Improving the characterization and comparison of football players with spatial flow motifs

Alberto Barbosa[1,2], Pedro Ribeiro[1,2], and Inês Dutra[1,3]

[1] Departamento de Ciência de Computadores, Faculdade de Ciências
Universidade do Porto, R. Campo Alegre s/n, 4169–007 Porto, Portugal
[2] CRACS & INESC-TEC, Portugal
[3] CINTESIS, Portugal
alberto.barbosa@fc.up.pt, pribeiro@dcc.fc.up.pt, dutra@fc.up.pt

Abstract. Association Football is probably the world's most popular sport. Being able to characterise and compare football players is therefore a very important and impactful task. In this work we introduce spatial flow motifs as an extension of previous work on this problem, by incorporating both temporal and spatial information into the network analysis of football data. Our approach considers passing sequences and the role of the player in those sequences, complemented with the physical position of the field where the passes occurred. We provide experimental results of our proposed methodology on real-life event data from the Italian League, showing we can more accurately identify players when compared to using purely topological data.

Keywords: sports analytics, subgraphs, network motifs, spatial data

1 Introduction

Association football, also known as soccer or simply football, is probably the world's most popular sport. It is therefore of no surprise that there has been an ever growing interest on collecting and analysing football data in order to inform players, coaches and management staff, trying to gain a competitive edge. Examples of related computer science research include a vast array of topics, such as team behaviour visualisation [16], talent discovery [3], injury forecasting [12], result prediction [11] or transfer market analysis [6].

In this work our focus is on providing a similarity metric for comparing football players in what concerns their role in the dynamics and passing behavior of the team. This is already useful as a rich characterization tool and could be further applied for instance to come up with suggestions of similar players in other teams that could potentially be good transfer targets.

Our main contribution is the concept of spatial flow motifs and a novel hybrid similarity metric, that incorporates both when and where passes occur in the game. We partition the football field into regions and we use temporal data to construct passing sequences that can be seen as small subgraphs where nodes are classified according to the region of the corresponding passing event. We also present experimental results on real life event data from the Italian League, showcasing how these spatial features can enrich and complement purely topological information, achieving a higher accuracy on the player comparison task.

2 Related Work

The amount of research work related to football analytics is too vast to be included in this paper [13,14]. Here, we will mainly focus on research that delved into studying motif based patterns in passing networks.

A passing network can be seen as a graph where the nodes represent players and directed edges represent successful passes between two players. Milo et al. defined network motifs as "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomised networks" [7]. Later, Gyarmati et al [4] defined flow motifs. Considering a passing sequence, a flow motif is a subsequence of the passes where labels represent distinct players without identity. In the context of this paper, all motifs are flow motifs and we will use both terms interchangeably. We will next make a short description of previous research on this topic.

The application of network motif methodology to football data has been a recent research topic among the fields of network science and sports analytics.

Using network motif methodology, Bekkers and Dabadghao [2] identified unique play styles for teams and players. Peña et al. [10] also applied network motif and clustering techniques to football data from the Premier League, La Liga and Champions League, concluding that Xavi Hernandez was the outlier in their analysis. Wiig et al. [15] use centrality measurements and PageRank to identify key passers and/or recipients in football teams.

Håland et al. [5] modelled sequences of passes as flow motifs and concluded that no connection between the ranking of a team and its distribution of flow motifs was clear.

Regarding team behaviour, Gyarmati et al. [4] present a quantitative method to characterise the passing behaviours of football teams, concluding that some unique styles of play do not consist of uncountable random passes but instead are finely structured.

In a previous work [1], we studied player similarity based on the topology of the passing networks. Flow motifs were extracted from sequences of passes that involved each player and the conclusion was that taking into account the specific position of the player in a motif, i.e. the orbit that represents a given player in a motif, was better in comparing players than just comparing the number and type of motifs they were a part of. Also, we provided a way to objectively measure the performance of the models generated, which is a great complement to the almost uniquely visual analysis that is made to evaluate player similarity algorithms.

Even though some work has been done in exploring the spatial dimension of football games [8], to the best of our knowledge there is no framework that entails the characterisation of the passing behaviour of a player through the study of spatial flow motifs.

3 Data Description

The data used in this work was retrieved from a public data set containing spatio-temporal events in association football [9]. We will be using event data from the 2017/2018 season of the top tier Italian league.

Events in the data were transformed into sequences of passes between different players. Players that did not participate in, at least, 80% of the games in the 2017/2018 season were not considered in our analysis. We excluded these players to guarantee that we had the most consistent and complete data as possible regarding the players we analysed.

4 Methodology

After pre-processing the data, the first step of our method was to extract spatial flow motifs from the generated passing sequences, as it is explained in this section. Our code is available at `https://github.com/BertoBoss/SpAn`.

In the context of this work, we will compute and count "flow motifs", as defined in [4], even when referring to them simply as motifs. We note however that flow motifs are not "network motifs" as classically defined in [7] since they do not take into account the statistical significance of each of the sequences extracted. This is mainly due to the lack of an appropriate null model for football spatial motifs, which could be an interesting future line of work.

We start by overlapping a grid over the space in which the events occurred. This grid will divide the space in m equal parts lengthwise and in n equal parts heightwise. So there will be $m \times n$ different rectangles in which an event can occur. In this work, we will only consider the case when $m = n = 3$, so there will always be $n^2 = 9$ different rectangles. All different cells originated by the divisions of the grid will be numbered in ascending order bottom-up from left to right, the smallest number being in the bottom cell of the leftmost column and the highest number in the top cell of the rightmost column.

The insertion of each sequence of passes in the grid is done by inserting each pass according to the coordinates in which it started, i.e., the coordinates of the origin of the pass. The only exception is the last point in the sequence that represents where the last pass ended. This will yield a grid where each cell contains a non-negative number of points and each point represents where a given pass was made, except for the last one that represents where the final pass ended. The points are inserted by order of occurrence in the game they happened and preserving that order is important, since we are studying the evolution of a passing sequence between members of the same team. We do so by storing such information in an auxiliary list for each grid we generate. An example of a grid filled with a size 3 passing sequence is presented in Figure 1.

Each of these generated grids now represents a spatial flow motif. For simplicity purposes, we decided to generate a string that represents each of these motifs in an unique way, so that each of these grids can be easily and uniquely identified by it. Such string needed to embody three crucial characteristics of

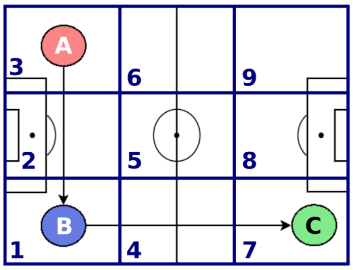

Fig. 1: Grid representing a passing sequence involving three players. First pass goes from player A to player B from cell 3 to cell 1. The second pass goes from player B to player C from cell 1 to cell 7. This sequence represents a 3A.1B.7C spatial flow motif.

these motifs: the topology, the relative order of occurrence of each pass and the space on the pitch where it occurred. So we transform a sequence into a string of the kind $K_1C_1.K_2C_2.\cdots K_N.C_N$, where each K_i is an integer representing the cell of the grid in which the pass it represents occurred, C_i is a character representing the player that made that pass, consistent with the definition of a flow motif, and the order in which the passes occurred is preserved by the order of appearance on the string from left to right. For the passing sequence in Figure 1, the corresponding string would be **3A.1B.7C**. For simplicity sake, whenever we want to make a reference to a single flow motif, we will use the string method and only present the visual aid if absolutely necessary.

For each player in our data set, we transformed every passing sequence he participated in into a spatial flow motif, using the methodology described above.

According to [1], the orbit a player occupies in each flow motif is very important in characterising the passing style of a player. In order to encapsulate that concept, we also consider the specific position a given player occupied in the passing sequence. Using the example in Figure 1 and extending the concept of spatial flow motifs to include orbits, we would say that player A participated in a ABC_A motif topologically and in a $3A.1B.7C.A$ spatial flow motif. A similar extension can be made with respect to nodes B and C.

The spatial flow motif concept as we present it here can be seen as a good complementary analysis to the purely topological flow motif analysis. Now, we will not only be able to see which types of passing plays a player is involved in more often, but also the areas on the pitch in which those plays tend to occur.

We will now exemplify how the inclusion of spatial data in motif extraction can give us better insight regarding the passing behaviour of each player. Unfortunately, due to the quantity of players analysed, it is not possible to perform this analysis for each player in our data set. Nonetheless, we will briefly analyse the passing behaviour of Jorginho, not only topologically, but also spatially.

Regarding topology, we present in Figure 2 a plot representing the relative frequency of the participations of Jorginho in different types of passing sequences.

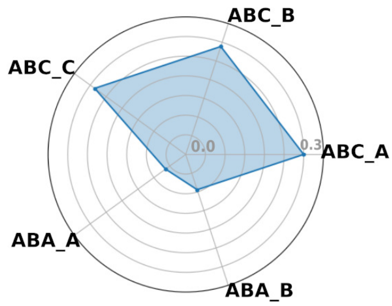

Fig. 2: Fingerprint of Jorginho passing behaviour with topological flow motifs

We can see that the player tends to be more involved in ABC types of plays (involving three different players) rather than in ABA (involving only two different players). Moreover, we can see that Jorginho seems to have a somewhat balanced participation among ABC plays, participating in similar quantities in all three possible roles of the play. This seems to go along with the intuition that a midfielder like Jorginho, being a playmaker, is involved in all stages of the offensive game of his team. Also, we can see that Jorginho, when participating in ABA motifs, tends to occupy the B position more often, which means that he tends to be the one that receives the pass from some player A and then passes the ball to that same player A.

Although the study of purely topological flow motifs proved to be a good way to characterise the passing behaviour of a player [1], more information can still be extracted from the same data if we look not only at the passing sequences' topology, but also at the spatial information encapsulated in that data, specially in football, where space is a very important aspect of the game.

To do that, we decided to study the distribution of spatial flow motifs grouped by their topology. We will consider a 3×3 division of the pitch, as presented in Figure 1. From Table 1, we can actually perform a more detailed analysis of the passing behaviours of Jorginho, when compared to the analysis of the frequency of each topological flow motif.

First, one can see that the majority of the highlighted values are on cells that represent the midfield area (cells 4, 5 and 6). Given that Jorginho is a midfielder, it is not surprising to see that a big part of the events he will be involved in will take place on central areas of the pitch. This is however an intuition that we can only confirm when incorporating the space dimension of the data, since it is not possible to do so by only analysing the topological motifs as in Figure 2.

Another important thing to notice is that the relative frequency of C in cells 7, 8 and 9 is lower in ABC_C flow motifs. This can be, at least partially, explained because he is a player that does not step into advanced areas of the pitch that much, given that he is not a forward, so when he is the last player on the passing sequence, that sequence tends to end in not so advanced areas of the pitch, like on cells 4, 5 or 6. This seems to be opposite to the behaviour when Jorginho occupies the A or B positions in ABC motifs, where position C tends to have higher frequency of appearance in cells 7, 8 or 9.

A noticeable aspect of the data in Table 1 is the fact that Jorginho apparently tends to impose some progression on the pitch with his passing. This becomes apparent by the fact that, when he makes a pass, the next player on the sequence tends to have a higher frequency of appearance in more advanced areas than Jorginho himself. For example, in ABC_B motifs, player A tends to be in cell 7 in higher frequencies than Jorginho (player B), which then passes the ball to a player C that, again, has higher frequency in cell 7. This seems to indicate a forward passing bias by Jorginho, meaning that the players that receive the ball from him are often in more advanced positions of the pitch than himself.

On the most defensive areas of the pitch, we can see that the only significant change between the three types of spatial flow motifs is related to the position Jorginho occupies on the passing sequence: when he occupies a given position p, the frequencies of that position on more retreated areas of the pitch are lower for p and higher for the other positions. For example, when Jorginho is on position B on ABC_B motifs, the frequencies of a player being on cells 1, 2 or 3 and on the B position are lower than when he occupies either position A or position C.

It also stands out that there seems to be a slight bias for these plays to winding themselves more on the right side of the pitch than on the left. That becomes more evident when comparing the values of the frequencies on cell 7 (right attacking side) to the values in cell 9 (left attacking side) for ABC_A, ABC_B and ABC_C flow motifs. This can either represent a team behaviour that somehow favours attacking plays to happen on the right side of the pitch when Jorginho is involved in them or this can represent a bias imposed by Jorginho to force the team to play through the right, probably also influenced by the position he occupies on the pitch and the fact that it is more likely that he will pass the ball to players near him. It can also be both, since often players positions and characteristics are highly correlated to the team macro behaviour.

All this domain specific knowledge regarding a single player can be acquired only when we combine the topological information with the spatial information contained in the raw event data.

5 Results

With the analysis of the spatial flow motifs Jorginho is involved in, we intend to show that incorporating an extra layer of spatial information in flow motif counting can result in better and more accurate knowledge extraction from event data. However, when we want to objectively measure if new useful information can be extracted from the processing of spatial data, we need to setup an experimental environment that allows us to take valid conclusions about our methodology.

In an ideal world, we would be able to not only check if the addition of spatial information results in a good complement to the purely topological information that flow motifs naturally provide and have a way to measure how good such potential complement is.

	ABC_A			ABC_B			ABC_C		
Cell	A	B	C	A	B	C	A	B	C
1	0.051	0.061	0.054	0.089	0.045	0.051	0.086	0.082	0.047
2	0.055	0.048	0.048	0.077	0.055	0.038	0.073	0.072	0.043
3	0.033	0.042	0.037	0.058	0.027	0.028	0.064	0.061	0.030
4	**0.299**	**0.251**	**0.238**	**0.245**	**0.307**	**0.220**	**0.244**	**0.264**	**0.297**
5	**0.287**	**0.189**	**0.141**	**0.142**	**0.279**	**0.197**	**0.128**	**0.131**	**0.279**
6	**0.158**	**0.133**	**0.129**	**0.160**	**0.159**	**0.112**	**0.148**	**0.157**	**0.150**
7	0.053	**0.155**	**0.176**	**0.126**	0.058	**0.197**	**0.147**	**0.123**	0.065
8	0.040	0.057	0.079	0.035	0.042	0.085	0.023	0.032	0.048
9	0.024	0.066	0.098	0.069	0.028	0.072	0.088	0.077	0.041

Table 1: Frequency of each player (A, B or C) on each cell on a 3×3 grid according to the different topological motif they participate in. Values are normalised by sub-column, meaning that, for example, in ABC_A motifs, player A (Jorginho) occupies a position in cell 1 0.051% of the times he participates in an ABC motif occupying position A. Highest values are on bold. Sub-columns representing Jorginho have a highlighted background.

We decided to build a similar experimental environment as the one on [1], but we have adapted it to also account for spatial flow motifs.

We first separate the matches in two different sets: one corresponding to the first half of the season and the other corresponding to the second half of the season (we also experimented a division into odd and even match weeks, but no relevant differences were found).

The idea behind this division is to be able to see how similar a player is to himself in different parts of the same season. Given that there is no ground truth in this domain, we believe that a good way to validate our approach is to exploit the idea that a player, of course with some exceptions, must have similar passing behaviour during the course of the same season. To achieve that, we designed a distance metric to measure the distance between two players, incorporating not only the topological difference between the different motifs that a player was involved in, but also the spatial dimension of the data.

After computing all the spatial flow motifs for every player on the dataset, we calculate the distance between each pair of players. The distance metric has two components: one incorporates the topological component of the flow motifs and the other incorporates the spatial dimension of the flow motifs.

The purely topological distance between two players is defined in Equation 1, where $D_{top}(A, B)$ represents the distance between players A and B according to the purely topological flow motifs that A and B participate in. M is the set of all flow motifs of size 3 and A_m and B_m represent the normalised frequency (between 0 and 1) of player A and B on motif m, respectively.

$$D_{top}(A, B) = \sqrt{\sum_{m \in M} (A_m - B_m)^2} \tag{1}$$

A big part of adding the spatial dimension is the ability to measure how distant two different sequences are in the pitch. We use cell centroids in order to compute the distance between two different spatial flow motifs. The centroid of a cell k in a grid is a point whose coordinates are $(min_x + max_x/2, min_y + max_y/2)$, where min_x (min_y) is the minimum value of x (y) that a point in a cell k can have and max_x (max_y) is the maximum value of x (y) that a point in a cell k can have. Then we calculate the distance between two motifs m and n by calculating the Euclidean distance between the centroid of the cell in which the first pass occurred in motif m and the centroid of the cell in which the first pass occurred in motif n. We then add it to the distance between the centroid of the cell in which the second pass occurred in motif m and the centroid of the cell in which the second pass occurred in motif n, and so on, until the motif is fully processed. In the context of this paper, we will call this distance $D_{centroids}(m, n)$.

The component of our metric that encapsulates the spatial information of the flow motifs is given in Equation 2. M represents the set of all flow motifs of size 3, M_k is a set of motifs that are topologically equivalent between themselves, m and n are two topologically equivalent spatial flow motifs, $D_{centroids}(m, n)$ calculates the Euclidean distance of the centroids of the cells in which the passing sequences occurred, $f_A(m)$ represents the frequency of player A in motif m and $f_B(n)$ represents the frequency of player B in motif n.

$$D_{space}(A, B) = \sum_{M_k \in M} \sum_{m,n \in M_k} D_{centroids}(m, n) * f_A(m) * f_B(n) \tag{2}$$

It is important to notice that Equation 2 is a distance metric thought to complement a merely topological setting, by trying to encapsulate some spatial information that would otherwise be discarded by a purely topological approach. Since both $0 \leq f_A(m) \leq 1$ and $0 \leq f_B(n) \leq 1$, this allows for the weight of the value of $D_{centroids}(m, n)$ to be, in some sense, proportional to the frequency of occurrence of motifs m and n in players A and B, respectively.

One could argue that a different approach, like using $f_A(m) - f_B(n)$, would encapsulate better the idea that "the higher the difference between the frequencies, the higher the weight the distance would have". Even though the intuition is correct, using the difference between the frequencies of each motif when complementing purely topological motif analysis would not take into account the individual frequencies of $f_A(m)$ and $f_B(n)$ for players A and B, respectively. For example, whether $f_A(m) = f_B(n) = 0.1$ or $f_A(m) = f_B(n) = 0.9$ the value of $f_A(m) - f_B(n)$ would be 0. However, using $f_A(m) * f_B(n)$, the second assignment would result in a higher value, that better mirrors the fact that m and n are highly frequent motifs for players A and B, respectively.

The final distance metric is a weighted mean of the two components we approached, as in Equation 3, where α is a constant between 0 and 1, influencing the weight of each component in the final distance metric.

$$Dist(A, B) = \alpha \times D_{space}(A, B) + (1 - \alpha) \times D_{top}(A, B) \tag{3}$$

Now that we have a distance metric that incorporates both topological and spatial information regarding flow motifs, we can use it to calculate the distance between players.

Let H_1 be the set of games that took place in the first half of the season and H_2 the set of games that occurred in the second half of the season. Also, let A_{H_i} be the set of spatial flow motifs that represent player A and were extracted from the set of games in H_i.

Our task will then be to calculate the distance $Dist(A, B)$ between every A_{H_1} and every B_{H_2}, i.e., we want to calculate the distance between each pair of players A and B such that the spatial flow motifs regarding player A occurred in games in the first half of the season and the spatial flow motifs regarding player B occurred in games belonging to the second half of the season.

Let L_A be a list of players such that the position j each player B occupies in the L_A represents that B is the j-th least distant player to A, according to the distance metric in Equation 3.

So, for each player P_1 in A_{H_1}, our job is to compute L_{P_1} such that each player P_2 in L_{P_1} belongs to A_{H_2}. When all L_{P_i} are calculated for every player P_i in A_{H_1}, we can see how well our method characterises a player, in the sense that we just need to measure, for every player P_i, the position P_i occupies in L_{P_i}. The lower the position, the better, since we want P_i to have similar passing behaviour to himself.

In [1], an arbitrary threshold was defined, stating that cases where P_i occupied a position $j \leq 10$ it would be considered a positive case, and a negative otherwise. All positive and negative cases were counted and the evaluation model was simply given by the percentage of positive cases that the model got right.

In this work we decided to extend that idea to a more continuous analysis of the distribution of the positions that each P_i occupies in L_{P_i}. A boxplot of those distributions is presented in Figure 3. Note that in that plot, the values of the constant α range from 0 to 1 in increments of 0.1, with the plot referring to $\alpha = 0$ is the one on the left and each successive plot represents a +0.1 increment.

Analysing the boxplot in Figure 3, it is noticeable that, for some values of α, the distribution is more skewed toward smaller position values than on others. Those values are $0.1 \leq \alpha \leq 0.6$. The mean values for each of the distributions presented in Table 2 confirm that on those values, the mean and standard deviation of the distributions are smaller when compared to $\alpha = 0$, that represents a distribution based on a merely topological distance metric.

The mean value in the distributions presented in 3 represents the average position a player P_i is in L_{P_i}. This means that, for example, with $\alpha = 0.2$, any player P_i is on average the 7th most similar player to himself, which is a really good improvement when compared to the 11th position, on average, that yields from $\alpha = 0$. It is also worth noting that, for $\alpha \geq 0.7$, the distributions seem to exhibit a worst behaviour than considering only the topological nature of the data. That seems to indicate that, on its own, the metric we designed to complement a merely topological study, is not better at characterising the passing behaviour of football players than a classical topologically reliant distance metric.

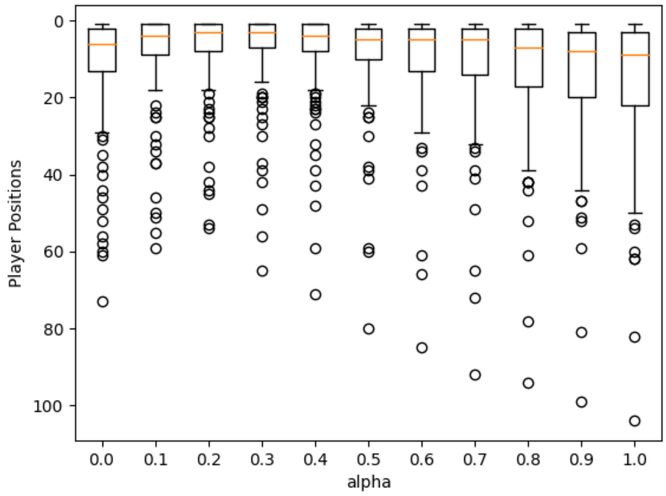

Fig. 3: Boxplot of different distributions generated by the variations in the constant α in $Dist(A, B)$. α varies from 0 to 1 in increments of 0.1

α	Mean	Standard Deviation
0	10.98	14.46
0.1	7.78	11.47
0.2	7.15	10.46
0.3	7.26	10.88
0.4	7.84	11.29
0.5	8.80	12.17
0.6	9.88	12.90
0.7	11.13	14.13
0.8	12.4	14.95
0.9	13.71	15.91
1	15.32	17.31

Table 2: Mean and Standard Deviation values for each distribution in Figure 3, according to the values of α

In Figure 4, we can see in more detail the curves representing the distributions for four different values of α. We can clearly see curves representing $\alpha = 0.2$ and $\alpha = 0.3$ have a much more compact look in the smaller position values, specially when compared to the distribution when $\alpha = 1$. Another thing we notice is that with $\alpha = 0.2$ and $\alpha = 0.3$ we have smaller ranges of values in positions and their concentration is higher in smaller position numbers.

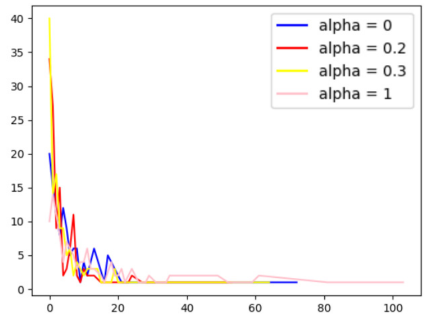

Fig. 4: Four different curve plots showcasing the different distributions caused by changing the value of α in the $Dist(A, B)$ distance metric.

6 Conclusions and Future Work

In this work we extended the concept of flow motif to incorporate the spatial dimension of football event data.

We were able to improve the characterisation of the passing behaviour of a player by encapsulating the spatial nature of the data. The distributions that represented the similarities between a given player in different halves of the season were proof of that increase in the capability to characterise a football player passing behaviour, for some values of α. In the future work, we aim at improving the way we count spatial flow motifs, since this approach has proven to be too much time consuming. This can be done either through the conceiving of a parallel algorithm or through improvements on the way we calculate the distance metric (through the exclusion of not relevant spatial flow motifs).

It also seems that it is possible to generalise this approach in order for it to be applied in different domains. In a more formal note, we are building graphs where the nodes are coloured based on the position they occupy in a two dimensional grid and the directed edges incorporate a time dimension in the sense that if node v has an outgoing edge that connects him to node u, then the event that represents node v happened before node u. This means that in a spatio-temporal domain, similar methodology may be applied to count k sized subgraphs (flow motifs) of those spatio-temporal graphs.

Acknowledgements: This work is partially financed financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project LA/P/0063/2020 and grant SFRH/BD/136525/2018.

References

1. Barbosa, A., Ribeiro, P., Dutra, I.: Similarity of football players using passing sequences. In: Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A. (eds.) Machine

Learning and Data Mining for Sports Analytics. pp. 51–61. Springer International Publishing, Cham (2022)
2. Bekkers, J., Dabadghao, S.: Flow motifs in soccer: What can passing behavior tell us? Journal of Sports Analytics 5(4), 299–311 (2019)
3. Fenner, J.S., Iga, J., Unnithan, V.: The evaluation of small-sided games as a talent identification tool in highly trained prepubertal soccer players. Journal of Sports Sciences 34(20), 1983–1990 (2016)
4. Gyarmati, L., Kwak, H., Rodriguez, P.: Searching for a unique style in soccer. arXiv preprint arXiv:1409.0308 (2014)
5. Håland, E.M., Wiig, A.S., Hvattum, L.M., Stålhane, M.: Evaluating the effectiveness of different network flow motifs in association football. Journal of Quantitative Analysis in Sports 16, 311 – 323 (2020)
6. Matesanz, D., Holzmayer, F., Torgler, B., Schmidt, S.L., Ortega, G.J.: Transfer market activities and sportive performance in european first football leagues: A dynamic network approach. PLoS ONE 13 (2018)
7. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science 298(5594), 824–827 (2002)
8. Narizuka, T., Yamamoto, K., Yamazaki, Y.: Statistical properties of position-dependent ball-passing networks in football games. Physica A: Statistical Mechanics and its Applications 412, 157–168 (2014)
9. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., Giannotti, F.: A public data set of spatio-temporal match events in soccer competitions. Scientific data 6(1), 1–15 (2019)
10. Peña, J.L., Navarro, R.S.: Who can replace xavi? a passing motif analysis of football players. arXiv preprint arXiv:1506.07768 (2015)
11. Razali, N., Mustapha, A., Utama, S., Din, R.: A review on football match outcome prediction using bayesian networks. In: Journal of Physics: Conference Series. vol. 1020, p. 012004. IOP Publishing (2018)
12. Rossi, A., Pappalardo, L., Cintia, P., Iaia, F.M., Fernández, J., Medina, D.: Effective injury forecasting in soccer with gps training data and machine learning. PLoS ONE 13 (2018)
13. Thakkar, P., Shah, M.: An assessment of football through the lens of data science. Annals of Data Science 8(4), 823–836 (2021)
14. Tuyls, K., Omidshafiei, S., Muller, P., et al.: Game plan: What ai can do for football, and what football can do for ai. Journal of AI Research 71, 41–88 (2021)
15. Wiig, A.S., Håland, E.M., Stålhane, M., Hvattum, L.M.: Analyzing passing networks in association football based on the difficulty, risk, and potential of passes. International Journal of Computer Science in Sport 18, 44 – 68 (2019)
16. Wu, Y., Xie, X., Wang, J., Deng, D., Liang, H., Zhang, H., Cheng, S., Chen, W.: Forvizor: Visualizing spatio-temporal team formations in soccer. IEEE Transactions on Visualization and Computer Graphics 25, 65–75 (2019)