Problem Chosen
D

2020
MCM/ICM
Summary Sheet

Team Control Number
2021062

In this paper, we utilize Network Sciences tools and machine leanring technique to study the complex interactions between football players.

First, we build up football passing networks where nodes are football players and directed edges represent football paths. We assign the weight of each edge after considering both the number of passes along this edge and the distance between two players. Then the average passing network of matches in the whole season is constructed. After cleaning up the network, we identified a heptagon-like array in Huskies' network. We further extracted and analyzed 2-nodes and 3-nodes motifs in that network. Mutually-connected motifs are of great importance in Huskies' passing networks.

Second, we assess the performance from the following aspects. The number of passes, shots, and goals reflect an overall performance of a team. Huskies has significantly less passes and shots than its opponents. This might explain why Huskies ranks 8-th among all the teams in the last season. The match can also be described by some spatial properties, e.g., centroid of players, their position dispersion, and advance ratio of passes. On average, Huskies players are farther from the opponent's goal, compare with other teams. We then utilize six network parameters which are broadly used in graph theory to indicate the clustering and connectivity of a graph. Comparing with other top teams, we find that Huskies' passing network is less clustered but well-connected.

Third, we define several team-level properties such as adaptability, flexibility, and tempo. The adaptability reflects how well the team is under different circumstances, such as the match location (home or away) and coaches. We find that Huskies performs much better in home-held matches and in those matches advised by Coach2. Flexibility refers to a team's mobility and the response time to its opponent. With the help of temporal passing networks, we studied one winning game from various aspects (response time t_{50}, spatial metrics, and network parameters) and concluded that a highly-clustered, well-connected, highly flexible team has greater chance to be success.

Fourth, taking the both dynamical and structural characteristics of a team into account, we build a model called H-indicator which could be used to predict the outcome of a game. Using the whole dataset from five European national soccer competitions and 2018 World Cup, we tested the validity of H-indicator on predicting the outcomes. It turns out that H-indicator is a robust indicator: the top team has larger H value on average. However, the accuracy of using H-indicator to predict Huskies' outcome is low. On top of that, we turn to machine learning and train a classifier with H-indicator and other dynamical and structural properties. The best accuracy we achieved is about 50%. We propose the possibility of using neural networks to improve.

Fifth, we further discuss the analysis above and summarized our findings on football teamwork. Effective strategies consist of a low averaged shortest path length, high clustering coefficient, high λ_1 and $\tilde{\lambda_2}$, high level of flexibility, high concentration, and high advance ratio.

Finally, our research on football teams shed lights on understanding group dynamics and other complex teamwork. We proposed several ways to help design an effective and collaborative team: creating more exchanges of information, developing adequate sub-groups, being active in communication, having a clear division of roles in a team, constructing supervision system, etc.

Contents

1 Introduction

1.1 Statement of the Problem

It is important to figure out the dynamics of a team to achieve more than simple sum-up of the abilities of individual members, which is also the case in football. We got data detailing match information of a foot ball team, Huskies. With the data, we need to quantify and formalize the structural and dynamical features of Huskies, extract strategies that have been successful and provide suggestions on improvements in the next season. On top of that, we need to extend our findings to more generalized organizations of society for designing more effective teams.

1.2 Data

The provided data consists 38 matches and all events in each match. Huskies played with 19 opponents and played twice with each opponent. There are 30 Huskies players appeared in all matches. We fully utilized the provided data in the analysis below. As stated in the problem, the dataset was processed from a larger dataset from five European national soccer competitions and 2018 World Cup. We retrieved the whole dataset from Pappalardo et al. (2019)[1].

The whole dataset consists of seven sub-datasets which are `Coaches`, `Competitions`, `Events`, `Matches`, `Players`, `Referees`, and `Teams`. The dataset include seven competitions matches record of England, European Championship, France, Germany, Italy, Spain, and the World Cup. We observe from the data that Huskies actually correspond to the team Everton in England league. We use scores and `opponent_id` to identify which team that opponent should be. Next, according to the sub-dataset `matches_England` which records the matches result, we work out a ranking of each team. Thus, with the help of the whole dataset, we have the ranking of Huskies (8-th) as well as the ranking of its 19 opponents in this season.

We analyze the network of other teams in Section 5 as comparisons. To do this, we need to deal with the sub-dataset `Events` in the same way that the ICM committee get the `passingevents` from `fullevents`. For example, an item A in `fullevents` records an event that *player_a* passed the ball from (x_1, y_1) to (x_2, y_2). If the next item B in `fullevents` also come up from *player_b* and *player_b* is in the same team as *player_a*, we identify *player_b* to be the target player item A. After doing so, we parse the raw dataset and build networks for other teams.

In this paper, we study the complex interactions between football players by quantifying the structural and dynamical features of the matches and the team, discuss effective strategies, and provide constructive suggestions to the coaches of Huskies team. The passing networks are constructed and investigated in Section 2. In Section 3, we study important indicators to assess the performance of football teams. We further analyze some team level processes in Section 4. A model comprising the above indicators is built in Section 5, and we use machine learning techniques to test the credibility of our model. As a result of these comprehensive analysis, we propose several suggestions towards the Huskies team to make it more competitive in the next season.

[1] https://figshare.com/collections/Soccer_match_event_dataset/4415000/2

2 Passing Networks

Network Science serves as a very helpful tool in the study of complex interactions between identities, which are quite typical in social sciences. In the study of football games, we create a football passing network (Buldú et al., 2019) and study how we can extract playing styles and strategies from the the network properties. The passing network is constructed from the data of passes, where nodes are players and directed edges are passes from one player to another. This directed network can be weighted.

Weights of edges are of great importance to the study of networks, since network parameters directly depends on weighting scheme of the network. Weight represents the significance of an edge. In our context, weight should be proportional to the passes between two players (as in Buldú et al. 2019). However, large distance between two players makes it harder to pass the football. Thus, we take these two factors into account, and assign the weight to an edge $A \longrightarrow B$ to be:

$$w = N_{A\to B}/\sqrt{d_{A\to B}},$$

where $N_{A\to B}$ is the passes and $d_{A\to B}$ is the distance between player A and B. The square root of distance makes it not a dominant factor in weighting scheme. We tested many other weighting schemes and find no significant differences regarding the results of this paper.

In the following, we construct the average passing networks, analyze both the macro and micro structure of the networks, and investigate time evolution of the networks using 50-passes network. In this paper, We construct networks with an open-source Python package NetworkX (Hagberg et al., 2008).

2.1 Average Passing Networks

To fully understand the networks formed by team players, we first take a **macro** view at the holistic networks formed by all players. We develop a graphic representation of the plays of a specific match (as well as the whole season). For example, Figure 1 shows the network structures that Huskies and its opponent developed in match 1 (Huskies won) with information compiled in different distributions to display features of the match. Nodes in the figures stand for players of the match. Node position is determined by the player's average field position and node number refers to the player's ID number. Nodes are color-coded by the relative size of the value which grows exponentially with the player's betweenness centrality. And the sizes of the nodes is determined by the play's eigenvector centrality. The width of the edges (arcs) reflects number of passes successfully completed between two players.

We could already identify several prominent local structures in the networks. From Figure 1, it's clear that Huskies has more passes than its opponents, since the edges in Huskies' network is much thicker.

Then we look into the long-term structure of Huskies' network. In Figure 2, we present the averaged structure of Huskies during the whole season. As we can see in the graph, for the past matches Huskies has adopted the strategy in a **heptagonally** distributed array. To illustrate the structure more clearly, we combine several nodes that are mostly in the same position (i.e. combine Huskies_D9 with Huskies_D10, combine Huskies_D8 with Huskies_D7, combine Huskies_D1, Huskies_D2, and Huskies_D3 to Huskies_D1, etc.). The heptagonally array becomes very prominent after combining nodes. The goalkeeper has very tight connection with D1, while D1 usually passes the ball to D7 and D4. D6, D9, F6, and M12 assist players in the midfield.

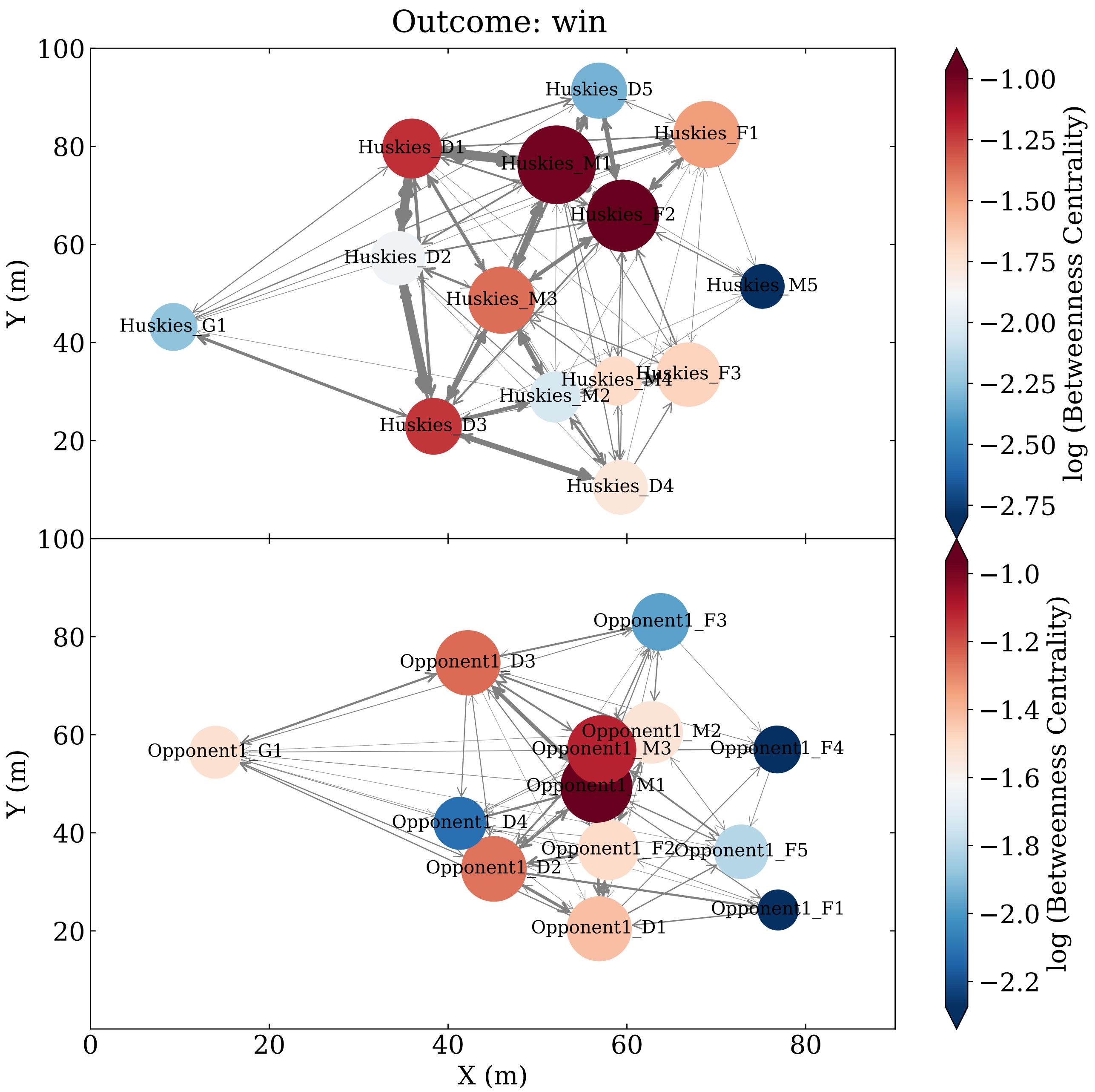

Figure 1: Average passing network of the two teams in Match 1 (Huskies won). Nodes are color-coded by the relative size of the value which grows exponentially with the player's betweenness centrality. The sizes of the nodes is determined by the play's eigenvector centrality. The width of the edges (arcs) is proportional to the number of passes successfully completed between two players.

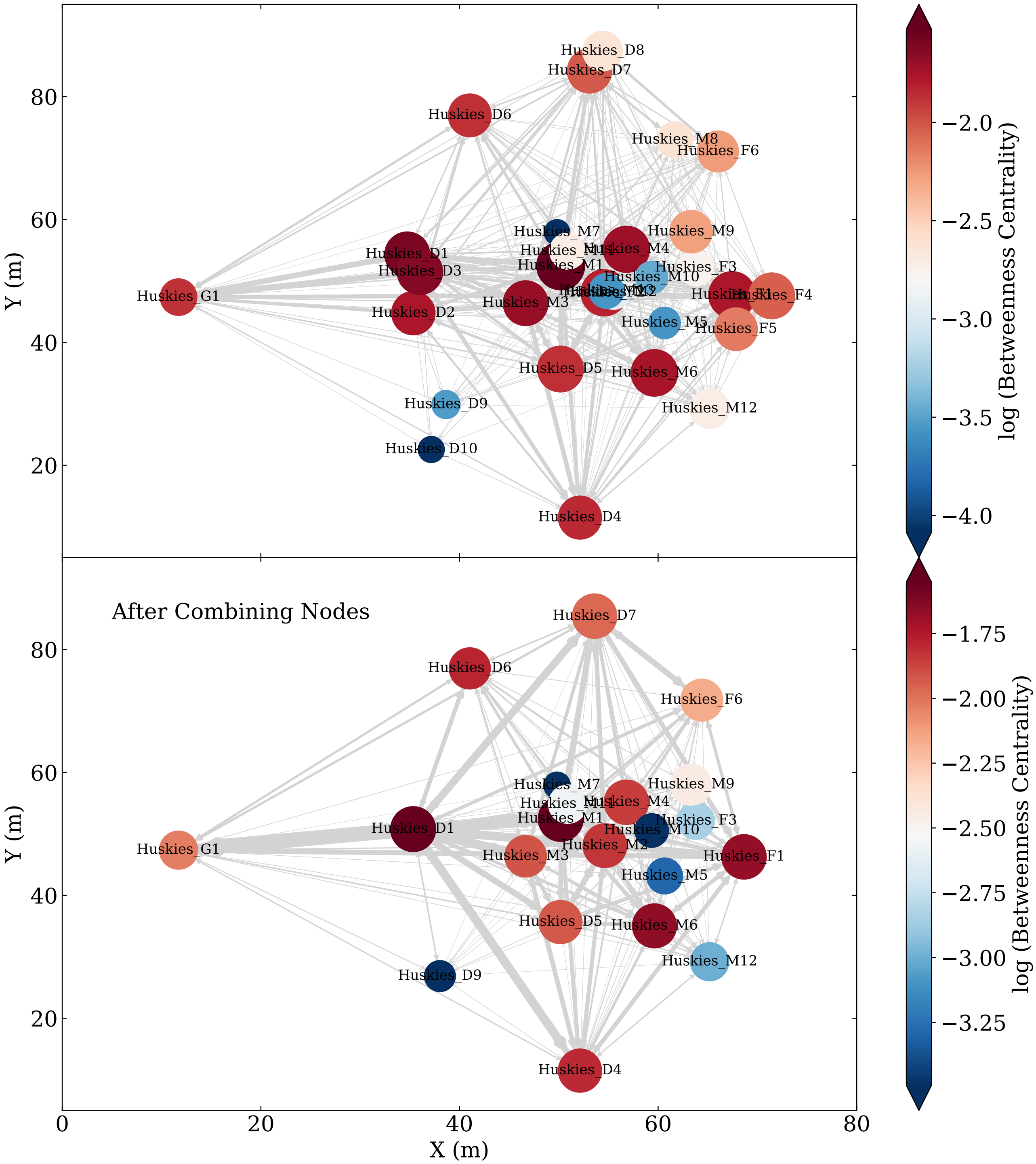

Figure 2: Average passing network of Huskies in the whole season. Huskies has adopted the strategy in a *heptagonally* distributed array, which is prominent after combining several pairs of nodes with the same average positions.

2.2 Motifs in Passing Networks

Motif refers to the statistically important local structure (sub-graph) of a network, typically the structure between several nodes. Motifs are building blocks of networks. They represent specific patterns of a network, hence inform us about the characteristic and function of a network. By studying the motifs in our passing networks, we will be able to understand the corporation style of the teams. We study both dyadic (2-nodes) and triadic (3-nodes) directed motifs in our passing networks.

The extraction of motifs could be computational expensive. Since we only care about dyadic and triadic motifs, we enumeratively count the number of motifs in the network. To account for the weights of edges, we also record the total weight (the sum of weights of each edge) of every identified motif.

- **Dyadic motifs**. There are only two directed 2-nodes motifs, i.e. $(1 \rightarrow 2)$ and $(1 \rightarrow 2, 2 \rightarrow 1)$. Without doubt, we find both motifs in our averaged network of Huskies for the whole season (see Figure 3). There are 314 mutually-connected motifs and 44 single-direction motifs. However, the total weight of motif A in Figure 3 overwhelms that of motif B. This can also be observed in Figure 1, e.g., D1 and D2 has very strong connection via a mutual interaction. It is comprehensible since mutual passes are very common in football games.

- **Triadic motifs**. According to GÜRSAKAL et al. (2018), there are 13 directed 3-nodes motifs. Using the same procedure as described above, we extracted the number and total weight of each motif. We select motifs with weights larger than the median weight and show them in Figure 4. Since their total weights are quite high, they represent dominant 3-nodes structures in our network. We find that motif A, B, and C are most frequent passing patterns for Huskies, in which three di-direction connections between three players is prominent in Huskies' playing style. This can also be proved in Figure 1, e.g., the triangle among G1, D1, and D6.

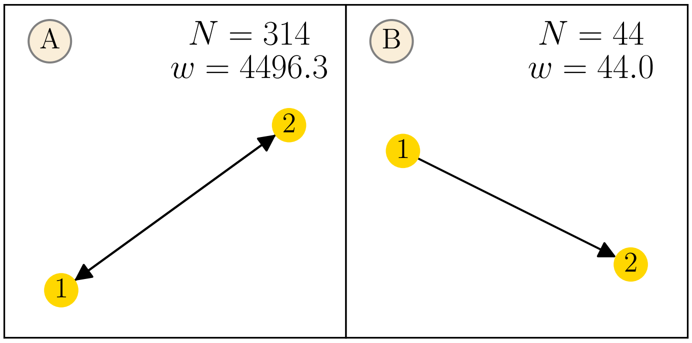

Figure 3: Dyadic motifs in Huskies' average passing network for the whole season. The number of motifs N and the total weights w are shown in the figure.

2.3 Temporal Passing Networks

A football match typically lasts 90 minutes, and the network of a team vary with time. To study how one team responds to the other team, we study the temporal network instead of an averaged network. As suggested by Buldú et al. (2019), we construct a 50-passes network which only includes the data of

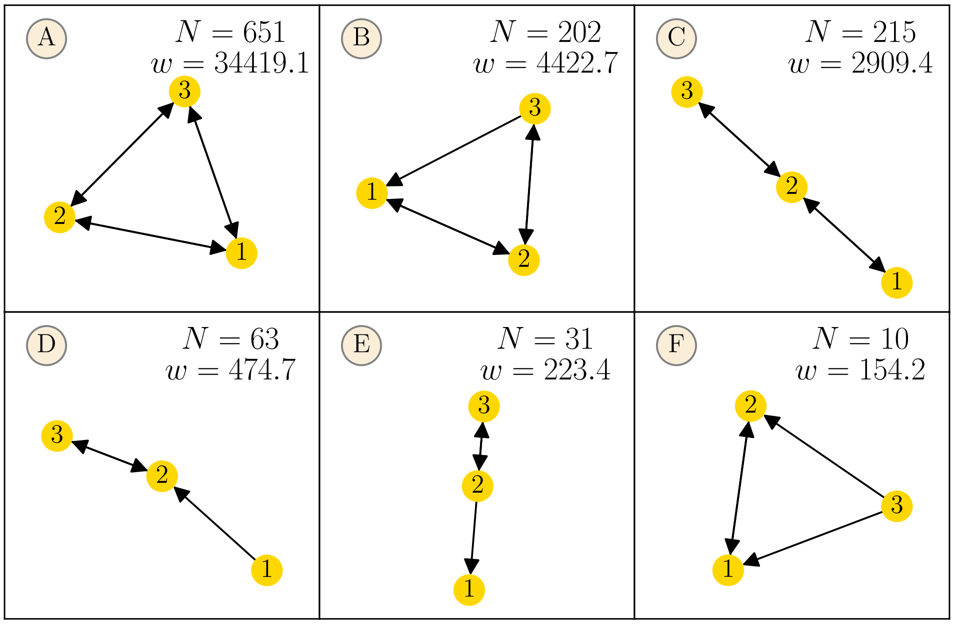

Figure 4: Triadic motifs in Huskies' average passing network for the whole season. The number of motifs N and the total weights w are shown in the figure. We only show motifs with relatively high total weights.

50 passes. The number of 50 is large enough to construct a stable and reliable network, but is small enough to tract the time evolution of the team. We consecutively construct 50-passes networks along time, and record the event time of the last pass in the 50-passes as the time mark of that network. We analyze the temporal networks further in Section 4.

3 Basic Performance Indicators

We depict properties of team performance in three main sub-divisions: overall performance, spatial properties, and network parameters. We study the team-level processes in Section 4.

3.1 Overall Performances

First, we analyzed the overall performance of Huskies versus its opponents during 38 matches. Three classical metrics of team performance are adopted, which are (a) the number of passes, (b) the number of shots, (c) the number of goals (Buldú et al., 2019). We summarize Huskies' and its opponents' performance firstly by match, then average the metrics by `OppponentID` in purpose of later comparison between Huskies and a specific opponent. For now, the opponents are averaged all together while Huskies is averaged with itself for observing differences in performance between Huskies and all other teams.

As is shown in Figure 5, left bars of all plots correspond to averaged metrics of Huskies' performance in terms of passes, shots, and goals, while right bars correspond to averaged metrics of all opponents' performance. Error bars account for the standard error of each metric. Plots in yellow represents statistically significant ($p < 0.05$) difference and plots in blue do not. As we can see in Fig. 5A, the number of passes of opponents is significantly higher than Huskies, indicating that Huskies did no have an intense or persistent control of the ball. And lower number of passes led to lower chances of being in charge, therefore resulting in lower number of shots (Fig. 5B). But we don't find significant difference between scored goals (Fig. 5C).

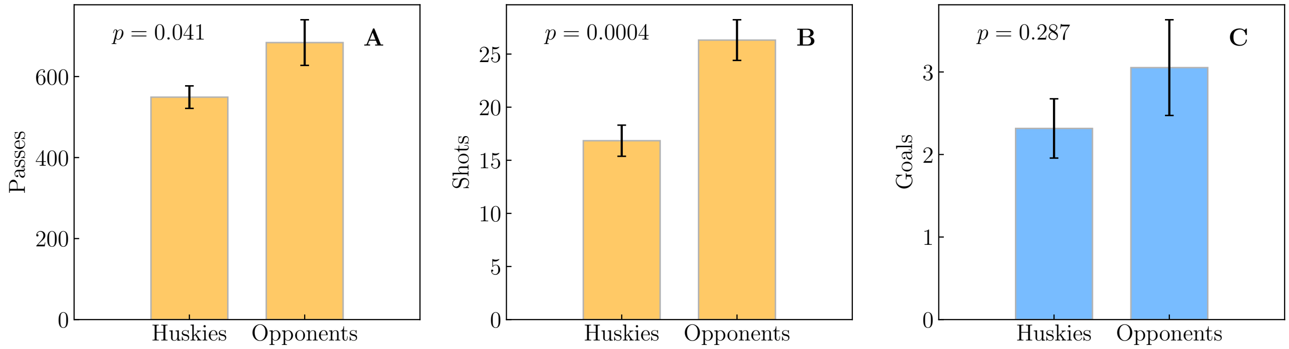

Figure 5: Average passes, shots, and goals differences between Huskies and its opponents. Plots in yellow represents statistically significant ($p < 0.05$) difference and plots in blue do not. We find that, on average, the number of passes and shots of Huskies is significantly lower than its opponents.

3.2 Spatial Properties

In addition to presenting classical parameters, we also identify metrics that reflect spatial characteristics of on-field networks of the teams: (d) x-coordinate of the network centroid $\langle X \rangle$, (e) y-coordinate of the network centroid $\langle Y \rangle$, (f) dispersion of players' distances from the network centroid, and (g) average advance ratio between total passing length of ΔY and ΔX. They are displayed in Figure 6, where plots in yellow represents statistically significant ($p < 0.05$) difference and plots in blue do not.

From the $\langle X \rangle$ and $\langle Y \rangle$ average coordinate of all passes' network centroid, we can observe that Huskies played more closer to its own goal (Figure 6A), which implies its disadvantaged situation in the matches or from which we can infer that a more defensive strategy was adopted. Difference in $\langle Y \rangle$ coordinate (Figure 6B) indicates that Huskies show a stronger preference for the middle part of the field while its opponents played more to the right side of the pitch facing Huskies' goal. No differences are found at dispersion and advance ratio in Figure 6C and Figure 6D, which suggests that the dispersion of players' position around the centroid as well as the direction of the passes are not statistically different between Huskies and other opponents.

3.3 Network Parameters

As for analyzing structures of average passing networks, we adopt six parameters related to topological characters of the passing networks (Buldú et al., 2019). In Figure 7, we plot the comparison of these parameters between Huskies (ranked 8) and opponents finally ranked top 1-6 this season. Rankings of the teams are obtained by analyzing the dataset from Pappalardo et al. (2019). Bars in each plot are color-coded according to the p-value of significance in hypothesis testing between parameters of Huskies and opponents: yellow for $p < 0.05$, green for $0.05 < p < 0.20$, and blue for $p > 0.20$.

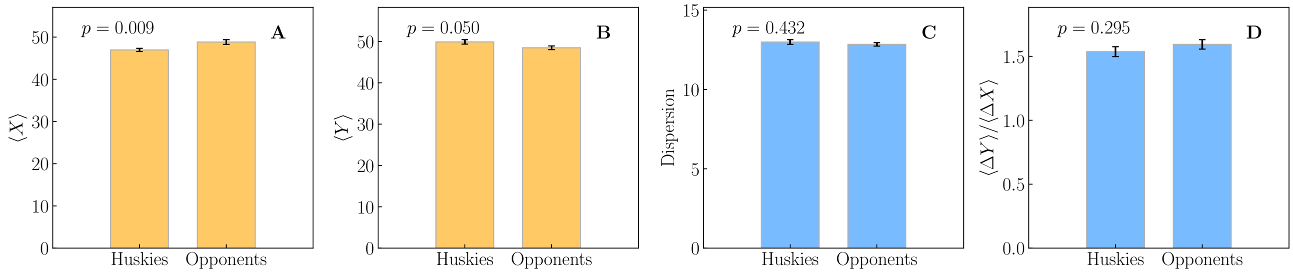

Figure 6: Spatial properties of Huskies' and the opponents' networks: centroid position, dispersion of players' distances to the centroid, and the average advance ratio. Huskies plays much closer to its own goal comparing with its opponents.

- Figure 7A shows the *clustering coefficient*, which describes the amount of triangles created between triplets of players. In football, the clustering coefficient measures the triangulation between three players. We calculate the clustering coefficient according to Onnela et al. (2005); Saramäki et al. (2007) [2]. As we can observe in Figure 7A, the value of opponents' clustering coefficient is higher than Huskies', which reveals that the connection between Huskies' three players are deficient compared to the rivals.

- The (weighted) *average shortest path length* of Huskies is an indicator about how well players are connected inside the team. From Figure 7B, we observe that Huskies' shortest path length is significant lower than that of opponents, indicating players of Huskies were better connected among them. Considering the weight of network arcs we discussed before, this fact could be the consequence of well-organized network that reduces distances between players and is not possible of being the result of passes, given Huskies' number of passes is rather low.

- Figure 7C shows the comparison between the *largest eigenvalue* λ_1 of the connectivity matrix. This indicator has been used as a quantifier of the network strength, and it increases with the number of nodes and links. Since Huskies has a low number of passes, it's not surprising to find the largest eigenvalue of Huskies is much lower than the corresponding values of opponents.

- *Algebraic connectivity*, which is computed as the second smallest eigenvalue $\tilde{\lambda}_2$ of the Laplacian matrix $\mathcal{L}$, can be interpreted as a metric for quantifying the division of a team. In this way, the higher the value of $\tilde{\lambda}_2$, the more interconnected the team is. We don't use the normalized algebraic connectivity here. In Figure 7D, we don't find significant difference between values of $\tilde{\lambda}_2$.

- Figure 7E and 7F show comparison of dispersion and maximum of the players' *eigenvector centrality*, revealing how players distributed inside the passing network. In both cases, differences are not statistically significant to suggest different centrality distribution between Huskies and its opponents.

Along the same thread, we compare the six network parameters (as described above) between Huskies and the opponents finally ranked bottom 1-6 this season in Figure 8. However, we only find

[2] https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.clustering.html#networkx.algorithms.cluster.clustering

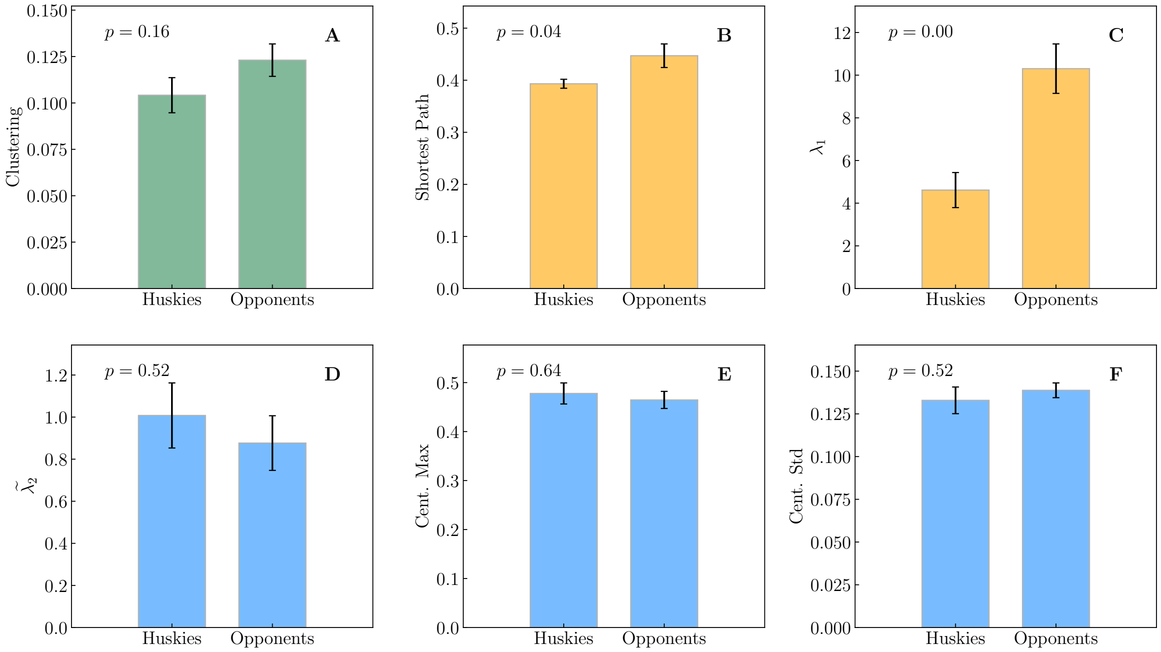

Figure 7: Comparison on six network parameters between Huskies (ranked 8) and top 1-6 teams in this season. Yellow represents a significant difference ($p < 0.05$), green represents $0.05 < p < 0.20$, and blue shows that no significant difference is found ($p > 0.20$). We find that Huskies has smaller shortest path length, smaller λ_1, and smaller clustering coefficient, which mean that the network of Huskies is less clustered but better connected. However, the network strength is much weaker than its opponents.

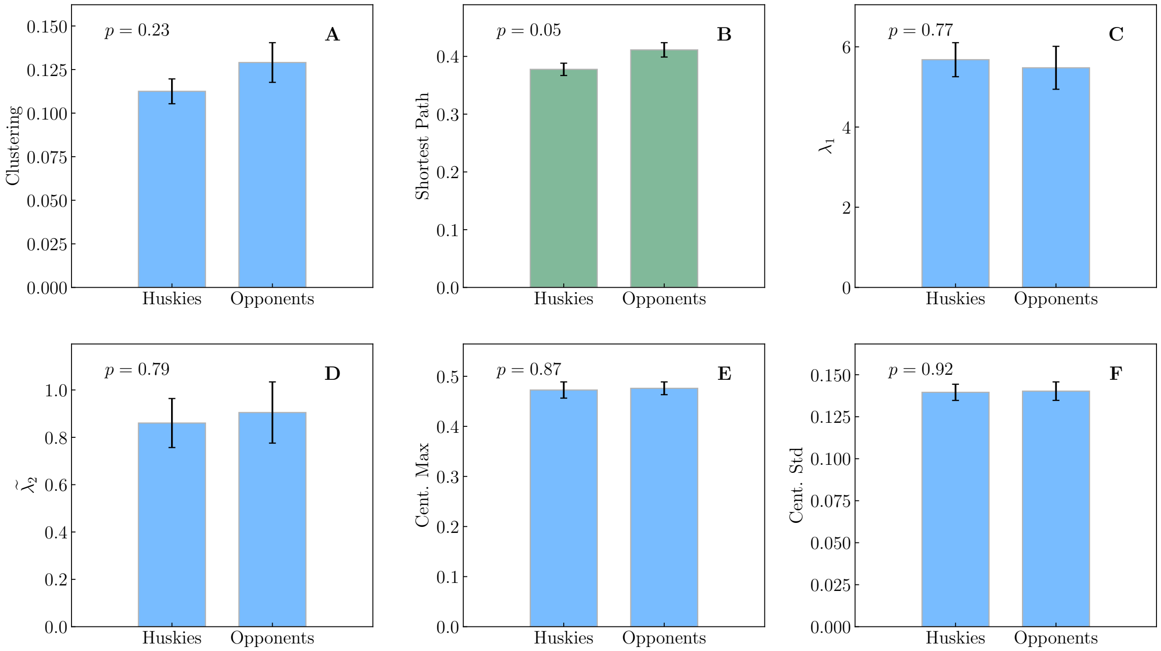

Figure 8: Comparison on network parameters between Huskies (ranked 8) and bottom 1-6 teams in this season. The plot is colored as in Figure 7. We don't find any strong evidence of the differences between Huskies and the bottom 1-6 teams regarding the network parameters.

that Huskies' shortest path length is smaller than the opponents. It is possible that Huskies has special strategies to make its networks more connected. Generally speaking, we don't find any strong evidence of the differences between Huskies and the bottom 1-6 teams regarding the network parameters. Since there might be a very large variation on playing styles and networking for the bottom 6 teams (since they don't played well), the averaged network parameters could be misleading and have large scatters. On the other hand, by comparing Figure 7 with Figure 8, we propose that successful teams might have some effective strategies which are universal, but unsuccessful teams have various reasons for their failures.

4 Team-level Processes

Now that we have already established a set of indicators to quantify performance of teams on the pitch, we move on to utilize these indicators for further investigation of team-level processes. Given those indicators we have discussed above, we inspect some other aspects of Huskies' performance in an effort to explore its attributes at the team level, including team coordination, adaptability, flexibility, tempo, etc. Some of the attributes are best shown in one specific match, such as flexibility, while others remain consistent through the matches and would only be explained more clearly in multiple matches all together, such as diversity, team coordination, and adaptability. So our analysis proceeds from studying some specific matches with various outcomes to all matches of the season.

Before that, we have to clarify our operational definitions for these team-level attributes. We defined *diversity in types of plays* as the standard deviation of spatial metrics and network parameters of the team across matches. As for *coordination*, we look at some of the network parameters of the team. *Adaptability* describes whether a team performs as well at home as play away. We also take into account the team's adaptation to different coaches. And *flexibility* refers to a team's mobility and speed of reaction on the field, so we adopt the time of achieving 50 passes t_{50} as a dynamic indicator of a team's flexibility during whole process of the match.

4.1 Flexibility: a winning case

We take the example of a winning match 18, where Huskies beat the opponent by 3-1, to illustrate attributes Huskies took on during the matches and discuss possible factors leading to its success. To present temporal changes of Huskies' team-level attributes, we construct temporal 50-passes networks (as described in Section 2.3), calculate their spatial metrics and network parameters, and plot their evolution in time series.

We first look at how *flexibilities* of both teams varies with time. Here we compare match 18 with match 26 (a lost game) to identify different representation of Huskies' *flexibility* against its opponents. As is shown in Figure 9, Huskies showed great mobility in match 18 with in-time response to opponent's action and eagerness to follow up the tempo of its rival. At the beginning, the opponent has a very intense pace, thus Huskies follows up since its response time t_{50} decreases quickly before $t = 20$ min. Then the paces of Huskies and the opponent slow down and nearly remain being at a constant level. On the whole, Huskies paced themselves according to the pace of the opponent and there is not a significant time lag between the motions of two teams.

However, the story was not the same in match 26, where Huskies were put in a great disadvantage at the very beginning: Huskies achieved 50 passes after more than 20 minutes. And the overall pace of

Huskies was much slower than the opponent, unavoidably resulting in its defeat. In this way we arrive at a hypothetical conclusion on **the importance of high flexibility** in a winning game.

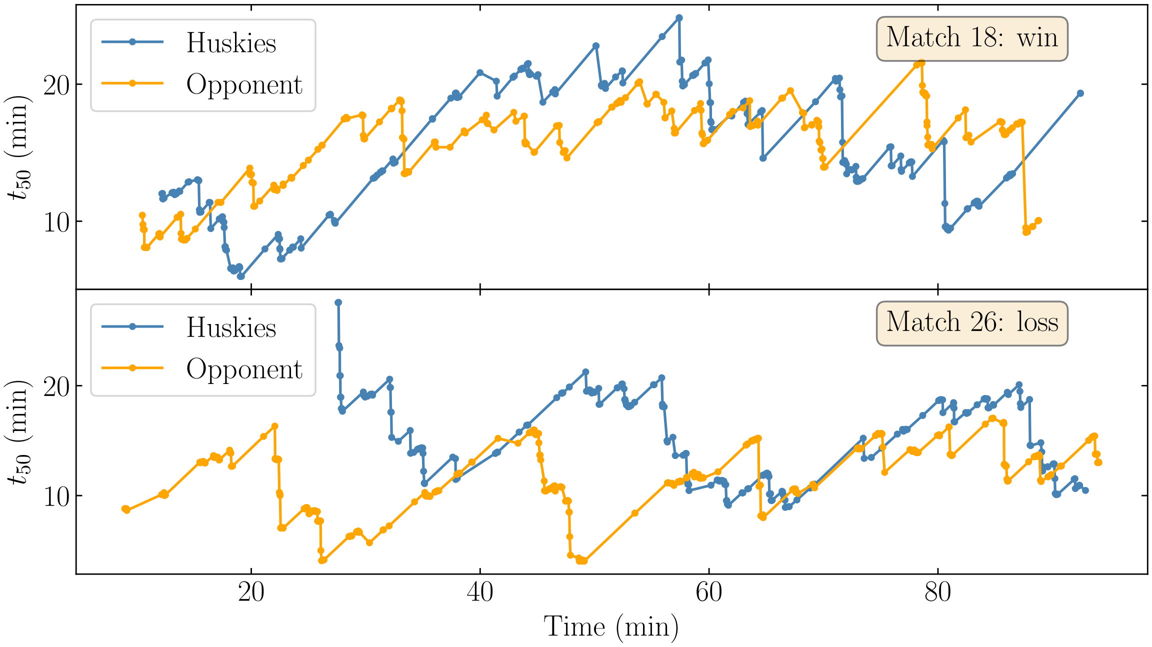

Figure 9: Flexibilities of the two teams vary with time. We define flexibility as the time of achieving 50 passes (t_{50}). Here we show how t_{50} evolves in match 18 (Huskies won) and match 26 (Huskies lost). In the upper panel, Huskies followed its opponent's pace tightly and won the match. In the lower panel, Huskies were put in a great disadvantage and lost the game.

Huskies has also taken on other characteristics in this winning game. We already noticed that Huskies tightly followed up with its opponent and won match 18. However, is the flexibility the only reason of this success? To answer this question, we further investigate how spatial metric as well as network parameters change with time in this game, as shown in Figure 10. Huskies advances to the opponent's goal from the beginning and keeps a relatively close distance to the opponent's goal; the opponent is less aggressive and is relatively far from Huskies' goal. With the game going on, the player dispersion of Huskies becomes smaller than that of the opponent, i.e., Huskies is more concentrated. At the same time, the advance ratio $\langle \Delta Y \rangle / \langle \Delta X \rangle$ of Huskies is prominently larger than that of the opponent, meaning that passes of Huskies are more parallel to the opponent's goal. Huskies might not so concerned about moving towards the goal, but advancing in parallel and waiting for better opportunities to advance. Indeed, Huskies has more passes than the opponent in this game.

During the time when Huskies' advance ratio is higher and dispersion is lower, its clustering coefficient is also higher, and its shortest path length is much lower than its opponents. These four independent evidences consistently suggests that Huskies played very well during 40 min to 70 min: the Huskies team is concentrated and well-connected. In this way, we arrive at a hypothetical conclusion on **the importance of a highly-clustered and well-connected network** in a winning game.

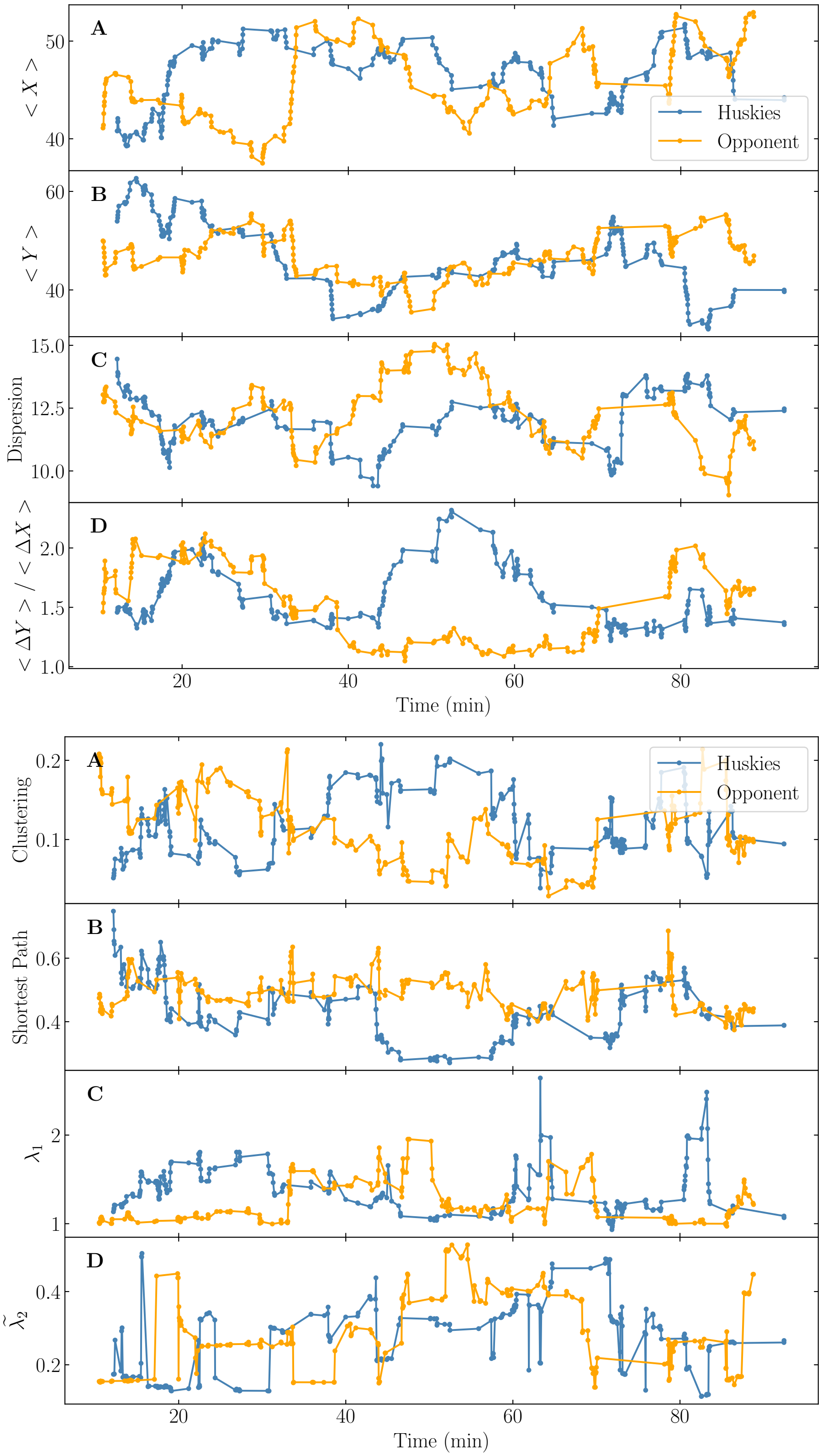

Figure 10: Time evolution of spatial metrics and network parameters in match 18. Blue line is Huskies, whereas orange line represents the opponent. During 40 min to 70 min, Huskies has high advance ratio, dispersion, clustering coefficient, and low shortest path length. As a result, Huskies won this game. This suggests that a highly-clustered and highly-connected network is crucial for the winning.

4.2 Over Matches: Adaptability and tempo

We have dynamically analyzed the role of flexibility, spatial properties, and network parameters in match 18 in great details. We proposed several strategies that might be keys to success. Now we discuss what are the universal strategies by analyzing all matches in this season. As two aspects of **adaptability**, we discuss how game locations and team coaches affect the performances.

Whether playing at home or away might be a critical factor for the outcome of a game. We group all matches into two categories (i.e., home and away), and compare four quantities for Huskies as shown in Figure 11. As we expect, it turns out that Huskies performs much better for home-based matches, while its opponents performed worse when the match is held by Huskies. No significant differences are found for response time t_{50} whether playing at home or not.

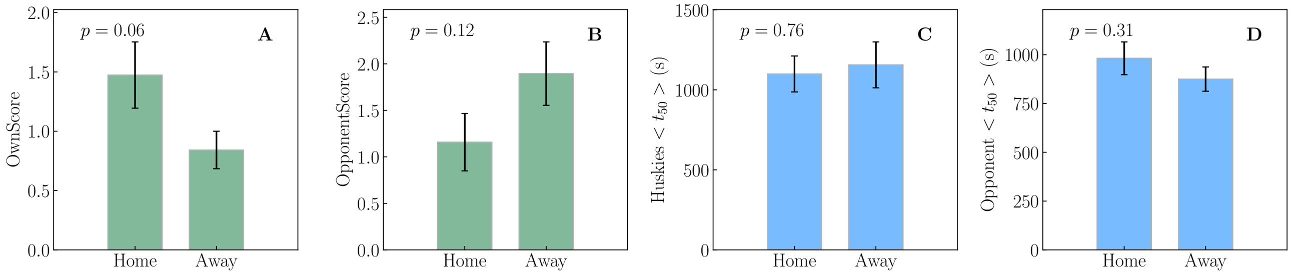

Figure 11: A comparison on the performances between two situations when Huskies is the host or guest. Huskies performs much better for home-based matches, while its opponents performed worse when the match is held by Huskies. No significant differences are found for response time t_{50} whether playing at home or not.

Different coaches has different mentoring style and playing strategies, and could directly affect the performance of a team. We also compare the scores of both Huskies and its opponents when team Huskies is mentored by three different coaches. As shown in Figure 12, significant differences are found in Huskies score: the matches advised by Coach2 got highest average score, whereas the matches advised by Coach1 got lowest scores. As expected, the coaches of Huskies don't directly affect the performance of the opponents. In Figure 12C, we find that the average response time (t_{50}) of Huskies shows the same ranking as the Huskies Score does in Figure 12A, but with less significance. It might be possible that Coach2 encourages less passes but emphasizes on other aspects such as the accuracy of passes and kicks, and the clustering or connectivity of the network. Indeed, Huskies' clustering coefficients of matches advised by Coach2 are higher.

As shown in Figure 9, the priority of match 26 is not on Huskies side, thus Huskies lost the game. The mean response time of Huskies is much larger than that of its opponent. Hence, the mean response time reflects the *tempo* of a game. In Figure 13, we compare the mean t_{50} of Huskies within matches with different outcomes. There is a significant difference on Huskies' mean t_{50}: Huskies responds much faster in those winning games.

In Figure 9, it seems that the curve of Huskies is slightly lagged behind the curve of its opponent. We quantify this time lag using cross-correlation method: we evenly interpolate the two curves and shift one curve with δt along time; then we calculate the Pearson correlation coefficient between one curve and the shifted curve. We identify the δt when the Pearson-r is largest to be the time lag between two curves.

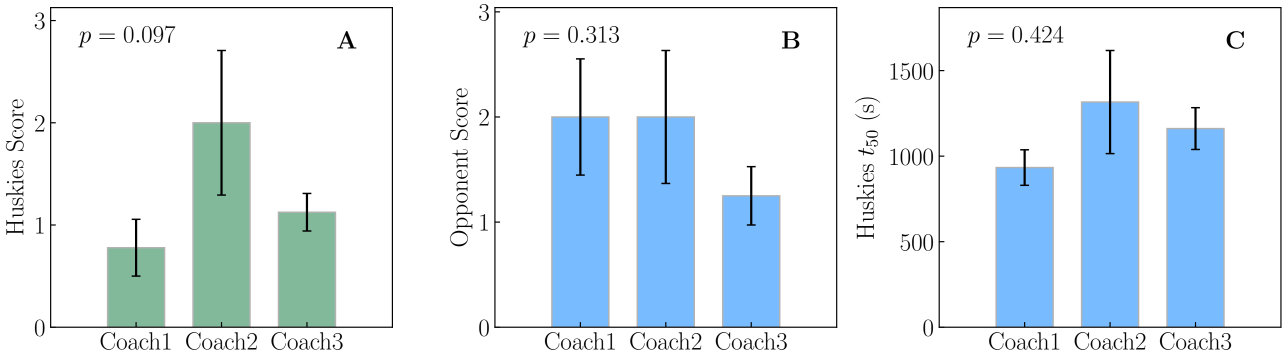

Figure 12: This figure shows how coaches affect the team performance. Significant differences are found in Huskies score: the matches advised by Coach2 got highest average score, whereas the matches advised by Coach1 got lowest scores.

We also study whether the outcome depends on the time lag between Huskies and the opponents. In the right panel of Figure 9, we see that winning matches has shorter time lag, losing matches has longer time lag. This might consolidate our hypothesis on **the importance of high flexibility** in a winning game.

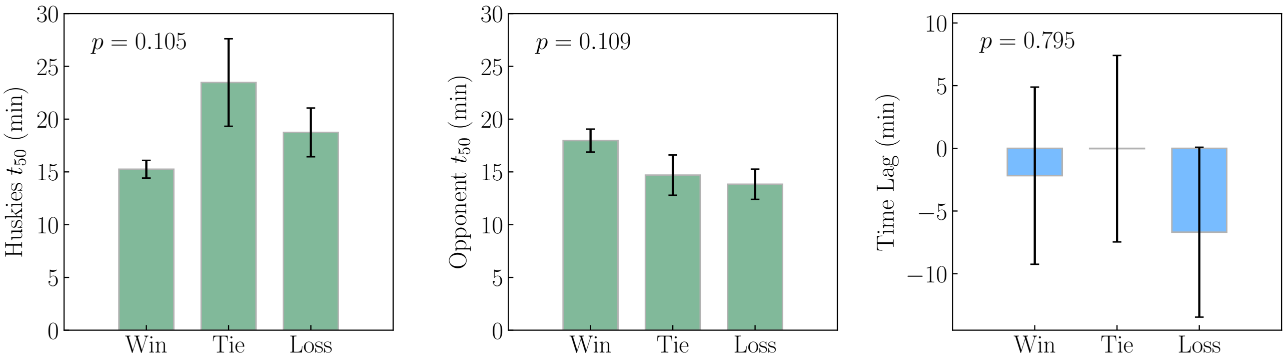

Figure 13: The effects of tempo and flexibility on the outcome of games. There is a significant difference on Huskies' mean t_{50}: Huskies responds much faster in those winning games. The time lag between Huskies and its opponent is also shorter in winning games.

5 Predictive Model

We build a comprehensive model taking both structural and dynamical properties into account. We describe the model below and discuss its usage.

5.1 H-indicator

Cintia et al. (2015) proposed an H-indicator as a proxy of match outcomes, which is defined as

$$H = 5/(1/w + 1/\mu_p + 1/\sigma_p + 1/\mu_z + 1/\sigma_z),$$

where w is the amount of passes, μ_p is the average amount of passes managed by players in the team during the game, σ_p is the amount of passes managed by players in the team during the game. Besides, we separate the playground into (100, 100) blocks. The distributions of passing count from one block to another block can be considered as a playing style, so it is worth being calculated. These features can be easily computed from the player passing network mentioned before. Thus, μ_z is the mean passing count between any two blocks, σ_z is the variance of passing volume of two blocks.

We test the validity of H-indicator using the ranking information from the whole dataset. We compute the H-indicator of each game for each team, and plot them in 14. The orange line represents the champions' H-indicators in four competitions (Barcelona of Spain, Manchester City of England, PSG of France, Juventus of Italy). It is obviously that champions' H-indicators are higher than others, making H-indicator a good predictor for the outcome.

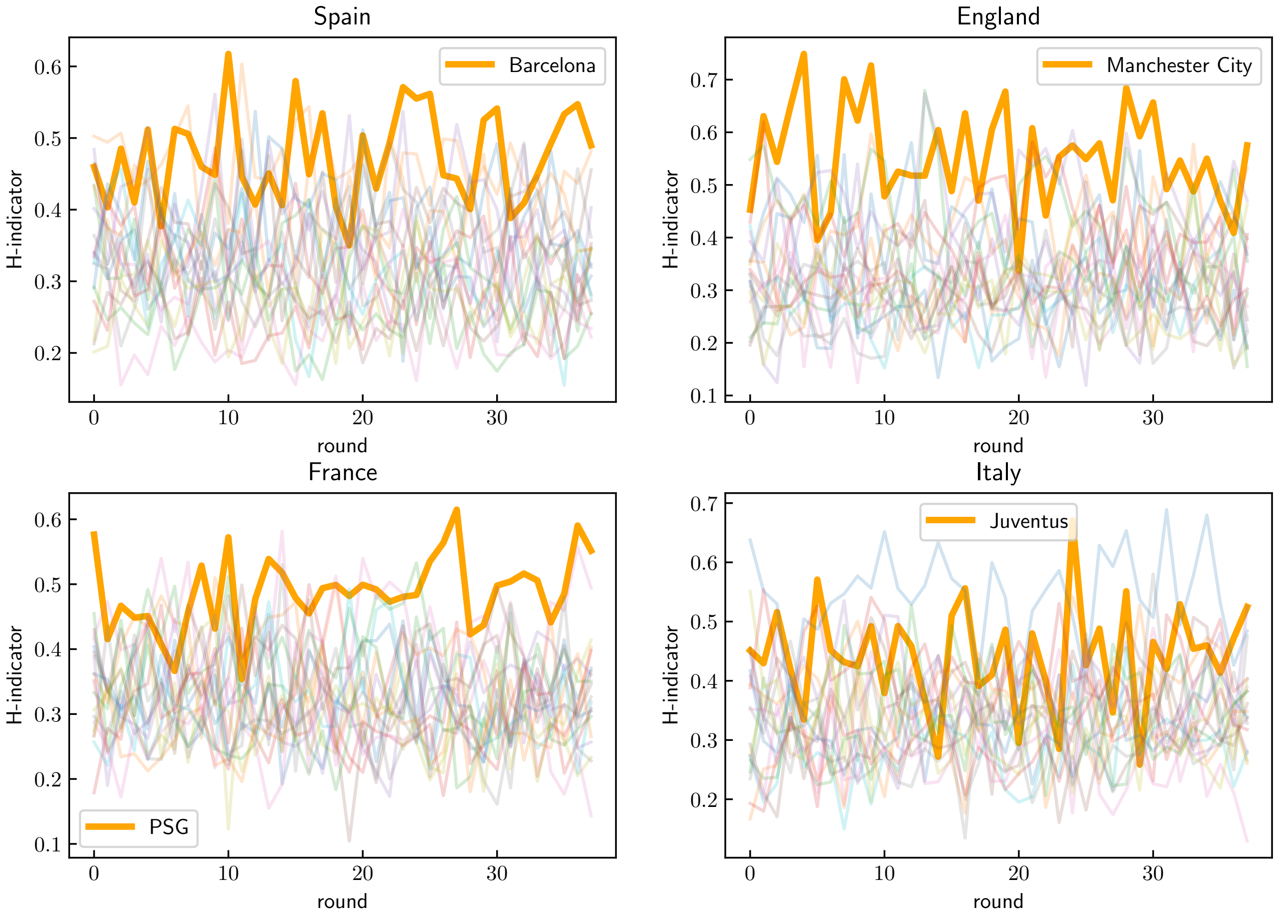

Figure 14: This figure shows H-indicators variation tendency during the season. The orange line represents the champions' H-indicators in four competitions (Barcelona of Spain, Manchester City of England, PSG of France, Juventus of Italy). It is obviously that champions' H-indicators are higher than others, making H-indicator a good proxy for the outcome.

Figure 15 shows the relation between the averaged H-indicators and the total scores in the whole season. Very tight correlations are found for different league, attesting that H-indicator performs well on average.

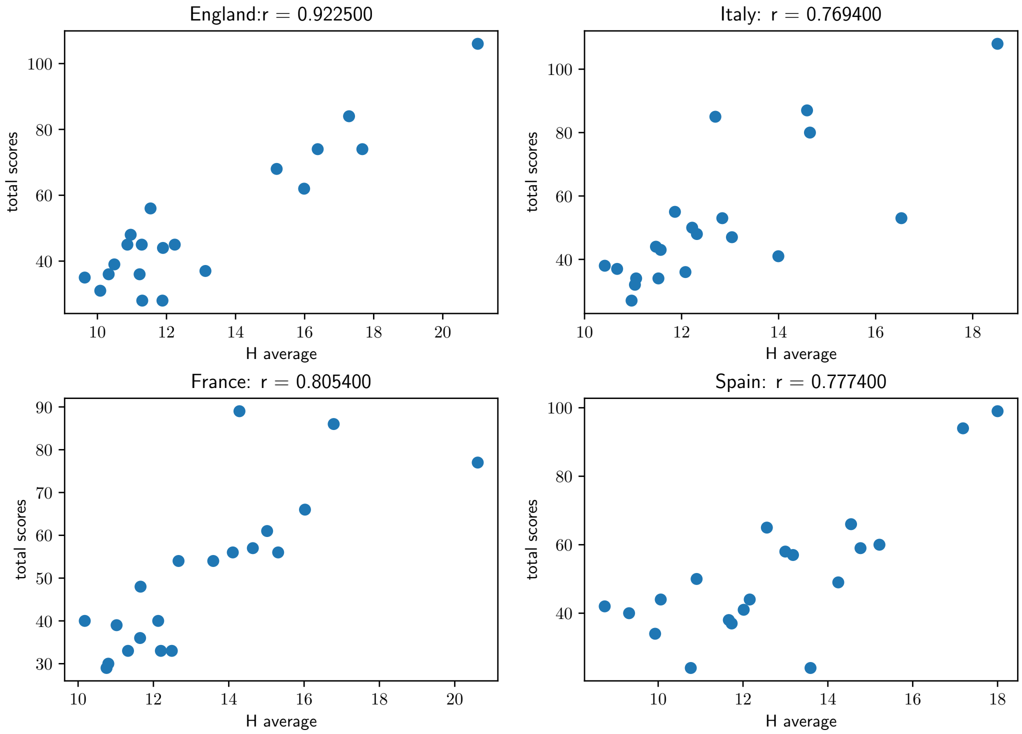

Figure 15: Very tight correlations between the averaged H-indicators and the total scores in the whole season are shown here.

5.2 Machine Learning

In a data science view, we could paraphrase the problem into a typical supervised machine learning task: we can use some features summarized from matches to predict the match outcome. Cintia et al. (2015) use six features $w, \mu_p, \sigma_p, \mu_z, \sigma_z, H$ to predict the outcome. Beside these features, we also have some additional features gained from network we build before which are clustering coefficient, shortest path length, λ_1, $\tilde{\lambda}_2$, algebraic connectivity, and eigenvector centrality.

To get a better prediction of match outcomes, we use machine learning techniques and train classifiers as follows. For each game, we compute twelve features for home team and away team. Then we use these features to construct a dataset. The label is 0 means loss, 1 means tie, 2 means win. To train the classifier with an extensive sample, we use the matches from the whole dataset (as described in Section 1.2). We tested several commonly-used classifiers: Support Vector Machine (SVM), decision tree, k-Nearest Neighbor (KNN), and random forest. Classifiers are constructed with the help of `sklearn` (Pedregosa et al., 2011). The training results are summarized in Table 5.2 below. Generally speaking, we achieved an accuracy of ~50%. It is possible that if we train an multi-layer neural network with the network and other dynamical properties as input, we might get a much higher accuracy on predict the outcomes.

Classifier	Italy	England	France	Spain
SVM	0.43	**0.47**	**0.41**	0.42
Decision Tree	0.42	0.46	0.40	0.43
KNearestNeighbor	0.42	0.47	0.41	0.36
Random Forest	**0.44**	0.45	0.39	**0.45**

6 Toward a more effective strategy

6.1 What proved to be effective?

Were there any types of play or attributes turn out to be effective? Now we may answer the question. The discussion above helps us to obtain a more detailed profile of Huskies and from that we may draw some common traits that persist in its successful history.

We may start with retrospection on how Huskies prevailed against its defeated rivals. From our analysis of network parameters between Huskies and bottom 1-6 teams, we don't find significant differences except for *shortest path*, therefore Huskies' being better connected between players on-field may have played a role. Besides, Huskies' winning cases may also have a say in extracting effective strategies. The spatial properties in match 18 shows that Huskies had a more interconnected and clustered formation, and its strength of connection was stronger, indicating inter-player connection as a key factor for success.

In match 18 the predominant feature of Huskies was its high level of *flexibility*, which strongly contrasted with that in match 26. So being mobile and flexible may be necessary in guaranteeing a win. This result can be further reinforced by our comparison of t_{50} in Figure 13. We can see Huskies displayed much faster motion in its winning games and rather short lag time. Lag time reflects how fast Huskies adapted to its opponent and sped the pace up. Shorter lag time implies faster adaptation speed. So Huskies has been very quick to respond and this ultimately led to winning, whereas Huskies lagged far behind of opponents in those lost matches.

As for spatial properties in match 18, we don't find significant discrepancy in $\langle X \rangle$ and $\langle Y \rangle$, but differences are more evident in terms of *dispersion* and *advance ratio*. Huskies showed more centered distribution and played more horizontal to the opponent's goal.

In terms of outer environmental factors, Huskies played better at home than away,scoring higher and showing more flexibility (Figure 11). Also it performed better under coach 2 with more averaged scores and shorter lag time. Therefore Huskies' success may partly attribute to certain coach.

We summarize the findings of all effective strategies here:

1. A low *averaged shortest path length*: players are well connected by network organization that minimize the distances between players.
2. High *clustering coefficient*: more triangles are created to connect three players, which makes way for more passes and forms a cooperation area.
3. High λ_1 and $\tilde{\lambda_2}$: strength of the network and the cohesion between groups of players is large.
4. High level of *flexibility* and short lag time: response quickly to opponent's pace.
5. Short t_{50}: achieve faster movement and spend less time to construct 50-pass networks.

6. Low *dispersion*: more centered distribution of players around the network centroid.
7. High *advance ratio*: play more parallel to the opponent's goal.

6.2 What can be improved?

We again turn back to the comparison of overall performances of the teams (Figure 5. It's evident that Huskies need to enhance its number of passes as the premises for more chances of shots, therefore more goals. So in football context, there should be more emphasis on keeping the ball as much as possible.

Then the comparison with top 1-6 teams (Figure 7 make it clear that although Huskies had the advantage in network connections over the defeated teams, it had not yet paralleled those top 6 teams, especially in *clustering coefficient* and λ_1. To augment *clustering coefficient*, more triadic motifs should be encouraged. Since λ_1 increases with the number of nodes and links, it's natural to improve by enhancing the number of passes. So the significance of more passes not only lies in more shots or goals, but also in the optimization of passing networks.

As we identify H-indicator as a robust predictor of game outcomes, we suggest Huskies make improvements in ways that can increase H, which mostly refers to increase the number of passes.

Finally, Huskies can also stick to Coach 2 to guarantee higher possibility of success.

7 Design an effective team

The approach we adopted and findings discovered above can be generalized to other forms of social groups. Taking the nodes as individual team members, passes as flow of information and ideas, we can also formalize such structures to depict dynamics of the group. In this part, We will discuss aspects of designing an effective team by drawing analogy between football team and general social group but not confining to football.

We previously emphasized on raising the number of passes, which is also apt for general teams. That is, in social context, raising the number of passes refers to more flows of information exchanged, and this will lead to more interaction between members.

Moreover, it is important to construct a interconnected working network as in a football match. There should be a point man to serve as the most important intermediate of all passes of information. A well-organized team (at a certain large scale) should also develop adequate sub-groups to deal with tasks and accommodate its size as well as its composition according to the constantly changing requirement of the task, so that the team will be high in *clustering coefficient*. The team should also establish a efficient system that one can easily get in reach with anyone else, which means a short *averaged shortest path*. And to increase λ_1 of a team, more 'nodes' and 'links' should be created, i.e. to ensure there are enough people in charge of tasks or responsible for transmitting information. As for enhancing algebraic connectivity $\tilde{\lambda}_2$, a mechanism should be invented to realize regular communications of different sub-groups and promote cohesion inside the team.

For a high level of *flexibility* and *tempo*, a team should keep fast pace with surrounding environment and actions of its rivals, coming up with instant counter-strategy towards emerging events accordingly.

Additionally, a team has to specify the division of jobs just as there are forward, midfielder, defender and goalkeeper in football. And the jobs come with corresponding responsibility incumbent on specific individual, so the incentive to work harder will be much stronger. But there also should allow room for variety in allocating jobs to suit the needs of the task and make the best use of its members' abilities.

It is also worth noticing the existence of principle-agent problem in some large forms of organizations. Unlike a football team, the difficulty in supervision increases as the size of the group grows. Therefore, constructing an effective supervision system should be well taken into consideration in designing larger generalized teams.

References

Buldú J. M., Busquets J., Echegoyen I., Seirul. lo F., 2019, Scientific Reports, 9, 13602

Cintia P., Giannotti F., Pappalardo L., Pedreschi D., Malvaldi M., 2015, 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp 1–10

GÜRSAKAL N., YILMAZ F. M., ÇOBANOĞLU H. O., ÇAĞLIYOR S., 2018, Turk J Sport Exe, 20, 263

Hagberg A. A., Schult D. A., Swart P. J., 2008, in Varoquaux G., Vaught T., Millman J., eds, Proceedings of the 7th Python in Science Conference. Pasadena, CA USA, pp 11 – 15

Onnela J.-P., Saramäki J., Kertész J., Kaski K., 2005, Physical review. E, Statistical, nonlinear, and soft matter physics, 71 6 Pt 2, 065103

Pappalardo L., Cintia P., Rossi A., Massucco E., Ferragina P., Pedreschi D., Giannotti F., 2019, in Scientific Data.

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Saramäki J., Kivelä M., Onnela J.-P., Kaski K., Kertész J., 2007, Physical review. E, Statistical, nonlinear, and soft matter physics, 75 2 Pt 2, 027105