Problem Chosen	2020	Team Control Number
D	**MCM/ICM**	**2022296**
	Summary Sheet	

In this paper, we study dynamical interactions among the Huskies' players and propose tactical modifications to improve their results. Due to the high complexity of the provided data, we develop methods which allow us to efficiently visualize team passing networks and calculate associated network quantities which represent each player's importance. We also define a metric with which we measure objective team success. With the metric and our ability to manipulate data, we identify and rank recurring team behaviors on a hierarchy of values. Our analyses point us to a prominent offensive triad involving Defender 1, Midfielder 1, and Forward 2. This triad has been significantly contributing to the Huskies' success, and we encourage their continued collaboration. However, we also find a tendency of Midfielder 1 and Forward 2 to form a more aggressive dyad at the expense of disconnecting from Defender 2 when the Huskies are disadvantaged. We recommend against this behavior, as it moves away from the superior triadic configuration and is often associated with losses.

Contents

1 Introduction

We the Intrepid Modeling Company have enjoyed a long history of successful modeling projects. Today, we are proud to accept the request from the manager of the Huskies to tackle a modeling problem in soccer —The Beautiful Game —in hopes of gaining enough insights about the players to suggest the appropriate changes for future improvements. In this paper, we present to the manager of the Huskies what we have learned from the Huskies' records from the past seasons and what we believe to be potentially beneficial tactical adjustments. Due to the unique nature of the problem, we utilize a combination of established methods and our own constructions to identify and place recurring team patterns on a hierarchy of values. Having tested our own methods against the provided data, we are confident in the validity of our findings and are strongly positive about our tactical proposals herein.

2 Objectives and Approach

Two main objectives of this paper are: (1) understanding the Huskies' team dynamics and finding key factors associated with certain outcomes via data analysis and visualizations, and (2) constructing concrete teamwork-rewarding metrics to appropriately rank the Huskies' current tactics on a hierarchy of success. We seek to suggest reinforcements and/or necessary changes that could bring about more desirable outcomes in future seasons. Broadly speaking, our approach to this problem consists of three main stages:

- **Data restructure and assessment:** Understanding data and data structure gives us the insights into capturing certain significant aspects of the Huskies' profile such as common statistics or patterns of play that are characteristic of the Huskies. We will design and use various algorithms and computer-scientific methods and metrics to parse the provided data for analysis through different perspectives.
- **Team Evaluation:** At this stage we look for common indicators of teamwork success in soccer, such as formations and dyadic and triadic configurations, in the Huskies' games, and use our metric to determine whether the Huskies can benefit from encouraging such formations. We can also rank the Huskies' most and least successful team efforts and find recurring patterns.
- **Application:** With our insights into the Huskies' game, we make informed tactical and personnel suggestions. We will also consider how one might apply our solutions to solving problems in sports analytics and beyond, especially in those contexts where similar dynamical systems are at play.

3 Preliminary Analysis

We consider the overall structure of the data and evaluate `matches.csv` and `fullevents.csv` to gain qualitative insights into the Huskies' seasonal performance on a game-to-game basis. A qualitative picture of the Huskies' team profile not only sets us up to define more concrete measures for rigorous quantitative analysis, but also helps us set appropriate overarching goals for which the Huskies should strive in the following seasons.

3.1 Data Overview

Soccer data is highly complex not only for their nonuniform time evolution (events spacings are not well-balanced), but also because each level of analysis is multi-faceted and multivariate. For example, across sets of events as local in time as sequences of consecutive passes, we can have classification problems (e.g. categorizing events by chance-creating sequences of passes or by player positions involved) or computational problems (e.g. computing for each sequences of passes a "score" that reflects the effectiveness of teamwork or objective offensive success). Each type of problems is associated with a number of variables. For example, it is easy to imagine that computing offense efficiency requires the knowledge of how frequently a team gets dispossessed, which, by itself, is another classification problem. Therefore, we believe that the ability to query the data with respect to *any* dimension is an absolute necessity for appropriately addressing the complexity of this problem. To this end, we rely on various computer scientific methods to decompose the provided data and make them query-able across any dimension of our choosing. We will provide on overview of our data-organization methods in the following sections.

Below is our summary and first impressions of the given datasets:

- `matches.csv` shows the Huskies' seasonal performance across three managerial changes. For our purposes, this dataset allows us to (1) construct a qualitative analysis of the Huskies' profile and (2) categorize match events in the subsequent datasets by outcome (win/loss/tie/goals) or by match location (home/away). We will also discuss whether new coaches could have had any significant effect on team performance.
- `fullevents.csv` is a highly detailed time-ordered set of tagged events across all 38 matches the Huskies played. Upon careful inspection, however, we realize the dataset is not without its limitations and errors. On the one hand, the main error is related to the "bunching" of spatial coordinates and time records of events that occur in quick successions (e.g. time separation can occasionally be zero). On the other, the main limitation has to do with the ambiguity in or lack of tags for key events such as goals, or dispossessions. To address these setbacks in further analyses, we create and adjust our own methods accordingly to remove ambiguity.
- `passingevents.csv` contains the rows in `fullevents.csv` that indicate completed passing events between two players. Its errors and limitations are thus the same as those in the passing data of `fullevents.csv`, and are most noticeable in long, consecutive passes with the exact same timestamp.

To effectively explore the multileveled and event-based structure of these datasets, we will prioritize tools that are efficient in the split-apply-combine paradigm of data manipulation. That is, splitting up data into groups based on specific attributes, applying some function to the data within each group, and then combining the transformed groups back into a single structure. Because the total size of the given data is on the order of 10 megabytes, it will also be important for our tools to be incredibly nimble on datasets that fit comfortably within a computer's memory. Based on these priorities, we will use the Pandas data analysis library in the Python programming language[1]. Although Pandas is quite memory-intensive, and thus less effective on very large datasets, it is very capable on relatively small datasets such as the given `csv` files. We will utilize the highly expressive syntax of Pandas to construct higher level tools to pick apart and explore the given data. Additionally, we will leverage the NetworkX library to construct intuitive visualizations of passing data[2].

3.2 Overview of Seasonal Performance

In this section we consider `matches.csv` to assess the Huskies' seasonal performance to characterize how the Huskies might compare to their domestic opponents.

	Win	**Loss**	**Tie**	**Goals (for)**	**Goal (against)**
Home-Away	10-3	5-10	4-6	28-16	22-36

Table 1: The Huskies's seasonal game stats

The Huskies' seasonal home and away statistics based on `matches.csv` are given in Table 1. They earned a total of 49 points, according to the conventional soccer league scoring system where teams earn +3 points for a Win, +1 for a Tie, and +0 for a Loss. Based on 2017/18 and 2018/19 final standings of major European national leagues, we placed the Huskies at the upper-mid table (rank ~10/20), which is far above the typical relegation zone, but not high enough to qualify for continental competitions.

While the Huskies are not a championship contender, it is with no doubt that they should be aiming for at worst a top-6/7 standing the following seasons in order to qualify for continental competitions or promotion. To this end, any improvement in their away performance would be hugely instrumental. Table 1 shows that even though the Huskies won more than half of their home games, they only managed to secure 3 points 16% of their time away. In addition, more than 66% of their losses were away games, which accounted for more than half of their away games. Hence, it would be interesting in the later section to look at how the team dynamics differ between home and away games.

In spite of three managerial changes throughout the Huskies' season, we have reasons to believe that these have had little effect on the overall team performance. Table 2 shows each coach's mean points collected per game and their standard deviations.

	Coach 1	**Coach 2**	**Coach 3**
Mean	0.89	1.40	1.42
St.dev	1.27	1.52	1.32
Home-Away	5-4	2-3	12-12

Table 2: The Huskies' coach stats

We reject the possibility that the home/away game differential is associated with each coach's mean points-per-game. Even when Coach 1 had a higher home/away match ratio, he won fewer points on average than Coach 2. Further, while the data does not provide us with enough conditions to construct a rigorous analysis of variance (ANOVA), we believe (within reasonable doubt) that the coaches did not bring about change in either the team's home/away performance or the overall team quality. The uncertainty in points won are too large for us to reject the hypothesis that there is any significant difference in the mean number of points won per match.

3.3 Passing Network Visualization

In order to intuitively view and analyze recurring team behaviors such as dyadic or triadic formations, we construct, via our computer scientific machinery, *passing networks* (which will be defined rigorously in Section 4.3). Fig. 1, 2 are samples of a large variety of graphs we are able to generate from the provided data. We pay special attention to Fig. 2 (left) where typical triadic behaviors manifest. Fig. 2 (right) shows a single possession with 23 passes. This is to demonstrate that we can visualize passing networks over any level of possession.

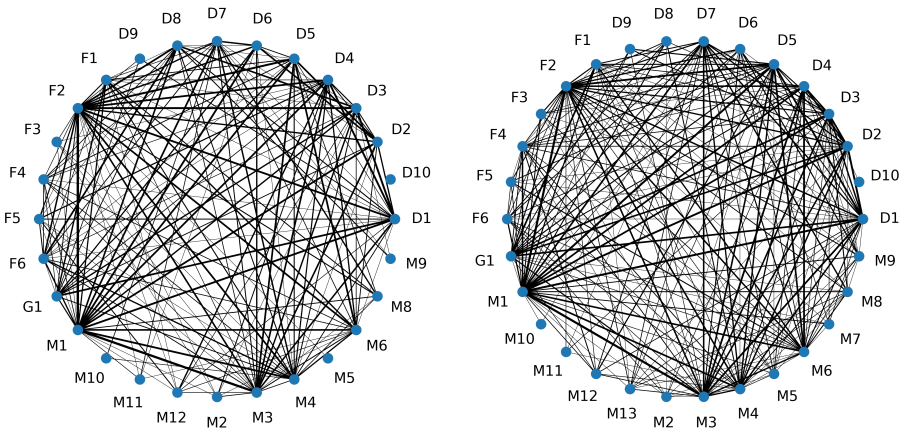

Figure 1: Passing Network, Wins (left), Losses (right)

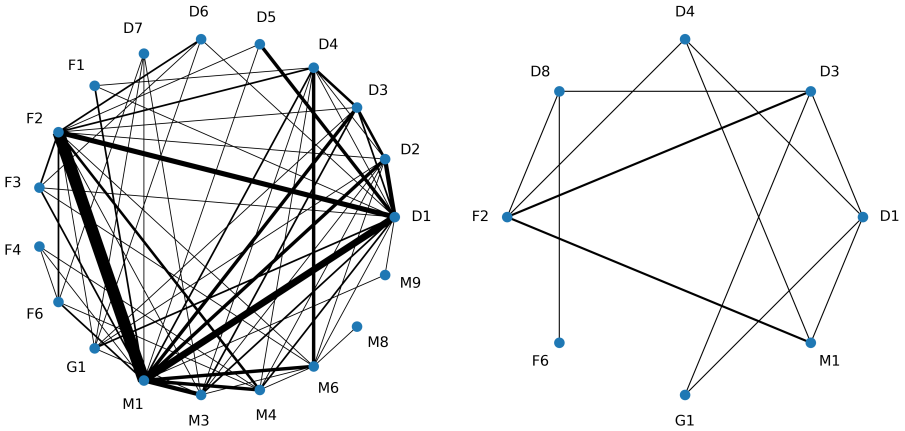

Figure 2: Passing Network, A Triadic grouping (left), A single possession (23 passes) (right)

3.4 Assumptions & Definitions

Unlike baseball or American football, soccer is generally a much more fluid sport in which a multitude of related events can occur over an extended period of time. This makes soccer data acquisition a highly demanding task. Thus, we begin this section by making assumptions that take into account the limitations in the given data. Working under these assumptions, we define our metrics for team success and use them along with established research methods to address our problems.

- *Data issues.* We observe 27 degenerate 2-pass sequences with first pass time exactly equal to second pass time (in 26 possessions). We also find instances where consecutive but significantly spatially separated events have recorded time separations on the order of milli- to nano-seconds, which we deem improbable in a real game. With this knowledge, we restrict our metric definitions from dependencies on the time components, as anomalies in time-tags can compromise our metric values.
- *Markers for passes*. Interceptions or pass completions are not well-defined in the data, in either time or event space. This makes sense practically because a pass can lead to a number of possibilities (apart from a completion) such as a foul, an offside, a duel, etc, but creates a challenge for accurately identifying possession and dispossession. To address this, we assume that a pass is incomplete unless it is confirmed as received by the next event tag.
- *Possession*. A sequence of more than one consecutive passes by members of the same team is called a *possession*. Note that the notion of "total possession," calculated as how much of the full match time a team has control of the ball, is different from our idea of a "singular possession."
- *Interception*. As discussed, there is a complex set of possible events that can occur after a pass or a shot is initiated. Our computer scientific decomposition of the data allows us to reliably create a definition for an *interception*, which we summarize in Figure 3. Consider two competing teams A and B. We define an *interception* on the basis that if Team A puts their ball possession in jeopardy by committing to an offside pass or a pass that results in a duel, then their pass is effectively intercepted. It is worth noting that the termination of A's possession does not correspond to an interception. For example, A taking a shot, scoring a goal, or clearing a ball out of bounds should not count as an interception. The event type "pass" is ambiguous because we do not know whether A might recover possession. To address this, we consider the subsequent layer of events and count not only `duel`, `offside` but also `pass` as interceptions. We believe this is justified because if possession remains ambiguous at this layer of events, Team A should be considered effectively dispossessed by B.

With these categorizations, we consider all nontrivial possessions throughout the entire season (excluding the 27 degenerate 2-pass sequences). Table 3 gives the frequencies of certain events following dispossessions.

As expected under our definitions, not all dispossessions are counted as interceptions. The *neutral* category contains all dispossessions that we deem neither offensive nor defensive. These include, for example, interruptions, substitutions, etc. This categorization will become useful for us in Section 4.2 where we define metrics to compute offense and defense "scores."

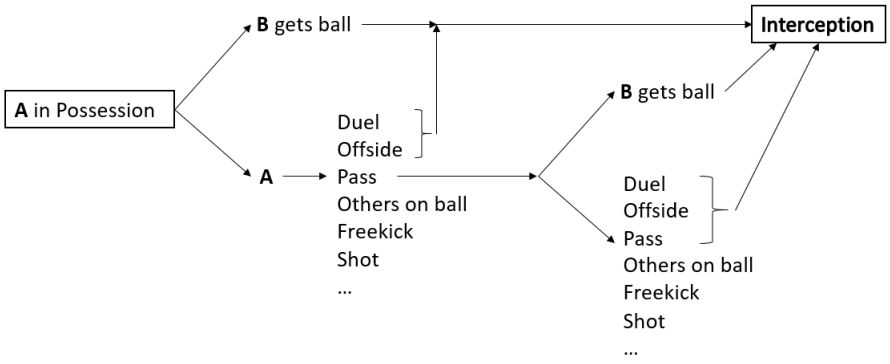

Figure 3: Flow chart characterization of an interception

Interception	Neutral	Shots	Freekicks	TOTAL
3648	1033	165	3	**4849**

Table 3: Events following dispossessions

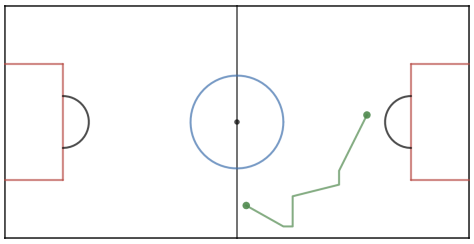

Figure 4: A chain of passes that leads to a shot, Match 26

3.5 Analysis of Key Games

In this section, we perform qualitative analyses of selected games which we deem reflective of the Huskies' strengths and weaknesses. These are games where either the Huskies or their opponent is dominated by a starkly stronger opposition. We are interested in seeing both teams in full action in addition to their weakest links and their level of adaptability under stress. Together, the analyses enable us to evaluate the team cohesiveness as a function of player position and the opponents' competence, from which we define measurable metrics for further quantitative analyses. We will study one win at home against `Opponent14` and two losses away against `Opponent9`. A secondary goal of this analysis is to verify that we are capable of correctly parsing the given data, and that we are able to uncover more information at deeper levels of analysis.

3.5.1 4-0 Home Win against `Opponent14`

We first consider the 4-0 win at home against `Opponent14, MatchID=14`.

	Goals	Possession	Shots	Passes
Huskies	4	50.1%	6	306
Opponent 14	0	49.9%	11	305

Table 4: Huskies' Match 14 (home) stats

Judging by the final score, we know that this is one of the biggest wins of the Huskies' season. Even though the summary statistics in Table 4 show very similar performances between the Huskies and `Opponent9`, the results suggest otherwise. This leads us to consider how *fluid* the Huskies' game was by considering the distributions, shown in Fig. 5, of the Huskies' and `Opponent14`'s passes per possession throughout the match.

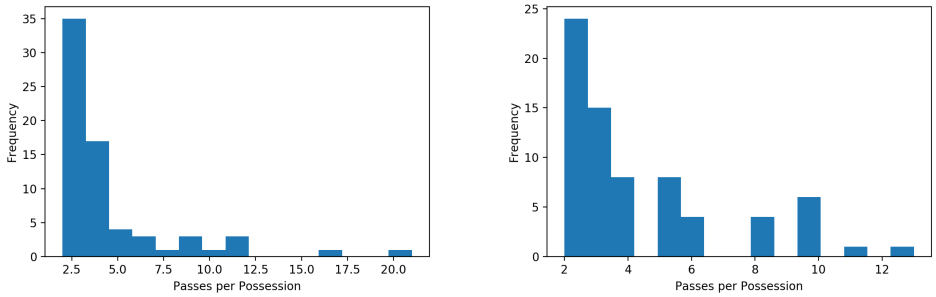

Figure 5: Game14, Passes per Possession, Huskies' (left) v. `Opponent9`'s (right)

It is clear through inspection that even though the passes were roughly equally shared between the teams, the Huskies dominated in both the low- and high-passing possessions. We also note that even though both teams have roughly equal numbers of passes, the area of the distributions in Fig. 5 are not the same. This is because we have chosen to exclude trivial 1-pass possessions. This also suggests that many of `Opponent14`'s possessions were interrupted, which implies that the Huskies were good at interceptions in this match. Of course, we cannot ignore the fact that the Huskies scored 4 goals.

3.5.2 5-1 Away Loss against `Opponent9`

Next, we consider a similar 5-1 loss away against `Opponent9, MatchID=26`. Based on goal difference and the total number of goals, this is one of the biggest losses in the Huskies' season.

	Goals	Possession	Shots	Passes
Huskies	1	39.8%	11	241
Opponent 9	5	60.2%	14	365

Table 5: Huskies' Match 26 (away) stats

Based on the summary statistics in Table 5, the Huskies were clearly dominated by `Opponent9` like before, but this time with only 40% possession. The Huskies also took their chances

poorly, as they only scored 1 goal with 11 shots. The largest number of consecutive passes by the Huskies was around 12, which was merely half that by `Opponent9`. Clearly, not only was the Huskies losing, they also were not showing signs of resistance.

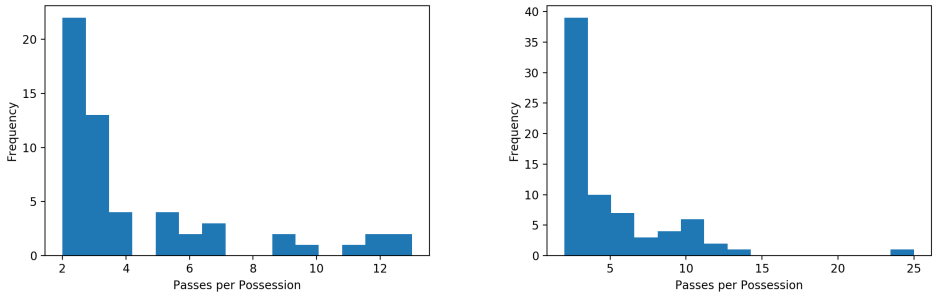

Figure 6: Game26, Passes per Possession, Huskies' (left) v. Opponent 9's (right)

4 Methods

In this section we lay out methods to efficiently query the provided data and quantitatively identify and analyze team formations, especially important patterns, recurring dyadic and triadic configurations, etc., and to create a hierarchy of values for team success. Relying on our computer scientific ability to decompose the provided data, we propose a general method herein that will let us quantify and visualize patterns and formations to various degrees. To quantify play "quality," we define a metric, called the *team success score*, and use it to assign to each sequence of possessions a score that characterizes offensive prowess. A full team success score for Team $\mathcal{A}$ is dependent upon their own offense score O_A and the inverse of their opposition's offense score O_B. However, throughout further analysis, we will focus only on the *offensive* component of the team success score.

4.1 Efficient Data-Query Framework

From our initial explorations of the given datasets, we determined that the passing data provides a useful view into the natural flow of soccer that the other discrete game event data cannot match. Chains of passes capture high level team formations, and they can also be tracked over time, thereby providing us with more potential avenues for exploration. Thus, we treat *possessions* as the fundamental unit of our data analysis, and utilize the other discrete game events as auxiliary data. To that end, we constructed a framework that provides an intuitive interface to query all the possessions in the data. We implemented the framework with the following properties:

- A user can construct queries on possessions by only thinking about high level properties of the data, and while giving little thought to the underlying structures. This allows users without programming experience to extract multifaceted groups of possessions, without diving deep into source code.

- For any property present in the framework, a user can include it arbitrarily in their query. A property can be used to filter other properties, to filter possessions based on its value (or any mutation of its value), and more; the opportunities are truly endless. We also include a possession's pass network as a property, allowing for queries on the underlying pass structure.
- It is intuitive for a programmer to implement new properties in the framework's source code; if it can be computed and associated with a possession, then it can be utilized arbitrarily alongside any other property.
- Any query on properties already in the framework is completed within milliseconds.
- The framework includes functions that operate across arbitrary possessions, which allows for predefined operations on arbitrary query results. For example, we very often utilize a function to combine adjacency matrices from queried possessions, and then plot the resulting aggregate pass network.

With this framework, extracting possessions with arbitrary properties is very intuitive and highly efficient, and it allows for fine-tuned explorations that would be too complicated or too time consuming with other tools.

As a contrived example to demonstrate the level of user-control provided by the framework, we can construct a query for *all possessions where Huskies intercepted an opponent pass in the first half of a home game, then in under twenty seconds, at least five different Huskies passed the ball up 20% of the field, then took a shot from within the penalty box*, we would find the start of the two qualifying possessions at rows 4718 and 12901 in `fullevents.csv`.

4.2 The Team Success Score

From our characterizations of events, we define a *score* with which we quantify aspects of the Huskies' game. The score measures collective dynamical aspects of the Huskies' style of play such as aggressiveness in offense and defense. By observing how the values of these scores change with respect to opponent, playing style, or any one of a multitude of dimensions, we can make informed judgments about how the Huskies perform in a variety of situations. We define two scores: the *offense score*, $\mathcal{O}_{\mathcal{T}}$, and the *defense score*, $\mathcal{D}_{\mathcal{T}}$. Then, we discuss how our definitions are justified in context and an appropriate tool to address our problem.

- *Offense score*. Over a set of $\mathcal{P}$ possessions, the offensive score $\mathcal{O}_{\mathcal{T}}$ is given by a combination of *aggressiveness*, how difficult it is for the opposition to *intercept* the offense, and a per-possession *bonus factor* which rewards clearer chances (shots and freekicks at smaller distances) and teamwork. High teamwork is characterized by a high total number of passes and a high number of players involved within a possession.

Qualitatively, the *offense score* is given by

$$O_{\mathcal{T}} = \overbrace{\frac{1}{\text{Interceptions} + \text{Possessions}}}^{\text{Interception coefficient}} \times \sum_{i=1}^{\text{Possessions}} \overbrace{\frac{\text{PassesToUpfieldMax} \cdot \text{MaxUpfieldDisplacement}}{\text{CrossfieldMaxMovement}}}^{\text{Aggressiveness coefficient}} \times \underbrace{\left\{1 + \frac{\text{critical distance}_i}{\text{distance}_i}\,(2\text{shot?}_i + \text{freekick?}_i)\right\}}_{\text{Bonus factor (shots, freekicks, teamwork)}}. \tag{1}$$

Symbolically, the offense score is given by

$$\boxed{O_{\mathcal{T}} = \frac{1}{\mathcal{I} + \mathcal{P}} \sum_{i=1}^{\mathcal{P}} \frac{P_i \cdot \Delta_{\text{up,max}}}{\Delta_{\text{cx,max}}} \left\{1 + \frac{d_c}{d_i}\,(2S_i + F_i)\right\}} \tag{2}$$

where

- $\mathcal{I}$ = total number of interceptions
- $\sum^{\mathcal{P}}$: summing over all $\mathcal{P}$ possessions
- P_i = number of passes to achieve Δ_{max}
- $\Delta_{\text{up,max}}$ = maximal up-field advancement achieved
- $\Delta_{\text{cx,max}}$ = maximal cross-field displacement
- S_i: shot indicator (takes value 0 for no shot or 1 for a shot) in each possession i
- F_i: freekick indicator (corner/penalty/freekicks/throw in/etc.)
- d_c: critical distance from goal, equivalent to 18 yards in field dimensions
- d_i: distance at which the shot/freekick is taken

- *Defense score*. The defense score for Team A, $\mathcal{D}_{\mathcal{T}A}$ is defined to be the inverse of the offense score $O_{\mathcal{T}B}$ of Team B, i.e.,

$$\mathcal{D}_{\mathcal{T}A} = \frac{1}{O_{\mathcal{T}B}}. \tag{3}$$

This definition makes sense because a high number of dispossessions by A leads to a lower $O_{\mathcal{T}B}$, hence higher $\mathcal{D}_{\mathcal{T}A}$. With our computationally efficient parsing of the data, we can compute this quantity for the Huskies and their opposition in any of their games, or for any arbitrary sets of possessions.

Roughly speaking, $O_{\mathcal{T}}$ is proportional to how aggressive the team is in moving upfield and is (approximately) inversely proportional to how many times the team gets actively dispossessed. While most possessions end with interception, our definition of an interception tells us that a set of $\mathcal{P}$ possessions can have fewer than $\mathcal{P}$ interceptions, i.e. $\mathcal{I} \leq \mathcal{P}$. With

this, the *possession factor* is a scaling factor which naturally rewards possessions with a low number of interceptions.

The *bonus factor* is designed so that a good chance created within a possession results in a boost in the offensive score $O_{\mathcal{T}}$. In order to appropriately account for shots and freekicks (which we set to be half as valuable as shots by the 2 strength factor on S_i), we normalize their individual values so that each factor is on the order of 1. We normalize the chances (shots and freekicks) by multiplying the indicators by the ratio between the critical shooting distance d_c and the actual distance d_i of the chance. This normalization also rewards shots near goal more than those taken from further away. We also intentionally designed the metric such that it only measures effectiveness in offense, regardless of whether it was a team or individual effort. We believe this is very important when evaluating whether having more dyadic and triadic configurations actually brings about a significant change in offense. It is possible that some teams are better off playing direct offense and relying on a few individuals rather than collective efforts.

We also note that the formula for $O_{\mathcal{T}}$ has been created to remain well-defined within the scope of the provided data. There is a possibility of taking the log of zero. However, since no possession can happen without at least two passes, this possibility cannot manifest. We also note that the bonus factor can never grow too large because the events `freekick` and `shots` are mutually exclusive. Within a single possession, no more than one shot or freekick occurs. In the case that no shot nor freekick occurs, the distance d_i is not defined. However, because the indicators F_i, S_i will be equal to zero, the shot/freekick bonus is simply zero. In addition, we have also designed the *offense score* in such a way that its reciprocal is the defense score of the opposition. With this definition, we only consider offense scores as the principal component of the team success metric.

4.3 Interaction Analysis via Passing Networks

Beside constructing visualizations such as Fig. 1 and 2 for insights, we consider various passing networks and look for qualitative patterns. In general, we consider (but may not necessarily analyze)

- sufficiently long chains of passes, to possibly extract team formation;
- chains of passes that result in a shot or a freekick, approximately at some critical distance from goal, for chance-creating formations;
- powers of the adjacency matrices associated with the passing networks to find dyadic, triadic, higher-order team configurations, and players' importance;
- other passing networks at more diverse levels of analysis (by possession, by game, by outcome, etc.) to assess team dynamics.

We first consider the distribution of the lengths of chains of consecutive passes. The distributions allow us to visualize and quantify what constitutes a "long" chain of consecutive passes. Team formations are most obvious when a team is playing under low pressure and is dominating. Thus, we select the Huskies' high-scoring winning matches where we

believe the team shows their formations most clearly, and find formations through visualizations from there. We also consider high-scoring losing matches where the Huskies might be dominated by a much stronger opponent. We will study whether there are significant changes in team formations as the Huskies transition from playing an easy to a difficult opponent. Not only does this analysis give us the Huskies' formations, it also shows the extent to which the Huskies could adapt to pressure. Further, because we have designed our computer-scientific methods such that data-query is very efficient, we can perform similar analyses for the Huskies' oppositions. By contrasting the Huskies' behavior to that of their rivals, we can remove uncertainties within our observations of only the Huskies.

Second, we consider a subset of chains of consecutive passes that result in a shot or a free-kick taken from around a critical distance from goal. We set this distance to be equivalent to 18 yards to opposition's goal in field dimensions. This distance translates to roughly 18 coordinate units from the oppositions' goal.

Third, we look for dyadic and triadic configurations by visualizing passing networks and computing powers of adjacency matrices associated with passing networks. We make use of the property that given an adjacency matrix $\mathcal{M}$ associated with a directed graph $\mathcal{G}$, the entry $(\mathcal{M}^n)_{ij}$ gives the number of chains of passes of length n to node from i to j. For example, given a directed graph $\mathcal{G}$ with associated adjacency matrix $\mathcal{M}$

$$\mathcal{M} = \begin{array}{c} \begin{matrix} P_1 & P_2 & P_3 \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} P_1 \\ P_2 \\ P_3 \end{matrix} \implies \mathcal{M}^2 = \begin{array}{c} \begin{matrix} P_1 & P_2 & P_3 \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} P_1 \\ P_2 \\ P_3 \end{matrix},$$

squaring $\mathcal{M}$ gives us the counts of 2-walks from P_i to P_j ($i, j \in \{1, 2, 3\}$).

In the final consideration, we compute quantities that are associated with various passing networks to identify key players and potentially weak links in the Huskies' formation. We rely on results from research on soccer network analysis and compute these values from our adjacency matrices. A *passing network* is defined as a weighted directed graph whose nodes correspond to the players on a team, joined together by arrows weighted with respect to the number of successful passes between the two players corresponding to an arrow's endpoints [3]. We will compute *closeness centrality* for each player within our passing networks.

The *closeness centrality* of a node in a passing network corresponds to how reachable a player is within the team. Specifically, given Players i and j in a passing network, define the weighted distance between Players i and j to be

$$d^w(i, j) = \min\left(\frac{1}{w_{ih}} + \cdots + \frac{1}{w_{hj}}\right), \tag{4}$$

where the minimum is taken over all directed paths connecting Player i to Player j. The

closeness centrality of Player i is then given by the formula

$$C_C^w(i) = \left(\sum_{j \neq i} d^w(j, i) \right)^{-1}. \tag{5}$$

Under this notion of centrality, players who receive large numbers of passes from a diverse set of teammates will have a high closeness score [3, 4].

The benefits of considering closeness centrality are threefold. First, it has been studied carefully and extensively in the network theoretical context, so we know with certainty that it will yield useful information about the Huskies' team performance. Second, given its general popularity among network research there are well-established algorithms to compute it from a given passing network with great efficiency. Third, the closeness centrality provides us with a quick and reliable way to assess the performance of individual players given their positions, which will be particularly important in our evaluation of the forwards and midfielders.

5 Primary Analysis & Validation

5.1 The Team Success Score

In this section we compute the scores for matches of certain outcomes. Our goal is to assess whether the offense scores reflect the team performance. We compute these scores for both the Huskies and their opponent for each selected game where one team is clearly dominated by the other. We will then compare these scores to validate the metric values. Note that we do not consider tie/goalless/close games in order to avoid ambiguity in our validation. We are only interested in the validity of our metric at this level of assessment. We believe there are more conflicting factors in tie/goalless/close games for which our metric cannot account.

We verify the validity and applicability of our metric (team success score) in a variety of situations (except those mentioned in the previous paragraph). For every match we consider herein, we compare the Huskies' and their opponents' scores in three regimes: *Interception*, *Neutral*, and *Shot*. Recall that each score is computed for a set of possessions, and so an *interception/neutral/shot score* computed is the score associated with the possessions that end with an interception/neutral/shot, respectively. A higher score in the *intersection* domain indicates better possession control. A higher score in the *shot* domain indicates higher pass-to-shot conversion.

	Interception	**Neutral**	**Shot**
Huskies	0.789	3.85	8.27
Opponent	1.46	2.86	14.1

Table 6: Metric values, Huskies 2-5 Opponent 9, Game 9

The shot and interception scores of `Opponent 9` in Table 6 show good agreement with the final result of the match (2-5). Indeed, all statistics in the subsequent Tables 7, 8, 9 follow a

	Interception	**Neutral**	**Shot**
Huskies	1.02	1.57	8.52
Opponent	0.95	1.74	16.1

Table 7: Metric values, Huskies 1-5 Opponent 9, Game 26

	Interception	**Neutral**	**Shot**
Huskies	0.97	1.86	10.6
Opponent	0.89	1.35	-1.40

Table 8: Metric values, Huskies 4-0 Opponent14, Game 14

	Interception	**Neutral**	**Shot**
Huskies	1.11	1.92	6.37
Opponent	1.05	0.60	13.6

Table 9: Metric values, Huskies 1-3 Opponent14, Game 38

similar pattern: winning teams tend to have higher *shot* scores, while *neutral* and *interception* scores are not often distinguishable due to variations that might not have been accounted for. On the basis of these agreements, we are confident that our metric, the offense score, is a valid measure of team success.

Using our validated metric, we proceed to look for the relationship between success score and the number of triads in team possession. Fig. 7 shows a strong, positive relationship between the log of the offense score and the trace of the third power of the adjacency matrix associated with the passing network. Since the log function is strictly increasing, and the trace of powers of adjacency matrices are directly related to the number of triads in a graph, the relationship shown in Fig. 7 is a strong indicator for the positive correlation between the offense score and triads. With this knowledge, we study specific triads within various passing network configurations in the next section.

5.2 Interaction Analysis via Passing Networks

In this section, we construct the Huskies' passing networks under various conditions (home/away, win/lose) and first visually look for patterns in team dynamics. Then, we calculate, for each player in each passing network, an associated *closeness centrality* in order to rank contribution to the network. From there, we select notable group(s) of players such as dyads and triads for further analysis. We note the player positions: *G*-Goalkeeper, *D*-Defender; *M*-Midfielder; *F*-Forward.

Visual inspection suggests that team dynamics is likely dependent on the whether the game is played at the Huskies' ground or away, which agrees with our qualitative seasonal assessment in Section 3.2. We also notice that there are a number of potential clusters of player with high interactions. These include but are not limited to players between lines (forward-midfield, or midfield-defense), or specific combinations of individual players. For example, defenders, especially in losing games, tend to have stronger connectivity.

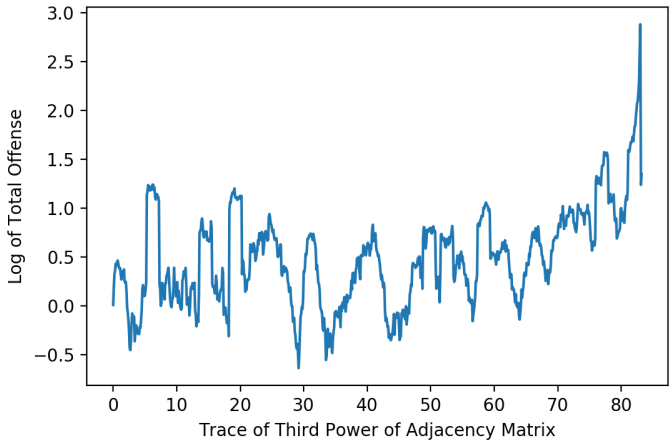

Figure 7: Log of Metric value vs. Triads value

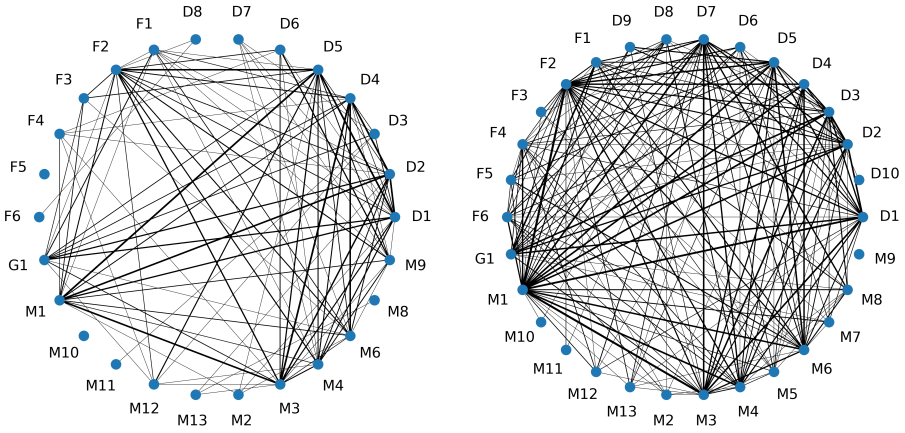

Figure 8: Huskies' Passing Network, **Losses-Home** (left) v. **Losses-Away** (right)

Upon *closeness centrality* calculations over the Huskies' possess over the entire season, we find that there exists a small group of players who are heavily involved in much of the Huskies' gameplay. Specifically, Midfielder 1 and Defender 1 have the perfect closeness centrality value of 1. We also find strong dynamics among Midfielder 1, Defender 1, and Goalkeeper 1, and not-as-strong dynamics among the triad Midfielder 1 - Defender 1 - Forward 2 (M1D1F2), from our visualizations. However, since we value combinations of players with higher offensive potential, we put more interest in the latter triad than the former. Fig. 10 shows the involvement of the M1D1F2 with the team and among themselves whenever all three appear on the pitch.

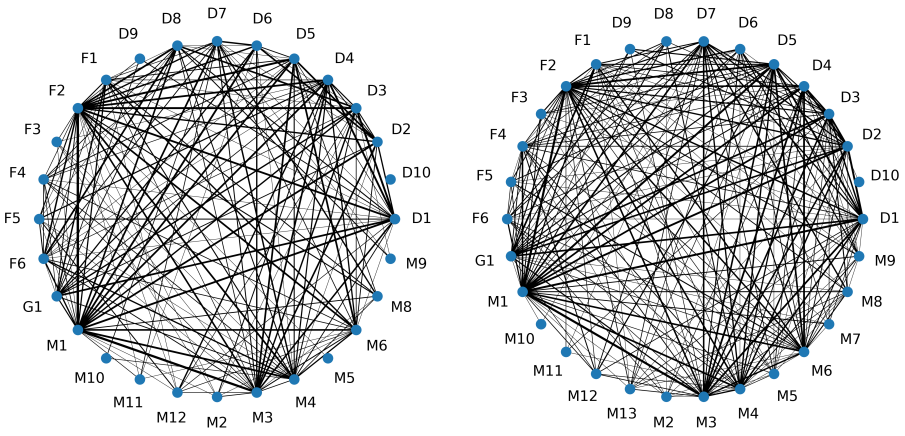

Figure 9: Huskies' Passing Network, **Wins-Home** (left) v. **Wins-Away** (right)

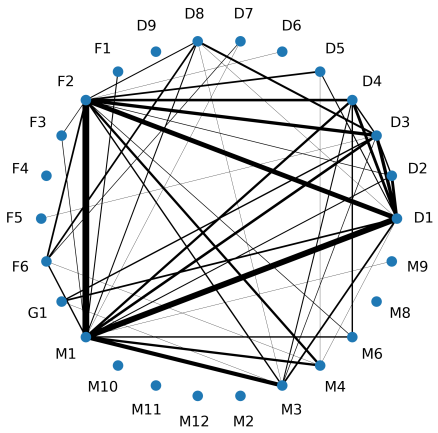

Figure 10: Huskies' Passing Network, with M1-D1-F2 triad

	M1-D1-F2	not all of M1-D1-F2	only M1-D1	only M1-F2	only D1-F2
Score	**1.40**	**1.15**	1.11	**1.55**	0.95

Table 10: Metric values under different combinations of M1, D1, F2.

Next, we seek to assess the importance of the M1-D1-F2 triad in team dynamics, to which end we calculate and compare the metric (offensive score) when the triad is involved in possessions versus a variety of situations where the triad is compromised. Table 11 summarizes our result. We see that when all the members of the M1-D1-F2 triad are present, the success score is always higher than when the triad is missing at least one member, except for when D1 is missing (1.55 > 1.40). It makes sense that the M1-F2 dyad, as compared to the M1-D1-F2 triad, simply scores higher because their aggressiveness coefficient (as only attacking players) is higher. However, whether removing D1 from the triad is a good

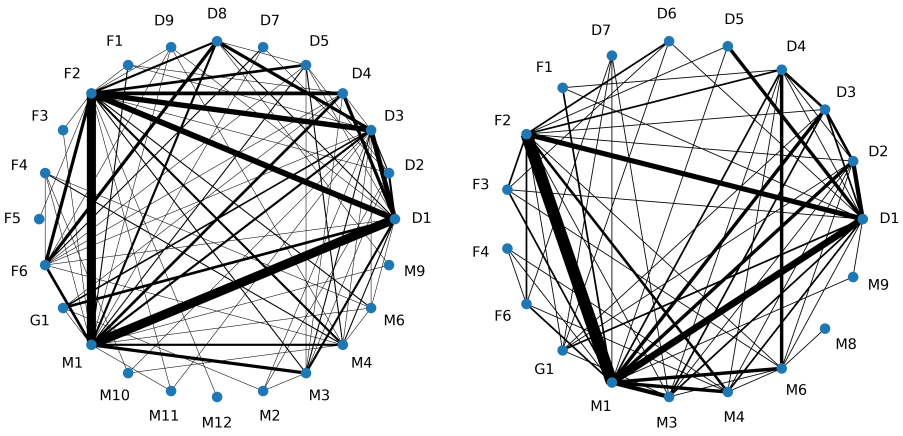

Figure 11: Huskies' Passing Network, with M1-D1-F2 triad, WIN vs. LOSE

strategy or not requires evidence from the data.

To answer this question we consider the passing networks of winning and losing matches, given in Fig. 11. In winning matches, there is clear triadic behavior among M1-D1-F2. However, in losing matches, the balance in the triad does not seem to be as strong and uniform. M1 and F2 strengthen their connection with each other while weakening their connected with D1. We believe that whenever the Huskies are losing, the dyad M1-F2 tries to be more aggressive/offensive (which we have shown through the metric) by disconnecting from the triad with D1. However, this clearly is not a good strategy because the results are losses. (Fig 12 shows an even stronger dyadic behavior in the absence of D1.)

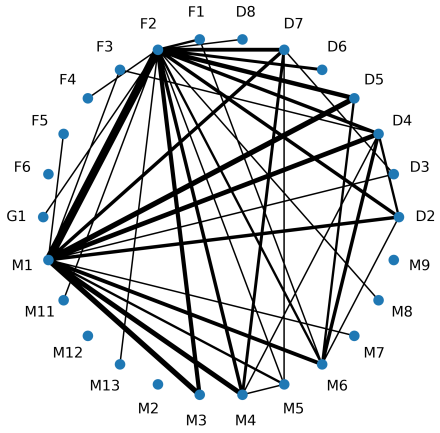

Figure 12: Huskies' Passing Network, M1-F2 only, no D1, Losing games

We conclude that the tendency for dyadic formation between M1 and F2 at the expensive of disconnecting from the triad (with D1) is **NOT** recommended, especially when the Huskies

are losing.

6 Results & Discussions

Based on our analysis of team patterns through passing networks and evaluating certain team behavior, we conclude that the triadic formation of M1-D1-F2 is generally beneficial for the Huskies and therefore encouraged. However, we warn the Huskies against the tendency of M1 and F2 to form a dyad in spite of D1, especially when the Huskies is on the losing side. While the metric tells us that the offense score in this case is higher, disrupting the triad is associated with negative results in the long run.

Our metrics and suggestions, while come from our best efforts, are not without limitations:

- We do not consider close/tie/goalless games. It is true that while our metric has been validated many times, it only performs consistently within certain regimes, such as clear wins or losses. The measure becomes ambiguous for tie/close/goalless games because we believe there are more conflicting factors at play that we have yet to account for in our metric.
- We have no knowledge of the Huskies' standings the season before, which means we don't know whether the Huskies were also involved in continental competitions during this season. This could have been a major fatigue factor that affected the Huskies' domestic performance. We are essentially building our model based on the assumption that the Huskies were only competing domestically.
- We have not considered all aspects of the provided data. For example, there are important factors of the game which we are ignoring in our approach. Some of these factors are fouls and offside. We have reasons to believe that aggressiveness and pace are positively correlated with how frequently these events occur.
- There is no event type `goal` in the provided datasets, which makes it difficult to correctly assess shots accuracy and offense effectiveness. In our current metric, a shot is as valuable as a goal, which clearly should not be the case.
- We cannot resolve the initiation and termination time of many events. In our analysis and metric, we only rely on the coordinates, displacements, and time-order of events because we were not able to obtain reliable results using the time stamps provided by the data. While it is possible to reconstruct some events based on recurring patterns in the data (e.g. a shot that forces a save, followed by a number of consecutive passes by the other team, often indicates a goal), it is beyond our capacity to complete within the given time frame.

7 Possible Applications

The process through which we have examined the Huskies' team dynamics may be generalized to other contexts where team collaboration is required to achieve a certain goal with efficiency. Consider the example of an automobile manufacturer that relies on the goods and services provided by a large number of intermediate suppliers to build its own brand

of vehicles. In order for the manufacturer to survive competition, it must seek to optimize the interactions among its suppliers. For example, it might seek to minimize the time and financial costs of shipping raw materials between factories, since suppliers in the network may at times need machines or the technical expertise from each other to complete certain tasks. At the same time, variables such as the strict environmental regulations or fierce competition will put the adaptability of the suppliers to test.

With some quantitative measure of the level of dependence between pairs of suppliers involved in the automobile business (such as the number of product exchanges between the two suppliers), it is possible to extract a “passing network" analogous to the one we have constructed for the Huskies. It might also be possible to devise a corresponding metric to measure the level of activity for the entire supply network as it adapts to an ever-changing legal and market environment. We expect that our method could then be adapted to analyze the new “passing network" along with the metric to yield useful suggestions for the manufacturer to survive and grow in the long run.

We note that a similar line of attack might be applicable for retail companies looking to boost their sales, for public library systems seeking to improve their interlibrary loan systems, and so forth.

References

[1] W. McKinney, "pandas: a foundational python library for data analysis and statistics," *Python for High Performance and Scientific Computing*, vol. 14, 2011.

[2] NetworkX developer team, "Networkx," 2014.

[3] J. Peña and H. Touchette, "A network theory analysis of football strategies," 06 2012.

[4] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks - SOC NETWORKS*, vol. 32, pp. 245–251, 07 2010.