ball passing within graph theory. Suppose we have a directed graph $G = (V, A)$ where V is a finite set of vertices and A is the ordered pairs of V called arcs. We then attribute the weight for each arc to form a weighted directed graph i.e. the network. In reality, we specify some of the graph theory notations with its actual meanings in this model.

- Each player is a node. Since the total number of players in a team is 14 in soccer competition, including 3 substitutes, $cardV$ is supposed to be 14.
- If there exists a successful pass between player $u \in V$ and $v \in V$, we draw an arc between these vertices with weight w_{uv}, where $(u, v) - passes$ denoting all pass events between nodes u and v. c_{pass} supposes to be 3 for long pass (> 40 m), 2 for medium pass (20 m ~ 40 m) and 1 for short pass (< 20 m).

$$w_{uv} = \sum_{(u,v)-passes} c_{pass} \tag{1}$$

- We define the outdegree ($Odeg$)(indegree ($Ideg$)) for each vertex to be the number of successful passes starting(ending) from this vertex.
- We call the maximum sum weight subgraph $G[V_i]$,

$$G[V_i] = \underset{|V_i|=i, V_i \in G}{\arg\max} \sum_{e_j \in G[V_i]} weight(e_j) \tag{2}$$

to be **i Order Best Cooperation Pair (iOBCP)**. For simplicity, 2OBCP is also called d.c. while triadic configuration is for 3OBCP.

Here, we present our network for the Huskies based on passing data events under Coach1 supervision in Fig.3 with detailed static analysis in Section 3.2 and dynamical analysis in Section 3.3.

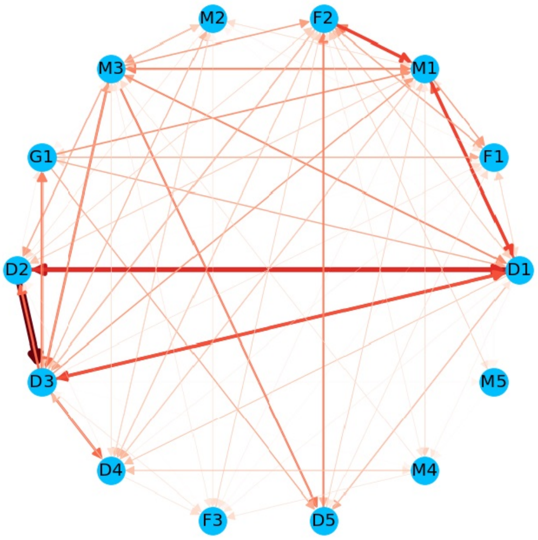


Figure 3: Network for Huskies. Blue nodes represent the 14 players labeled with positions (F-Frontier, M-Midfielder, D-Defender, G-Goalkeeper). Directed red arcs between these nodes denote the effective ball passing between the two connected players with deeper color for larger weight. The triangle between player D1, D2, and D3 is the overall triadic configuration which indicates the teams defending structure.

3.2 Static Analysis of the Huskies

The philosophy of the static analysis is aimed to gain a deep insight into the **cooperation itself**, regardless of the time dependence which could exert significant influences on the players' physical conditions and the variational competition strategies. Obviously, a successful soccer cooperation relys on excellent personal performance to some extent. The team members should have relative high individual capabilities, such as precisely passing balls between teammates and quickly receiving balls without being intercepted. To measure such quantities, we define intercepting (I_{int})(losing (L_{los})) for each team member to be the amount of getting ball(losing ball) events from(to) their opponents. Similarly, holding (H_{hod}) is defined to be the passes within their own team.

Back to the Huskies, we first give out a detailed data analysis to each team member in Table.2.

Player	H_{hod}	L_{los}	I_{int}
D1	47	5	10
D2	33	3	4
D3	42	6	3
D4	14	3	2
D4	12	5	2
F1	10	1	2
F2	34	4	8
F3	8	1	1
G2	10	2	2
M1	45	7	5
M2	14	6	4
M3	40	5	3
M4	8	1	3
M5	2	1	0

Table 2: The Huskies' players ball passing performance

Corresponding to this table, the defenders are all have **relatively strong** passing ball abilities. Their wrong action rates are below 10%. For midfielders, this ability is not as good as the defender (especially M2), but it is still strong enough to create chances for the forwarders to making attacks. However, the forwarders perform **less than satisfactory** in keeping and passing balls which may lead to the missing of scoring oppotunities.

In order to gain more objective quantities in measuring the mutual correlations between players. We uses more complex variables and some key feature parameters in graph theory. We list our formulas below with simple explanations. In Fig.4, we calculate the normalized value of such variables to player D1, D2, M1, and M2 with distinctive features. The overall performance further demonstrates their **relatively good** cooperation between team members. However, as for the whole team, the cooperation ability is a bit **lower** than their opponents. We specify this in Section 4.

- Closeness Centrality (Cc):measuring how the players are close to each other.

$$Cc(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(u,v)} \tag{3}$$

- PageRank (Pr): Popularity (getting more passes) among other players (see [11]).

- Betweenness Centrality (Bc): representing the interactions and the control of team power among all other players.

$$Bc(u) = \sum_{s,t \in V} \frac{\sigma(s,t|u)}{\sigma(s,t)} \tag{4}$$

- Eigenvector Centrality (Ec): providing information on the regulatory relevance between players (see [12]).
- Clustering (Clust): measuring the possible triadic configurations with this player participated in.

$$Clust(u) = \frac{1}{Deg(u)(Deg(u)-1)} \sum_{uw \in V} (\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw})^{\frac{1}{3}} \tag{5}$$

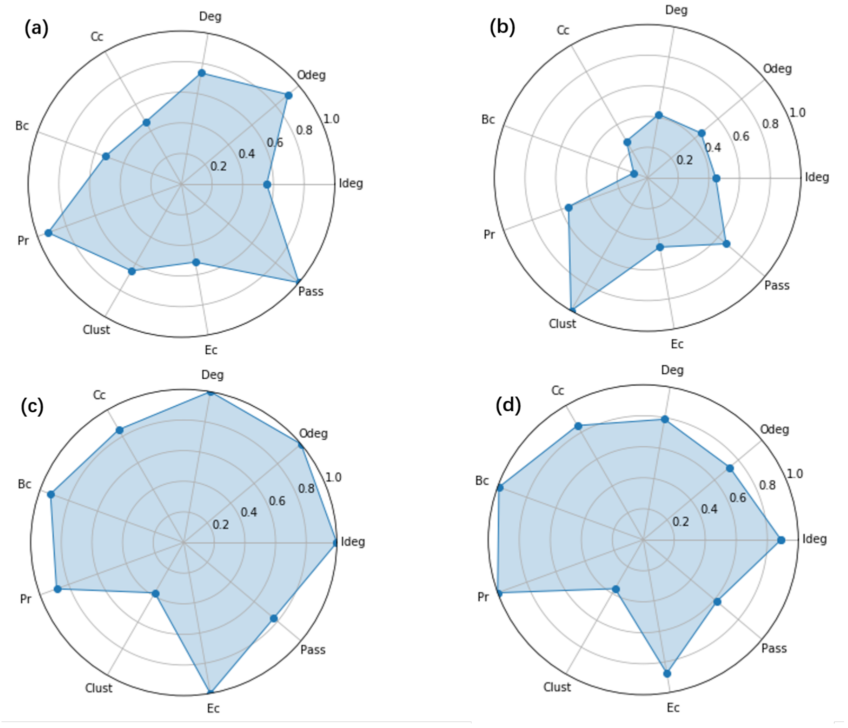


Figure 4: Different feature parameters for four players. (a)D1 (b)D2 (c)M1 (d)F2

We then present a whole team analysis for the Huskies. As for such senario, it is convenient for us to investigate the **team vector**, an eigenvector with key elements describing all players combined together. We list the basic elements and our estimation in Table.3:

Vector elements	Values	Definitions or Explanations	Estimation
Total Links	504	Sum of all ball passing events	Medium
Avg. Deg	15.857	Average Deg to all the vertices	Medium
Avg. Deg Centralization	1.220	Fraction of connected notes	Medium
Diameter	3	The maximum eccentricity	Good
Avg. Path Length	22.999	Average holding ball length	Good
Avg. Clustering Coefficient	0.707	mean fraction of exsiting triangles	Low

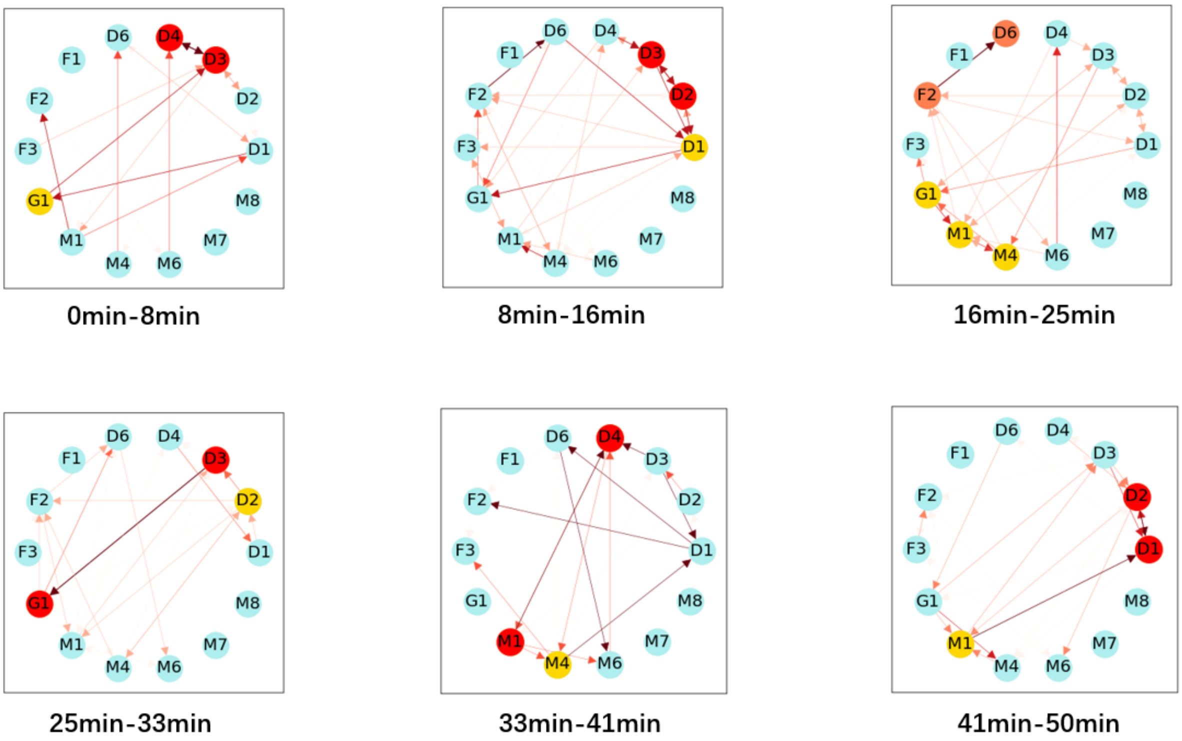
Table 3: Team vector elements and its estimation

Furthermore, we also abstract out the key cooperation mode when the Huskies makes competitions. In seeking dyadic and triadic configurations, we again look at Fig.3. For 2OBCP, they are **(F2,M1)** and **(M1,D1)** with the same sum weight. This two pairs form a passing route from the back field to the front field. From Fig.4, we can again check this conclusion because these three players all have large Pr value. M1 and F2 also have large Bc, Cc and Ec values. This can reveal that **(M1,F2)** pair is an ideal attacking pair. When M1 receives balls from D1, he can directly pass it to F2. When F2 gets the ball, he will quickly run into the goal area and kick the ball to get scores. For 3OBCP, it is **(D1,D2,D3)**, the defender pair. They form a strong defending structure which could minimize the scores they lose. However, our current model can only extract one team's configuration. We will tackle this problem in Section 6.

3.3 Dynamical Analysis of the Huskies

Dynamical analysis basicly means the **time dependent analysis**. We exam the team performance according to different duration of time, even keeping an eye on the whole competition season.

We first analyze the network variation according to different time period in the whole competition. The shift of the dyadic and triadic configurations can be extracted from the time dependent graph, indicating the team's changing strategies. In Fig.5, we show this phenomenon within time period 8 min in the first half in Match3.



0min-8min

8min-16min

16min-25min

25min-33min

33min-41min

41min-50min

Figure 5: Network time dependence in Match3. Orange and yellow nodes represent the 2OBCP and 3OBCP respectively, while the red nodes are for players both in these two catagories.

We can clearly see the differences between these six time periods. In the first 16 min, the Huskies' players focus extensively on defendance, fully using the 3OBCP **(D4,D3,G1)** and **(D3,D2,D1)** pairs. When entering the second 16 min, they struggle to start attackings to the opponents. The involved midfielders 2OBCP **(M1,M4)** and forward **F2** strongly demenstrate this point. During the last 16 min, they pay more attention to the midfield without losing great defendance. We name this time dependent strategy **'Defend-Attack-Preserve' (DAP)**. However, we ignore the strategies of the opponents which may also pose influences on the strategies and the final results. We will try to analyze this effect in Section 5.

This DAP strategy can also be viewed though time variant players' **relative positions**. With the data in Match3, we calculate the time average positions and make this analysis. The results are shown in Fig.6.

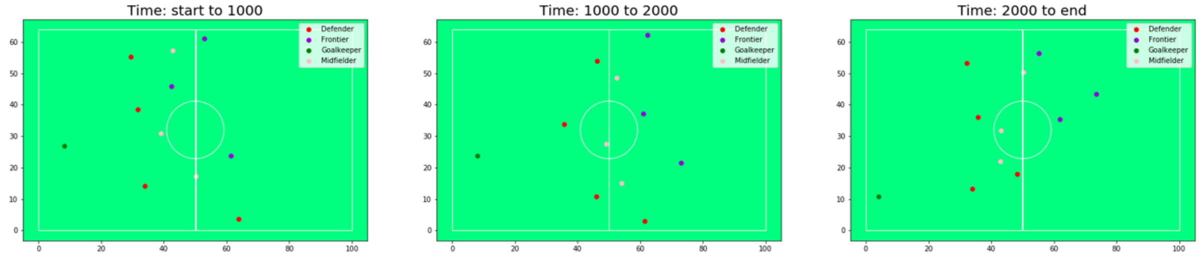


Figure 6: The relative positions in the first half of Match3 with time period 1000 s for each panel starting from the left. Red, purple, green and pink dots represent defender, frontier, goalkeeper and midfielder respectively.

Finally, we keep an eye on the time period extended to the **whole section**. Since in this season, the Huskies participates in 38 competitions. The winning percentage is about 34.2%, a relative low rate. We then plot the frequency of soccer position in Fig.7. It is obvious that the ball almost stays at the backfield, indicating the Huskies' **defects** in attacking abilities and cooperation skills.

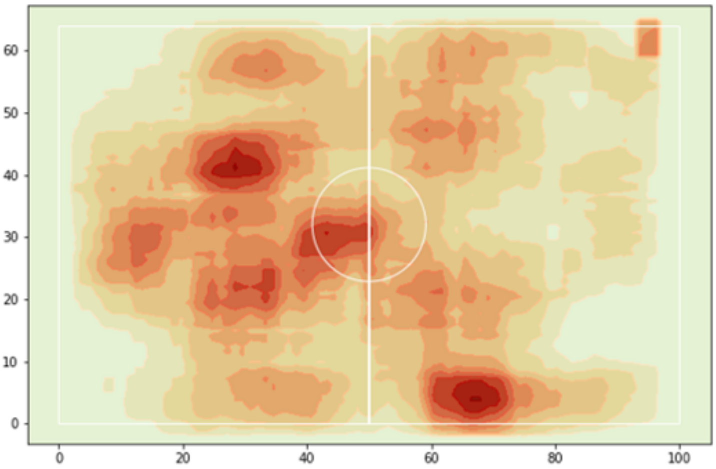


Figure 7: The frequency of soccer position during the whole season. Darker area indicates higher frequency.

4 The Competitive-Oriented Index Estimation System (COIES)

In this section, in order to quantify our model parameters, we set up a new Competitve-Oriented Index Estimation System (COIES) to present a reasonable analysis for **current team situation** and put forward our suggestions on **universal, long-term strategies**.

4.1 Quantitative Performance Indicators (QPI)

In this system, the first step is to identify the key performance indicators in certain teamwork. These indicators not only reflect the team members' personal and overall performance, but also capture the structural and even the dynamical aspects of the cooperation network.

In order to conduct the quantitative analysis, we give our clear definitions of **six Quantitative Performance Indicators (QPI) indices** as follows:

- Coordinate Coefficient: Weighted successful passes divided by losing ball events.
- Technical Flexibility: Average player contributions (contributions: sum of weighted scores for different event types).
- Contribution Difference: Standard deviation for player contributions.
- Holding Capability: Ave.Path Length times holding ball time.
- Attacking Capability: Total fraction of winning duel events over the losing ones.
- A.W.C. Coefficient: Average weighted Clustering Coefficient.

As an example, we analyze these indices within Match1 (winning 1:0) and Match13 (losing 1:4) to show its rationality. The results are summarized in Fig.8. In the left panel, although the opponent's attacking capability is more stronger than the Huskies, its other index value is relative small thus accounting for the final failure to the Huskies. In the right panel, most of the opponent's indices are higher than the Huskies'. This eventually leads to Huskies 1:4 losing.

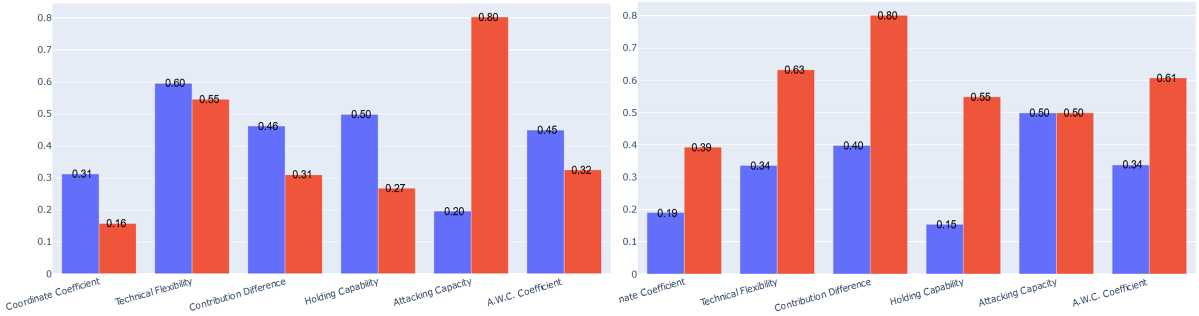


Figure 8: QPI indices for Match1 (left panel) and Match13 (right panel). Blue bars and Red bars denote the Huskies and the Oppenents respectively.

4.2 Modified Logistic Regression (MLR) for Weighting Coefficients

Currently, we need to quantify how important each QPI is in reflecting the overall strongness or weakness of a certain team. In this project we choose the Modified Logistic Regression (MLR) methods.

Before showing our detailed analysis, we first explain why MLR is good. Since we have three competition outcomes, namely win,tie and loss, we need to add a new 0.5 label to the standard 1 and 0 label. Although support vector machine (SVM) or softmax regression (SR) may achieve better regression results, the MLR additionally indicates **strong possibility meaning** which better measues how successful could a team be.

We first normalize each original QPI index P_i into range [0.05, 0.95] with fomula

$$Q_i = 0.9 \times \frac{P_i - min\{P_i\}}{max\{P_i\} - min\{P_i\}} + 0.05 \tag{6}$$

where Q_i is the normalized ith QPI index and $max\{P_i\}(min\{P_i\})$ represents the maximum(minimum) value of the original ones among all the teams. After this, we can construct the properties eigenvector for the Huskies $H^j \in \mathbb{R}^6$ and for Opponents $O^j \in \mathbb{R}^6$, $j = 1, 2, \cdots 38$. Here, j is the match number and O^j is the opponent to the Huskies in match j.

Then we construct the difference eigenvector (DEV) $X^j \in \mathbb{R}^6$which describes the overall performance differences between the Huskies and the Opponents.

$$X^j = H^j - O^j \tag{7}$$

According to logistic regression, we assume a function $h_\theta(X^j)$

$$h_\theta(X^j) = \frac{1}{1 + \exp(-\theta^T X^j)} \tag{8}$$

and the loss function $J(\theta)$

$$J(\theta) = \frac{1}{2} \sum_j (h_\theta(X^j) - y^j)^2 + \lambda \sum_{k=1}^{n} \theta_k^2 \tag{9}$$

where y^j is the related label function and λ is the regularization constant. Here we have a added complexity penalty factor to avoid overfitting. We use the update rule

$$\theta_i := \theta_i - \alpha \frac{\partial}{\partial \theta_i} J(\theta) \tag{10}$$

Since we have relative small data sets, it is crucial for us to add a process called 5-Fold Cross Validation (5FCV). This is a widely applied method in machine learning (see [13, 14] for more details). We also stress that when testing our models, we attribute tie equal to loss for simplicity. Our model accuracy is defined as the fraction of correct predicted competitions over the whole cases.

4.3 System Evaluation

Based on COIES system, we give out our evaluation to the Huskies teamwork . In Table.4, we present the **fitting coeffecient** for each QPI index which could reflect how important each index is in the overall perforance and the competition results.

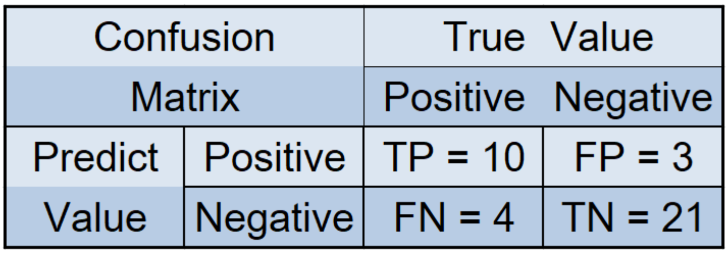
The Contribution Difference weighs most, followed by A.W.C Coefficient, Coordinate Coefficient, Technical Flexibility, Attacking Capacity and Holding Capability. It fits our intuition that all variables related to cooperation are very **crucia**l while those regarding personal abilities weighs much less. Additionally, there is also an anomaly. It seems that the smaller the Contribution Difference is, the better a team will be. However, that's not true. This anomaly comes from **different positions** the player will be. If we have very tiny difference in contribution,

QPI index	Fitting Coeffecient (Weight)
Coordinate Coefficient	0.81
Technical Flexibility	0.76
Contribution Difference	0.99
Holding Capability	0.29
Attacking Capacity	0.38
A.W.C. Coefficient	0.82

Table 4: Fitting coeffecient for QPI indices. Larger weight indicates greater importance.

that means all the players are the same without forming real team configurations. However, in reality, this will not happen. For the whole team, they need to take advantage of some superstars, arranging other players to corperate with him to make more scores. Of course, this difference can not be so larger to destroy the suitable cooperation structure. In a word, the Contribution Difference should be **suitable** for the team.

We then use this system to judge how successful the Huskies could be. From the MLR's 5FCV, we present our results in Fig.9 ,the confusion matrix. Our model's **prediction accuary** is 81.6%, and the **overall estimation score** is 76.6. Details are shown in shown in Fig.10.



Confusion		True Value	
Matrix		Positive	Negative
Predict	Positive	TP = 10	FP = 3
Value	Negative	FN = 4	TN = 21

Figure 9: Confusion matrix. Positive (negative) predict value is for win(lose) prediction. True values are the real competion results.

We extract the three strong feature QPI indices from Table.4 and draw a plane in this 3D parameter space. According to the MLR analysis, if the related three indices in X^i falls above this plane, we predict that the Huskies will win,vice versa. We then plot the actural competition results in colored dots. Even though we have some errors in the judgement, we can clearly see that the competition level of the Huskies is **a bit below the average**. Also, almost all the teams that win the Huskies have relatively higher Coordinate Coefficient and Contribution Difference. This conclusion indicates how the Huskies can improve themselves. We detail this in Section 4.4.

4.4 Adding Hyperplanes in COIES

In this section, we give some **quantitative advice** for the coach of the Huskies on how to improve team success before the next season. We reduce this problem to find the **minimum length** from a certain point to fixed hyperplanes in $\mathbb{R}^6$. To preserve the rationality, we make another assumption that the cost to improve any QPI index one percentage up is the same.

With the fitting coefficients given by MLR, we define a brand new QPI combined variable —— **overall performance indicator (OPI)** ψ in Eq.(11), where Q_i has been defined in Eq.(6)

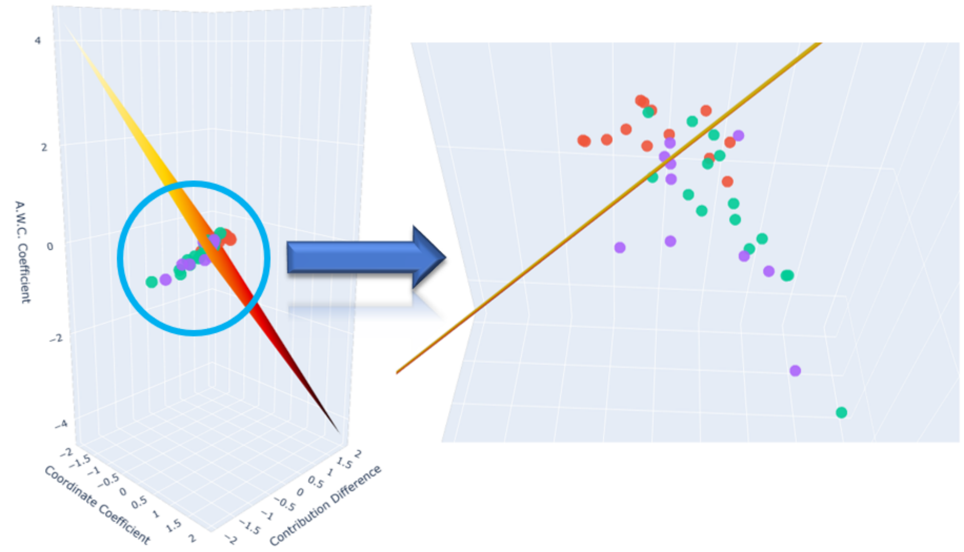


Figure 10: The three strong feature (Coordinate Coefficient, Contribution Difference and A.W.C Coefficient) of QPI indices and our prediction plane. Red, green and purple dots are for the Huskies' win, loss and tie competition results in reality. The right panel shows the enlarged part in the circle.

and ω_i are the related fitting coefficients shown in Table.4.

$$\psi = \sum_{i=1}^{6} \omega_i Q_i \tag{11}$$

For a fixed ψ value, we can define a hyperplane Σ_ψ with normal vector in this 6D parameter space to be $(\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6)$. For the Huskies, the hyperplane equation can be writen as $\psi_H = \sum_{i=1}^{6} \omega_i h_i$. (Since we are considering the DEV in Fig.10, , the related hyperplane equation is $\sum_{i=1}^{6} \omega_i x_i = 0$. The 3D projection has been shown out.)

Now, we want to improve the OPI index of the Huskies to reach the **average level**. In Table.5, we show the current OPI value. Since we all know that the shortest distance from a point to a hyperplane is the length of vertical line segment. For intuitive view, we show the 3D case in Fig.11. As in our situation, we need to consider similar things in $\mathbb{R}^6$.

Team	Ave.ψ	Std.ψ
Huskies	1.58	0.36
Ave.Opponents	1.89	0.56

Table 5: Current OPI value for the Huskies and the Opponents average. Ave.ψ and Std.ψ represent the average and standard deviation for all related competitions.

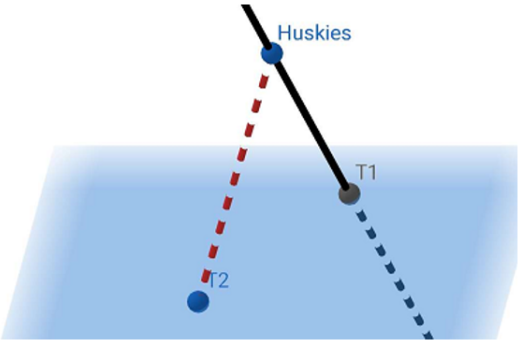


Figure 11: Schematic diagram for the 3D shortest distance from a point to a plane. T1 is the foot point and T2 is another random point in that plane.

After deriving the vertical line equation to the hyperplane in Eq.(12), we can easily get the length of vertical line segment. Its axis projection could give us the actual **percentage** the Huskies need to improve before the next season. We summarize our results in Table.6 and Table.7

$$\frac{(h_1 - H_1)}{\omega_1} = ... = \frac{(h_6 - H_6)}{\omega_6} \tag{12}$$

QPI Index	Percentage of Improvements
Coordinate Coefficient	39.2%
Technical Flexibility	15.3%
Contribution Difference	21.1%
Holding Capability	7.30%
Attacking Capacity	8.00%
A.W.C. Coefficient	24.0%

Table 6: Percentage of Improvements for the Huskies to reach average OPI level.

QPI Index	Percentage of Improvements
Coordinate Coefficient	53.1%
Technical Flexibility	20.7%
Contribution Difference	28.6%
Holding Capability	9.91%
Attacking Capacity	10.8%
A.W.C. Coefficient	32.6%

Table 7: Percentage of Improvements for the Huskies to reach OPI $\psi = 2$.

From long-term viewpoint, the Huskies have to **imporve their cooperation by strenghening the structure of 2OBPC and 3OBCP** to a large extent as well as **preserve the proper differences through cultivating more talented players and superstars**. Meanwhile, they do not need to care too much about the personal holding ball and attacking abilities. We will summarize this in a letter to the coach in our last section.

5 The Optimal Passing Route Model (OPRM)

In this section, we further construct a more concrete and useful model based on the properties we get before. This OPRM serves as a modification to the previous BPGM and COIES for not only preserving their advantages but also reveal more details inside the team cooperation. We also give out an **on-field strategy** —— the best ball passing route for competing with **certain** opponents in the next season.

5.1 The Mathematical analysis of OPRM

We create this nice model based on the classical reinforcement learning (RL) method. The **Markov decision processes (MDP)** provide the formalism in which RL problems are usually posed [15, 16].

- **Senario**
  MDP provides a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker. That suits our soccer competition background very well.
- **Tuple** $(S, A, \{P(s, a)\}, \gamma, R)$
  - S is a set of states. Its elements consist of the players on the field and the two goals which describe where the ball belongs. This is quite different from the previous work based on pitch [17, 18].
  - A is a set of actions. It describes all the passes between two states (players) as well as the shooting action.
  - $P(s, a)$ is the state transition probability which is crucial to determine ball passing direction.
  - $\gamma = 0.95$ is the discount factor to guarantee the algorithm's convergence.
  - $R : S \times A \mapsto \mathbb{R}$ is the reward function which reflects the contribution of certain type of ball passing.
- **Policy**
  A policy is any function $\pi : S \mapsto A$ mapping from the states to the actions. In our case, that is one of the possible ball pass routes.

Here we need to specify the formula of state transition probability $P(s, a)$. First, we create the whole ball passing network for all players on the field in Fig.12. Then we get a neibouring matrix $A \in \mathbb{R}^{(n+2)\times(n+2)}$ which is the weighted adjacent matrix. n is the total amount of players on the field. Based on the analysis in Section 4, we construct a player capability vector $F \in \mathbb{R}^{1\times n}$. With the help of two other parameters β the shooting factor and θ the probability factor, we define $P(s, a)$ as

$$Q(s, a) = (\psi \cdot A_s,\ F,\ [0\ \cdots\ 0\ \beta_s\ \ -\beta_s], 1(\pi(s) = s')) \cdot (\omega_1, \omega_2, \omega_3, \theta)^T + \Lambda \tag{13}$$

$$P(s, a) = \frac{Q(s, a)}{\sum\limits_{a_i} Q(s, a_i)} \tag{14}$$

where A_s is the sth line in matrix A. ψ has been defined in Eq.11 which shows our consideration of unit performance of a team. Since morale encouraged by **team strength** affects personal performance to some extent, this term here is reasonable. Also, it is obvious that factors related to players' **personal ability** and **different roles**(G,D,M,F) have been taken into account. Besides, we define Λ, a random matrix obeying standard normal distribution, which takes **random noise** like weather, injurys into consideration. The term $1(\pi(s) = s')$ bases on the fact that clear **passing target** has an impact on some, if not all, transition matrix. $\omega_1, \omega_2, \omega_3$ is the weight factor for different actions.

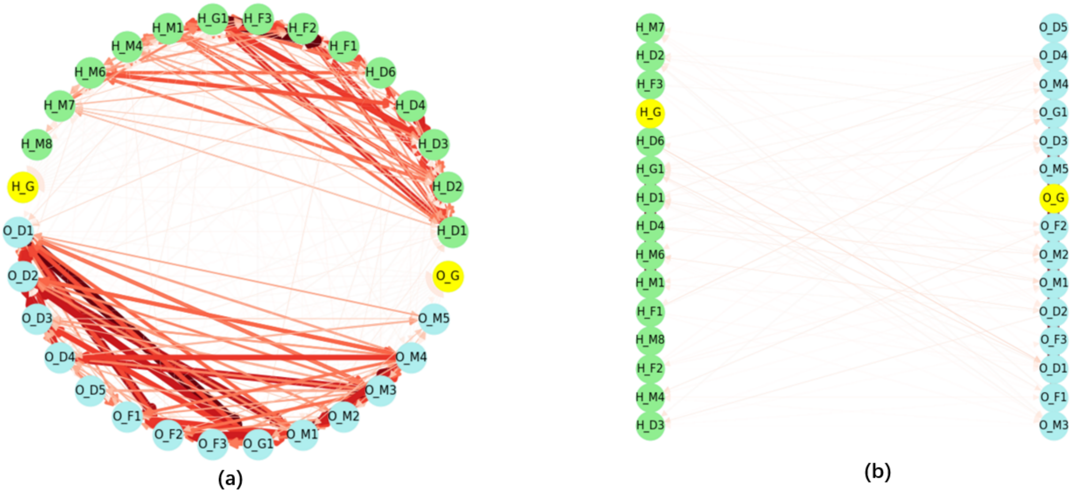


Figure 12: The whole ball passing network. (a) Passes between different states. (b) Passes between different teams.

5.2 The Process for Optimal Searching

In order to search for the best ball pass route, we define the value function $V^{\pi}(s,a)$for a policy π according to

$$V^{\pi}(s,a) = E[R(s_0,a_0) + \gamma R(s_1,a_1) + \gamma^2 R(s_2,a_2) + ...|s_0 = s, \pi] \tag{15}$$

For a given policy, V^{π} satisfies the **Bellman equation**:

$$V^{\pi}(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s',a') V^{\pi}(s',a') \tag{16}$$

Then we define the optimal value function according to

$$V^{*}(s,a) = \underset{\pi}{max}\, V^{\pi}(s,a) \tag{17}$$

To optimize the value function, we adopt the value iteration method.

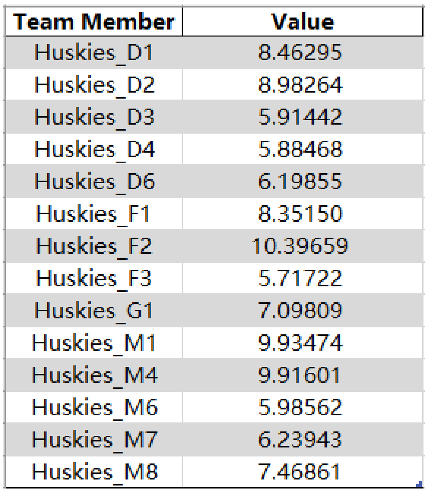
- For each state s, and action a initialize $V(s,a)$:=0.
- Repeat until convergence{

  For every state, update $V(s,a) := R(s,a) + max_{a' \in A} \gamma \sum_{s'} P(s',a')V(s',a')$}

5.3 Route Visualization

As an example, we conduct our above analysis to Opponent3 and give out the **optimal value** for each player in the Huskies. We show this in Fig.13

Aimed at achieving excellent visual effect, we also plot the recommanded ball pass routes in a real soccer field. From Fig.14, we can easily see that there are several routes to pass the ball. For example, if D6 gets the ball, the best choice for him is to pass it to D1 immediately. Then D1 tries to give a high kick, directly passing the ball to **F2**. Finally, F2 breaks the opponent's defendance and starts shooting. Through this way, without any improvement as we mentioned in Section 4, the Huskies have an on-field strategies to win the game. In a word, this OPRM analysis is a modern method for making concrete strategies before or within the competition.



Team Member	Value
Huskies_D1	8.46295
Huskies_D2	8.98264
Huskies_D3	5.91442
Huskies_D4	5.88468
Huskies_D6	6.19855
Huskies_F1	8.35150
Huskies_F2	10.39659
Huskies_F3	5.71722
Huskies_G1	7.09809
Huskies_M1	9.93474
Huskies_M4	9.91601
Huskies_M6	5.98562
Huskies_M7	6.23943
Huskies_M8	7.46861

Figure 13: The optimal value for the Huskies.

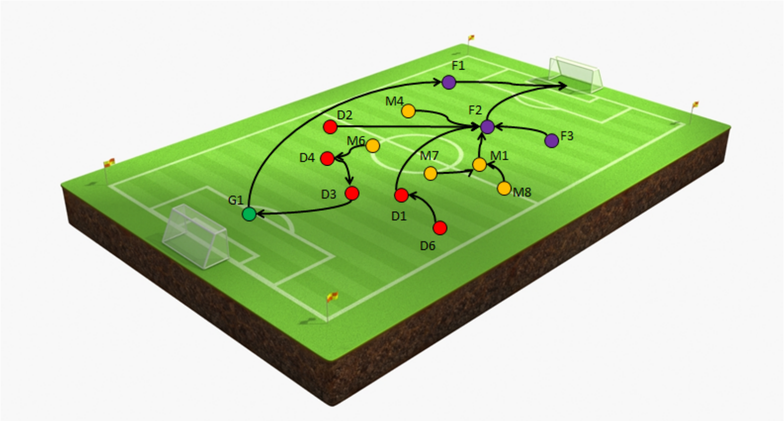


Figure 14: The recommanded ball pass routes for the Huskies. Green, red, yellow, and purple colors are for G,D,M, and F players respectively.

6 Sensitive Analysis

Sensitive analysis is quite important in mathematical modeling. Not only can it examine our predictions, but also it helps us make modifications to the current model. In this section, we focus on the sensitivity of OBCP identification and optimal ball passing route.

6.1 OBCP identification Sensitivity

In the end of section 3.2, we have mentioned the problem of failing to extract more team configurations. Now, through making some technical imporvements in our OBCP definition, we can greatly improve the model identification sensitivity.

- All the subgraphs $G[V_i]$ with sum weight larger than threshold f are called iOBCP.

We then compare the 3OBCP capture difference between the definitions within Match3. The results are shown in Fig.15. Obviously, this improvement works well. We successfully identify three 3OBCP at last, which is quite helpful to our future analysis.

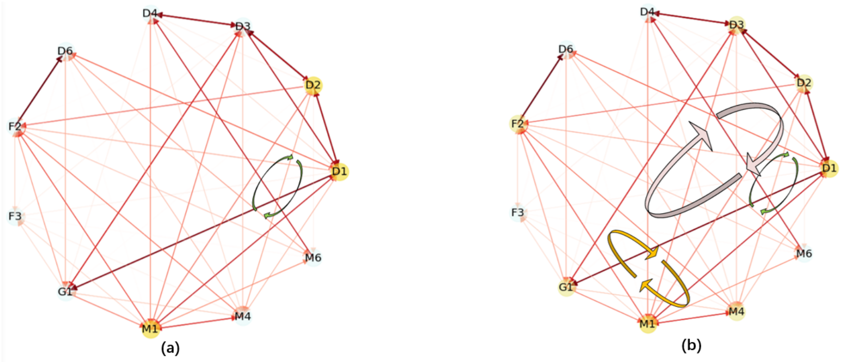


Figure 15: Two 3OBCP results. (a) One 3OBCP from previous definition (b) Three 3OBCP from new definition $f = 32$.

6.2 Optimal Ball Passing Route Sensitivity

In order to realize our recommanded route in Fig.14, F2 plays the dominant role because of its highest popularity among the other players. However, in the calculation, we have some parameters without strict definitions, like **the weight of players' skills**. If the value function of F2 is sensetive to this parameter, then this optimal route may break down. Luckily, our model is relative stable at this point. The results are shown in Fig.16. We can see that the value functions of all these three players are **stable** enough in range $0.175 \sim 0.300$. That demonstrates our model's good stability.

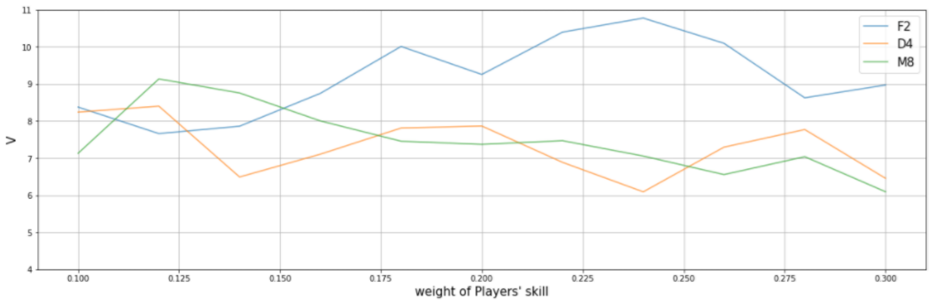


Figure 16: The dependence of value function with the weight of players' skills. Blue, orange, and green denote F2, D4, and M8 respectively. We choose 0.225 in the previous calculation.

7 Strength and Weakness

In this section, we briefly summarize the strength and weakness in our model.

- **Strength**
  - We make full use of the graph theory to measure many meaningful parameters, such as centrality and clustering coefficient. This work not only provides the mathematical foundations of ball passing networks, but also enables us to extract useful information for future analysis.
  - The logistic regression analysis checked by 5FCV enables us to calculate the fitting coefficients based on a small data sets without losing much accuracy.
  - The OPRM based on Markov process simplifies the competition into what we can calculate by computers. The optimal passing route for a certain game is quite important for the soccer teams.
- **Weakness**
  - In the passing network, we do not consider the additional effects of head pass and high pass etc.
  - Since logistic regression is linear, we may achieve better results through non-linear methods.
  - The Markov process assumes the ball passing strategies are relatively stable, which leads to lack of universality. Also, we cannot make predictions on those who are not on the field.

8 Generalized Models for Improving Team Efficiency

In this section, we take a first step to generalize our findings into social background. Since most sectors in our societies increasingly rely on cooperation work, such as food production, car manufacturing, policy development and business management, it is necessary for us to create relatively universal strategies to help the organization leaders make decisions and improve the team efficiency.

The most important findings of our model is that, as for the team work, people's **mutual cooperation** is far more crucial than the personal abilities. But that does not mean every person plays the same role. Conversely, we need to cultivate different kinds of people to take various duties, preserving the **proper differences** in their positions. This conclusion can be fully revealed from Section 4.

As an example, we explore the business management industry. In large companies, the relation between different sectors are very similar to Fig.17. The ultimate goal of such team work is to quickly react to the changes in the market. Thus, our generalized findings guarantee the following **'R-T-O' procedure**. Research committees should conduct their investigations before the others, analyzing the current situation before asking managing committees to make decisions. After research, the decision process should be very fast, leaving much time for operation committees to make the actual plans. Another 'R-T-O' key feature is that any sectors cannot intervene the others decision in order to preserve the differences we have mentioned.

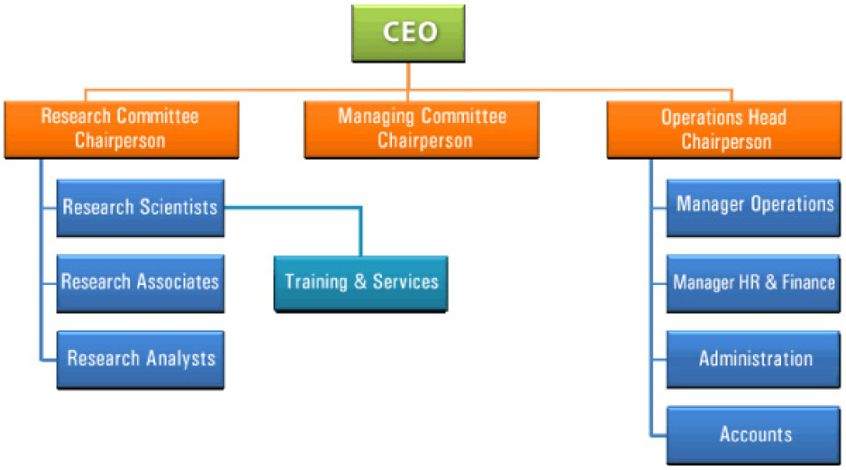


Figure 17: Company structure

However, we still lack some charateristics like the salary for different committees, the CEO's power in decision making and the general friendship between these three sectors. These are the **new aspects** of teamwork that need to be captured to further test and complete our models.

9 A Letter to the coach of the Huskies

Dear distinguished coach,

Thank you for your invitation. It is an honor for us to do this project. In this email, we will briefly summarize your team's current situation based on our model, and then provide concrete suggestions.

Among the 19 teams in the season, your team's level is a bit lower than the average. In order to make a big improvement, your need to greatly enhance the players' coordinate abilities and cultivate the more talented athletes rather than keep on training their average skills. A superstar may be more valuable for this team.

Defenders basically perform very well. They have strong keeping and passing ball capabilities. As for the forwarders, their related skills are not good enough, which is a possible bottleneck for your team to upgrade to a higher level. It may be a good idea to spend some time to train them alone rather than in the whole team.

As for a specific competition, we analyze the match between your team and Opponent3. If you meet them again, we recommand you to make full use of F2. The detailed information can be checked in Section 5 of this paper.

We all wish your team perform better in the next season.

Sincerely yours,

Your friends in ICM

References

[1] Donati A V, Montemanni R, Casagrande N, et al. Time dependent vehicle routing problem with a multi ant colony system[J]. European journal of operational research, 2008, 185(3): 1174-1191.

[2] Lee J D, Kantowitz B H. Network analysis of information flows to integrate in-vehicle information systems[J]. International journal of vehicle information and communication systems, 2005, 1(1-2): 24-43.

[3] Xie, Yuanchang, Dominique Lord, and Yunlong Zhang. "Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis." Accident Analysis & Prevention 39.5 (2007): 922-933.

[4] Ceder, A., & Wilson, N. H. (1986). Bus network design. Transportation Research Part B: Methodological, 20(4), 331-344.

[5] Israeli, Y., & Ceder, A. (1995). Transit route design using scheduling and multiobjective programming techniques. In Computer-aided transit scheduling (pp. 56-75). Springer, Berlin, Heidelberg.

[6] Leskovec, Jure, Kevin J. Lang, and Michael Mahoney. "Empirical comparison of algorithms for network community detection." Proceedings of the 19th international conference on World wide web. 2010.

[7] Šubelj, Lovro, and Marko Bajec. "Robust network community detection using balanced propagation." The European Physical Journal B 81.3 (2011): 353-362.

[8] Lusher, Dean, Garry Robins, and Peter Kremer. "The application of social network analysis to team sports." Measurement in physical education and exercise science 14.4 (2010): 211-224.

[9] Johnson, Thor. "Onsite fantasy sports game using onsite and network-based data collection and processing." U.S. Patent Application No. 11/091,197.

[10] Buldú, Javier M., et al. "Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game." Frontiers in psychology 9 (2018): 1900.

[11] Xing, Wenpu, and Ali Ghorbani. "Weighted pagerank algorithm." Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.. IEEE, 2004.

[12] Bihari, Anand, and Manoj Kumar Pandia. "Eigenvector centrality and its application in research professionals' relationship network." 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE). IEEE, 2015.

[13] McLachlan, Geoffrey J.; Do, Kim-Anh; Ambroise, Christophe (2004). Analyzing microarray gene expression data. Wiley.

[14] "Elements of Statistical Learning: data mining, inference, and prediction. 2nd Edition". web.stanford.edu. Retrieved 2019-04-04.

[15] Filar, Jerzy, and Koos Vrieze. Competitive Markov decision processes. Springer Science & Business Media, 2012.

[16] Puterman, Martin L. Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons, 2014.

[17] Cheung, Ricky. "Stochastic based football simulation using data." (2018).

[18] Damour, Gabriel, and Philip Lang. "Modelling Football as a Markov Process: Estimating transition probabilities through regression analysis and investigating it’s application to live betting markets." (2015).

10 Appendix

10.1 Sum Weight Subgraph

Part of Sum Weight Subgraph Source:

```
# The function to calculate the max weights subgraph of 3 vertexs.
# To fit the model, any of the three edge weights
# should be at least fifth of total weight
# G: the origin graph, which should be networkx’s graph instance.
def get_maxweight_subgraph(G):
    n = 3
    get_weight = lambda n1, n2: \
        ((float(G.edges[(n1, n2)][’weight’]) if (n1, n2) in G.edges else 0) \
        + (float(G.edges[(n2, n1)][’weight’]) if (n2, n1) in G.edges else 0))
    comb = []
    weights = []
    for c in combinations(G.nodes, 3):
        # calculate the total weight
        weight = get_weight(c[0], c[1]) + get_weight(c[0], c[2]) \
                 + get_weight(c[1], c[2])
        if np.min([get_weight(c[0], c[1]), get_weight(c[0], c[2]),
                   get_weight(c[1], c[2])]) > weight/5:
            comb.append(c)
            weights.append(weight)
    weights = np.array(weights)
    comb = np.array(comb)
    comb = comb[weights.argsort()[::−1]]
    weights.sort()
    weights = weights[::−1]
    return np.c_[comb, weights]
```

10.2 Modified Logistic Regression

Part of Logistic Regression Source:

```
# Part of logistic regression
# 5−Fold Cross Validation
import gc
from sklearn.model_selection import KFold,StratifiedKFold
seeds = [3, 11 ,12, 1024, 322, 2048, 1564]
num_model_seed = 5
score = []
coef = []
```

```
X_train = dif.iloc[:,1:-1]
y = dif.iloc[:,-1]
for model_seed in range(num_model_seed):
    print(model_seed + 1)
    skf = StratifiedKFold(n_splits=5,
                          random_state=seeds[model_seed], shuffle=True)
    for index, (train_index, test_index) in enumerate(skf.split(X_train, y)):
        train_x, test_x, train_y, test_y = X_train.iloc[train_index], \
                X_train.iloc[test_index], y.iloc[train_index], y.iloc[test_index]
        gc.collect()
        model = LogisticRegression()
        model.fit(train_x, train_y)
        score.append(model.score(test_x, test_y))
        coef.append(model.coef_)
        gc.collect()
```

10.3 Markov Desicion Procession

Part of Markov Desicion Procession Source:

```
# Part of Markov Decision Procession
# nodes: the nodes(all players as well as two gates)
while True:
    for i in range(len(nodes)):
        if nodes[i][0] == ’G’:
            V[i] = reward.iloc[:-2, -2].max()
        else:
            max_value = -np.inf
            max_dest = -1
            max_policy = None
            if nodes[i][0] == ’H’:
                for j in range(len(nodes)):
                    p = p_mat[i].copy()
                    p[j] += w4
                    cur_value = (p * V).sum()
                    if max_value < cur_value:
                        max_value = cur_value
                        max_dest = j
                        max_policy = j
            else:
                for j in range(len(nodes)):
                    p = p_mat[i].copy()
                    p[j] += w4
                    cur_value = (p * V).sum()
                    if max_value < cur_value:
                        max_value = cur_value
                        max_dest = j
                        max_policy = j

            V[i] = reward_dict[(nodes[i], nodes[max_dest])] +\
                   gamma*max_value
            policy[i] = max_policy

    if np.sum(np.square(lastV-V)) < e:
        break
    lastV = V.copy()
```