Controlling Ball Progression in Soccer

Zoey Drassinower[1], Ari Fialkov[2], Daniel Forestell[3], Haozhi Hong[4], Emily Hunter[5]
Project Advisor: Catherine Pfaff[6]

[1]**Queen's University, 17nd7@queensu.ca**
[2]**Queen's University, 19anf@queensu.ca**
[3]**Queen's University, daniel.forestell@queensu.ca**
[4]**Queen's University, 19hh4@queensu.ca**
[5]**Queen's University, emily.hunter@queensu.ca**
[6]**Queen's University, c.pfaff@queensu.ca**

ABSTRACT

This paper focuses on how a soccer team can progress the ball up the field from the defensive third to the attacking third. We define a "safe configuration" of soccer players as one in which the ball possessor is part of a collection of players on the team that are pairwise connected by open passing lanes. We then study how teams can shift between these configurations in progressing the ball up the field. We provide some evidence of the benefits of "safe configurations."

Keywords: Soccer, football, pitch control

1 INTRODUCTION

In this paper, we examine how soccer players can use their spatial relationships to control parts of the field and safely move play up the field via chains of "safe configurations," i.e. configurations of players on a team ensuring the ball possessor has a collection of open passing options all connected by open passing lanes. An underlying philosophy informing our work is that it is most difficult to disrupt an attacking team's progression forward (with the ball) when this attacking team has multiple "good" options of how to proceed at each moment in time. Our philosophies are not new, in that teams have for years focused on forming triangles. What is new is that we use a physics-based notion of a passing lane, record configurations of players containing these triangles (or larger cliques), and record how teams could move between these configurations. In so doing, we in fact encode two steps of options in that the player with the ball can safely pass to each other player representing a vertex in the clique and then the player who receives the ball can again pass to any other player representing a vertex in the clique. We provide some evidence in §4 of the benefits of having these kinds of "passing cliques."

We achieve our aims by constructing (in §2) a directed weighted graph recording these "safe configurations" and how the team we studied has historically moved between these configurations. The nodes of a safe configuration graph are "clusters" of safe configurations. A clustering algorithm partitions a set in a manner aiming that the elements of a partition element are more similar to each other than to those of distinct partition elements. The directed edges reflect the frequency with which teams have shifted between configurations within the initial and terminal node clusters.

We also provide in §2.6 and §3 examples of how construction of (expanded versions) of the safe configuration graph could allow one to analyze a team's progression into the 18 yard box or movements leading to a loss of possession.

We believe our work can serve as a launching platform for significant further investigation into how teams can "safely progress" the ball up the field and focus training. In particular, we aim to construct a framework for creating and comparing new successful patterns of movement and identifying configurations on the field often resulting in turnovers or allowing for a lot of strategic flexibility (also impeding defensive containment), and hence focusing training.

One of the largest early advances in soccer analytics was that of Rudd [4], where she used chains of events (such as dribbles or passes) within a soccer game to place values on individual actions. Van Roy, Yang, De Raedt, and Davis [17] built upon this work and incorporated it into defensive strategy. One of the most sophisticated notions building upon Rudd's notion of an "expected threat" was the "expected possession value" developed by Fernandez, Bornn, and Cervone [10, 14]. There are also the g+ [15] and VAEP [8] models, and the work of [9]. With the arrival of tracking (spatial-temporal) data and applications of physics principles, Spearman [5] introduced "pitch control," a model quantifying the probability that a given team will have possession of the ball if it were at a location x. [7] provides a version adapted to individual player attributes. More recent papers have looked at team motion, including that [16] studies the motion of a team as a whole, with a focus on team synchronization. Along a slightly different tack, [13] focused on spatial movement pattern detection.

To the best of our knowledge, while building off such work of others, our work is unique in studying the evolution of configurations of players on an attacking team as they move from the defensive to attacking third, particularly with the flexibility our framework provides for developing new strategies. The purpose of this paper is to introduce a new framework and methodology for studying soccer configurations and ball progressions. However, to provide examples of implementation, we use some of the least problematic of the data we had access to (namely that from Mjällby in the 2021 Alsvenskan season). We look forward to seeing further examples of uses as better tracking data becomes more widely accessible.

It is noteworthy that, apart from our clustering of safe configurations, our work presented here avoids use of machine learning or artificial intelligence, with the benefit that one can look directly at situations to analyze what occurs within them and whether the computations may have included error or bias.

1.1 Outline

As a first step toward defining "safe configurations" (Definition 2.3), we build on Spearman's notion of pitch control ([6, 5]) to define "passable regions" where a particular team is most likely to control the ball upon arrival (please see §2.1). For a configuration of players on an attacking team to be considered "safe," we require that a suitable combination of passing lanes (please see §2.3) is within the attacking team's passable region. We next define in §2.5.1 a directed weighted graph, which we will call the

"safe progression graph," where a node will represent a cluster of safe configurations, a directed edge between two nodes will represent a transformation between configurations represented by the nodes, and weights indicate a transformation's frequency. In §2.6 we introduce the "Lose the Ball Node," provide data regrading the frequency of configurations leading to such a possession loss, and provide examples of configurations leading to a possession loss. In §3 we shift our focus to sequences of safe configurations leading into the 18 yard box. We provide both numeric information on these sequences (and the configurations they involve) and analysis of a concrete example. In §5 we discuss some potential further investigations and coaching applications of our work presented here.

1.2 Acknowledgements
The authors of this paper are grateful to the co-supervisory role played by Timothy Chan. The authors of this paper are also indebted to Kieran Doyle, Randy Ellis, Christian Muise, Ilya Orlov, Devin Pleurer, Luke Steverango, and David Sumpter for helpful discussions in the process of developing this work, as well as their ongoing help and support. Finally, they are grateful to the reviewers for their helpful comments in revising the paper.

2 THE SAFE PROGRESSION GRAPH $\mathscr{G}_{\text{SAFE}}$

In this section we define the safe progression graph $\mathscr{G}_{\text{safe}}$. The aim of the safe progression graph is to record means of moving between "safe configurations" of players on a field in a manner allowing one to build and study new ways to progress up the field.

The process for defining the safe progression graph is as follows:

1. We define in §2.1 a "passable region," i.e. a physics-determined region of the field that the possessor of the ball could have confidence in passing to.

2. The aim of §2.3 will be to define what we call a "safe configuration" of players. This will be a configuration of players on the field so that the possessor of the ball is part of a collection of players in which any pair of players in the collection will have betwixt them a passing lane entirely within their passable region.

3. From the safe configurations we define in §2.4.1 "passing matrices," on which we will perform a clustering algorithm in §2.4.2 to form the nodes of $\mathscr{G}_{\text{safe}}$. Each entry of the matrix represents the presence of a player or the ball in a square of the field. We Gaussian blur these matrices to help encode relative proximity of these squares in the field.

4. In §2.5.1 we define the safe progression graph $\mathscr{G}_{\text{safe}}$ itself and in §2.5.2 we give a matrix representation of this graph.

2.1 Passable Regions
In order to define safe passing lanes and cliques, to then define safe configurations, we will need a model dividing the soccer pitch into probabilistic zones of control for each team. For this purpose we define the passable regions model, similar to Spearman's pitch control model in that it assigns points on the field to certain teams by determining which team has a player who could be first to reach that location. This takes into account the players' current velocity, direction, and maximum acceleration (please see §6). Unlike Spearman's pitch control model, which assumes the ball is at a particular location (x,y) and then assigns that location to a particular team, the passable regions model also calculates the time required for the ball to reach (x,y) from its current location. This will be important for our computation of "safe passing lanes." To illustrate the significance, one can imagine the following: It looks like there is a clear passing lane from player A to player B and player B controls a region for "receiving" the ball. However, the ball is intercepted because an opposition player could reach a location on the passing lane before the ball could. The region player B controls is only useful for aerial passes. This kind of situation is taken into consideration by Spearman et al in [6], but not included in the application of the Pitch Control Function [6, 5]. See Definition 2.1 for the full definition of a team's passable region.

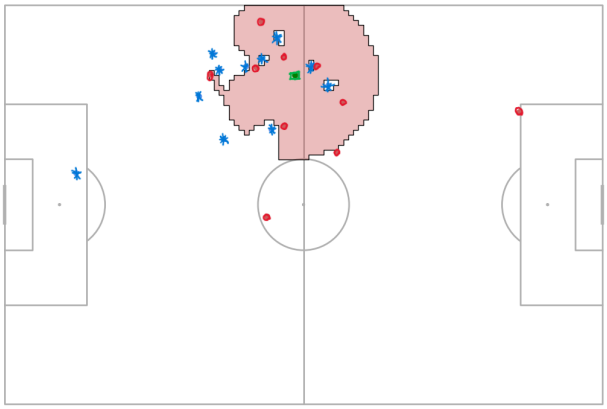


Figure 1. This figure illustrates the passable region (shaded in red) for the red team at a given moment in time in a game. The two teams are shown as red dots and blue stars, and the ball is a green square.

2.1.1 The passable region definition

In this section we define a "passable region,' which is a version of a dominance region adapted specifically for computating passing lanes. The precise physics computations are provided in the appendix.

Each 1x1 metre square on the field has been assigned a colour, namely red if the attacking team can reach that square first or the ball can reach there first (and actually reach that location) and white otherwise. Using these equations, and the resulting coloured squares, we define a "passable region" as follows.

Definition 2.1 (Passable region)**.** The (representation of the) *passable region* for a given team at a given moment in a game is the union of the following sets of squares (where squares refer to the squares as in Figure 2, but representing 1x1 metre areas of the pitch):

- the set of squares that a player from that team can reach before any other player and
- the set of squares that the ball reaches before any player.

In terms of the visual representation, this is the union of all red squares.

In short, the passable region for a team represents the area of the field where that team can play the ball and be reasonably confident of maintaining possession.

Remark 1. Our passable regions model additionally gives a refinement of the space controlled by a team. For example, if the ball is in one of the corners, it would not matter much which team "controlled" the opposite corner according to the model, as a player would not be able to reach there before the ball. If we had counted this "unreachable" space, it might falsely inflate the space a team is said to control.

We note that the passable region is defined using the data of a single moment of time. Most of this data is recorded in a single "frame," as described in the next section, §2.2. Although, as also mentioned in §2.2, two consecutive frames are needed to decompose velocities into their x and y components.

2.2 Data

The data used in this paper is Signality-provided tracking data for games from the men's Allsvenskan 2021 season. Tracking data for each game is organized into rows, where each row is a *frame*. There are 25 frames of data captured per second. We used the following data elements from each frame in our analysis:

- Whether the frame belongs to the first half or second half of the game.
- The frame index, which depends on the match time, in milliseconds.
- Data for each of the players on the home and away team, including their jersey number, (x, y) field location, and their current speed. (Please see Figure 2.)
- Information about the (x, y) coordinates of the ball and which team is in possession.

Our analysis required each player's velocity in both the x and y directions for each frame. This data was derived using their speed and the change in their position from the previous frame. It is important to note at this point that there are large gaps in the data (primarily pertaining to the location of the ball). While the missing ball location is unlikely to significantly impact the player velocity computations, frames missing player locations did occur and could impact velocity computations (please see §4.1).

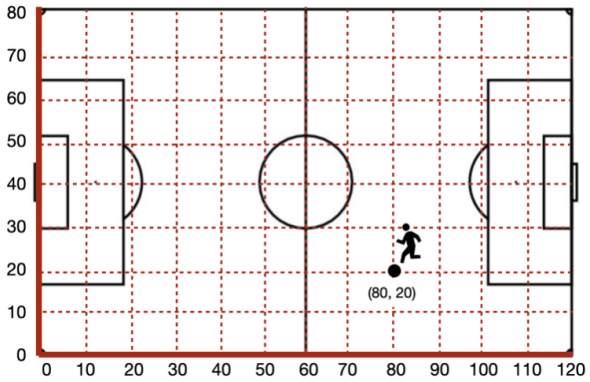


Figure 2. The coordinate system found in the tracking data.

2.3 Identifying passing cliques & safe configurations of players

Next we define in Definition 2.2 a line segment connecting the locations of two players on the same team to be a "safe passing lane" when it only passes through that team's passable region. Safe passing lanes represent passes which have a high probability of success. We did not include passes into empty space for a player to run into or aerial passes, although a more advanced model could incorporate these.

Definition 2.2 (Safe passing lane)**.** Consider player locations at a given point in time (in a single frame) to be as in the coordinate system of Figure 2. Suppose p_1 and p_2 are player locations at a given point in time for two players on the same team. We call the line segment between p_1 and p_2 a *safe passing lane* (*between* p_1 *and* p_2) if it is contained entirely within the passable region of the team containing the players represented by p_1 and p_2.

We now identify configurations of players whose passing lanes form a complete graph, i.e. there is an open passing lane betwixt any two of these players. To do this, for each instance of time we define a graph, called the "passing graph," and from this "passing cliques" and "safe configurations," according to Definition 2.3. (Please see Figure 3.)

Definition 2.3 (Passing graph, passing clique, safe configuration of players)**.** We are again considering players locations to be as in the coordinate system of Figure 2. Given a team and their passable region at a moment in time, we define a *passing graph* as the graph which has:

- a vertex for the location of each player on that team and
- an edge between 2 vertices if the line segment between the players represented by those 2 vertices is a safe passing lane, in the sense of Definition 2.2.

By a *passing clique* we will mean a subgraph of the passing graph that

- has ≥ 3 vertices,
- is a complete graph in a graph-theoretic sense, and
- satisfies that one of the vertices represents a player in possession of the ball (i.e. the ball is within 1m of this player).

We say that a collection of players on a team is in a *safe configuration* if one of these players is in possession of the ball and that player in possession of the ball is part of a passing clique formed in conjunction with other players from that collection.

Remark 2 (Passing clique interpretations)**.** We mentioned in the introduction how having a complete graph leads to multiple steps of multiple options, making it hard for the opposition to halt progression. As

another interpretation, when a graph component is incomplete rather than complete it implies an opposing player is standing within the graph component or can block a number of the passes between the players.

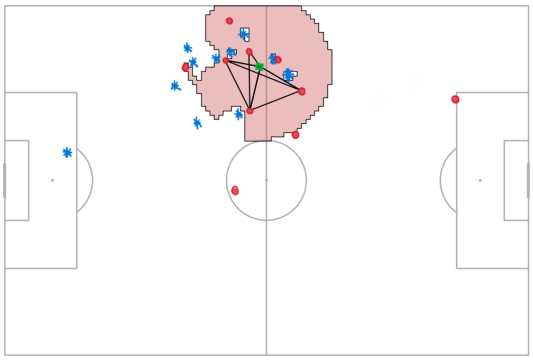


Figure 3. Using the passable regions model, we defined "safe passing lanes" between players and then "passing cliques." Red team players are represented by red circles, the blue team by blue stars, and the ball by a green box. The image depicts all passing cliques for the red team. Note that there is a vertex for each player of the red team that is part of a passing clique and these vertices are connected by an edge precisely when that edge only runs through the red team's passable region. The locations of the blue players, and those red team players not in passing cliques, are not represented by vertices but do impact the red team's passable region, and thus the possible "safe passing lanes."

2.4 Clustering safe configurations of players

The goal is now to identify and categorize configurations of players in games leading to passing cliques. In practice this will mean applying a clustering algorithm to the "Gaussian-blurred passing matrices," as defined in §2.4.1.

2.4.1 The (Gaussian-blurred) passing matrix

As a first step in the identification and categorization of player configurations, we transform each data point into a matrix that encodes the proximity of players, the location of the ball, and the passing cliques. For each safe configuration of players, we define in Definition 2.4 a 120×80 "passing matrix" representing the players and their spatial relationships. The values for the entries were chosen using an "eye test."

Definition 2.4 (Passing matrix)**.** Given a passing graph, we define the associated *passing matrix* $[x_{ij}]$ as the 120×80 matrix with the i, j entry x_{ij} corresponding to the $1m \times 1m$ square $(i, i+1) \times (j, j+1)$ on the field. This entry is:

- 0 if there is no attacking team player of a passing clique nor the ball in the corresponding square.
- 5 if either there is an attacking team player within a clique but without the ball or the ball without an attacking team player.
- 10 if the square contains the attacking team player with the ball and they are part of a passing clique.

With the grid of Figure 2, a matrix entry would correspond to a 10m$\times$10m square. In our computations we use a finer grid where a matrix entry corresponds to a 1m$\times$1m square.

Remark 3 (Passing matrix entries)**.** We chose the values of 0, 5, and 10 in Definition 2.4 because those values (of those we tried) created clusters that best matched our intuitive sense of when two configurations of players would play similar roles in a soccer game. We chose not to perform an extensive comparative analysis of techniques for clustering configurations of soccer players in this particular paper, but instead to focus on introducing our framework for analysis. On the other hand, we would be very interested to see such a comparative analysis of techniques for clustering configurations of soccer players and encourage an enthusiastic reader to perform this analysis.

Note that entries of the passing matrix associate high values to areas of the field with players in passing cliques. Also, it is possible that the player possessing the ball is not in the same grid square as the ball (as possession is computed using a radius of 1 instead of a same-square requirement). Consequently, the ball and possessor of the ball may add weight to different matrix entries.

If one clustered the vectors determined by the passing matrices, information would be lost as to how close the regions represented by grid squares were to each other on the field. To add this information we apply a Gaussian filter to the matrix. What the Gaussian filter achieves is to "smear" the value in a square (as determined by Definition 2.4) to increase the value assigned to adjacent squares. The smearing is consistent with the fact that the area adjacent to a player is under some level of control by that player.

The process of applying the Gaussian filter to the passing matrix is described in Definition 2.5:

Definition 2.5 (Gaussian-blurred passing matrix)**.** Suppose $P = [x_{i,j}] \in M_{120,80}$ is a passing matrix. Let N be the Gaussian distribution with mean $(11, 11)$ and covariance $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$. The Gaussian blur operator $G : M_{120,80} \to M_{120,80}$ is defined as the discrete convolution with $[n_{a,b}]$, where

$$n_{a,b} = N(a,b)$$

for each (a,b) in $[1,21] \times [1,21]$.

Then the *Gaussian-blurred passing matrix* is defined by $G(P)$.

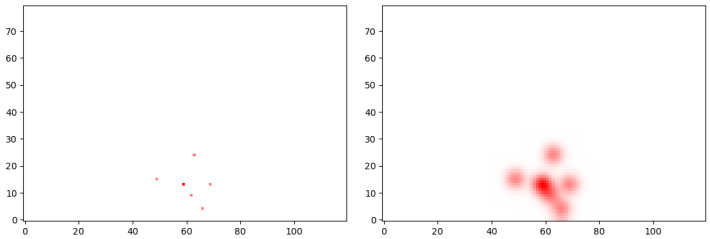


Figure 4. Information about a passing clique is encoded in a passing matrix, as in Definition 2.4. For simplicity, these images represent the (Gaussian blurred) passing matrix for a single maximal passing clique. (a) Red players are on the attacking team. The players of the maximal passing clique are slightly darker as they have larger entries in the matrix. The player with the ball is the darkest. These images are visual representations of an 80×120 matrix where the colours show the relative values of the matrix entries. (b) This image represents a Gaussian-blurred passing matrix with covariance matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$. In (a) there were exactly 6 coloured cells in the matrix. Now, the cells around the players have also been coloured according to their proximity to different players as a result of Gaussian blurring. For example, at roughly $(60, 15)$ there is a player with the ball. This player has the largest associated matrix entry and so has the darkest and biggest associated dot.

An example of a Gaussian-blurred passing matrix is shown in Figure 4 b, where a colour scale is used to show the matrix values. The decision to maximize the passing matrix entry that had the ball results in all entries around the ball increasing, as seen in Figure 4 b.

This function, G, was implemented in Python using the `transforms.GaussianBlur()` function from the `pytorch` library [12]. For more information about the Gaussian filter, one can reference [3].

2.4.2 Clustering the safe configurations

We cluster safe configurations to be able to retrieve from the data when it is possible to move from one category of safe configurations to another category of safe configurations (in contrast to recording only when a team had moved from the exact positions of players in one moment in time to the exact positions of players in another moment of time). This gives us a discretization (via the graph $\mathscr{G}_{\text{safe}}$) of possibilities for the motion of a team moving between safe configurations.

Before clustering the configurations, to reduce the computational resources needed for clustering, we partition the frames containing safe configurations into 8 sets. Given a frame X, let $n(X)$ denote the total number of players involved in at least one of the passing cliques in that frame. We remove from consideration all frames X for which $n(X) > 6$, as they appeared to mostly only occur during set pieces or at the start of play directly after a goal (and not during active play, which is the focus of this paper). We then partition the set of all remaining frames using $n(X)$ and whether or not the goalkeeper is contained in a passing clique. Since a passing clique must contain ≥ 3 players, this gives 2 partition elements for each

value in $\{3,4,5,6\}$, namely one for when the goalkeeper is included in a clique and one for when the goalkeeper is not included in a clique. We then apply the clustering algorithm separately to each partition element, obtain a list of clusters for each partition element, and combine each of these lists of clusters into a single list.

Each of the 6 partition elements consists of a collection of frames of data, with each frame representing a safe configuration. Each of these frames has associated to it a Gaussian blurred passing matrix. We applying the clustering algorithm to these matrices (hence giving a partitioning of each of the 6 partition elements into safe configurations of players that resemble each other). We use OPTICS clustering with the L^1 metric giving the distance between Gaussian-blurred passing matrices. We use the "Elbow Method" to determine how many clusters to use.

As mentioned with regard to choices made in the Gaussian-blurred passing matrix (§2.4.1), we would be very interested in seeing a comparative analysis of techniques for clustering configurations of soccer players and encourage the enthusiastic reader to form and publish this analysis. For example, one could view player locations in the field as spikes in a probability distribution and then use the Wasserstein distance between two probability distributions as the distance between configurations. One could also alter the preclustering methodology or parameters for describing the configurations, such as the angles in a clique, orientation of a clique, or passing lane length.

Definition 2.6 (Safe configuration cluster)**.** Given a set of tracking data on which we have applied a suitable clustering algorithm, we will refer to the clusters returned from the algorithm as the *safe configuration clusters*.

2.5 The safe progression graph and its adjacency matrix

In this section we define the "safe progression graph" $\mathscr{G}_{\text{safe}}$, which encodes how frequently teams have historically moved between which clusters, i.e. shifted between the configurations within 2 clusters and, in so doing, provides a discretization of possibilities for the motion of a team in moving up the field.

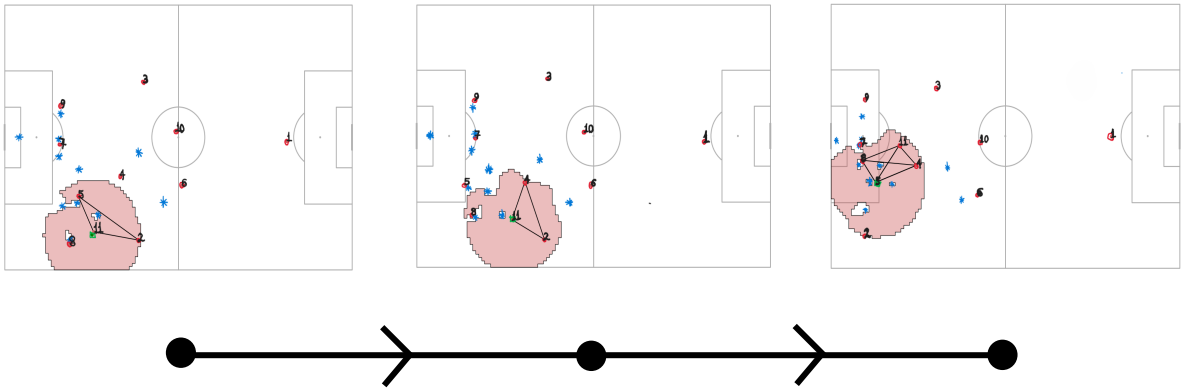
2.5.1 The safe progression graph $\mathscr{G}_{safe}$

Definition 2.7 (The safe progression graph $\mathscr{G}_{\text{safe}}$)**.** Given a set of tracking data and parameters defining the passing matrix distance, the *safe progression graph* $\mathscr{G}_{safe}$ will have:

- A *node* for each safe configuration cluster.
- A *set of directed edges* $E(A,B)$ defined by there being a directed edge from Node A to Node B if the team in question has ever in the data set within 10 seconds moved from the safe configuration cluster represented by A to the safe configuration cluster represented by B.
- A *set of weights* $w(A,B)$, one for each edge and defined as follows. The weight $w(A,B)$ of a safe progression graph edge $E(A,B)$ is the number $x \in (0,1]$ that is the fraction of the transformations from the safe configuration cluster of Node A which go to the safe configuration cluster of node B out of all the transformations originating in A. The sum of the weights of all out-going edges at a given node must be 1. For example, suppose that the team transforms from the safe configuration cluster node A to the B,C, and D safe configuration clusters. Suppose further that, historically, the team has shifted from configuration A to configuration B eight times, from A to C twelve times, and from A to D twenty times. Then $w(A,B) = \frac{8}{40}, w(A,C) = \frac{12}{40}$, and $w(A,D) = \frac{20}{40}$.

One may wish to remove all nodes of degree 0 (as they represent a cluster not involved in any safe configuration transformation) and we do so in the examples below. Such clusters may have occurred when the team gained possession only briefly, made a long pass, or moved in a way not captured by our model.

Example 1 ($\mathscr{G}_{\text{safe}}$ edge example)**.** In this example we provide one example of how an edge can form. The configurations represented by the vertices are in fact Figures 11, 12, and 13 in §3 below. Note that, while we only show one configuration per cluster in the figure, each node represents an entire cluster of configurations (close to the representative configurations shown). The first edge results from an attacking player running into the box (drawing a defender with them) and then a second attacking player becoming open. The players changed the passable region and safe passing lanes. The second edge results from the motion of several players, combined with a pass. Red team players are represented by red circles, the blue team by blue stars, and the ball by a green box.



Remark 4 (Markov chain interpretation)**.** With an assumption that configuration transformations do not depend on preceding transformations, $\mathscr{G}_{\text{safe}}$ can be viewed as a Markov Chain where the safe configuration clusters represent the states and the weights represent the transition probabilities.

We propose in §5 several uses for the safe progression graph.

2.5.2 The adjacency matrix

As the graph itself is rather unwieldy, we present $\mathscr{G}_{\text{safe}}$ as an adjacency matrix where a row represents the start configuration node of a configuration transformation and the column is the end configuration node of the transformation. As such, there is a 1-to-1 map from the set of clusters to the set of rows (or columns) in the adjacency matrix of $\mathscr{G}_{\text{safe}}$. Additionally, the clusters have been sorted according to the x-value of the cluster's center-of-mass, as defined in Definition 2.8. This sorting means that clusters which generally occur closer to the defensive net come before clusters which generally occur closer to the attacking net.

Definition 2.8 (Cluster center-of-mass)**.** Suppose that C is a given cluster composed of n passing matrices $\mathscr{M} = \{m_1, ..., m_n\}$. For each passing matrix m_i we denote the positions of the l attacking players in the passing cliques of m_i by $\{(x_{i,1}, y_{i,1}), ..., (x_{i,l}, y_{i,l})\}$. Then the center-of-mass of C, denoted $COM(C)$, is the pair $(\text{COM}_x(C), \text{COM}_y(C))$ defined by

$$\text{COM}_x(C) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{l}\sum_{j=1}^{l} x_{i,j}\right) \qquad \text{and} \qquad \text{COM}_y(C) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{l}\sum_{j=1}^{l} y_{i,j}\right). \tag{1}$$

Definition 2.9 (Adjacency matrix)**.** Given a safe progression graph $\mathscr{G}_{\text{safe}}$ with d nodes, the adjacency matrix $A(\mathscr{G}_{\text{safe}}) = [a_{ij}]$ is the following $d \times d$ matrix: The d clusters (represented by the nodes of $\mathscr{G}_{\text{safe}}$) are labeled C_k with the indices k determined by a monotonically increasing function $\{1, \ldots, d\} \to \{\text{COM}_x(C)\}$ sending k to the cluster C_k with the k^{th} smallest value of COM_x. Then $[a_{ij}] = w(C_i, C_j)$ for each $1 \leq i, j \leq d$.

Example 2 (The safe progression graph adjacency matrix)**.** In this example we give the $\mathscr{G}_{\text{safe}}$ adjacency matrix $A(\mathscr{G}_{\text{safe}})$ (Figure 5), computed using OPTICS clustering on the passing matrices for half the games of the data set of §2.2 in which Mjällby AIF played. More precisely, we used OPTICS in scikit-learn [1]. For clarity of the image, for the sake of Figure 5, we iteratively removed all nodes of $\mathscr{G}_{\text{safe}}$ of outgoing valence < 2 (not counting contributions to the valence resulting from loops). The axes in the image are flipped from that of the adjacency matrix and the weights are indicated by intensity of coloring (as indicated on the right-hand side of the image), instead of by a number.

We first observe that most of the transformations in Figure 5 occur along, or near, the diagonal. Transformations directly on the diagonal correspond to transformations within the same cluster (i.e. the start and end configurations of the transformation are in the same cluster, so that the transformation forms a loop in $\mathscr{G}_{\text{safe}}$). Transformations near the diagonal correspond to transformations between safe configurations that *tend* to be located in the same general area of the field. It makes sense that such transformations make up the majority of the transformations in Figure 5, as most ground passes do not cover large distances. Transformations above the diagonal represent progression up the field.

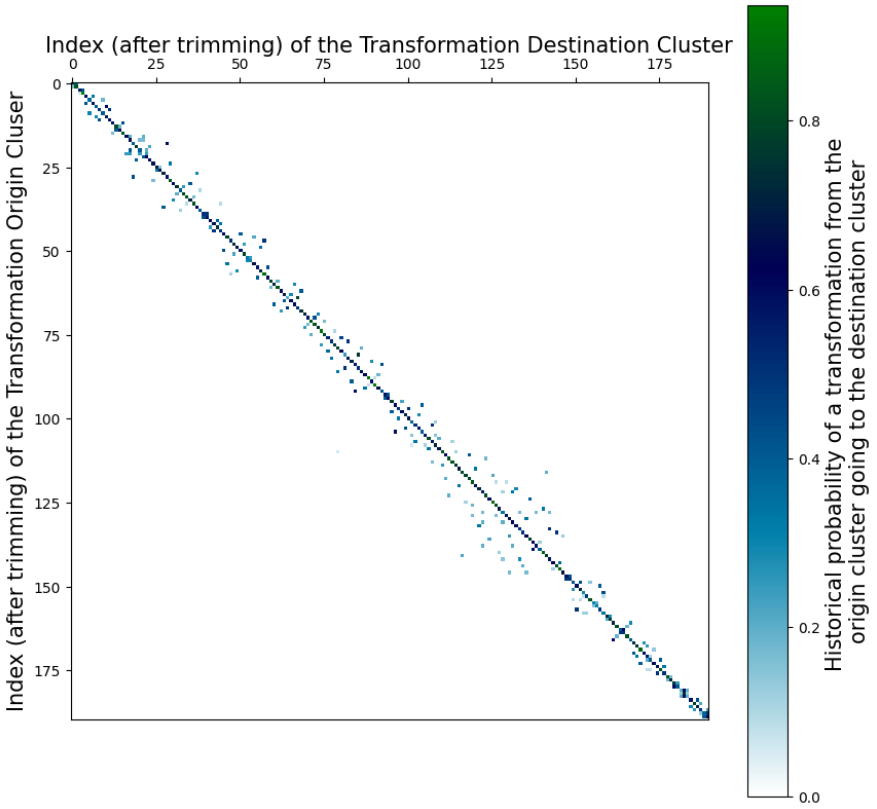


Figure 5. This figure displays the adjacency matrix of the safe progression graph obtained from half the Mjällby AIF games in the §2.2 data set. The axes represent the same set of clusters, which have been ordered by the mean x value from the centers of mass of the passing cliques in each cluster. For example, passing cliques within the cluster referenced by 160 on the two axes tend to be further up the field (toward the attacking net) compared to passing cliques in the 100 cluster. A value of v at entry (x_i, y_j) indicates that $v \times 100\%$ of the transformations originating from cluster y_j go to cluster x_i.

2.6 The Lose the Ball Node

In the calculation of $\mathscr{G}_{\text{safe}}$, we also recorded which clusters occurred directly before a loss of possession, highlighting which configurations may ideally be avoided. This was encoded in $\mathscr{G}_{\text{safe}}$ by adding a new node, called the *Lose The Ball (LTB) node*, and then keeping track of the transformations during which the ball was lost via edges from a safe configuration cluster node to the LTB node. Thus, the LTB node is the unique node that does not represent a particular cluster of configurations, but merely a loss of possession. The weight on a directed edge entering the LTB node indicates the frequency of a loss of possession after a team is in a configuration of the configuration cluster of the initial node of that edge. The LTB node has no outgoing valence.

Framing this into our Markov chain interpretation of $\mathscr{G}_{\text{safe}}$, we have added a new state (LTB node) and adjusted the transition probabilities accordingly. The LTB node can be viewed as an absorbing state.

Example 3 (The LTB Node and the Mjällby AIF data)**.** We include in the following image 3 configurations that with high frequency lead into the LTB node. Red team players are represented by red circles, the blue team by blue stars, and the ball by a green square.
The locations on the field are different, but one may notice that in each of these 3 frames:

1. there is only a single maximal passing clique and
2. the players in the clique without the ball give no progressive pass options to the player with the ball.

We did not run an analysis on whether (1) and (2) are typical for clusters directly preceding the LTB node, though one could particularly see how (2) could lead to an increased likelihood of losing possession, particularly if the players were eager to progress the ball and tried an option outside the clique.

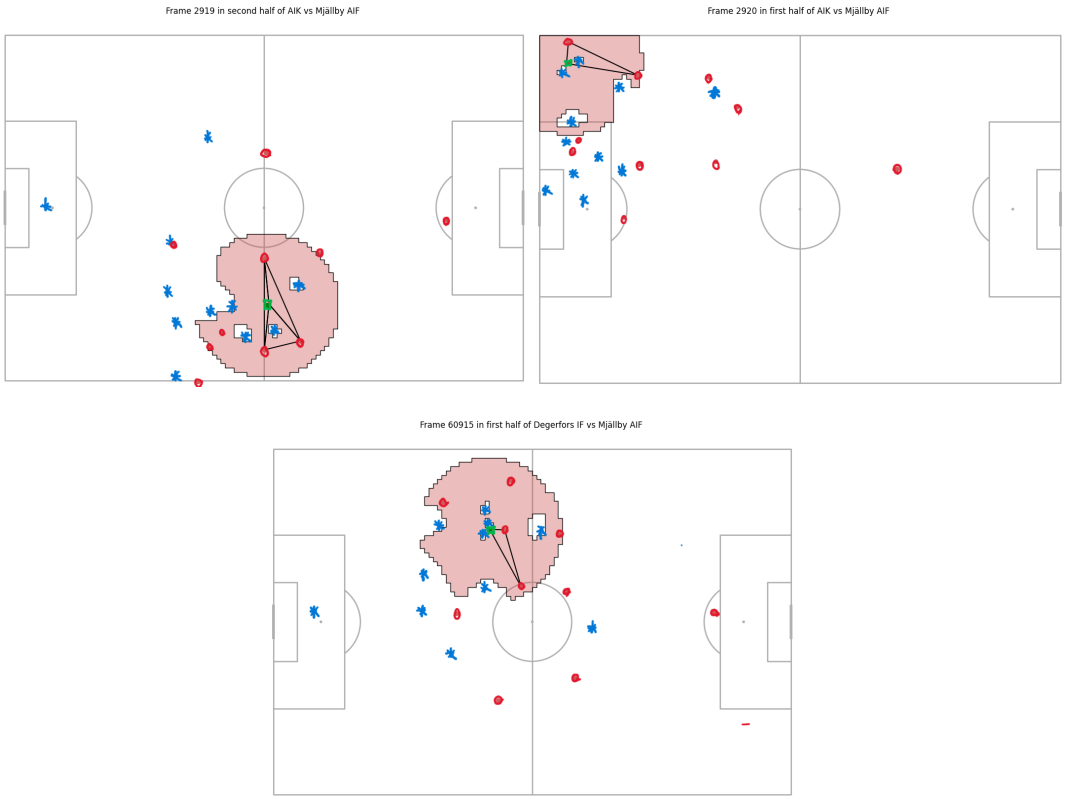


Figure 6. Configuration clusters leading to possession loss (so having edges leading to the LTB node).

In each cluster the ball carrier is mostly surrounded by opposing blue players, and the radius of the passing cliques and passable region areas in these clusters are relatively small. Based off this it appears the defence is making a collective aggressive movement, closing in on the player with the ball and moving up-field to cover opposing players in less dangerous positions, but closer to the ball. This strategy is particularly visible in the last configuration pictured. The blue team's back line has moved up halfway to midfield despite the ball being on their side of the field. The right back in particular is very visibly leaving space between his defensive assignment on the weak wing, in order to provide added pressure on the ball. In addition, the blue team has three players closing in on the ball, taking away any forward passing option. This strategy is aggressive and can lead to an opponent's scoring chance if not timed properly, as it leaves the wide areas very vulnerable. However, we know this cluster is one of the most probable to reach the LTB node in the next progression, indicating that the blue team was mostly successful in pressing high on the ball.

The following is the sequence leading from the first configuration cluster of Figure 6 to a loss of possession. This would be represented by a single directed edge from that 1st configuration cluster into the LTB node.

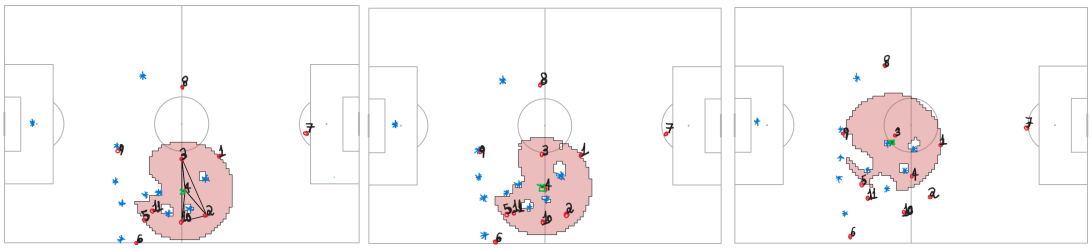
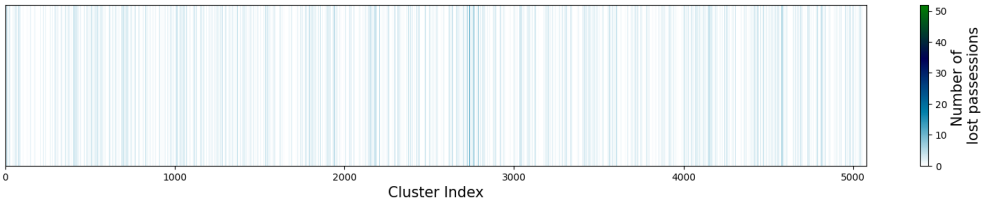
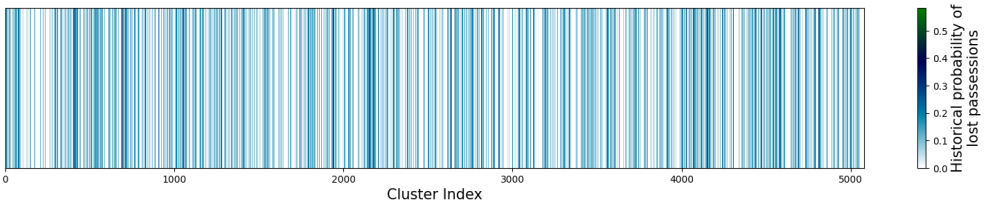


Figure 7. Sequence leading from Figure 6's 1st configuration cluster to possession loss. This sequence is represented by a single directed edge from the sequence's 1st configuration cluster to the LTB node.

We shift to representations of trends in the data, but still focusing in on the LTB node. If in $\mathscr{G}_{\text{safe}}$ we only consider transitions to the LTB node or state, we are left with the transitions in Figure 8. Figure 8a shows the absolute number of transitions from each cluster to the LTB node and Figure 8b shows the transitions to the LTB node per visit to each state. In this example we did not require that nodes have an outgoing valence ≥ 2.



(a) The adjacency matrix of all incoming edges to the Lose The Ball node. In this single-row matrix, a value of v at cluster x indicates that v of the plays in cluster x led to a loss of possession over the course of the entire season.



(b) The adjacency matrix with the probability that a cluster transforms to the Lose The Ball node. A value of v at cluster x indicates that of all the plays in cluster x, we see $v \times 100\%$ of them resulting in a possession loss.

Figure 8. This diagram records clusters occurring directly before a loss of possession, i.e. directly before the LTB node. One can identify likely "problematic" configurations occurring via darker lines.

Question 1. Can one use the safe configuration graph to provide further insights into why certain configurations lead frequently into the LTB node? For example, do configuration cluster nodes with low outgoing valence have high weight on their outgoing edge directed into the LTB node?

3 SAFE CONFIGURATIONS PRECEDING IN-BOX POSSESSION

In addition to the safe configuration graph, one could apply the notion of chains of (clusters of) safe configurations to study progression into the opponent's 18 yard box. We do so in what follows, with an example following the explanation. A cluster was considered to occur directly before an in-box possession if it includes a safe configuration (that occurred 1-10 seconds before the team entered the box with the ball (either a player dribbled into the box, or received a pass in the box). If multiple clusters had such safe configuration data points, we considered the one that was closest in time to when they entered the box. That cluster was called *Cluster 0*. From there, if a cluster had a safe configuration data point that was 1-10 seconds *before* Cluster 0, it was called Cluster 1. This continued, forming chains of configuration clusters occurring directly before the team had an in-box possession. The clusters belonging to these chains are shown in Figure 14. The clusters are organized according to their mean center of mass, where clusters with larger indices correspond to clusters further up the field.

Example 4 (Chain leading into the 18 yard box)**.** This example includes a chain of safe configurations leading into the 18 yard box. Please see Figures 9 - 13. As in Figure 14, we required that the time (in seconds) between each safe configuration was in $[1, 10]$. Red team players are represented by red circles, the blue team by blue stars, and the ball by a green box.

An interesting observation is that in all but one stage of the chain there is a player in the passing clique who is closer to the 18 yard box than the player with the ball. Even though that player is not always passed to, open passes not used can draw the attention of the opposition's defence, opening other options. We now provide a more thorough analysis.

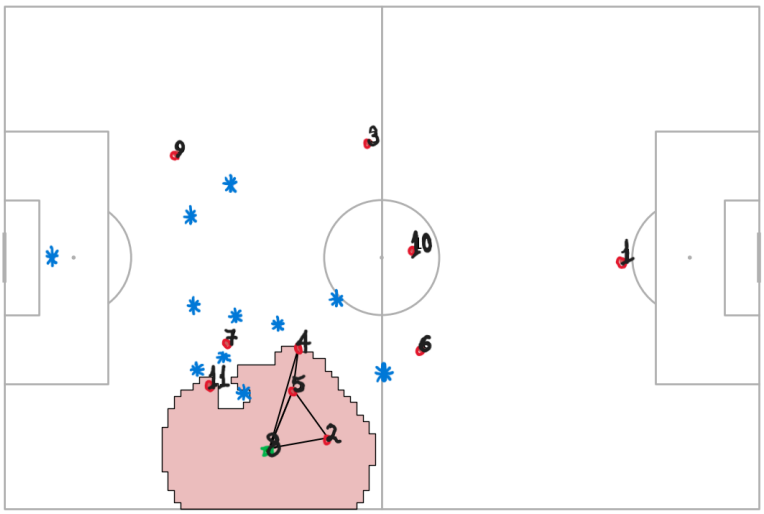


Figure 9. The first safe configuration in the chain.

The possession chain begins in midfield with the ball carrier having 3 passing options (2 contained in the clique). Upon initial analysis it appears that the red team is completing a slow progression up the field, playing possession ball by bringing the majority of their midfield and attacking players close to the ball, opening up maximal passing options. This playing style is highlighted throughout the progression, as the attacking red team only uses short passes and quick movements to get possession in the 18-yard box. It is noteworthy that this offensive playing style requires the whole team to be very skilled at short accurate one touch passes and dribbling, as progressions usually involve most of the attacking players, and the ball must be moved around quickly since the defenders have shorter distances to cover.

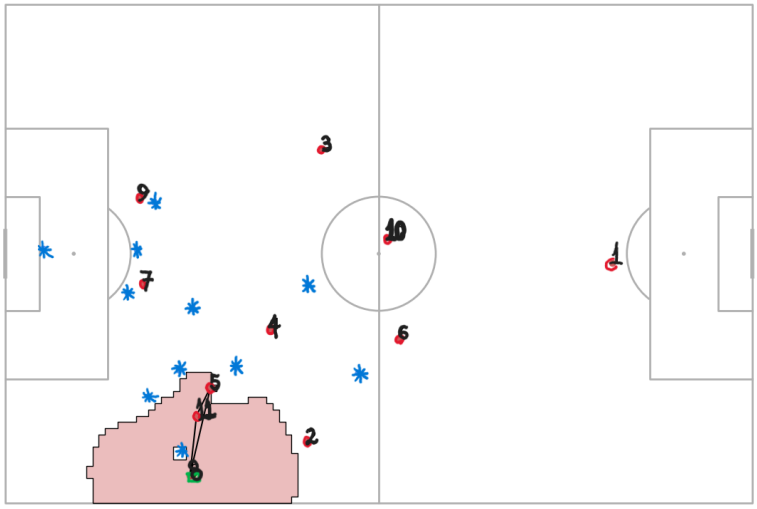


Figure 10. The second safe configuration in the chain.

During the configurations of Figures 9-10, the ball carrier moves down the sideline, drawing pressure from the defender on that side. In the first configuration (Figure 9), there are 6 blue defenders in the area of the passing clique. While this clique is a complete graph, any of the passing options receiving the ball are in a position to be immediately surrounded by 3 defenders.

This clique is not stable (is likely not represented by a node with a loop in $\mathscr{G}_{\text{safe}}$) and so the team needs to shift to another configuration, which they do: To dissipate this concentration of defenders, the furthest passing option of Figure 9 has moved away from the ball into the middle of the field, out of range of the passing clique, but successfully creating space for the other passing options in that area.

This extra space proves beneficial, as the ball carrier passes to one of the open players from the configuration of Figure 10, who now has space to run, look for a ball into the box, or continue to move it around the outside. With the extra space the ball carrier waits and allows for other attacking players to make dangerous runs into the box.

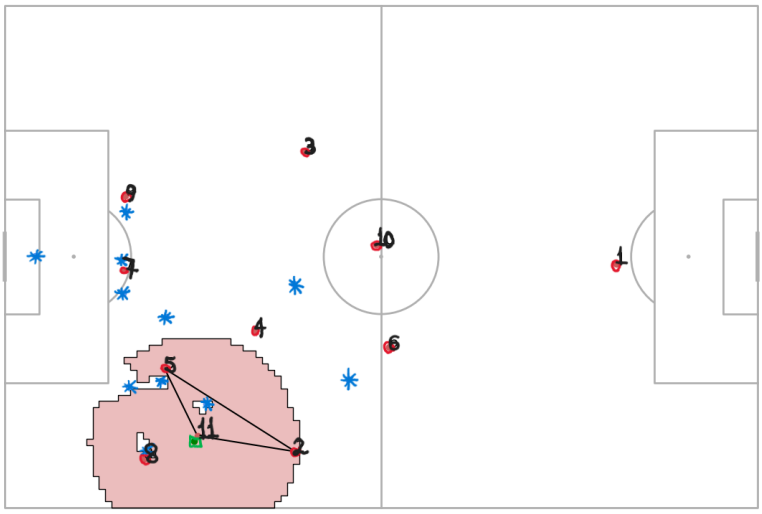


Figure 11. The third safe configuration in the chain.

In Figure 12 a defender pressures the ball carrier leaving space behind him on the edge of the box exposed. The player who just passed off the ball takes advantage and runs into the space, receiving the ball just inside the box and with their momentum in the direction of the goal as shown in Figure 13. This usually generates a high quality scoring opportunity, as the ball carrier is in range of the net, and has multiple passing options nearby.

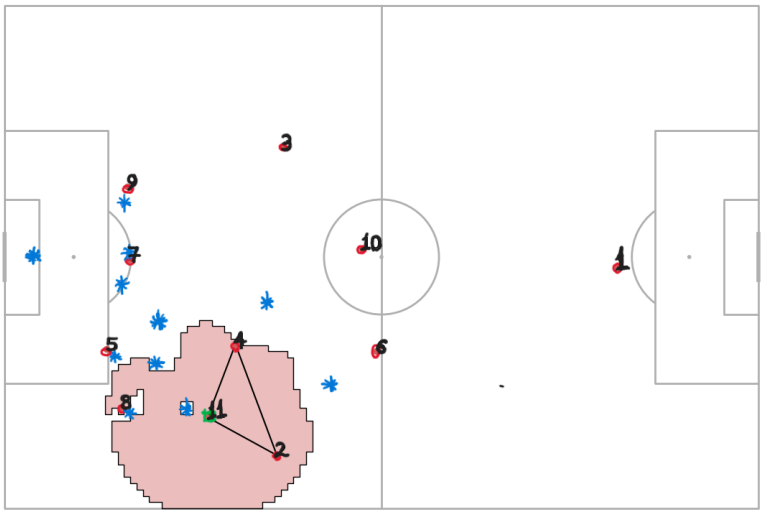


Figure 12. The fourth safe configuration in the chain (i.e. the safe configuration occurring directly before an in-the-box possession).

The attacker who started behind the ball carrier in Figure 12 also makes a run down the sideline into the space created by the player who received the ball in Figure 13, giving the passer another dangerous option.

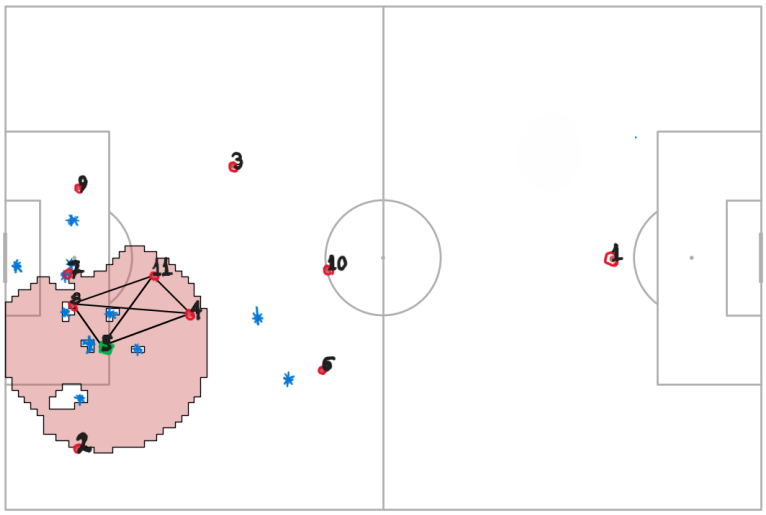


Figure 13. The fifth safe configuration in the chain (i.e. the first of the chain with possession in the box)

It is noteworthy that the constant movement of all players in this progression is what makes it so effective. When playing within a small radius of passing options, the opponent's defensive shape tightens as more passes are made, hence movement is required to keep breaking the defensive shape and cause miscommunication on defensive assignments. This rapid progression through different configurations would be visible in the safe progression graph $\mathscr{G}_{\text{safe}}$ as a path without loops.

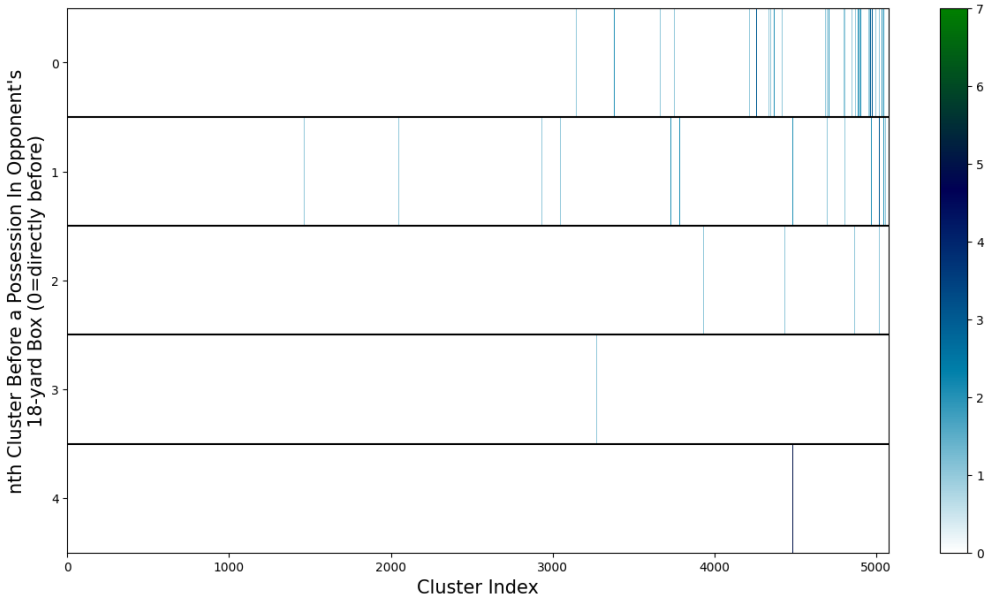


Figure 14. Visual representation of the clusters occurring leading up to a possession in the opponent's box. The y-axis corresponds to the temporal order of the cluster, where $y = 0$ corresponds to a cluster occurring directly before an in-box possession. The x-axis corresponds to the same clusters as Figure 8.

4 DATA LIMITATIONS AND FUTURE POSSESSION PROBABILITY

Many of the most interesting tests and applications of our methodology involve studying chains of configurations (i.e. paths in the safe configuration graph). For example, it would be interesting and valuable to study the frequency of progressions from the defensive to attacking thirds arising from paths

in $\mathscr{G}_{\text{safe}}$ (i.e. from sequences of safe progressions). It would further be interesting to compare the value (computed in various ways) of paths in $\mathscr{G}_{\text{safe}}$ or into the 18 yard box, or of player configurations. In §5 we discuss some such means for valuing (sequences of) player configurations.

4.1 Data Limitations

The position of the ball, and then consequently also the team in possession, were missing from large sections of the data (in particular, approximately 36% of the frames were missing the ball location). Please see Figure 15 for a visualization of how frequently the ball location was missing and for how long. Unfortunately, without the ball location, we could not determine the passing cliques in the moments of time represented by these frames. As higher quality tracking data becomes more available, one could (for example) better track the efficacy and ubiquity of the use of paths in the safe progression graphs, as well as get a more accurate computation of the graphs. Unfortunately, any reasonable such analysis would have required more reliable data. We explain in this section the limitations of the data we had access to and one test we did perform.

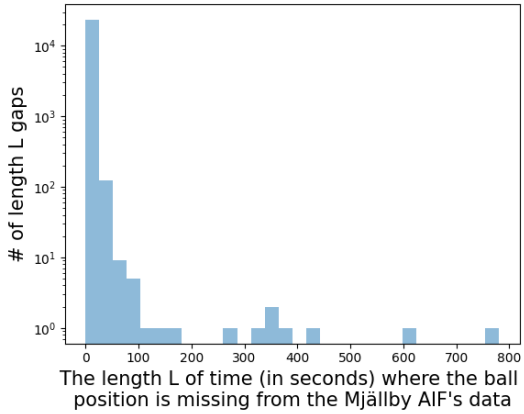


Figure 15

4.2 Future possession probability

As explained in §4.1, the data we had access to at the time had limitations that made studying sequences of configurations particularly difficult. We hope to see the interested reader with access to less "wonky" data perform some of this evaluation. In this meantime, we here provide some evidence that safe player configurations, i.e. those with passing cliques, have a higher probability of maintaining possessions than player configurations without passing cliques.

For our test, we considered the following 3 sets of selection criteria for selecting a random frame:
Criteria 1: possession is defined[1] and the team in possession has a passing clique,
Criteria 2: possession is defined and the team in possession does not have a passing clique, and
Criteria 3: possession is defined with no preference for existence of a passing clique.

For each set of selection criteria, we took 3000 samples of random frames satisfying the selection criteria. To select each sample, we first randomly selected a game, then randomly selected a frame from that game which satisfied the given criteria.

For each sample frame, we identified the team in possession as Team 1, and the opposition as Team 2. We then looked at a t-second window, starting T seconds in the future (where T ranges from 1-39 and $t \in \{1,2,3,4\}$), and logged it as a success/failure according to the following scheme:
• If Team 1 had possession at least once in the t-second window, and Team 2 never had possession in that window, then the sample was logged as a success.
• If the Team 2 had possession at least once during the t-second window, and Team 1 never did, then the sample was logged a failure.
• If both teams had possession during the t-second window, or if possession was undefined for the entire window, the sample was rejected and not counted as one of the 3000 samples taken.

[1] From our observations, it appeared the possession was considered undefined when either the ball location was missing from the data or there was no clear ball possessor.

The columns in the following tables represent:
time (s): the parameter T, in seconds,
diff%: (success rate for a random frame with a clique) - (success rate for a random frame without a clique) divided by (the success rate for a random frame without a clique)

The results of this test are displayed in Table 2 - Table 5. The main positive result is that "diff%" is consistently positive, indicating that the existence of the clique improves one's chances of having possession in the future.

Table 1. Future Possession Probability, with 1-4 second windows (full tables in the appendix)

time T (s)	t=1 diff %	t=2 diff %	t=3 diff %	t=4 diff %
1	-0.612	0.358	0.288	2.346
3	-1.693	1.297	-0.395	0.197
5	-0.924	0.500	0.042	1.354
7	0.350	4.014	1.455	2.118
9	1.134	0.555	0.140	3.603
11	2.748	0.287	1.812	0.624
13	5.497	1.670	6.074	2.869
15	4.573	4.146	6.920	5.882
17	8.481	5.016	5.398	5.692
19	6.171	4.348	4.729	3.243
21	7.200	9.218	8.311	5.504
23	9.858	5.543	4.087	6.798
25	9.918	4.030	3.555	8.424
27	4.654	3.970	8.030	1.078
29	6.775	5.430	3.580	3.642
31	7.816	-0.426	2.542	4.287
33	6.224	3.889	6.235	3.195
35	1.641	4.603	0.063	2.222
37	4.290	9.638	5.584	0.588
39	-1.136	5.729	4.400	-2.295

One may be surprised that the difference between success rates with and without cliques stays positive even after 40 seconds. One possible explanation is that a team that had one clique is more likely to have another during those 40 seconds, leading to an increased likelihood of maintaining possession (based on the statistics for lower T values).

5 FUTURE RESEARCH AND POTENTIAL COACHING APPLICATIONS

We discussed above possible future research in comparing techniques for clustering configurations. We focus here on four (of many) possible applications of the constructions of this manuscript. Regrettably, a satisfying study of these applications would require more complete data. In what follows, by a *progression path* we will mean either a path through $\mathscr{G}_{\text{safe}}$ from the attacking to defensive third or one leading into the attacking 18 yard box, depending on the focus at the time. All nodes and edges referred to are in $\mathscr{G}_{\text{safe}}$. In particular, the nodes will always be (clusters of) safe configurations.

I. Determine configurations of players to prioritize practising moving into, between, and from:

We present two evaluation metrics for configurations:

1. For a node C, we define the strategic frequency $F(C)$ as

$$F(C) := \sum freq(p) \tag{2}$$

where p ranges over the progression paths containing C as a node and $freq(p)$ is the number of times the team used that path. Thus, $F(C)$ records how key the configuration C is to the progression of a particular team during an attack.

2. One could also define the *Attacking Configuration Robustness Index (ACRI)* for a node A in a safe progression graph G, denoted $ACRI(A)$, as the sum of the weights of all out-going edges at A that do not terminate in the LTB node (i.e. in a loss of possession):

$$ACRI(A) := \sum_{B \in V(\mathscr{G}_{\text{safe}})} w(A,B) \tag{3}$$

The inspiration behind the ACRI is as follows. Given a vertex representing a configuration C of attacking players, an outgoing edge (from C to C') of high weight represents a likely transformation from C to C'. Thus, a vertex C having high weighted outgoing valence indicates that the corresponding configuration C likely has multiple good options for transforming into other "safe configurations." By removing the weights and simply looking at the outgoing valence, a large ACRI would indicate multiple potential means of progressing out of the configuration, which it may not be possible to simultaneously defensively prevent or "contain."

II. Determine new and "optimal means of progressing from the defensive to attacking third:

Viewing safe configurations in the defensive third as start states and in the attacking third as end states, paths in $\mathscr{G}_{\text{safe}}$ from start to end states are (possible) sequences of safe configurations leading from the defensive to attacking third. Any such sequence not previously used is a new potential strategy. One can place a value on such a path/strategy as follows (note, one can similarly assign a value to configuration sequences leading into the 18 yard box). We assign to each edge in $\mathscr{G}_{\text{safe}}$ a "cost" computed by:

a. Multiply the probability of possession loss during an attempt at the edge's configuration shift by the difference between the probabilities the opposition and attacking team will score from the opposition gaining possession in that situation. Then

b. multiply the value computed in (a) by a function (yet to be determined) that reflects the impact on progression success of the time it takes to transition between the configurations that are the endpoints of the edge.

The cost of a path will be the sum of the costs of the edges it traverses. And a low cost will be considered preferable, as it indicates a means of progressing from the defensive to attacking thirds with minimal risk of the opposition profiting from its attempt.

Without access to the necessary information for such a valuation, one could possibly use the robustness index of the nodes, as described in (I) above to place a high value on paths that at each stage have a strong likelihood of success.

III. Determine configurations and patterns of play possibly to be avoided:

We anticipate teams preferring to avoid configurations X frequently leading to a loss of possession, i.e. those with a high $E(X,LTB)$ weight (equivalently low weighted $ACRI(X)$ value). Those configurations X with low $F(X)$ value would also have lower flexibility, and thus likely be less useful and reliable.

IV. Construct effective defensive strategy/training:

We also anticipate coaches, from a defensive strategy viewpoint, focusing on disrupting (or avoiding the formation of) configurations frequently used by one's opponent(s) in successful progressions, i.e. those configurations of high $F(C)$ values.

6 APPENDIX: COMPUTATIONS

We computed the times for all players and the ball to reach a particular location (x,y) using an initial velocity and constant deceleration for the ball, as well as estimates of acceleration and the maximum velocity for a player. As we found it did no have a significant enough impact on our computations, we left out consideration of the time it takes a player to change direction (turn). We could then determine whether it would be possible for a player to reach this location (x,y) before the ball. When the model is applied to a moment in the game we obtain images as seen in Figure 1.

Since we found little supporting research for the physics parameters involved in passing the ball, we determined the physics parameters as follows (using the data of Mjällby AIF's games as the sample set).

To obtain the ball's initial speed and deceleration, using the start and end time of a pass given in the event data, we:

• estimated the ball speed by the central difference method, and then

• changed the start time of each pass to where it reached its maximal speed, and then

• determined a "good" subset of the passing data by imposing the following 3 condition on passing data:

1. we required that the data of each pass (i.e. the data in the interval from the start time to the end time) contains ≥ 20 frames (0.8s) and

2. required that the data satisfies smoothness, i.e. that the speed difference between any 2 adjacent frames is always < 2 m/s and

3. required that the average speed during the 1st half of the data during a pass is larger than during the 2nd half.

The initial ball speed is then taken to be the average of the set of maximal speeds of the sets of "good" passing data (there is at most 1 "good set" for each pass and then 1 maximal speed for each "good set", we average these maximal speeds). For the deceleration, we use linear regression on the "good" passing data and take the slope of a fit as the deceleration of a pass. The ball deceleration is taken as the average deceleration of all of the "good" passes. Through this methodology we obtain an initial ball speed of 16.20 m/s and ball deceleration of 7.93 m/s^2.

We now explain how we found the maximal speed and maximal acceleration of a player that we use in our computations. We use the values of a maximal speed of 10 m/s^2 and a maximal acceleration of 4.98 m/s (taken from [11]), as they are close to those we determined using the following methodology. We use the central difference method to estimate the acceleration of a player (the speed of a player is taken from the data). The speed and acceleration in all Mjällby AIF's games is plotted and we obtain that the top 1% of the accelerations are above 4.779 m/s^2, and the top 0.5% of the accelerations are above 5.718 m/s^2. The maximal speed is 11.13 m/s and the top 1% of this speed is above 7.12 m/s.

Finally, the horizontal distance covered by the ball while in contact with the foot is approximately 2/3 the diameter of the ball [2]. As the average diameter of a soccer ball is $22cm$, this means that the ball travels 14.67cm while in contact with the foot.

Supposing the ball must travel a total distance of d, we set $d_r = d - 0.1467$m as the remaining distance to be travelled after the ball leaves the foot. The time t_b taken for the ball to travel the remaining distance d_r is then calculated via the 1-dimensional calculus computation given below with the given initial velocity v_b and deceleration $-a_b$ computed above. The time is set to infinity if the ball's velocity reduces down to zero before travelling the distance of d_r. The time is set to 0 if $d < 1$m.

We now give the calculus computations we used. We let $d_{remaining} = max(d - 0.1467,\ 0)$ and then define

$$\Delta := v_b^2 + 2a_b d_{remaining} \tag{4}$$

The time required for the ball to reach the given location was then:

$$t_b = \begin{cases} 0 & d < 1 \\ \frac{\sqrt{\Delta} - v_b}{a_b} & \Delta \geq 0 \text{ and } d \geq 1 \\ \infty & \text{otherwise} \end{cases} \tag{5}$$

Next we calculate the time required for a player to reach a particular location on the field. We assume that the player turns around and then moves toward the final location $\vec{r}$. The initial velocity $\vec{v_0}$ of a player can be taken from the the tracking data. We set the velocity toward the final location to be

$$v_{toward} = ||\vec{v_0}|| \cos\theta$$

where θ is the angle that the player needs to turn. The time t_p taken for the player to travel is then calculated via the 1-dimensional calculus computation of the time it takes a player to reach a given distance $||\vec{r}||$ with initial velocity v_{toward} and then the maximal speed $\tilde{v}_p$ and maximal acceleration $\tilde{a}_p$ computed above. We set the time to be 0 if $||\vec{r}|| < 1$.

We defined v_{max} as

$$v_{max} = max(v_{toward},\ \tilde{v}_p) \tag{6}$$

and then set

$$d_{max} = \frac{(v_{max})^2 - (v_{toward})^2}{2\tilde{a}_p}. \tag{7}$$

The time required for the player to reach the given location is then

$$t_p = \begin{cases} 0 & ||\vec{r}|| < 1 \\ \frac{\sqrt{(v_{toward})^2 + 2\tilde{a}_p||\vec{r}||} - v_{toward}}{\tilde{a}_p} & 1 \leq ||\vec{r}|| < d_{max} \\ \frac{v_{max} - v_{toward}}{\tilde{a}_p} + \frac{||\vec{r}|| - d_{max}}{v_{max}} & \text{otherwise} \end{cases} \tag{8}$$

REFERENCES

[1] M. Ankerst et al. "OPTICS: Ordering Points to Identify the Clustering Structure". In: *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. SIGMOD '99. Philadelphia, Pennsylvania, USA: Association for Computing Machinery, 1999, pp. 49–60. ISBN: 1581130848. DOI: 10.1145/304182.304187. URL: https://doi.org/10.1145/304182.304187.

[2] T. Asai et al. "The curve kick of a football I: impact with the foot". In: *Sports Engineering* 5.4 (2002), pp. 183–192. DOI: https://doi.org/10.1046/j.1460-2687.2002.00108.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1460-2687.2002.00108.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1460-2687.2002.00108.x.

[3] R. Fisher et al. *Image Processing Learning Resources: Gaussian Smoothing*. https://homepages.inf.ed.ac.uk/rbf/HIPR2/gsmooth.htm. 2003.

[4] S. Rudd. "A framework for tactical analysis and individual offensive production assessment in soccer using Markov chains". New England Symposium on Statistics in Sports. 2011. URL: https://www.metacafe.com/watch/7337475/2011_nessis_talk_by_sarah_rudd/.

[5] W. Spearman. *Quantifying Pitch Control*. Feb. 2016. DOI: 10.13140/RG.2.2.22551.93603.

[6] W. Spearman et al. "Physics-Based Modeling of Pass Probabilities in Soccer". In: *Proceedings of the 11th MIT Sloan Sports Analytics Conference*. Mar. 2017.

[7] U. Brefeld, J. Lasek, and S. Mair. "Probabilistic movement models and zones of control". In: *Machine Learning* 108.1 (2019), pp. 127–147.

[8] T. Decroos et al. "Actions speak louder than goals: Valuing player actions in soccer". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 1851–1861.

[9] U. Dick and U. Brefeld. "Learning to rate player positioning in soccer". In: *Big data* 7.1 (2019), pp. 71–82.

[10] J. Fernandez, L. Bornn, and D. Cervone. "Decomposing the Immeasurable Sport: A deep learning expected posession value framework for soccer". In: *Proceedings of the 13th MIT Sloan Sports Analytics Conference*. Mar. 2019.

[11] I. Loturco et al. "Maximum acceleration performance of professional soccer players in linear sprints: Is there a direct connection with change-of-direction ability?" In: *PLoS ONE* 14.5 (May 2019). DOI: https://doi.org/10.1371/journal.pone.0216806. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0216806.

[12] A. Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[13] J. Beernaerts et al. “Spatial movement pattern recognition in soccer based on relative player movements”. In: *Plos One* 15.1 (Jan. 2020). DOI: 10.1371/journal.pone.0227746.

[14] J. Fernandez, L. Bornn, and D. Cervone. “A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions”. In: *arXiv preprint arXiv:2011.09426* (2020).

[15] M. Kullowatz. *Goals Added: Deep dive methodology*. Accessed: 2022-11-14. 2020. URL: https://www.americansocceranalysis.com/home/2020/5/4/goals-added-deep-dive-methodology.

[16] F. R. Goes et al. “The tactics of successful attacks in professional association football: large-scale spatiotemporal analysis of dynamic subgroups using position tracking data”. In: *Journal of Sports Sciences* 39.5 (2021), pp. 523–532.

[17] M. Van Roy et al. “Proceedings of the AI for sports analytics (AISA) workshop at IJCAI 2021”. In: (2021).